ATENEO DE MANILA UNIVERSITY

CAUSAL INTERVENTIONS FOR ROBUST VISUAL QUESTION ANSWERING

A THESIS SUBMITTED TO
THE GRADUATE FACULTY OF
THE SCHOOL OF SCIENCE AND ENGINEERING
IN CANDIDACY FOR THE DEGREE OF
MASTER OF SCIENCE IN
COMPUTER SCIENCE

DEPARTMENT OF INFORMATION SYSTEMS
AND COMPUTER SCIENCE

BY
RYAN CAESAR C. RAMOS

QUEZON CITY, PHILIPPINES
JUNE 2023

The THESIS entitled:

CAUSAL INTERVENTIONS FOR ROBUST VISUAL QUESTION ANSWERING

submitted by Ryan Caesar C. Ramos has been examined and is recommended for **Oral Defense**.

PATRICIA ANGELA R. ABU, Ph.D.
Chair

RAPHAEL B. ALAMPAY, Ph.D.
Adviser

PATRICIA ANGELA R. ABU, Ph.D.
Co-Adviser

RAPHAEL A. GUERRERO, Ph.D.
Dean
School of Science and Engineering

The Faculty of the Department of Information Systems and Computer Science, School of Science and Engineering, Ateneo de Manila University ACCEPTS THE THESIS entitled:

CAUSAL INTERVENTIONS FOR ROBUST VISUAL QUESTION ANSWERING

submitted by Ryan Caesar C. Ramos in partial fulfillment of the requirements for the degree of Master of Science in Computer Science.

RAPHAEL B. ALAMPAY, Ph.D.
Member

JOHN PAUL C. VERGARA, Ph.D.
Member

JANN RAILEY E. MONTALAN, M.Sc.
Member

RAPHAEL B. ALAMPAY, Ph.D.
Adviser

PATRICIA ANGELA R. ABU, Ph.D.
Co-Adviser

RAPHAEL A. GUERRERO, Ph.D.
Dean
School of Science and Engineering

Grade: A-
Date: JUNE 29, 2023

ABSTRACT

Contemporary visual question answering (VQA) models have been shown to exhibit poor out-of-distribution (OOD) generalization ability due to their tendency to learn superficial statistical correlations from training data as opposed to more reliable underlying causal features. This can be addressed by widening the training distribution through data augmentation, but though recent advances have been made in generative modelling and training large foundation models, the application of these methods for data augmentation targeting robust VQA remains underexplored. This study proposes a novel approach to ensembling foundation models in order to generate OOD datapoints to widen the distribution of a training dataset. In particular, this study proposes a novel token sampling method to perturb existing image captions into OOD captions, which can then be used to steer a pretrained text-to-image model. The resulting images along with the original questions and answers can then be used to finetune a VQA model that has only been trained on the original training dataset. This method is empirically shown to lead to robustness improvements; with a BLIP pretrained on VQA v2.0, finetuning with the study's generated data introduces a 7.59% accuracy drop reduction on AQUA and a 1.43% accuracy drop reduction on VizWiz.

ACKNOWLEDGMENTS

This study would like to thank the following for their significant contributions in this study:

Dr. Raphael B. Alampay and Dr. Patricia Angela R. Abu, the advisors of this study, for their immense patience, encouragement, guidance, and support through every step of this conducted study;

Dr. John Paul C. Vergara and Mr. Jann Railey E. Montalan, the panelists of this study, for their patience and kindness, and for their detailed insight in elevating the quality of this study's contributions;

Ateneo de Manila University, for financially supporting the study's researchers through the generosity of their provided scholarship;

The Department of Information Systems and Computer Science, especially Ms. Joannie J. Ereño, for their assistance in the logistics of completing this study;

and Patrick John C. Ramos, for his nearly incomparable technical and moral support in helping this study come to fruition.

TABLE OF CONTENTS

CHAPTER

LIST OF FIGURES

Figures Page

LIST OF TABLES

Tables Page

CHAPTER I

INTRODUCTION

Visual question answering (VQA) is the task of providing a natural language answer to an open-ended question about a given image [6]. For example, provided an image of a cow, a model should be able to answer questions like ”What animal is this?”. Training models capable of VQA creates machines that can see and reason about the world around them. Such machines have potential applications for visually-impaired people, surveillance authorities, virtual or robot assistants, etc [4].

