SK-VQA: Synthetic Knowledge Generation at Scale for Training Context-Augmented Multimodal LLMs

Anonymous ACL submission

Abstract

Synthetic data generation has gained significant attention recently for its utility in training large vision and language models. However, the application of synthetic data to the training of multimodal context-augmented generation systems has been relatively unexplored. This gap in existing work is important because existing vision and language models (VLMs) are not trained specifically for context-augmented generation. Resources for adapting such models are therefore crucial for enabling their use in retrieval-augmented generation (RAG) settings, where a retriever is used to gather relevant information that is then subsequently provided to a generative model via context augmentation. To address this challenging problem, we generate SK-VQA: a large synthetic multimodal dataset containing over 2 million question-answer pairs which require external knowledge to determine the final answer. Our dataset is both larger and significantly more diverse than existing resources of its kind, possessing over 11x more unique questions and containing images from a greater variety of sources than previously-proposed datasets. Through extensive experiments, we demonstrate that our synthetic dataset can not only serve as a challenging benchmark, but is also highly effective for adapting existing generative multimodal models for context-augmented generation.

1 Introduction

Recent advances in Multimodal LLMs (MLLMs) have extended the impressive capabilities of LLMs to the vision domain, enabling advanced reasoning and chat capabilities over multimodal input queries consisting of both text and images (Achiam et al., 2023; Liu et al., 2024c). While MLLMs have demonstrated promising results, they suffer from the same hallucination and reliability issues as LLMs (Li et al., 2023; Zhou et al., 2023; Liu et al., 2024a). This motivates the need to incorporate MLLMs into retrieval-augmented generation (RAG) systems, where retrieved documents can ground answer generation in factually-correct information via context augmentation (Lewis et al., 2020; Ram et al., 2023). However, context augmentation for MLLMs presents unique challenges. Generated answers must be conditioned on both multimodal input queries as well as retrieved contexts which potentially span multiple modalities. Existing MLLMs have not been trained with context-augmented generation in mind, which makes them ill-suited for use in a RAG system.

Context: The Beverly Hilton is a luxury hotel located in Beverly Hills, California. Opened in 1955, the hotel was designed by architect Welton Becket and is owned by the Oasis West Realty, LLC. This iconic hotel has hosted numerous prestigious events over its history, most notably the annual Golden Globe Awards. The Beverly Hilton features more than 500 guest rooms, including 101 suites, and a variety of ...

Q: When was the building in the image opened?

A: 1955

Q: Who owns the building in the image?

A: Oasis West Realty, LLC

Q: What type of events is the location in the image known for hosting annually?

A: Golden Globe Awards

Q: How many suites are available in the building shown in the image?

A: 101

Figure 1: Example from our dataset. Given an input image, we prompt GPT-4 to generate a context document and QA pairs which require multimodal reasoning.

Adapting MLLMs for use in a RAG system requires extensive datasets which can support the training of models with multimodal queries and relevant context documents. Unfortunately, naturally-occurring data of this kind is relatively scarce; unlike other common types of internet data (e.g., text, image-text pairs), input queries consisting of both images and text with associated context

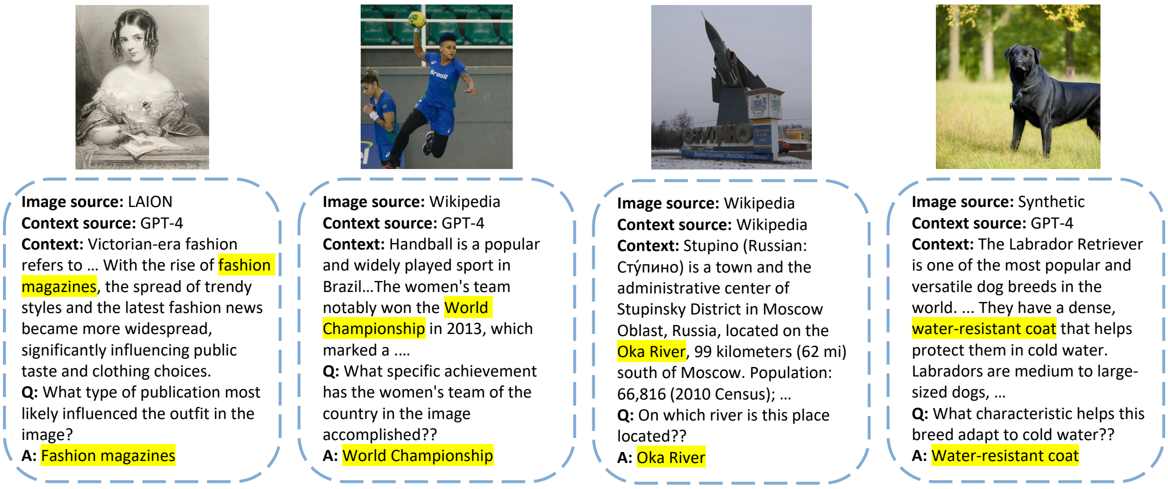


Image source: LAION
Context source: GPT-4
Context: Victorian-era fashion refers to … With the rise of fashion magazines, the spread of trendy styles and the latest fashion news became more widespread, significantly influencing public taste and clothing choices.
Q: What type of publication most likely influenced the outfit in the image?
A: Fashion magazines

Image source: Wikipedia
Context source: GPT-4
Context: Handball is a popular and widely played sport in Brazil…The women's team notably won the World Championship in 2013, which marked a ….
Q: What specific achievement has the women's team of the country in the image accomplished??
A: World Championship

Image source: Wikipedia
Context source: Wikipedia
Context: Stupino (Russian: Сту́пино) is a town and the administrative center of Stupinsky District in Moscow Oblast, Russia, located on the Oka River, 99 kilometers (62 mi) south of Moscow. Population: 66,816 (2010 Census); …
Q: On which river is this place located??
A: Oka River

Image source: Synthetic
Context source: GPT-4
Context: The Labrador Retriever is one of the most popular and versatile dog breeds in the world. … They have a dense, water-resistant coat that helps protect them in cold water. Labradors are medium to large-sized dogs, …
Q: What characteristic helps this breed adapt to cold water??
A: Water-resistant coat

Figure 2: Examples of QA pairs in our dataset created for images and contexts collected from different sources.

documents are not readily available at the scale needed for training MLLMs. Alternative methods of data collection must therefore be explored in order to enable the development of context-augmented MLLMs at scale.

Synthetically generated data has recently grown in popularity as a solution for cases where sources of naturally-occurring data are scarce or have been exhausted in existing training datasets. In the context of training MLLMs, synthetic data has played an important role in the visual instruction tuning required to develop such models (Liu et al., 2024c). A limited number of knowledge-based visual question answering (KB-VQA) datasets suitable for training MLLMs in a context-augmented setting have been constructed by synthetically producing question-answer (QA) pairs for real images and related text documents (Chen et al., 2023; Lerner et al., 2022; Mensink et al., 2023). However, these existing resources are mostly limited to images which can be linked to context documents sourced from Wikipedia, focus on entity-specific knowledge, and lack question diversity due to their reliance on templates for constructing QA pairs.

In this work, we introduce an alternative paradigm for collecting natural and diverse data at scale which is suitable for training context-augmented MLLMs. Rather than relying on template-based methods to construct QA pairs for real data, we propose a fully automated synthetic multimodal data generation approach which utilizes a strong foundation model (GPT-4) to produce relevant context documents and multiple QA pairs for a given image (Figure 1). This enables the acquisition of data which spans diverse sources of images (Figure 2), even allowing for the generation of fully-synthetic data which includes synthetic images, contexts, and QA pairs. Using this approach, we construct SK-VQA: the largest KB-VQA dataset to-date, containing over 2 million QA pairs associated with synthetic context knowledge and images sourced from LAION (Schuhmann et al., 2021), WIT (Wikipedia images) (Srinivasan et al., 2021), and the synthetic COCO-Counterfactuals dataset (Le et al., 2024).

Through comprehensive analyses, we show that our dataset is much more diverse compared to existing datasets. To further demonstrate its utility, we first conduct zero-shot experiments on six state-of-the-art MLLMs, showing that it is a challenging benchmark for these powerful models. We then fine-tune MLLMs on our dataset and compare their performance to models fine-tuned on existing KB-VQA datasets. Our experiments show that our dataset enhances the generalization capabilities of MLLMs, whereas other datasets result in poor generalization performance. We attribute this improved generalization capacity to the diversity of our dataset. In summary, our dataset is not only a challenging benchmark for KB-VQA, but also a valuable resource for improving the context-augmented generation capabilities of MLLMs. Our dataset and code will be made publicly available.

2 Related Work

Synthetic Data Generation Synthetic data has grown in popularity lately as an effective strategy for data augmentation, particularly in the multimodal domain where data is often more scarce. Advances in text-to-image diffusion models (Nichol et al., 2021; Rombach et al., 2021; Saharia et al., 2022; Ramesh et al., 2022) have enabled the generation of synthetic data for a variety of use cases such as image classification (He et al., 2022; Trabucco et al., 2023; Vendrow et al., 2023) and image-text counterfactuals (Le et al., 2024; Howard et al., 2023). In the domain of NLP, augmenting prompts with LLM-generated context documents has been demonstrated to be competitive with retrieving real text documents for context augmentation in RAG systems (Yu et al., 2022). Synthetic data has also been shown to be useful for training text embedding models for retrieval (Wang et al., 2023). To the best of our knowledge, our work is the first to explore fully synthetic datasets for training MLLMs.

