# Analyze, Generate, Improve: Failure-Based Data Generation for Large Multimodal Models

Supplementary Material

## 1. Related Work

**Synthetic datasets for training LMMs.** Chen et al. [4] introduced ALLaVA, a 1.3M-sample dataset of real images with annotations and QA pairs from a frontier LMM, but it lacks failure-driven data generation and synthetic images. Li et al. [14] generate synthetic QA pairs for real chart images, focusing on chart VQA. Yang et al. [26] use code-guided generation (e.g., LaTeX, HTML) to create text-rich synthetic images. In contrast, our approach is broadly applicable across domains and leverages text-to-image diffusion models for greater image diversity.

**Synthetic data generation from model failures.** Prior work has explored leveraging model failures for synthetic data generation. Jain et al. [8] identify failure-related directions in a vision model's latent space to guide diffusion models in generating corrective images. Chegini and Feizi [3] use ChatGPT and CLIP to generate text prompts for diffusion models based on vision model failures. In language models, DISCERN [18] iteratively describes errors for synthetic data generation, while Lee et al. [13] uses incorrect answers from a student LLM finetuned on specific tasks as input to a teacher LLM which generates new examples to use for training. Unlike prior work, which generates single-modality data (text-only or image-only) and focuses on classification tasks, our approach generates multimodal image-text datasets aimed at training models for open-ended text generation.

**Generating synthetic data from frontier models to teach new skills.** AgentInstruct [19] is an agentic framework for generating synthetic data from a powerful frontier model (e.g., GPT-4) to teach new skills to a weaker LLM. Similarly, Ziegler et al. [27] utilize few-shot examples annotated by humans and retrieved documents with produce synthetic data from LLMs for teaching specialized tasks to models. Prompt-based methods for synthetic data generation from LLMs without seed documents [5, 22] as well as knowledge distillation from a teacher model [9] have also been proposed. Unlike our work, these prior studies focus on language-only data generation and use seed documents (e.g., raw text, source code) or prompts as a basis for data generation rather than an analysis of model failures.

## 2. Dataset generation

### 2.1. Compute Infrastructure

To generate our dataset, we queried GPT-4o through the Azure OpenAI API and deployed Qwen2-VL on Nvidia RTX A6000 GPUs. Using Intel® Gaudi 2 AI accelerators from the Intel® Tiber™ AI Cloud, we generated 1.024 million images from the VizWiz failed samples and 535k images derived from OK-VQA.

### 2.2. Dataset statistics

Our synthetic dataset is derived from the MFS of LLaVA-1.5-7B on four benchmark training sets: VizWiz [6], InfoVQA [17], ScienceQA [15], and OK-VQA [16], selected to cover diverse visual and reasoning challenges. VizWiz consists of real-world images captured by visually impaired users, often requiring detailed scene understanding. OK-VQA focuses on visual question answering which requires external knowledge. InfoVQA involves text-rich images where reading comprehension is crucial, assessing the model's ability to extract and interpret textual information from images. ScienceQA includes multimodal scientific reasoning questions which require both spatial and logical reasoning, making it valuable for evaluating complex reasoning capabilities. To generate the synthetic images, we utilized FLUX.1-schnell Labs [11] text-to-image model with a resolution of 1024×1024 pixels and guidance scale range of 3 to 13.

Table 1 provides statistics detailing the quantity of synthetic examples in our dataset which were derived from reasoning failures on different benchmarks. Additional discussion of the dataset composition is provided in Section **??**.

Our filtering approach successfully removes poor-quality samples, with the following removal rates across benchmarks: for VizWiz, 81% of synthetic-image samples and 34% of real-image samples were removed. This indicates that generating entirely synthetic samples is more challenging than generating synthetic text alone for real images. OK-VQA had a lower removal rate of 29% for synthetic images, possibly resulting from simpler and less ambiguous visual content. Among real-image-based samples, ScienceQA experienced a similar removal rate (29%), likely due to the complexity of spatial and scientific reasoning tasks. In contrast, InfoVQA exhibited a significantly lower removal rate of only 5% with an average filtering score of 2.9 (out of 3), indicating the strong capability of GPT-4o in handling text-based images.

| Dataset$_{\text{Image Type}}$ | Original | Failures | Filtered |
|---|---|---|---|
| VizWiz$_{\text{real}}$ | 20,523 | 7,785 | 100,280 |
| VizWiz$_{\text{syn}}$ | 20,523 | 7,785 | 190,172 |
| InfoVQA$_{\text{real}}$ | 10,074 | 5,250 | 95,783 |
| ScienceQA$_{\text{real}}$ | 5,585 | 1,562 | 39,090 |
| OK-VQA$_{\text{syn}}$ | 9,009 | 607 | 128,667 |

Table 1. Dataset statistics across benchmarks, including original training set size, number of failure samples (LLaVA-1.5-7b: 0, GPT-4o: 1), and synthetic samples with filtering score 3.