State-of-the-art performance for this task is led by deep learning models, particularly large foundation models finetuned on the task [5, 35, 39, 41, 42]. Common datasets for training and testing include VQA v2.0 [14], GQA [17], VG [21], and VizWiz [15]. With the exception of zero-shot models, the standard practice for training a VQA model is to train a model on the train split of a dataset and validate on the validation split of the same dataset.

1.1 Statement of the Problem

Though progress is continuously made on standardized datasets, many train and test sets are sampled from similar environments. Thus, recent progress only measures the ability of a model to perform on independent-and-identically distributed (IID) data, or data sampled from a similar distribution to the train set. When used to make inferences on out-of-distribution (OOD) data, or data outside the training distribution, VQA models exhibit poor generalization ability [3]. For example, if the training distribution mostly includes cows in fields, a VQA model may fail to answer questions about a cow on a beach. This significantly affects the practicability of deploying VQA models in real-life scenarios where environments may not be as constrained as training environments.

This weakness can be attributed to the tendency to train deep learning models, and in extension VQA models, as statistical models, which causes them to rely on input features statistically correlated with target outputs (e.g. the presence of grass) regardless of whether or not these correlations would persist under distribution changes (e.g. moving a cow to a beach) [10]. In response to this shortcoming of statistical learning, a paradigm shift to causal learning has been proposed [33]. In contrast to a statistical learning approach, a causal learning approach would uncover the underlying causal structures of the phenomena being modelled (e.g. the features of a cow cause the presence of the cow). A resulting causal model would thus be robust to distribution shifts, as it relies on consistent causal features as opposed to spurious correlations or confounders.

For this reason, causal learning principles have already been applied to VQA, with one of the most common methods of doing so being counterfactual or interventional image augmentation [1, 8, 36]. Augmentations (e.g. flipping, saturation changes) have been shown to simulate interventions on environments, effectively inducing distribution shifts to expose underlying causal mechanisms [18, 38]. Augmentation has been shown to be a process that can be automated without the need for extensive manual labor and labelling [1]. However, these methods only simulate a minimal number of interventions per sample, limiting the strength of the distribution shifts that could be used in training.

1.2 Research Objectives

This study aims to extend the application of causal learning principles to VQA. Specifically, this study proposes a novel method for widening the distributions represented by VQA training data through pretrained language and generative image models, as well the pre-existing data and metadata. To evaluate the usefulness of training with this augmentation method, this study plans to follow Agrawal et al. [3] and train vision-language models on the proposed dataset, and evaluate the trained models on the test sets of separate datasets to gauge the model's ability to maintain performance in new domains.

1.3 Research Questions

1. How can data augmentation methods be used to intervene on a VQA dataset and widen its distribution?
2. How does widening the range of artificially-induced distribution shifts in a train set improve generalization in the OOD setting?

1.4 Scope and Limitations

This study only plans to augment images from a single dataset as opposed to mixing datasets in order to more effectively gauge out-of-distribution performance; other datasets will be preserved for generalization probing. The expressiveness of this study's augmentations will be limited by the available text descriptions for each image; and the quality of available pretrained open-source text-to-image synthesis models and language models.

1.5 Significance of Study

VQA models have the potential to aid visually-impaired people become more independent and mobile, assist in educational or cultural preservation contexts, and extend the capabilities of robot or virtual assistants to the visual domain [4]. However all of these practical real-life settings are dependent on the ability of VQA models to be robust and invariant to distributional shifts and environmental changes, as these applications are not guaranteed to follow the same constraints and conditions of the data the model was trained on. Hence it becomes important to not only focus on increasing the performance of VQA models on standard benchmarks, but to also focus on improving their ability to generalize well outside the training domain.

CHAPTER II

REVIEW OF RELATED LITERATURE

This section provides a review of related works and relevant topics, including previous techniques for robust VQA and an overview of causal representation learning and its applications to robust VQA.

2.1 Robust VQA

In parallel to the standard VQA task, an active area of research is the training of specifically robust VQA models. Standard VQA models have been shown to generalize poorly to new domains as a result of an exploitation of spurious statistical correlations in the training data [2, 3]. In response, various methods have been proposed to train models that do not rely on dataset biases but can instead adapt to new domains and distribution shifts.

These methods usually target specific types of superficial correlations. To counter the effect of answer priors in the training data (e.g. a model may learn to answer “tennis” to the question ”What sport is being played?” regardless of the input image if a majority of the training samples were answerable by guessing ”tennis”) [2, 3], proposed methods include explicitly training the model to base its answer on specific regions of the input image [2, 40], ensemble-based methods [16, 30], and the collection of more diverse training data to decrease the presence of the answer priors [14]. To address overfitting to the linguistic formatting of questions (e.g. a model may be able to answer ”Is it safe to turn left?” but fail when asked the same question in a novel rephrasing such as ”Can one safely turn left?”) [3, 34], Shah et al.