Despite its demonstrated benefits, several risks have been noted in utilizing synthetic data for model training. Shumailov et al. (2023) showed that training language models on data that is contaminated with increasing amounts of model-generated content leads to model collapse, while Gerstgrasser et al. (2024) found that accumulating model-generated content without replacing original content can avoid this phenomenon. In the context of training vision-language models, synthetic image data has been shown to scale similarly in effectiveness of CLIP (Radford et al., 2021) training as real images, while significantly under performing real data in training supervised image classifiers (Fan et al., 2023). Improving out-of-domain generalization by training on synthetic data has also been shown to be sensitive to the ratio of real and synthetic data (Howard et al., 2022; Le et al., 2024).

Knowledge-Based Visual Question Answering Datasets Marino et al. (2019) introduced OK-VQA, a KB-VQA dataset of 14k crowdsourced questions for COCO images which are designed to require external knowledge to answer, but are not associated with ground truth context documents. Lerner et al. (2022) introduced the ViQuAE dataset, which consists of 3.7k questions about named entities paired with images and text articles from Wikipedia. Chen et al. (2023) found that many questions in OK-VQA and ViQuAE can be answered without external knowledge; motivated by this finding, they introduced the InfoSeek dataset containing over 1.3 million information-seeking questions paired with images from existing image classification and retrieval datasets which have been grounded to Wikipedia articles. Although they curate a smaller set of 8.9k human-written questions for testing, the vast majority of InfoSeek is automatically constructed by populating human-authored templates from Wikidata triples.

Encyclopedic VQA (Mensink et al., 2023) is another recently-proposed KB-VQA dataset consisting of 221k unique QA pairs which are each associated with up to 5 images from the iNaturalist (Van Horn et al., 2021) and Google Landmarks (Weyand et al., 2020) datasets. They utilize the WIT dataset (Srinivasan et al., 2021) to link images with Wikipedia text documents and employ templates along with a question generation model to automatically construct 1 million question-answer pairs. SnapNTell (Qiu et al., 2024) also contains KB-VQA questions requiring entity-specific external knowledge to answer, but contains fewer QA pairs (75.6k) and was not publicly available at the time of writing. Other knowledge-intensive VQA datasets have been proposed for more specific domains of multimodal documents, including technical engineering requirements (Doris et al., 2024) and scientific journal articles (Ding et al., 2024).

Multimodal RAG systems In the domain of KB-VQA, augmenting transformer-based generators with retrieved multimodal documents has been shown to be effective in architectures such as RA-CM3 (Yasunaga et al., 2022) MuRAG (Chen et al., 2022), and REVEAL (Hu et al., 2023). More recently, LLMs augmented with vision encoders such as LLaVA (Liu et al., 2024c) and GPT-4 (Achiam et al., 2023) have demonstrated state-of-the-art performance on a variety of image-to-text generation tasks, motivating the investigation of retrieval-based context augmentation for such models. Re-ViLM (Yang et al., 2023) augments Flamingo with retrieved multimodal documents, while Wiki-LLaVA (Caffagni et al., 2024) augments LLaVA model with Wikipedia documents.

Wei et al. (2023) proposed UniIR for multimodal retrieval, utilizing score-level and feature-level fusion approaches with pre-trained CLIP (Radford et al., 2021) and BLIP (Li et al., 2022) models. Sharifymoghaddam et al. (2024) showed that UniIR can improve the performance of large multimodal language models on image captioning and image generation tasks. UniMur (Wang et al., 2024) embeds multimodal inputs and retrieves multimodal outputs via frozen LLMs. Our work differs from these studies in that we focus on how to adapt MLLMs for context-augmented generation in a RAG system rather than training the retriever.

3 Methodology

3.1 Dataset generation

Motivated by recent advances in MLLMs, we use a fully automated approach to generate synthetic context documents and question-answer (QA) pairs for a given image with GPT-4. This provides several distinct advantages over alternative approaches. First, the powerful language abilities of GPT-4 allow us to acquire more natural and diverse questions than previous datasets which rely on templated construction of question-answer pairs. Second, generating context documents enables the use of a much broader range of images than what has been considered in previous datasets, where images are typically restricted only to those that can be linked to Wikipedia passages.

Given an input image, we prompt GPT-4 to generate a context document[1] related to the image and QA pairs which require reasoning over both the image and the context document. The complete prompt we use for this purpose is provided in Figure 3 (see Figure 7 of Appendix E for additional discussion). Importantly, we generate both the context document and QA pairs in a single inference step. In doing so, the generation of the context is conditioned on the task of producing questions that require both the image and the context. This helps ensure that the context associated with each image is suitable for the creation of the style of QA pairs we seek, which is not necessarily the case when context documents are acquired automatically from existing sources such as Wikipedia. Following generation, we parse the output of GPT-4 to extract the context document and QA pairs (see Appendix E for details). We then apply two stages of filtering to create separate filtered subsets.

3.2 Image Reference (ImRef) filtering

In manual evaluation of generated context documents, we found that GPT-4 sometimes directly references the input image that was provided. For example, the context documents may include references such as “In the image, ...” or “As shown in the picture, ...”. In such cases, the information contained in the context document is more similar to an extended caption or image description than a knowledge-intensive document which is related to the image. While this may not necessarily be detrimental to the training of multimodal RAG systems, it is unlikely in practice for RAG systems to require the retrieval of image-specific context documents. We therefore create a filtered subset of our dataset which excludes these cases by identifying the presence of the words `picture`, `photo`, `image`, or `painting` in the generated context document. We refer to this subset as the SK-VQA$_{\text{IR}}$ subset.

```
Write a Wikipedia article related to this image
without directly referring to the image. Then
write question answer pairs. The question answer
pairs should satisfy the following criteria.

1: The question should refer to the image.
2: The question should avoid mentioning the name
of the object in the image.
3: The question should be answered by reasoning
over the Wikipedia article.
4: The question should sound natural and concise.
5: The answer should be extracted from the
Wikipedia article.
6: The answer should not be any objects in the
image.
7: The answer should be a single word or phrase
and list all correct answers separated by commas.
8: The answer should not contain 'and', 'or',
rather you can split them into multiple answers.
```

Figure 3: Our prompt for generating synthetic data.

3.3 Context Answer Presence (CAP) filtering

In existing datasets for KB-VQA, it is common for the answer to be explicitly stated in the associated context document. This is not necessarily required in order for a QA pair to be valid since the answer could sometimes be inferred indirectly rather than being explicitly stated in the context. Nevertheless, the presence of the answer in the context document provides an indication that the question can indeed be answered from information contained in the context document. Therefore, we create an additional filtered subset of our dataset which only contains QA pairs where (1) at least one of the answer candidates is contained in the context document, and (2) the context does not directly reference the image (as described previously). We refer to this subset as the SK-VQA$_{\text{IR+CAP}}$ subset

4 Dataset Analysis

4.1 Dataset composition

In order to acquire synthetic data which spans a broad range of different domains, we utilize images from multiple sources during generation. These sources include LAION-400m, Wikipedia images contained in the Wiki dataset, and synthetically generated images from the COCO-Counterfactuals

[1] See Limitations for a discussion of hallucination impact

Image source	Context source	SK-VQA	SK-VQA$_{\text{IR}}$	SK-VQA$_{\text{IR+CAP}}$
LAION	GPT-4	908,116	584,126	371,936
Wikipedia	GPT-4	702,332	585,768	354,244
Wikipedia	Wikipedia	181,554	167,352	137,160
COCO-CFs	GPT-4	214,487	193,226	121,284
		2,006,489	1,530,472	984,624

Table 1: Total number of QA pairs in our dataset by image and context source, computed separately for SK-VQA, SK-VQA$_{\text{IR}}$, and SK-VQA$_{\text{IR+CAP}}$.

(COCO-CFs) dataset. We use the entirety of the COCO-Counterfactuals dataset along with a sub-sample of images from LAION and Wikipedia to generate context documents with QA pairs using our prompt. We also generate only QA pairs for a sub-sample of Wikipedia images paired with Wikipedia context documents from the Wiki dataset, which enables us to compare the effect of using real context documents to synthetically generated contexts (see Appendix E for details).

Table 1 provides a breakdown of the total number of QA pairs in our dataset by image and context source. Our full SK-VQA dataset contains over 2 million QA pairs, making it the largest KB-VQA dataset created to-date. Of these 2 million QA pairs, 45% are associated with images sourced from LAION, 44% are associated with Wikipedia images, and the remainder are paired with synthetic images from COCO-Counterfactuals. The SK-VQA$_{\text{IR}}$ subset contains 24% less QA pairs than the full SK-VQA dataset, while the SK-VQA$_{\text{IR+CAP}}$ subset contains approximately half the number of QA pairs as the full SK-VQA dataset.

Our full dataset contains 290,266 unique image-context pairs, which each have 7 QA pairs on average. GPT-4 generated context documents are associated with 7.1 QA pairs on average, whereas Wiki context documents are only associated with 5.7 QA pairs. This indicates that having GPT-4 generate both the context document and QA pairs simultaneously enables the acquisition of more QA pairs, which could be attributable to how the generation of the context document is conditioned on the subsequent task of producing QA pairs.