### 2.3. Data generation prompt

Figure 1 shows the prompt used to generate fully synthetic question-answer, image samples based on the failure modes of an LMM. To enhance data diversity, we use a variation of our prompt, expending step 4 to generate examples in different domains. Figure 5 compares fully synthetic samples with generated images, within similar and non-similar domain of the original failed sample. we notice that domain-similar samples preserve the original theme, while non-similar samples cover a more diverse contextual range to improve generalization. Additionally, we created samples where we both enforced and relaxed constraints on question format (e.g., multiple-choice, true/false) and instructions (e.g., requiring responses like "Unanswerable" when information was insufficient or limiting answers to short responses, see the Shiba Inu example from Figure 8).

### 2.4. Filtering prompt

Figure 2 provides the prompt which we used for the filtering stage of our synthetic data generation pipeline. See Section **??** of the main paper for additional filtering details.

## 3. Training hyperparameters

To train our model, we used 8 Nvidia RTX A6000 GPUs using the hyperparameters from Table 2. We employed DeepSpeed ZeRO stage 3 [1] for distributed training.

| | |
|---|---|
| Batch Size/GPU | 16 |
| Number of GPUs | 8 |
| Gradient Accumulation | 1 |
| Number of epochs | 1 |
| LLaVA Image Size | 576 |
| Optimizer | AdamW |
| Learning Rate | $2e-5$ |
| BF16 | $True$ |
| LR scheduler | $cosine$ |
| Vision Tower | openai/clip-vit-large-patch14-336 |
| Language Model | lmsys/vicuna-7b-v1.5 |

Table 2. Hyperparameters to train our model.

```
You are analyzing the performance of
a vision-language model (called Model
A). Model A's answer could deviate from
the ground truth due to limitations in
visual understanding, interpretation, or
reasoning.
Step 1: Describe the image.
Step 2: Given a question, the Ground
truth answer, and Model A's generated
answer, describe any key visual
elements that might influence Model A's
interpretation.
Step 3: Analyze the reasoning steps Model
A might have used to generate its answer,
considering both the visual and textual
information. Identify any weaknesses,
errors, or gaps in Model A response
compared to the ground truth.
Step 4: Suggest 10 additional challenging
detailed examples to address these
limitations.
Step 5: Transform each example into a
detailed prompt designed to generate
a clear and realistic image using a
text-to-image generation model.
```

Figure 1. Prompt used to generate fully synthetic image-text samples based on the failure modes of an LMM (Method 2).

```
Given sample containing an image, a
question, and an answer, your task is
to grade the sample from 1 to 3 based
on the following criteria:
Score 1: The answer is incorrect.
Score 2: The answer is correct, but
it is one of several possible valid
answers.
Score 3: The answer is correct,
specific, and the only valid answer.
The image provides all the necessary
context for the answer.
```

Figure 2. Filtering prompt

## 4. Additional Analysis

### 4.1. Human evaluation of dataset quality

Three of the authors of this work conducted a human evaluation by assessing three different aspects of our generated samples: (1) the alignment of the question and answer in relation to the image prompt, (2) the alignment between the image prompt and the generated image, and (3) the correctness of the answer given the question and image. The first evaluation reflects the quality of reasoning, the second evaluates the fidelity of the image generator's output, and the third combines both aspects. Scores range from 1 to

3, where 1 indicates an irrelevant alignment, 3 signifies a relevant alignment, and 2 represents a partially relevant or ambiguous alignment. We evaluated 200 samples in total, with 101 containing real images and 99 being fully synthetic. The overall correctness score for answers was 2.78, with real-image-based samples scoring 2.75 and fully synthetic samples scoring 2.81, indicating that fully synthetic samples achieve a level of fidelity equal to or even slightly exceeding that of real-image-based samples. For the synthetic samples specifically, we also measured the alignment between the image prompt and the generated image (2.66), and the alignment of the generated question and answer with the image prompt (2.84), indicating the high quality of reasoning in the generated responses.