[34] propose training with a cycle consistency-based task wherein a model is trained to generate a question from its inferred answer, which must be answered to match its original inferred answer. To address brittleness to semantic variations in visual input (e.g. a model may correctly answer the question ”What color is the keyboard?” but fail when an irrelevant area of the mouse is removed), Agarwal et al. [1] increase their training data by performing image augmentations that affect the answer in a predictable manner, creating new image-question-answer triples. This last category of spurious correlations is of particular interest to the present study.

2.2 Causal Representation Learning for VQA

A subset of these techniques fall under the framework of causal representation learning. Causal representation learning has been proposed as a possible framework to address poor generalization issues that arise from the use of statistical models, as well as explain the effectiveness of current methods for increasing model performance on OOD data. Causal representation learning focuses on building models that do not rely simply on statistical observations in training data but mine the underlying causal structure of the phenomena being modelled [33]. The resulting causal graph, whether explicit or implicit, can be leveraged to understand novel compositions during inference, effectively increasing generalization ability. Techniques to build causal models include the use of architectural inductive biases that factorize the inference process through the use of independent mechanisms [12, 13, 22, 28] as well as meta-learning related training schemes or objectives [7, 26]. These methods have been shown to increase OOD performance. For this reason, causality principles have already been applied to robust VQA. Chen et al. [8] train a model to answer counterfactual questions, a key element of causal learning, by asking the model to answer questions about edited versions of the same image where regions relevant to the question are masked. Teney et al. [36] also leverage parallel counterfactual image-question-answer pairs whose gradients could be used as an additional training signal.

Causal structures can be learned not just through architectural structures and training schemes, but also solely through data seen during training [33]. One well documented method is to simply increase the scope of the training data to ensure that multiple distributions are seen during training, reducing the amount of unknown domains the model would have to adapt to during inference. This is the approach leveraged by large language models and other foundation models [20, 29] which have shown unprecedented performance on new domains with little to no fine-tuning data. A second approach would be to train a model to build rich representations through a self-supervised learning task, which could be finetuned on downstream tasks such as VQA. This is usually performed in conjunction with the previous method, and is the approach behind the most performant VQA models [5, 35, 39, 41, 42].

A third approach which is commonplace in computer vision would be to augment the data. This method ties into the causality concept of interventions. An intervention is defined as an action that changes the joint distribution of variables in an environment (e.g. moving all cows from fields to beaches reduces the overlap of images with cows in them and images with grass in the background) [19]. Applying interventions essentially creates multiple training environments under different distributional shifts, exposing which elements of certain phenomena stay invariant under environmental changes. Traditional image augmentations (e.g. image flipping, saturation editing) have been shown to simulate the act of performing interventions (e.g. different saturation levels can be considered different training domains) [18]. Interventions can also be simulated through unshuffling the train set into different partitions with unique priors, essentially creating multiple unique training distributions [37].

In the use of interventions for creating robust VQA models, the most relevant works to this study are the IV-VQA and CV-VQA datasets [1]. The former leverages pre-existing COCO image and text annotations to determine which elements of an image should not affect the answer to a given question, and removes the element using an off-the-shelf GAN-based inpainter; the latter performs the same procedure but targets relevant entities in the image and changes the answer to the question accordingly. Training with these augmented image-question-answer triples exposes models to distributions outside the standard domain where spurious correlations between visuals and answers may no longer hold. A similar technique was leveraged by Gokhale et al. [11], who also augmented questions and used all augmented image-question-answer triples with a specialized architecture designed to align latent encodings of images and answers.

The study's proposed technique for generating OOD data points in terms of the both the general approach as well as breadth of the interventions applied. Firstly, this method does not rely on directly augmenting preexisting images, but focuses on simply learning their underlying distribution and sampling from it with a text-conditioned generative image model, with interventions being simulated on the marginal distribution of images.

Secondly, pre-existing works primarily focus on altering targeted regions of images to break visual correlations. The current work makes no such restrictions, other than ensuring the question-answer pairs still apply to the generated image. Furthermore, these previous studies only target specific question types (e.g. ”How many”, ”What color”) whereas this work seeks to overcome this restriction and generate data regardless of the questions involved.