4.2 Question diversity

Table 2 provides statistics on the diversity of questions in our dataset and other existing KB-VQA datasets. In addition to having nearly 50% more questions then the next-largest dataset, our dataset also exhibits significantly greater question diversity. The only two existing datasets of comparable size, InfoSeek and Encyclopedic-VQA (Enc-VQA), contain significantly fewer unique questions; less than 1% of questions in InfoSeek are unique, while fewer than 17% are unique in Enc-VQA. In contrast, over 96% of our questions are unique, which corresponds to 11x more unique questions than Enc-VQA. Questions from our dataset also exhibit a greater number of unique POS sequences, total vocabulary size, and mean word length. This points to the value of leveraging powerful MLLMs for synthetic data generation over simpler techniques (e.g., populating templates).

Dataset	Total Qs	Unique Qs	Unique POS	Vocab Size	Length
ViQuAE	3,700	3,562	2,759	4,700	12.4
InfoSeek	1,356,000	1,498	267	725	8.9
Enc-VQA	1,036,000	175,000	91,945	40,787	11.6
SK-VQA	**2,006,489**	**1,928,336**	**926,817**	**138,372**	**12.7**

Table 2: Comparison of question (Q) diversity in KB-VQA datasets. ViQuAE, InfoSeek, and Enc-VQA values are previously reported in Lerner et al. (2024).

	Mean	Standard Deviation
SK-VQA	0.77	0.02
SK-VQA$_{\text{IR}}$	0.77	0.02
SK-VQA$_{\text{IR+CAP}}$	0.87	0.03

Table 3: Performance of human annotators on different filtered subsets of 100 sampled QA pairs from SK-VQA, calculated using semantic evaluation.

4.3 Knowledge classification

We apply an unsupervised topic modeling technique to categorize the knowledge contained in our dataset's context documents (see Appendix G for details). Figure 4 depicts the distribution of major topic categories identified in this analysis. Whereas previous KB-VQA datasets have focused primarily on entity-specific knowledge, our dataset spans a broader range of topics such as art, fashion, sports, events, and music. This demonstrates the diversity of external knowledge required by questions in our dataset and suggests that it can serve as a complementary resource to existing datasets which focus on entity-specific knowledge.

4.4 Human evaluation

We randomly sample 100 QA pairs from our dataset for human labeling by three of the authors of this work, ensuring that the 100 QA pairs are equally distributed across the four image & context source types shown in Table 1. For each QA pair,

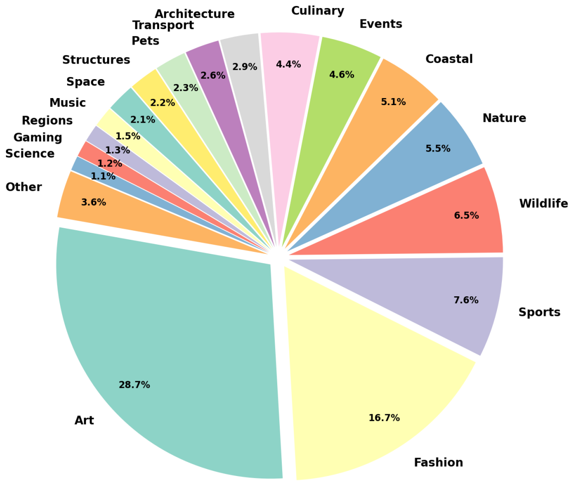


Figure 4: Distribution of 18 topic categories identified for context documents in SK-VQA.

the annotators were presented only with the image, context document, and question. They were then instructed to write an answer to the question, and optionally note any deficiencies that were evident.

Table 3 provides the mean and standard deviation of annotator accuracy, calculated using the Enc-VQA semantic evaluation method (see Section 5.2 for details). The overall mean accuracy of the three annotators was 77% for the 100 QA pairs sampled from SK-VQA and the subset of those which belong to SK-VQA$_{\text{IR}}$. For the subset of annotated QA pairs which belong to SK-VQA$_{\text{IR+CAP}}$, the mean human accuracy increases to 87%, which is consistent with previously reported human accuracy for other VQA datasets (Hudson and Manning, 2019; Sheng et al., 2021). The relatively low standard deviation indicates the annotators achieved similar performance.

To understand potential failure cases in our dataset, we categorized common annotator comments by identifying those for which at least two annotators recorded the same category of issue for a question. The most common issue reported by annotators were cases in which the question could be answered solely using the context document, assuming that the context document was provided at inference time. While this concern was noted for 9% of evaluated QA pairs, these examples may still require multimodal reasoning in a broader RAG system in which the question and image are necessary to retrieve the relevant context document. A small number of examples (5%) were identified as being answerable solely by looking at the image, while 1 question was noted as having insufficient information in the context and image to answer (see Figure 12 for examples).

5 Experiments

5.1 Experimental Setup

We conduct zero-shot and fine-tuning experiments on MLLMs using both our dataset and existing KB-VQA datasets. For the zero-shot experiments, we test the following popular MLLMs: PaLIGemma-3B [2], LLaVA-v1.5-7B (Liu et al., 2024c), LLaVA-1.6-7B/34B (Liu et al., 2024b), Qwen-VL-7B (Bai et al., 2023), and Idefics2-8B (Laurençon et al., 2024). For the fine-tuning experiments, we utilize LLaVA-v1.5-7B and PaLI-Gemma-3B. To demonstrate the effectiveness of our synthetic data, we train models using various subsets of our dataset generated from different sources, as described previously in Section 3.1. Additionally, we train two baseline models on existing KB-VQA datasets: InfoSeek and Enc-VQA. For InfoSeek, we use a 140K subset of the training data processed by Wei et al. (2023), where only external textual knowledge is required for the given questions and images (denoted as as task 6 by Wei et al. (2023)). We use the original Enc-VQA training set, but since each question can be paired with multiple images, we select only the first image from the original annotations for the training set, which results in approximately 220K training samples. For a fair comparison, we down-sample our dataset subsets to 200K samples each. Additional experimental details are provided in Appendix C and Appendix D.

5.2 Evaluation Datasets and Metrics

Datasets We use three existing KB-VQA datasets for evaluation: InfoSeek, Enc-VQA, and ViQuAE. Additionally, we use a subset of our dataset (10,744 ImRef-filtered QA pairs associated with images from LAION) for model evaluation. For InfoSeek, similar to the training dataset described in Section 5.1, we use a subset of its validation set processed by Wei et al. (2023), which includes 11,323 samples where only external textual knowledge is required for the given questions and images. For Enc-VQA, we use its official test set, which contains 5,750 samples. Due to the small size of the ViQuAE test set, we combine the original train, validation, and test sets to create a larger testing set of 3,625 samples.

[2] From Google Research

Model	Infoseek	Enc-VQA	ViQuAE	SK-VQA
PaliGemma-3B	25.66	32.89	47.72	25.51
LLaVA-v1.5-7B	42.82	53.69	78.41	40.99
LLaVa-v1.6-7B	41.94	57.92	72.00	46.68
Qwen-VL-7B	39.48	53.67	69.46	42.55
Idefics2-8B	**44.33**	67.92	**82.43**	38.08
LLaVa-v1.6-34B	38.81	**77.73**	79.17	**50.02**

Table 4: Zeroshot evaluation of SOTA MLLMs on existing KB-VQA and our datasets. Both our dataset and InfoSeek present a harder testing set for MLLMs.

Image Source	Context Source	Infoseek	Enc-VQA	ViQuAE	Avg.
LAION	GPT-4	44.32	65.44	79.22	62.99
Wiki	GPT4	47.00	53.98	78.58	59.85
Wiki	Wiki	47.75	**66.67**	77.95	64.12
COCO-CFs	GPT4	**48.00**	65.42	**79.23**	**64.22**

Table 5: Compare the performance of LLaVa-v1.5-7B trained on different source. The data are sample from data Wikihout any filtering.

Evaluation Metrics For Enc-VQA, we utilize the official semantic evaluation method, BEM (Bulian et al., 2022), where a BERT-based model (Devlin et al., 2018) is fine-tuned on a question, a ground truth answer, and a generated answer. The model outputs a value indicating whether the generated answer is semantically correct. Following the official settings, we use a threshold of 0.5 to determine if the generated answer is correct or not. For other datasets, we use exact string matching.

5.3 Zero-shot Results

Table 4 provides the results for zero-shot evaluations. From the results, it is evident that all tested state-of-the-art MLLMs perform better on Enc-VQA and ViQuAE compared to InfoSeek and our SK-VQA dataset. This suggests that SK-VQA and InfoSeek present significant challenges to these models. Unlike with Enc-VQA and ViQuAE, larger models do not always yield better performance on InfoSeek and SK-VQA. This indicates that simply relying on model size may not be sufficient to address the reasoning challenges presented by these dataset.

5.4 Fine-tuning Results

We evaluate fine-tuned MLLMs on their ability to generalize to other datasets (i.e., out-of-domain performance). Figure 5 shows that for LLaVA-7B, fine-tuning with the InfoSeek and Enc-VQA datasets improves performance in only 1 out of 6 cases compared to the zero-shot baseline. Specifically, the model trained on InfoSeek and tested on our synthetic dataset shows improvement. In contrast, models trained on our SK-VQA dataset significantly improve zero-shot performance in 6 out of 9 cases. The remaining three cases, tested on the ViQuAE dataset, show no significant improvement but also no severe performance degradation.

Fine-tuning PaliGemma-3B using InfoSeek results in performance degradation in 2 out of 3 settings compared to the zero-shot baseline. However, fine-tuning with Enc-VQA consistently improves performance. Again, models trained on our dataset show significant performance improvements in all 9 cases and achieve the best out-of-domain performance. Overall, these results indicate that fine-tuning MLLMs with SK-VQA effectively improves out-of-domain performance in most cases. Even when there is no improvement, SK-VQA does not result in performance degradation as with fine-tuning on other datasets.