### 4.2. Training data substitution vs. augmentation and impact of synthetically generated images

Our previous experiments augmented an existing 624k sample training dataset (LLaVA-Instruct) with our synthetic data. In domains where data is scarce, training datasets of this size may not be available. To investigate the utility of our synthetic data in such low-resource settings, we conducted experiments in which we randomly substituted different quantities of examples from the original dataset with our synthetically generated data[1]. The results of this experiment are provided in rows 2-3 of Table 3. Even when up to 25% of the original dataset is substituted with our synthetic data, we achieve performance that is either as good or better than the baseline LLaVA model across a broad range of downstream tasks. This is despite the fact that the original LLaVA training dataset utilizes real images, whereas our synthetic data used in this experiment contained only synthetically generated images. The fact that our synthetic data achieves similar or better performance than an existing real data source is significant, as prior studies have shown that training on synthetically generated image data is often much less efficient than training on an equivalent amount of real image data [7]. Table 3 also shows the impact of using real vs. synthetic images in our pipeline. Specifically, we compare the effectiveness of our synthetic data derived from Vizwiz reasoning failures when paired with real images (from Vizwiz) or synthetically generated images. In the training data augmentation setting, we observe that synthetic images generally achieve similar results as utilizing real images. Synthetic images even surpass the performance of real images in TextVQA, OK-VQA, and MMBench. This demonstrates the high quality of our synthetic images and their potential to serve as replacements for real images in low-resource settings where data is scarce.

[1]We used synthetic data derived from Vizwiz failures in this setting.

### 4.3. Impact of filtering on data quality

To investigate the impact of filtering on the quality of synthetically generated data, we repeated our in-domain evaluation experiments for ScienceQA and OK-VQA using raw unfiltered data. In the maximum synthetic data augmentation setting (last row of each section in Table **??**), using unfiltered data reduces EM from 73.0 to 72.2 on ScienceQA and from 63.3 to 58.8 on OK-VQA. This shows that our filtering approach improves model performance when using our synthetic examples for training data augmentation. Furthermore, using only synthetic examples which were assigned the lowest rating in our filtering process decreases the EM score on OK-VQA to 57.5, which highlights the difference in quality between the lowest-scoring and highest-scoring synthetic examples identified during filtering.

### 4.4. Comparison of LLM synthetic data generators

We compared two frontier LMMs, GPT-4o and Qwen2-VL-7B [24], for generating synthetic data grounded in LLaVA-7B failures. Qwen2-VL-7B was selected due to its high accuracy on vision-language benchmarks. Our results show that using samples generated by Qwen2-VL leads to reduced downstream performance compared to those produced by GPT-4o, with a decrease of 2% on InfoVQA and 6.5% on OK-VQA. Additionally, samples generated by Qwen2-VL received lower filtering scores: 1.9 (Qwen2-VL) vs. 2.5 (GPT-4o) for OK-VQA, and 2.6 (Qwen2-VL) vs. 2.9 (GPT-4o) for InfoVQA. Based on our manual analysis, we hypothesiize that these differences may result from the detailed and precise reasoning provided by GPT-4o, resulting in synthetic samples that are better tailored to address identified reasoning failures. In contrast, samples generated by Qwen2-VL-7B sometimes demonstrate lower diversity, which could limit their effectiveness in addressing the broad range of failure modes. Figure 4 provides an example of these observed differences in the reasoning processes of GPT-4o and Qwen2-VL-7B models, as well as the corresponding generated fully synthetic samples.

### 4.5. Correcting specific types of reasoning failures

Our synthetic data generation approach explicitly identifies different types of LMM reasoning failures. To systematically categorize these failures, we encoded each reasoning explanation using sentence transformers [21] and clustered them using k-means. Figure 3 presents the resulting clusters, highlighting prevalent failure modes such as optical character recognition (OCR) and object detection errors. Based on this analysis, we further investigated whether targeted synthetic data can effectively address these specific failure cases and enhance LLAVA's reasoning capabilities.

Specifically, we augmented LLaVA-Instruct with 10,579 synthetic samples from our $\text{VizWiz}_{\text{syn}}$-MFS addressing object detection reasoning failures and repeated the second

| Train Data | $N$ | $N_{syn}$ | TextVQA | OCR-Bench | InfoVQA | OK-VQA | ScienceQA | MMBench | MMMU |
|---|---|---|---|---|---|---|---|---|---|
| Baseline | 624,610 | 0 | 47.0 | 31.9 | 26.7 | 57.0 | 70.7 | 52.3 | 36.4 |
| Substitute w/ | 624,610 | 62,461 | 47.1 | 31.6 | 26.5 | 56.9 | 70.8 | 52.3 | 35.3 |
| syn images | 624,610 | 156153 | 46.9 | 31.1 | 27.0 | 57.0 | 70.6 | 51.2 | **37.9** |
| Augment w/ | 645,222 | 20,612 | 46.7 | 32.0 | 27.0 | 57.4 | **71.2** | **53.4** | 34.9 |
| syn images | 687,071 | 62,461 | **47.7** | 31.8 | 25.8 | **59.4** | **71.2** | 52.3 | 36.4 |
| Augment w/ | 645,222 | 20,612 | 46.9 | **32.5** | **27.2** | 57.4 | 70.6 | 53.1 | 35.0 |
| real images | 687,071 | 62,461 | 47.2 | 32.2 | **27.2** | 56.9 | **71.2** | 52.3 | 33.8 |

Table 3. Ablation experiments comparing baseline LLaVA to LLaVA models trained with synthetic data generated from VizWiz failures. We investigate substitution and augmentation strategies for synthetic data, as well as the use of synthetic vs. real images.