CHAPTER III

METHODOLOGY

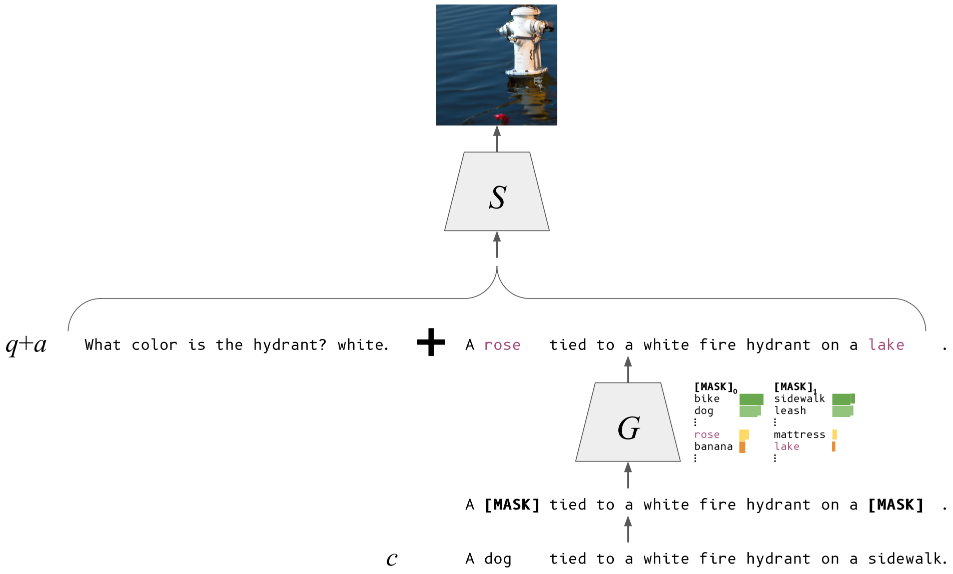

S
q+a
What color is the hydrant? white.
+
A rose tied to a white fire hydrant on a lake .
G
[MASK]$_0$
bike
dog
⋮
rose
banana
⋮
[MASK]$_1$
sidewalk
leash
⋮
mattress
lake
⋮
A [MASK] tied to a white fire hydrant on a [MASK] .
c
A dog tied to a white fire hydrant on a sidewalk.

Figure 3.1: Proposed method for artificially creating OOD image-question-answer triples.

The proposed method, shown in Figure 3.1, consists of two stages: a training stage where the distribution to be considered in-distribution is modelled, and a sampling stage where the trained model is used to steer data generation away from the original distribution. In the training stage, a language model G is first fitted on the distribution of images via their captions. During the sampling stage, captions are masked, with the scoring from G used to sampled low-probability replacements to create new captions. These new captions are then concatenated with the question and answer, with the final prompt used to steer a generative text-to-image model S.

3.1 Training stage

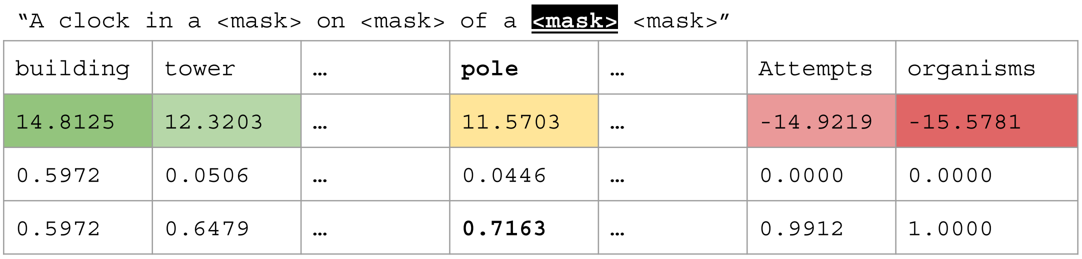
Given a distribution of image-question-triples $\langle i, q, a\rangle \in D$, a naive approach to generate new data would be to use a q and an a to steer a large pretrained text-to-image model S to generate a corresponding i. This method however lacks any form of information about the original distribution D and is not optimized to maximize the distance of the generated samples from it.

To address this, the study proposes to first partly model D, specifically I, the marginal distribution of $i \in D$. A model G fitted onto the the distribution of i can be used as an energy function for how close or far an image is to to the distribution. This can provide a signal to an optimization method for sampling with S.

Particularly, the study models the distribution of i as text, due to the relative computational efficiency of modelling standard text sequences compared to full images, as well as the predominance of generative image models that are conditioned on text. Each i is associated with a caption c, and a language model G is fitted onto the distribution of c's. This study in particular models the distribution with a bidirectional text encoder via a masked-language modeling objective.