Table 7 provides additional evaluation results for in-domain performance. In this setting, where models are trained and tested on their respective training and testing sets within the same dataset, both the LLaVA-7B and PaliGemma-3B models show significant improvements compared to the zero-shot baseline, as expected. These models not only surpass the zero-shot baseline performance but also outperform models trained on other datasets.

5.5 Ablation Studies

Impact of Generation Source We explore the performance of models trained on data generated from different images and context sources across three external datasets. Table 5 shows that the best combination involves using images from COCO-CFs and context documents from GPT-4. Notably, this combination even surpasses using images from Wiki with their real context. This indicates that our generated dataset can offer advantages for fine-tuning MLLMs compared to real data.

Additionally, when using GPT-4 to generate context, comparing the use of images from LAION and Wiki reveals that models fine-tuned on data generated from LAION images perform better on Enc-VQA and ViQuAE, whereas models fine-tuned on data from Wiki images perform better on infoSeek. This suggests that combining images from different sources may be necessary to achieve better generalization across all external datasets, as we have shown in Section 5.4

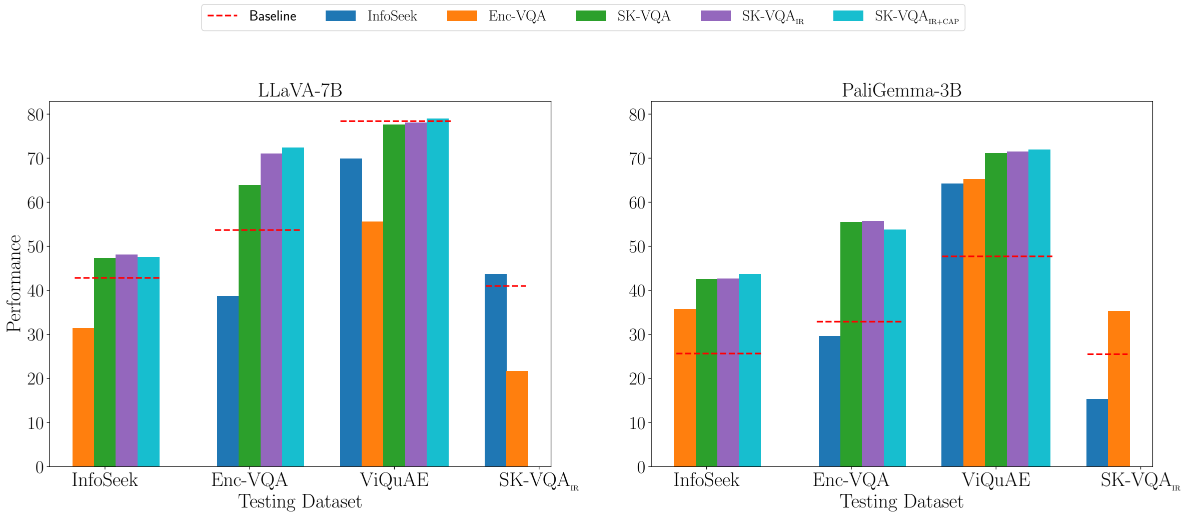


Figure 5: Generalization performance of MLLMs fine-tuned on different KB-VQA datasets.

Impact of Filtering Techniques To further explore the impact of different filtering methods, we fix the image source for data generation and compare the performance of models trained on data filtered by various methods across three out-of-domain datasets, as shown in Table 6. We sequentially apply IR and CAP filtering on SK-VQA. For data from LAION, SK-VQA$_{\text{IR}}$ retains 64% of the original data, while SK-VQA$_{\text{IR+CAP}}$ retains 40%. For data from Wikipedia, SK-VQA$_{\text{IR}}$ retains 83%, and SK-VQA$_{\text{IR+CAP}}$ retains 50%.

The results indicate that with LAION as the image source, SK-VQA$_{\text{IR+CAP}}$ achieves the best average performance across the three datasets, though SK-VQA outperforms SK-VQA$_{\text{IR}}$. For data from Wikipedia, SK-VQA$_{\text{IR}}$ outperforms both SK-VQA$_{\text{IR+CAP}}$ and SK-VQA overall, while certain datasets (InfoSeek) benefit most from the full SK-VQA dataset. This suggests that specific filtering methods may improve performance for certain domains or datasets, while the full unfiltered SK-VQA dataset might be more valuable for others, demonstrating the versatility of our dataset and filtering techniques. The various filtered subsets of our dataset can be treated as a hyperparameter to find the best performance for specific tasks. We also note that a significant benefit of filtering is the ability to achieve similar or better performance with significantly fewer samples, which holds true for both filtering methods across all datasets.

Training Data	Image Source	Infoseek	Enc-VQA	ViQuAE	Avg.
SK-VQA	LAION	44.32	65.44	**79.22**	62.99
SK-VQA$_{\text{IR}}$	LAION	44.43	63.08	75.50	61.00
SK-VQA$_{\text{IR+CAP}}$	LAION	**45.85**	**69.88**	78.18	**64.64**
SK-VQA	Wiki	**47.00**	53.98	78.58	59.85
SK-VQA$_{\text{IR}}$	Wiki	45.99	**67.36**	79.37	**64.24**
SK-VQA$_{\text{IR+CAP}}$	Wiki	46.48	64.55	**79.83**	63.62

Table 6: Impact of filtering techniques by image source. All results utilize context documents from GPT-4.

6 Conclusion

We presented a methodology to automatically acquire high-quality synthetic data suitable for training and evaluating MLLMs in a context-augmented generation setting. Using our approach, we constructed SK-VQA: a large dataset of 2 million QA pairs over images from multiple different sources. SK-VQA is the largest and most diverse resource of its kind, possessing 11x more unique questions than similar datasets for KB-VQA. Our evaluations of popular MLLMs showed that our dataset can serve as a more challenging benchmark than existing resources. Additionally, training MLLMs on our dataset leads to greater improvements in out-of-domain generalization than other datasets.

These results point to not only the utility of SK-VQA, but also the effectiveness of our approach for acquiring synthetic multimodal data at scale. Opportunities for future work in this direction include leveraging larger amounts of synthetic image data with our approach to produce fully-synthetic data for domains of images which are under-represented in existing KB-VQA datasets. Leveraging our dataset for training multimodal retrieval models could be another promising direction to aid in the development of multimodal RAG systems.

Limitations

In this work, we explored the generation of synthetic data from GPT-4 due to its demonstrated state-of-the-art performance in a broad range of multimodal reasoning tasks. While our data generation approach could be used with other MLLMs, we leave the investigation of such applications to future work. Our dataset is limited to English language QA pairs and context documents. While the images in our dataset were collected from a diverse range of sources, they may not be representative of all images domains which might be relevant to users of our dataset.

Due to the scale of our automatically constructed dataset, we are unable to fully validate the accuracy all examples that it contains. We believe that our empirical results provide strong evidence of its quality and usefulness for training MLLMs. While human annotators did not explicitly validate the accuracy of all information contained in evaluated context passages, no obvious cases of hallucination were identified during the annotation process. However, the synthetic nature of the data introduces the possibility that it contains fallacies. Since our primary aim is to train MLLMs to ground generated answers in context documents, we believe such errors pose relatively low risk to the intended use of our dataset. Nevertheless, caution should be exercised when utilizing our dataset, including validation of the performance of any models which are trained on it.

Ethical Considerations

Our dataset was generated from GPT-4 using the Azure OpenAI API, which includes a content filter for multiple risk categories (e.g., hate speech, fairness, sexual language, violence). As this filter automatically removes potentially offensive content that is generated by GPT-4, we believe that the likelihood of our dataset containing such content is relatively low. However, content filtering models are not infallible and we are unable to manually inspect our entire dataset for the presence of offensive content due to its large scale. It is also possible that potentially harmful biases possessed by GPT-4 which do not trigger content filters are reflected in our dataset. Users should carefully consider these risks relative to the benefits of our synthetic dataset before deploying systems which are built using it.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

AI@Meta. 2024. Llama 3 model card.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*.

Jannis Bulian, Christian Buck, Wojciech Gajewski, Benjamin Börschinger, and Tal Schuster. 2022. Tomayto, tomahto. beyond token-level answer equivalence for question answering evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 291–305.

Davide Caffagni, Federico Cocchi, Nicholas Moratelli, Sara Sarto, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. 2024. Wiki-llava: Hierarchical retrieval-augmented generation for multimodal llms. *arXiv preprint arXiv:2404.15406*.

Wenhu Chen, Hexiang Hu, Xi Chen, Pat Verga, and William W Cohen. 2022. Murag: Multimodal retrieval-augmented generator for open question answering over images and text. *arXiv preprint arXiv:2210.02928*.

Yang Chen, Hexiang Hu, Yi Luan, Haitian Sun, Soravit Changpinyo, Alan Ritter, and Ming-Wei Chang. 2023. Can pre-trained vision and language models answer visual information-seeking questions? *arXiv preprint arXiv:2302.11713*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Yihao Ding, Kaixuan Ren, Jiabin Huang, Siwen Luo, and Soyeon Caren Han. 2024. Mvqa: A dataset for multimodal information retrieval in pdf-based visual question answering. *arXiv preprint arXiv:2404.12720*.