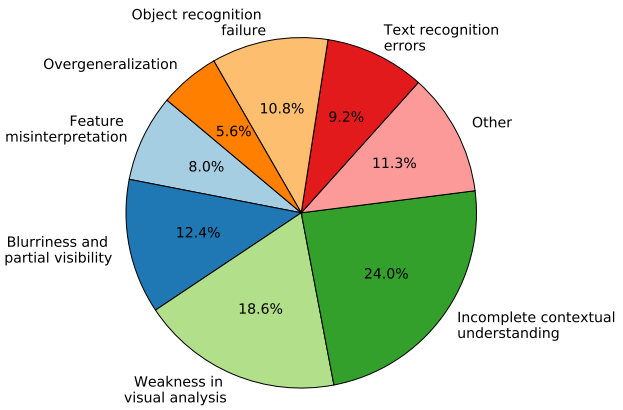



Figure 3. Figure shows the clusters of LLAVA reasoning failures described by GPT-4o.

| Dataset | LLAVA | $\text{LLaVA}_{\text{syn}}$ |
|---|---|---|
| CIFAR-10 [10] | **82.1** | 81.2 |
| Food-101 [2] | **13.4** | 13.2 |
| iNaturalist [23] | 20.6 | **52.0** |
| MNIST [12] | 75.1 | **80.5** |
| F-MNIST [25] | 9.8 | **10.0** |
| Oxford-pets [20] | 39.6 | **96.4** |

Table 4. Image classification accuracy of LLaVA and a $\text{LLaVA}_{\text{syn}}$ model augmented only with synthetic examples corresponding to object recognition failures.

stage of LLaVA finetuning. The model was then evaluated on CIFAR-10 [10], Food-101 [2], iNaturalist [23], MNIST [12], Fashion-MNIST [25], and Oxford-Pets (Binary) [20] by formatting samples as multiple-choice questions. Table 4 presents a comparison of LLaVA and $\text{LLaVA}_{\text{syn}}$. The results show that $\text{LLaVA}_{\text{syn}}$ surpasses LLaVA on four out of six datasets, with particularly notable improvements on iNaturalist, MNIST and Oxford-Pets. This demonstrates the significant impact of our synthetic dataset in addressing specific reasoning failures within LLAVA. By systematically incorporating targeted synthetic samples, we can mitigate common failure cases, leading to measurable performance improvements across multiple benchmarks. Our findings highlight the effectiveness of leveraging targeted synthetic data to refine model reasoning and suggest that incorporating such data-driven interventions can significantly enhance the robustness and generalization of LMMs.

## 5. Detailed OOD results for models fit to different subsets of synthetically generated data

Table 5 provides additional evaluation results for models trained individually on real and synthetic data derived from Vizwiz, InfoVQA, ScienceQA, and OK-VQA. All reported values are the official evaluation metrics corresponding to each dataset. The first two rows of each section in Table 5 provide a direct comparison of the efficiency of our synthetic data to real data; we observe that augmenting the LLaVA-Instruct dataset with our synthetic data achieves as good or better performance across most settings as augmenting with real domain-specific data. Furthermore, significant performance gains are achieved relative to the LLaVA baseline when our synthetic data is derived from a dataset in the same domain as the benchmark. For example, synthetic data generated from reasoning failures on InfoVQA significantly improve LLaVA's performance on tasks which require fine-grained text understanding such as OCR-Bench and InfoVQA.

## 6. Examples from our dataset

In this section, we present examples from our dataset and highlight its weaknesses and limitations. Figure 5 shows a comparison of fully synthetic similar vs non-similar samples. Figures 6 show sampled examples from VizWiz and InfoVQA highlighting the diversity of question types and demonstrating the overall quality of generated images and text. Figure 8 shows our synthetic data generated from the OK-VQA dataset, while Figure 9 corresponds to the VizWiz

| Train Dataset | $N$ | $N_{syn}$ | TextVQA | OCR-Bench | InfoVQA | OK-VQA | ScienceQA | MMBench | MMMU |
|---|---|---|---|---|---|---|---|---|---|
| Baseline | 624,610 | 0 | 0.47 | 0.32 | 0.27 | 0.57 | 0.71 | 52.30 | 0.36 |
| Vizwiz | 645,133 | 0 | 0.47 | 0.28 | 0.26 | **0.59** | **0.71** | 51.74 | **0.38** |
| | 687,071 | 62,461 | **0.48** | **0.32** | 0.26 | **0.59** | **0.71** | 52.25 | 0.36 |
| | 749,532 | 124,922 | 0.47 | **0.32** | **0.27** | **0.59** | 0.70 | **53.02** | 0.37 |
| InfoVQA | 634,684 | 0 | 0.47 | 0.32 | 0.32 | 0.58 | 0.70 | 52.50 | 0.37 |
| | 634,684 | 10,074 | 0.47 | 0.33 | 0.31 | **0.59** | 0.70 | 52.16 | 0.36 |
| | 687,071 | 62,461 | 0.47 | **0.34** | 0.33 | 0.57 | **0.71** | **52.69** | 0.37 |
| | 710,610 | 86,000 | **0.48** | 0.33 | **0.34** | 0.56 | **0.71** | 52.53 | **0.38** |
| ScienceQA | 630,195 | 0 | 0.47 | 0.29 | 0.26 | **0.58** | 0.70 | 52.88 | 0.36 |
| | 630,195 | 5,585 | 0.47 | **0.32** | **0.27** | 0.57 | 0.72 | **53.19** | 0.37 |
| | 646,594 | 21,984 | 0.47 | **0.32** | 0.26 | 0.56 | **0.73** | 53.12 | **0.38** |
| OK-VQA | 633,619 | 0 | **0.47** | 0.30 | 0.27 | 0.54 | **0.71** | **53.35** | **0.36** |
| | 633,619 | 9,009 | **0.47** | **0.33** | **0.28** | **0.61** | **0.71** | 52.68 | 0.35 |
| | 687,071 | 62,461 | **0.47** | **0.33** | 0.27 | **0.61** | **0.71** | 51.96 | 0.35 |

Table 5. Training data augmentation experimental results. $N$ denotes the total number of examples used for training, while $N_{syn}$ denotes the number of synthetic examples in the training dataset which were generated using our approach.


Figure 4. Comparison of GPT-4o and Qwen2-VL for generating Failure-Grounded Synthetic Datasets: GPT-4o demonstrates stronger reasoning capabilities, identifying multiple reasoning failures such as missed contextual cues and a lack of correlation between visual elements and the question. In contrast, while Qwen2-VL correctly answering the original question, identifies fewer failure modes and is less accurate in diagnosing LLaVA's reasoning failures, sometimes focusing on less relevant aspects, such as the kite in the sky. As a result, Qwen2-VL's generated samples are less diverse, often repeating the same question, whereas GPT-4o's samples provide broader coverage of identified reasoning failures. Note: GPT-4o's reasoning is 2–3 times longer than Qwen2-VL's; only a portion of GPT-4o's reasoning is shown here, while Qwen2-VL's reasoning is presented in full.

dataset. These are fully synthetic examples, including generated image, along the question and answer. Figure 7 provides additional fully synthetic text & image examples derived from VizWiz and OK-VQA.

Figure 10, 11 and 12 illustrate examples derived from the ScienceQA, InfoVQA and VizWiz benchmarks respectively, where the images are real but the questions and answers are synthetically generated.

Lastly, Figures 13 and 14 show some incorrect examples for each benchmark.


Figure 5. Comparison of fully synthetic similar and non-similar samples. Similar samples maintain a children's characters-based theme like the original sample, while non-similar samples address the failure modes by introducing diverse contexts.

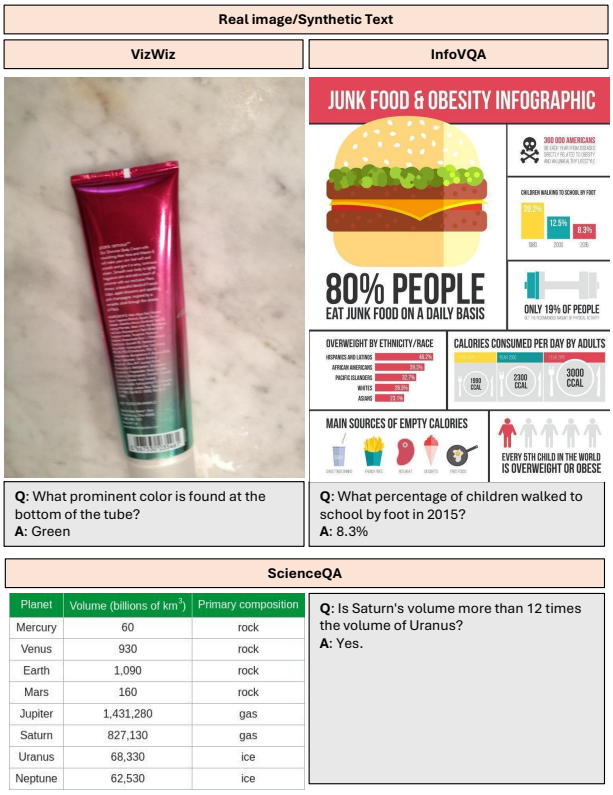



Figure 6. Examples of generated synthetic question-answer pairs for real images from VizWiz, InfoVQA, and ScienceQA.