Anna C Doris, Daniele Grandi, Ryan Tomich, Md Ferdous Alam, Hyunmin Cheong, and Faez Ahmed. 2024. Designqa: A multimodal benchmark for evaluating large language models' understanding of engineering documentation. *arXiv preprint arXiv:2404.07917*.

Lijie Fan, Kaifeng Chen, Dilip Krishnan, Dina Katabi, Phillip Isola, and Yonglong Tian. 2023. Scaling laws of synthetic images for model training... for now. *arXiv preprint arXiv:2312.04567*.

Matthias Gerstgrasser, Rylan Schaeffer, Apratim Dey, Rafael Rafailov, Henry Sleight, John Hughes, Tomasz Korbak, Rajashree Agrawal, Dhruv Pai, Andrey Gromov, et al. 2024. Is model collapse inevitable? breaking the curse of recursion by accumulating real and synthetic data. *arXiv preprint arXiv:2404.01413*.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *Preprint*, arXiv:2203.05794.

Ruifei He, Shuyang Sun, Xin Yu, Chuhui Xue, Wenqing Zhang, Philip Torr, Song Bai, and Xiaojuan Qi. 2022. Is synthetic data from generative models ready for image recognition? *arXiv preprint arXiv:2210.07574*.

Phillip Howard, Avinash Madasu, Tiep Le, Gustavo Lujan Moreno, Anahita Bhiwandiwalla, and Vasudev Lal. 2023. Probing and mitigating intersectional social biases in vision-language models with counterfactual examples. *arXiv preprint arXiv:2312.00825*.

Phillip Howard, Gadi Singer, Vasudev Lal, Yejin Choi, and Swabha Swayamdipta. 2022. Neurocounterfactuals: Beyond minimal-edit counterfactuals for richer data augmentation. *arXiv preprint arXiv:2210.12365*.

Ziniu Hu, Ahmet Iscen, Chen Sun, Zirui Wang, Kai-Wei Chang, Yizhou Sun, Cordelia Schmid, David A Ross, and Alireza Fathi. 2023. Reveal: Retrieval-augmented visual-language pre-training with multi-source multimodal knowledge memory. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 23369–23379.

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*, pages 6700–6709.

Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. 2024. What matters when building vision-language models? *Preprint*, arXiv:2405.02246.

Tiep Le, Vasudev Lal, and Phillip Howard. 2024. Coco-counterfactuals: Automatically constructed counterfactual examples for image-text pairs. *Advances in Neural Information Processing Systems*, 36.

Paul Lerner, Olivier Ferret, and Camille Guinaudeau. 2024. Cross-modal retrieval for knowledge-based visual question answering. In *European Conference on Information Retrieval*, pages 421–438. Springer.

Paul Lerner, Olivier Ferret, Camille Guinaudeau, Hervé Le Borgne, Romaric Besançon, José G Moreno, and Jesús Lovón Melgarejo. 2022. Viquae, a dataset for knowledge-based visual question answering about named entities. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3108–3120.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

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*, pages 12888–12900. PMLR.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*.

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

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024b. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024c. Visual instruction tuning. *Advances in neural information processing systems*, 36.

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

Thomas Mensink, Jasper Uijlings, Lluis Castrejon, Arushi Goel, Felipe Cadar, Howard Zhou, Fei Sha, André Araujo, and Vittorio Ferrari. 2023. Encyclopedic vqa: Visual questions about detailed properties of fine-grained categories. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3113–3124.

Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. 2021. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*.

Jielin Qiu, Andrea Madotto, Zhaojiang Lin, Paul A Crook, Yifan Ethan Xu, Xin Luna Dong, Christos Faloutsos, Lei Li, Babak Damavandi, and Seungwhan Moon. 2024. Snapntell: Enhancing entity-centric visual question answering with retrieval augmented multimodal llm. *arXiv preprint arXiv:2403.04735*.

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*, pages 8748–8763. PMLR.

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. In-context retrieval-augmented language models. *Transactions of the Association for Computational Linguistics*, 11:1316–1331.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2021. High-resolution image synthesis with latent diffusion models. *Preprint*, arXiv:2112.10752.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494.

Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. 2021. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*.

Sahel Sharifymoghaddam, Shivani Upadhyay, Wenhu Chen, and Jimmy Lin. 2024. Unirag: Universal retrieval augmentation for multi-modal large language models. *arXiv preprint arXiv:2405.10311*.

Sasha Sheng, Amanpreet Singh, Vedanuj Goswami, Jose Magana, Tristan Thrush, Wojciech Galuba, Devi Parikh, and Douwe Kiela. 2021. Human-adversarial visual question answering. *Advances in Neural Information Processing Systems*, 34:20346–20359.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Yarin Gal, Nicolas Papernot, and Ross Anderson. 2023. The curse of recursion: Training on generated data makes models forget. *arXiv preprint arXiv:2305.17493*.

Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. 2021. Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2443–2449.

Brandon Trabucco, Kyle Doherty, Max Gurinas, and Ruslan Salakhutdinov. 2023. Effective data augmentation with diffusion models. *arXiv preprint arXiv:2302.07944*.

Grant Van Horn, Elijah Cole, Sara Beery, Kimberly Wilber, Serge Belongie, and Oisin Mac Aodha. 2021. Benchmarking representation learning for natural world image collections. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12884–12893.

Joshua Vendrow, Saachi Jain, Logan Engstrom, and Aleksander Madry. 2023. Dataset interfaces: Diagnosing model failures using controllable counterfactual generation. *arXiv preprint arXiv:2302.07865*.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2023. Improving text embeddings with large language models. *arXiv preprint arXiv:2401.00368*.

Ziyang Wang, Heba Elfardy, Markus Dreyer, Kevin Small, and Mohit Bansal. 2024. Unified embeddings for multimodal retrieval via frozen llms. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1537–1547.

Cong Wei, Yang Chen, Haonan Chen, Hexiang Hu, Ge Zhang, Jie Fu, Alan Ritter, and Wenhu Chen. 2023. Uniir: Training and benchmarking universal multimodal information retrievers. *arXiv preprint arXiv:2311.17136*.

Tobias Weyand, Andre Araujo, Bingyi Cao, and Jack Sim. 2020. Google landmarks dataset v2-a large-scale benchmark for instance-level recognition and retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2575–2584.

Zhuolin Yang, Wei Ping, Zihan Liu, Vijay Korthikanti, Weili Nie, De-An Huang, Linxi Fan, Zhiding Yu, Shiyi Lan, Bo Li, et al. 2023. Re-vilm: Retrieval-augmented visual language model for zero and few-shot image captioning. *arXiv preprint arXiv:2302.04858*.

Michihiro Yasunaga, Armen Aghajanyan, Weijia Shi, Rich James, Jure Leskovec, Percy Liang, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2022. Retrieval-augmented multimodal language modeling. *arXiv preprint arXiv:2211.12561*.

Wenhao Yu, Dan Iter, Shuohang Wang, Yichong Xu, Mingxuan Ju, Soumya Sanyal, Chenguang Zhu, Michael Zeng, and Meng Jiang. 2022. Generate rather than retrieve: Large language models are strong context generators. *arXiv preprint arXiv:2209.10063*.

Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. 2023. Analyzing and mitigating object hallucination in large vision-language models. *arXiv preprint arXiv:2310.00754*.

	Training Data	**Infoseek**	**Enc-VQA**	**ViQuAE**	**SK-VQA$_{IR}$**
LLaVA-7B	-	42.82	53.69	78.41	40.99
	Infoseek	68.68*	38.68	69.90	**43.69**
	Enc-VQA	31.39	88.75*	55.59	21.60
	SK-VQA	47.35	63.89	77.60	68.30*
	SK-VQA$_{IR}$	**48.11**	70.99	78.04	68.35*
	SK-VQA$_{IR+CAP}$	47.55	**72.33**	**78.95**	68.61*
PaliGemma-3B	-	25.66	32.89	47.72	25.51
	Infoseek	66.63*	29.58	64.22	15.27
	Enc-VQA	35.78	83.30*	65.19	**35.29**
	SK-VQA	42.50	55.53	71.12	65.02*
	SK-VQA$_{IR}$	42.69	**55.67**	71.48	65.26*
	SK-VQA$_{IR+CAP}$	**43.72**	53.72	**71.97**	64.69*

Table 7: Performance of Generator fine-tuned on ~200K from different datasets. * denotes the in-domain results. Encyclopedia evaluated on semantic matric, others are Exact Matching scores.

	Training Data	All	Hard
LLaVA-7B	-	40.99	34.74
	InfoSeek	43.69	33.82
	Enc-VQA	21.60	19.84
	SK-VQA	68.30*	52.37*
	SK-VQA$_{IR}$	68.35*	51.38*
	SK-VQA$_{IR+CAP}$	68.61*	51.81*
PaliGemma-3B	-	25.51	23.52
	InfoSeek	15.27	15.77
	Enc-VQA	35.29	30.18
	SK-VQA	65.02*	49.35*
	SK-VQA$_{IR}$	65.26*	49.07*
	SK-VQA$_{IR+CAP}$	64.69*	50.23*

Table 8: Model accuracy on SK-VQA$_{IR}$, calculated for all questions and the "hard" subset. * denotes the in-domain results.

A Additional Fine-tuning Experimental Results

In-domain performance Table 7 provides additional in-domain performance results of fine-tuned models from the experiments discussed in Section 5.4. The results indicate that models trained on specific datasets perform best on their corresponding test sets. However, as noted in Section 5.4, good in-domain performance for models trained on InfoSeek and Enc-VQA does not guarantee good out-of-domain performance. In contrast, our models, trained on our dataset, achieve strong performance both in-domain and out-of-domain.