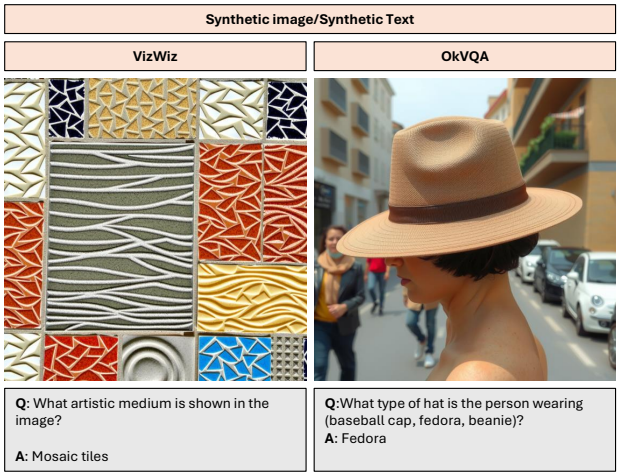
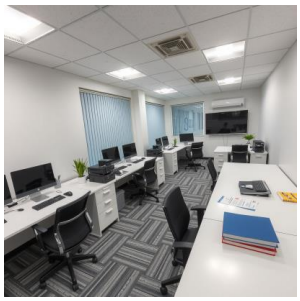



Figure 7. Examples of fully synthetic samples, using Method 2 as described in **??**, both question-answer pairs and images were generated.

**OkVQA– Synthetic Image/Synthetic Text**

**Q**: What equipment is the person using to catch fish?
**A**: Fishing rod
**Prompt:** A person standing on a boat under a clear sky, casting a fishing rod into the water, with fishing gear and a cooler in the background

**Q**: Which breed is generally smaller and has a more fox-like appearance? When the provided information is insufficient, respond with 'Unanswerable'. Answer the question using a single word or phrase?
**A**: A Shiba Inu
**Prompt:** A Shiba Inu with a smaller, fox-like appearance and a curled tail, walking in a Japanese garden.

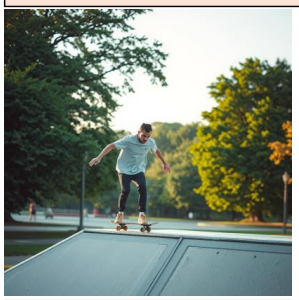


**Q**: Is it more likely to find a coffee mug or a pillow in this room?
**A**: Coffee mug
**Prompt:** A modern office space with desks, chairs, computers, and office supplies, arranged in a professional and organized manner, with some papers and stationery items on the desks.

**Q**: What type of clothing is this?
**A**: Kimono
**Prompt:** A traditional Japanese kimono displayed on a mannequin, with intricate designs and vibrant colors. The background is minimal, with a soft-focus effect. The kimono's patterns and texture are detailed and realistic

Figure 8. Examples of generated samples from OK-VQA with synthetic images and synthetic text.

**VizWiz – Synthetic Image/Synthetic Text**



**Q**: What sport is this individual engaged in?
**A**: Skateboarding
**Prompt:** A person skateboarding in a park, mid-action on a skateboard ramp. The background includes trees, a clear sky, and other park elements. The skateboarder is wearing casual streetwear and protective gear

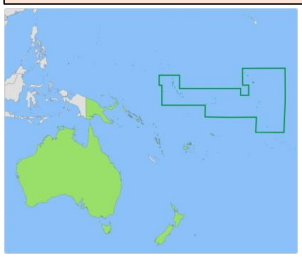
**Q**: What musical instrument is this?
**A**: Violin
**Prompt:** A violin lying in its case with a bow next to it. The background is a wooden surface. The image is detailed, showing the strings, tuning pegs, and fine tuners of the violin.

**Q**: What type of animal skin is shown in the image?
**A**: Reptile scales
**Prompt:** A close-up image of reptile scales, showing their overlapping structure and textured surface.