Model performance on more challenging synthetic test set Our human analysis in Section 4.4 demonstrates that our synthetic dataset includes questions answerable solely with the provided context. To increase the difficulty, we created a more challenging test subset (hard) and utilized the state-of-the-art language model LLAMA-3-70B-Instruct (AI@Meta, 2024) for filtering. We applied the LLAMA-3 model to the synthetic data test set used in Figure 5 and Table 7, offering only the context to answer the questions. We then filtered out questions that LLAMA-3 could answer using just the context, resulting in a subset of 2853 samples. The remaining questions predominantly require combined reasoning from both the image and the context to derive the correct answer.

As shown by the results in Table 8, there is a significant performance drop in 11 out of 12 cases on our hard subset, indicating that this subset indeed presents a greater reasoning challenge. This also demonstrates the potential future uses of our generated dataset.

Retrieval Augmented Generation Results In addition to generating gold passages as context, as seen in Table 4 and Table 7, we use the CLIP Score Fusion model from Wei et al. (2023) to retrieve knowledge from external text knowledge bases as context to simulate a real RAG setup. In a real RAG setup, the model faces more challenges as it needs to identify relevant parts from unrelated information, and sometimes the entire context may be irrelevant.

In this experiment, we focus on using the Paligemma-3B model. For constructing external knowledge bases, we use the InfoSeek dataset knowledge base processed by Wei et al. (2023), which includes 611,651 passages. For the other three datasets, Enc-VQA, InfoSeek, and ViQuAE, we create synthetic external knowledge bases by merging the corresponding gold passages for each test set question. The sizes of the knowledge bases for Enc-VQA, ViQuAE, and our synthetic dataset are 3,859, 71,985, and 1,514 passages, respectively.

For each question, we retrieve the top 10 most relevant passages. During inference, we combine each of these 10 retrieved passages with the question and perform inference. The final answer is determined by selecting the most frequently occurring answer among these 10 inferences. The specific model results are presented in Figure 6.

The results show that even when using retrieved contexts, the model trained on our dataset performs strongly both in-domain and out-of-domain. Notably, in out-of-domain performance, it surpasses the baseline zero-shot performance and models trained on the other two datasets in all 9 cases.

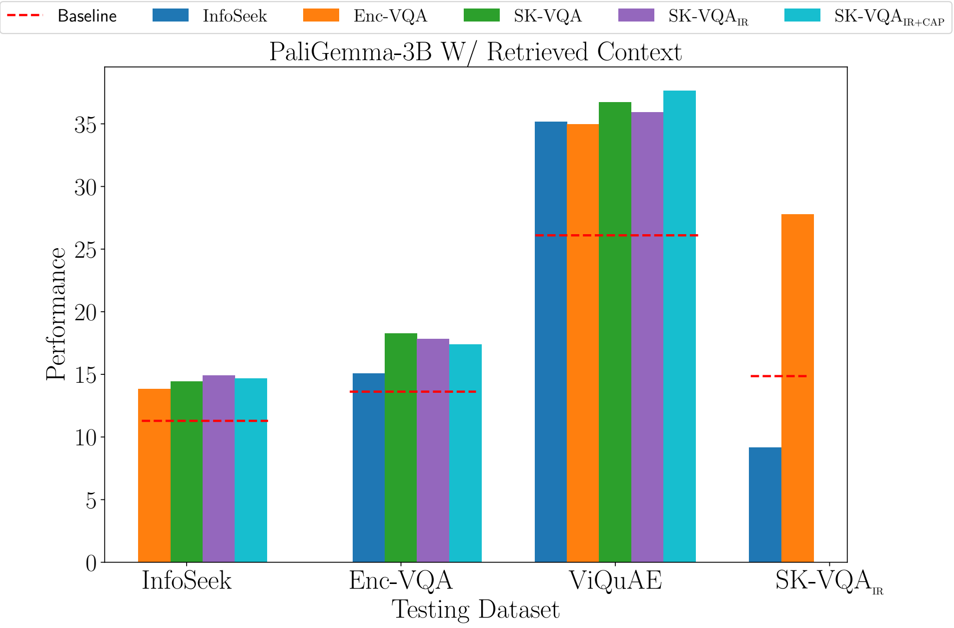


Figure 6: Generalization Performance of fine-tuned PaliGemma in RAG setup. Our models achieve the best generalization.

B LLM Evaluation

Previous studies have shown that a significant proportion of questions in OK-VQA and ViQuAE can be answered by an LLM when prompted with only the question (Chen et al., 2023). To investigate whether this is the case for our dataset, we generate answers from LLaMA-3-70b-Instruct for all 2 million QA pairs in our dataset using the following prompt:

```
Write a single word or phrase which
answers the question.
Question: [QUESTION]
```

where [QUESTION] is populated with questions from our dataset at query time. We found that the exact match accuracy of LLaMA-3-70b is only 9.92% for our dataset, indicating that the vast majority of questions do indeed require reaosning over the associated images and context documents.

C MLLMs Zero-shot Prompts

For all MLLMs in Table 4, we use the following text prompt when conducting zero-shot prompting, in addition to each model's specific image token:

```
Context {context} Based on the context,
{question} answer the question using
a single word or phrase.
```

D MLLMs Fine-tuning Hyperparameters

We use the official codebase[3] from LLaVA-1.5 to fine-tune the llava-v1.5-7b model[4] and the Trainer from Huggingface Transformers library[5] to fine-tune the paligemma-3b-mix-224 model [6]. For the llava-v1.5-7b model, we use a batch size of 16 and a learning rate of 2e-5, training the model for one epoch using bfloat16. Similarly, for the paligemma-3b-mix-224 model, we use a batch size of 64 and a learning rate of 2e-5, also training for one epoch using bfloat16. The inputs to the models are a combination of the question, image, and context, and the outputs are the answers to the questions.

E Additional details of GPT-4 generation

Prompts Figure 7 provides an explanation of the motivation for each condition which we include in our prompt. The numbered explanations correspond to the numbered conditions in our prompt which are shown in Figure 3. The instruction and conditions in our prompt were derived through manual prompt engineering, where different prompts were tested and iteratively up-

[3] https://github.com/haotian-liu/LLaVA

[4] https://huggingface.co/liuhaotian/llava-v1.5-7b

[5] https://github.com/huggingface/transformers

[6] https://huggingface.co/google/paligemma-3b-mix-224

```
Write a Wikipedia article related to this image
without directly referring to the image. Then
write question answer pairs. The question answer
 pairs should satisfy the following criteria.

1: Guide the model to generate questions using
image information.
2: Avoid questions that can be answered without
looking at the image.
3: Guide the model to generate questions using
external context rather than simple visual
information from the image.
4: Since GPT-4 tends to generate unnecessarily
lengthy questions that do not sound natural,
this condition helps to prevent such questions.
5: Guide the model to utilize the context and
also make answer evaluation straightforward.
6: GPT-4 tends to ask questions where the answer
 is an object in the image. For such questions,
context is not needed, which is not of our
interest.
7: We can split multiple correct answers into a
list to make the evaluation easier.
8: This condition is also for making the
evaluation easier.
```

Figure 7: Explanation for each prompt condition. The numbered explanations correspond to the numbered conditions in our prompt (Figure 3)

dated in response to issues that were identified in the synthetic data produced by GPT-4.

As discussed in Section 4.1, we also generated only QA pairs for a sub-sample of Wikipedia image-context pairs sourced from the WIT dataset. Figure 8 provides the prompt that we used with GPT-4 for this generation setting. In this prompt, `[CONTEXT]` is a placeholder where the actual Wikipedia context document is inserted at inference time. Other conditions in this prompt are identical to those in our main prompt specified in Figure 3.

Output parsing Here we describe the process of extracting context, question, and answer pairs from the text output generated by GPT-4. We first segment the entire output into two parts: the Wikipedia article and Question Answering pairs, identified by the line containing "question", "answer", and "pair". For both chunks, we remove extra symbols like hashes, stars, and consecutive spaces. In the Wikipedia article, we also remove the words "Wikipedia article" at the beginning. For the Question Answering pair chunk, we segment it by line and extract the question and answer by splitting each line using the symbol ":", retaining only the sentences after the ":".

```
Here is a Wikipedia article related to this
image:

[CONTEXT].

Write question answer pairs which require both
the image and the Wikipedia article. The question
answer pairs should satisfy the following
criteria.

1: The question should refer to the image.
2: The question should avoid mentioning the name
of the object in the image.
3: The question should be answered by reasoning
over the Wikipedia article.
4: The question should sound natural and concise
5: The answer should be extracted from the
Wikipedia article.
6: The answer should not be any objects in the
image.
7: The answer should be a single word or phrase
and list all correct answers separated by commas.
8: The answer should not contain 'and', 'or',
rather you can split them into multiple answers.
```

Figure 8: Prompt used to generate only QA pairs for an existing image-context pair using GPT-4.

API We accessed GPT-4 via the Azure OpenAI API and collected our entire dataset between the dates of May 24, 2024 and June 5, 2024. We used the `gpt-4o-2024-05-13` version of GPT-4 for all of our synthetic data generation.

F Details of compute infrastructure used in experiments

We utilized 24 Intel Gaudi2 AI Accelerators to obtain LLaMA-3-70b predictions for our dataset, which were used to create the 'hard' version of our test dataset (Appendix A) and perform LLM evaluation using only questions from our dataset (Appendix B).