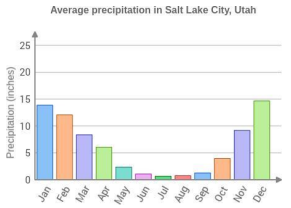
**Q**: What type of clothing is this?
**A**: Kimono
**Prompt:** A traditional Japanese kimono displayed on a mannequin, with intricate designs and vibrant colors. The background is minimal, with a soft-focus effect. The kimono's patterns and texture are detailed and realistic

Figure 9. Examples of generated samples from VizWiz with synthetic images and synthetic text.

**ScienceQA– Real Image/Synthetic Text**



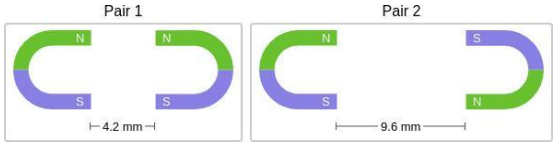
**Q**: Which country's maritime boundary is outlined in green?
**A**: Kiribati




**Q**: Which month has the highest average precipitation in Salt Lake City?
**A**: December

| Aquarium | Initial temperature (°C) | Final temperature (°C) |
|---|---|---|
| Covered aquarium | 21.7 | 16.2 |
| Uncovered aquarium | 21.7 | 16.2 |

**Q**: Based on the temperature data, did the aquariums lose or gain thermal energy?
**A**: Lose thermal energy

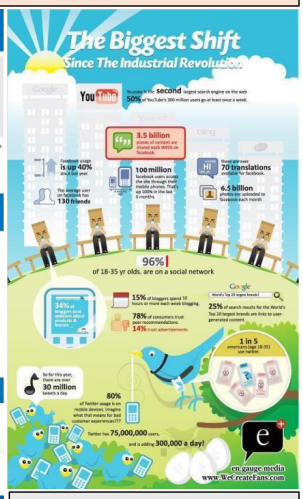



**Q**: In Pair 1, what color represents the North pole?
**A**: Green

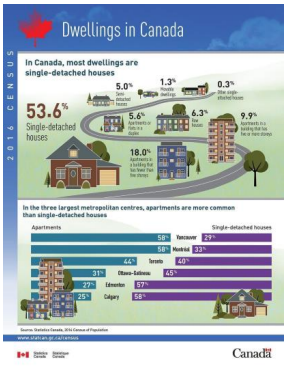
Figure 10. Examples of generated samples from ScienceQA with real images and synthetic text.

**InfoVQA– Real Image/Synthetic Text**


**Q**: What is the recommended distance to maintain from others after an earthquake?
**A**: 1 meter




**Q**: What percentage of bloggers spend 10 hours or more each week blogging?
**A**: 15%




**Q**: What percentage of Canadian dwellings are row houses?
**A**: 6.3%


**Q**: What fraction of the grain grown in the US is used to feed farmed animals?
**A**: More than 70%

Figure 11. Examples of generated samples from InfoVQA with real images and synthetic text.

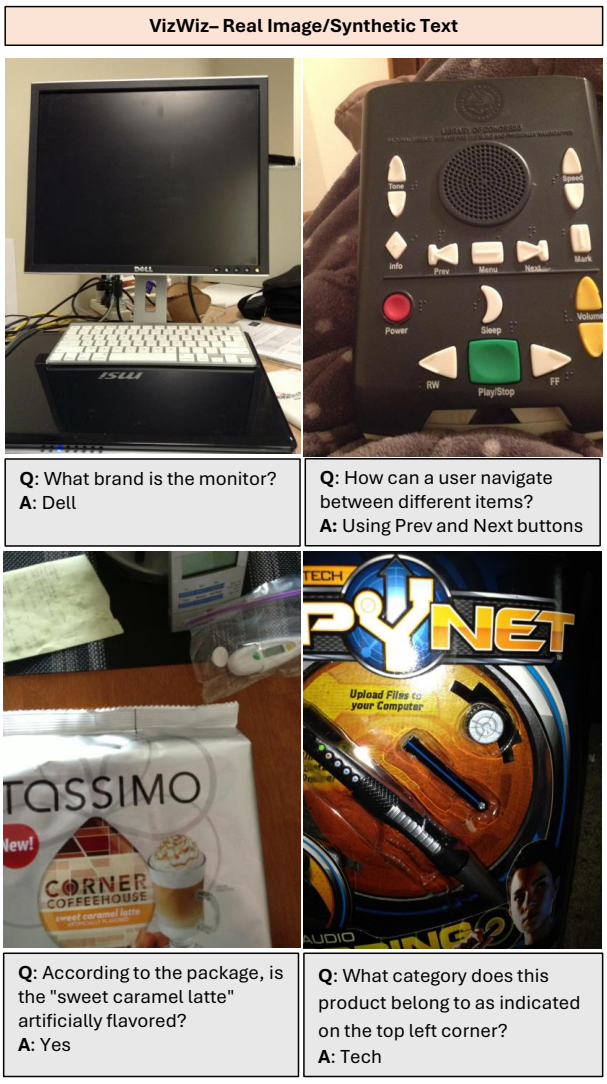



Figure 12. Examples of generated samples from VizWiz with real images and synthetic text.