For our zero-shot MLLM evaluation and MLLM training experiments, we used an internal linux slurm cluster with Nvidia RTX 3090, Nvidia A6000, and Nvidia A100 GPUs. We used up to 48 GPUs to parallelize various experiments on this cluster. Each parallelized worker was allocated 14 Intel(R) Xeon(R) Platinum 8280 CPUs, 124 GB of RAM, and 1 GPU. The total comptue time for job varied between 6-48 hours depending upon the model, dataset, and evaluation setting.

G Topic model details

We removed stop words from context and applied BERTopic (Grootendorst, 2022) to apply categorical TF-IDF on context embeddings created with all-MiniLM-L6-v2 sentence transformer model(Reimers and Gurevych, 2019). We initially reduced the number of topics to 40 using agglomerative clustering. Subsequently, we manually merged semantically related clusters, resulting in the following 25 topics, listed from most to least frequent: general, design, fashion, sports, wildlife, nature, coastal, events, culinary, architecture, transport, pets, structures, space, music, regions, gaming, science, politics, biology, military, postal, entertainment, economics, and religion. The "general" category could not be easily interpreted because it contained a mixture of many different unrelated topics; to improve visual clarity of the figure, it was therefore excluded, and categories representing less than 1% of the dataset were grouped under 'Other' category.

H License information

We abide by the licenses and intended uses of all models and datasets which were employed in this study. License information for models and datasets are provided below.

Models used in our study The PaliGemma-3B model is available under the Gemma license. The LLaVA-v1.5-7B is available under the Llama 2 Community License. The LLaVa-v1.6-7B, LLaVa-v1.6-34B, and idefics2-8b are available under the Apache License, Version 2.0. Qwen-VL-7B is available under the Tongyi Qianwen License.

Existing datasets used in our study The WIT dataset is available under the Creative Commons Attribution-ShareAlike 3.0 Unported license. The ViQuAE datset is available under the MIT license. The COCO-Counterfactuals dataset is available under the CC BY 4.0 license. The InfoSeek dataset is available under the Apache 2.0 license.

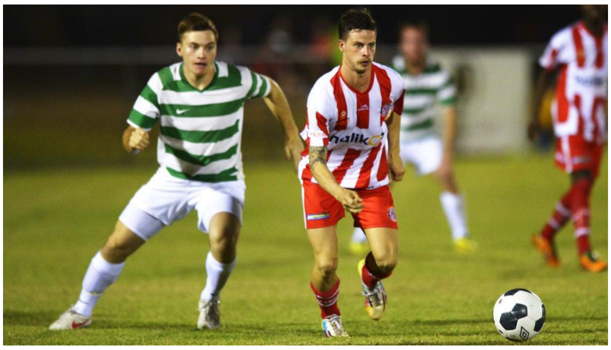
Our dataset We will make our dataset publicly available under the MIT license. In addition to the terms of this license, use of our dataset should abide by the OpenAI terms of use.

I Additional examples

Comparison of context documents sourced from GPT-4 and Wikipedia A subset of our dataset contains Wikipedia images for which we obtained context documents both from GPT-4 and from Wikipedia (via the image-text associations provided in the WIT dataset). Figures 9, 10, and 11 provide examples of the GPT-4 and Wikipedia-sourced context documents for identical Wikipedia images. The example in Figure 9 shows how context documents obtained from Wikipedia tend to have more entity-specific knowledge, whereas GPT-4 often generates more general knowledge which is related to what is depicted in the image. In Figure 10, the context document generated by GPT-4 is more specific to what is depicted in the image (a vineyard) than what is discussed in the Wikipedia context document (a specific type of wine). Finally, Figure 11 shows how the GPT-4 generated context documents can be longer and more detailed than the Wikipedia context documents which are linked to the image via the WIT dataset.

Examples of failure cases identified by human annotators Figure 12 provides examples of failure cases which were identified by human annotators. As described previously in Section 4.4, the three common categories of synthetic data generation failures identified by human annotators were cases where the question is answerable without looking at the image (Figure 12a), cases where the question is answerable without the context document (Figure 12b), and a single case where insufficient information was provided to answer the question (Figure 12c).

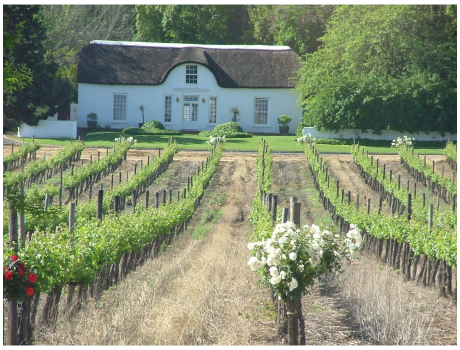
In all three cases depicted in Figure 12, the generated context documents are highly relevant to he image and also contain the answer. However, the question for for example depicted in Figure 12a does not reference the image, and could potentially be answered solely by uni-modal retrieval of the context document based on the question. The question associated with Figure 12b could potentially be answered solely through visual perception capabilities, although the associated context document would increase the chance of generating the correct answer. Finally, the example depicted in Figure 12c is for a question which references a "classical approach", which is not described in the associated context document.



Context source: GPT-4
Context: In the sport of association football, teams are distinguished by the unique colors and designs of their kits. Kits usually consist of a shirt, shorts, and socks, and teams have both a home and an away version. The home kit is typically in the team's primary colors, while the away kit is in a color that contrasts with their home kit to avoid clashes with other teams. Stripes are a common feature in football kit designs, with many teams opting for vertical or horizontal stripes in two or more colors. Red and white striped kits are popular among several football clubs around the world, a design often associated with clubs like Sunderland AFC and Atlético Madrid. Similarly, green and white horizontally striped kits are also traditional for teams such as Celtic FC. Kits often feature the team logo, the manufacturer's …

Context source: Wikipedia
Context: Alice Springs Celtic FC is an Australian soccer club based in Alice Springs, the Northern Territory. The club was one of the first football clubs to be founded in the FFNT Southern Zone Premier League. The club is part of the Football Federation Northern Territory. The club has seen great success in recent years, becoming Premiers in 2015 and Champions in 2016, the club has competed in the annual FFA Cup since 2015 where it reached the Seventh Qualifying Round, eventually losing 5–0 to Darwin Olympic. The club was knocked out in the Alice Springs final in the 2016 edition of the FFA Cup losing 1–0 to Gillen Scorpions and narrowly lost 2-0 in the 2017 edition of the Sports Ministers Cup to Darwin Rovers. At the beginning of the 2017 season prolific striker Joseph Worrall transferred to MPH Vikings ….

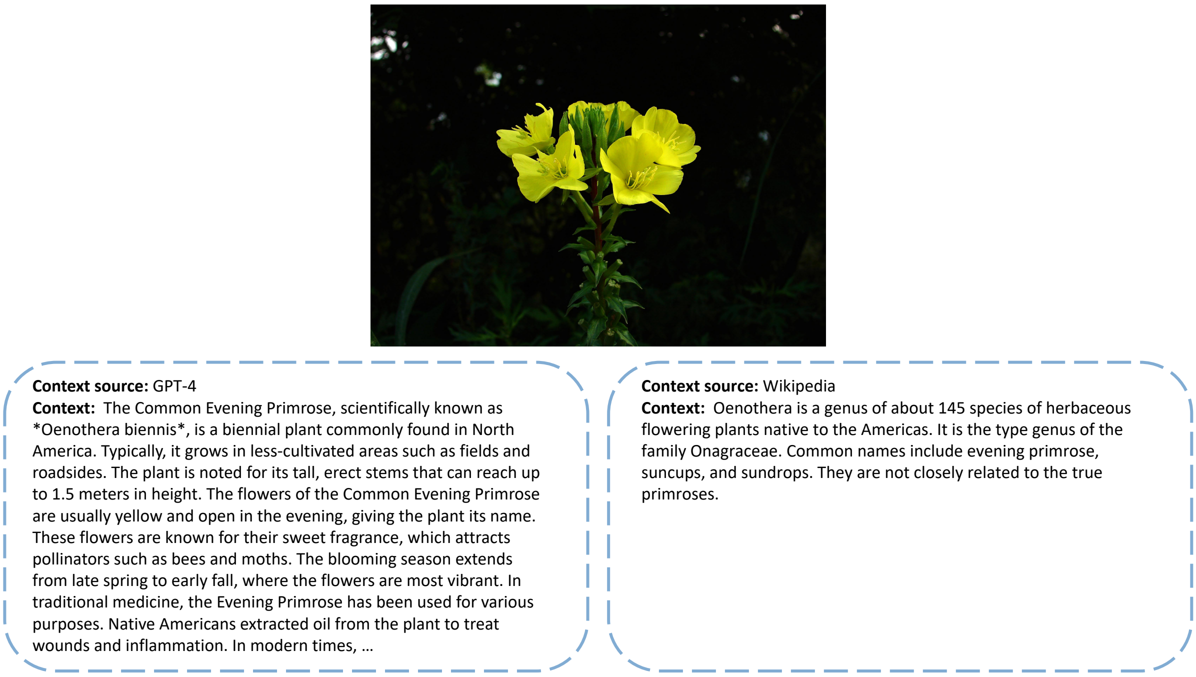
Figure 9: Comparison of context documents sourced from GPT-4 and Wikipedia for the same Wikipedia image. The context document from Wikipedia contains more entity-specific knowledge, whereas the GPT-4 context document contains more general knowledge about what is depicted in the image.



Context source: GPT-4
Context: A vineyard is an agricultural area where grapevines are grown for the purpose of wine production. The history of viticulture dates back to ancient times, with early evidence of grape cultivation and wine-making discovered in regions such as Mesopotamia, Egypt, and the Mediterranean. Vineyards are typically located on slopes with good sun exposure and well-draining soil, which are ideal conditions for growing grapes. Vineyards are often characterized by rows of trellised vines that maximize sunlight exposure and air circulation, benefiting the health and productivity of the grapevines. The rows of vines are usually planted with companion plants like roses at the end of each row, which serve as an early warning system for diseases that can affect the vines. Wine estates, also known as wineries, are ….

Context source: Wikipedia
Context: New World wines are those wines produced outside the traditional winegrowing areas of Europe and the Middle East, in particular from Argentina, Australia, Canada, Chile, Mexico, New Zealand, South Africa and the United States. The phrase connotes a distinction between these "New World" wines and those wines produced in "Old World" countries with a long-established history of wine production – most notably, France, Italy, Germany, Spain and Portugal. In 1595 the Spanish Crown banned the establishment of new vineyards in the Americas, but this order was largely ignored. The ban sought to protect Iberian wine from competition by Peruvian wine and can be considered an example of commodity mercantilism. Moreover, the Spanish Crown banned the export of Peruvian wine to Panamá ….

Figure 10: Comparison of context documents sourced from GPT-4 and Wikipedia for the same Wikipedia image. The context document sourced from Wikipedia contains knowledge that is less specific to the image than the GPT-4 generated context document.



Context source: GPT-4
Context: The Common Evening Primrose, scientifically known as *Oenothera biennis*, is a biennial plant commonly found in North America. Typically, it grows in less-cultivated areas such as fields and roadsides. The plant is noted for its tall, erect stems that can reach up to 1.5 meters in height. The flowers of the Common Evening Primrose are usually yellow and open in the evening, giving the plant its name. These flowers are known for their sweet fragrance, which attracts pollinators such as bees and moths. The blooming season extends from late spring to early fall, where the flowers are most vibrant. In traditional medicine, the Evening Primrose has been used for various purposes. Native Americans extracted oil from the plant to treat wounds and inflammation. In modern times, …

Context source: Wikipedia
Context: Oenothera is a genus of about 145 species of herbaceous flowering plants native to the Americas. It is the type genus of the family Onagraceae. Common names include evening primrose, suncups, and sundrops. They are not closely related to the true primroses.

Figure 11: Comparison of context documents sourced from GPT-4 and Wikipedia for the same Wikipedia image. The GPT-4 generated context document is significantly longer and more detailed than the context document sourced from Wikipedia (via the WIT dataset).

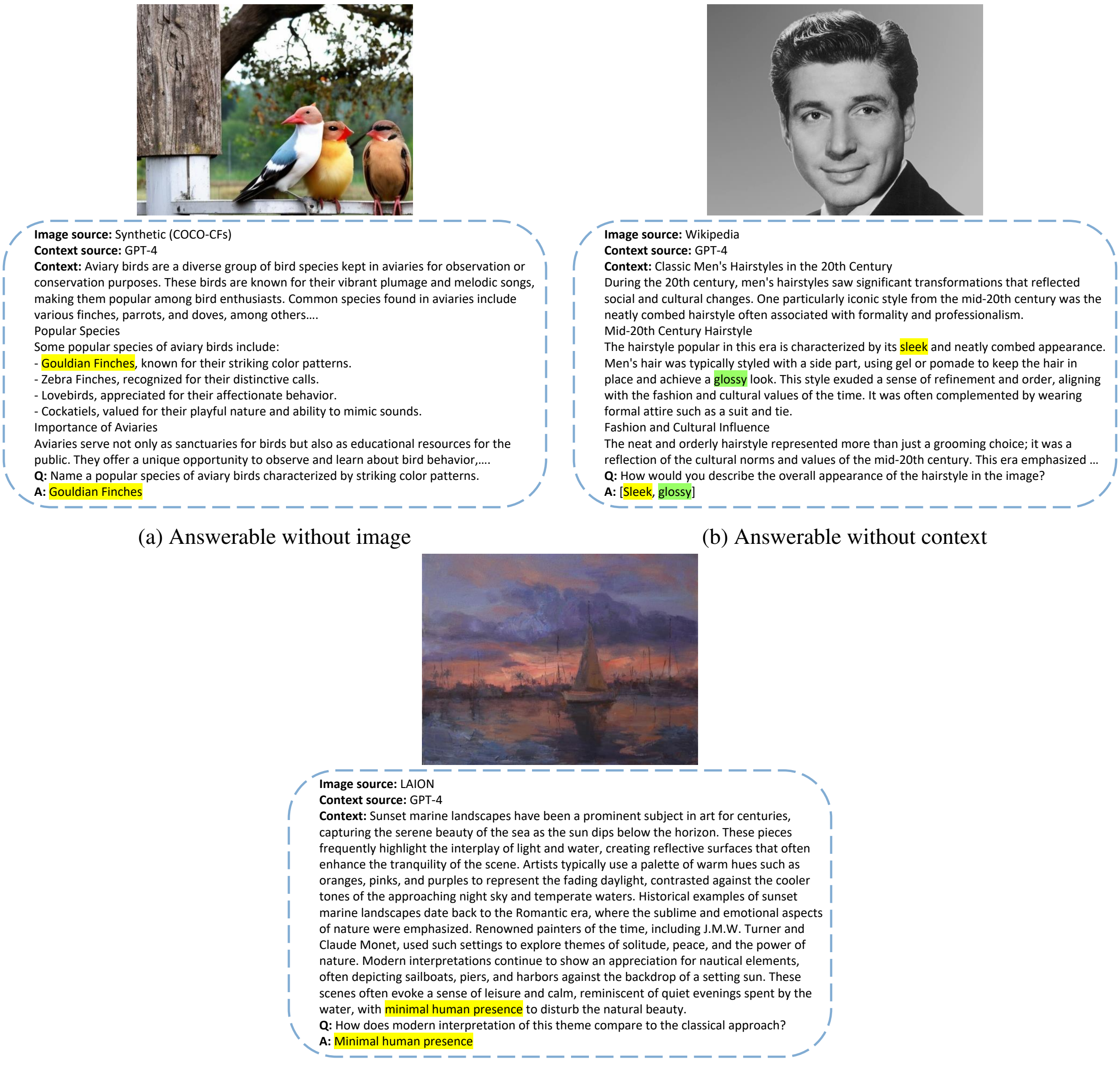


Image source: Synthetic (COCO-CFs)
Context source: GPT-4
Context: Aviary birds are a diverse group of bird species kept in aviaries for observation or conservation purposes. These birds are known for their vibrant plumage and melodic songs, making them popular among bird enthusiasts. Common species found in aviaries include various finches, parrots, and doves, among others....
Popular Species
Some popular species of aviary birds include:
- Gouldian Finches, known for their striking color patterns.
- Zebra Finches, recognized for their distinctive calls.
- Lovebirds, appreciated for their affectionate behavior.
- Cockatiels, valued for their playful nature and ability to mimic sounds.

Importance of Aviaries
Aviaries serve not only as sanctuaries for birds but also as educational resources for the public. They offer a unique opportunity to observe and learn about bird behavior,....
Q: Name a popular species of aviary birds characterized by striking color patterns.
A: Gouldian Finches

(a) Answerable without image

Image source: Wikipedia
Context source: GPT-4
Context: Classic Men's Hairstyles in the 20th Century
During the 20th century, men's hairstyles saw significant transformations that reflected social and cultural changes. One particularly iconic style from the mid-20th century was the neatly combed hairstyle often associated with formality and professionalism.
Mid-20th Century Hairstyle
The hairstyle popular in this era is characterized by its sleek and neatly combed appearance. Men's hair was typically styled with a side part, using gel or pomade to keep the hair in place and achieve a glossy look. This style exuded a sense of refinement and order, aligning with the fashion and cultural values of the time. It was often complemented by wearing formal attire such as a suit and tie.
Fashion and Cultural Influence
The neat and orderly hairstyle represented more than just a grooming choice; it was a reflection of the cultural norms and values of the mid-20th century. This era emphasized ...
Q: How would you describe the overall appearance of the hairstyle in the image?
A: [Sleek, glossy]

(b) Answerable without context

Image source: LAION
Context source: GPT-4
Context: Sunset marine landscapes have been a prominent subject in art for centuries, capturing the serene beauty of the sea as the sun dips below the horizon. These pieces frequently highlight the interplay of light and water, creating reflective surfaces that often enhance the tranquility of the scene. Artists typically use a palette of warm hues such as oranges, pinks, and purples to represent the fading daylight, contrasted against the cooler tones of the approaching night sky and temperate waters. Historical examples of sunset marine landscapes date back to the Romantic era, where the sublime and emotional aspects of nature were emphasized. Renowned painters of the time, including J.M.W. Turner and Claude Monet, used such settings to explore themes of solitude, peace, and the power of nature. Modern interpretations continue to show an appreciation for nautical elements, often depicting sailboats, piers, and harbors against the backdrop of a setting sun. These scenes often evoke a sense of leisure and calm, reminiscent of quiet evenings spent by the water, with minimal human presence to disturb the natural beauty.
Q: How does modern interpretation of this theme compare to the classical approach?
A: Minimal human presence

(c) Insufficient information to answer

Figure 12: Examples of synthetic data generation failures noted by human annotators.