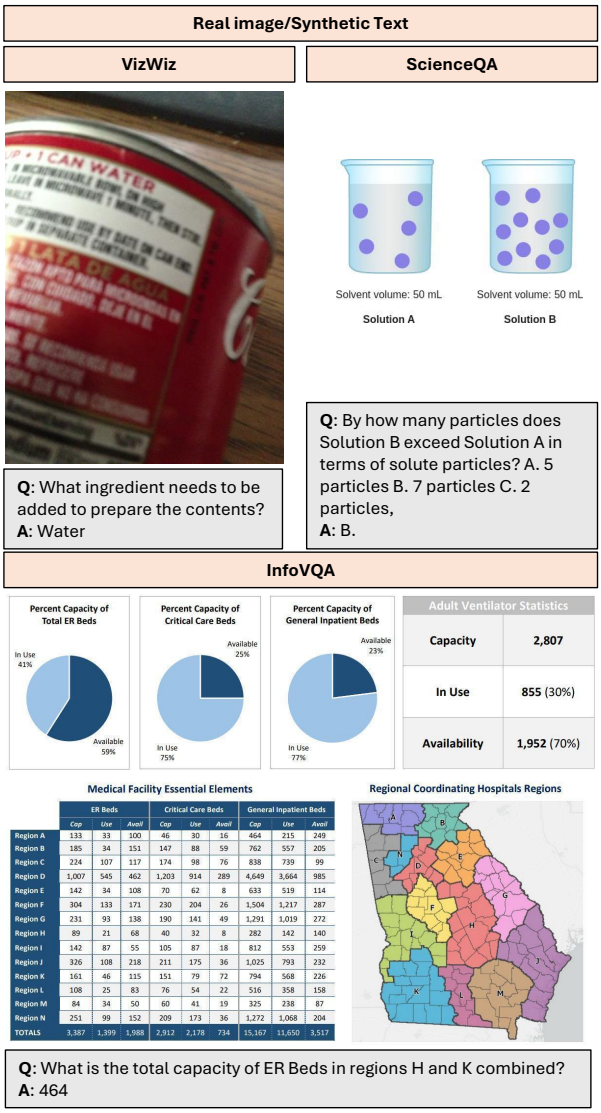



Figure 13. Examples from our dataset real image-synthetic text, where the sample is ambiguous or incorrect.

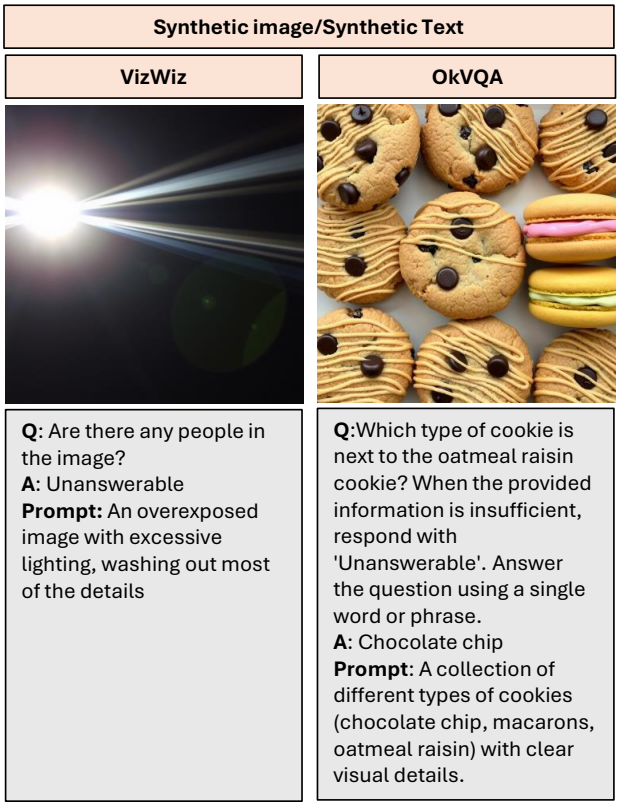



Figure 14. Examples from our dataset synthetic image-synthetic text, where the sample is ambiguous or incorrect. For readability, the ScienceQA image was cropped to focus on the region of interest.