3.2 Sampling stage

Captions are first sampled from the distribution of c's. Spacy is used to randomly mask out 70% of nouns and adjectives from a given c. Through G, replacements can be sampled for these masked tokens. However, to generate captions away from the original distribution, low-likelihood replacements are identified instead of the traditional approach of sampling the



“A clock in a <mask> on <mask> of a **<mask>** <mask>”

building	tower	…	**pole**	…	Attempts	organisms
14.8125	12.3203	…	11.5703	…	-14.9219	-15.5781
0.5972	0.0506	…	0.0446	…	0.0000	0.0000
0.5972	0.6479	…	**0.7163**	…	0.9912	1.0000

Figure 3.2: An example of the study’s proposed sampling method.

most likely tokens. Note that the absolute least likely tokens may lead to grammatically incorrect captions. To address this, this study selects tokens t that satisfy $\arg\max_t \{L_t \mid C_t \geq 0.7\}$, where L_t is the logit assigned to t and $C_t = \sum_{L_i \in \{L_i | L_i < L_t\}} L_i$. This identifies the least likely token replacement following the restriction that its likelihood falls within the top 70% of the total probability mass. This balance encourages generations to be both away from the original distribution but sensible enough to be valid image captions. An illustration of this sampling method is show in Figure 3.2.

The generated image captions t' are then concatenated with the questions and the answers creating image prompts of the form "{q} {a}. {t'}" which are then passed to S.

CHAPTER IV

EXPERIMENTS

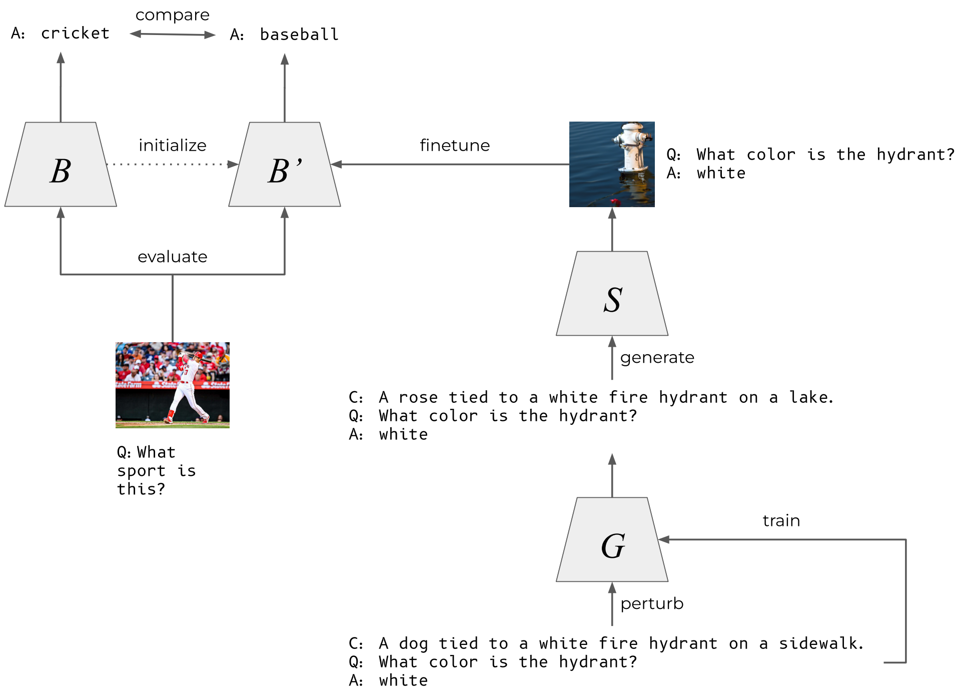

compare
A: cricket
A: baseball
initialize
B
B'
finetune
Q: What color is the hydrant?
A: white
evaluate
S
generate
C: A rose tied to a white fire hydrant on a lake.
Q: What color is the hydrant?
A: white
Q: What sport is this?
G
train
perturb
C: A dog tied to a white fire hydrant on a sidewalk.
Q: What color is the hydrant?
A: white

Figure 4.1: Overview of overall experimental set-up.

Figure 4.1 shows the overall set-up for data generation, fine-tuning, and evaluation. As discussed in Section 3, a finetuned language model G along with text-to-image model S are used to generate OOD image-question-answer triples. This data is then used to finetune a pretrained VQA model B', whose performance is compared to its baseline non-finetuned counterpart B for potential robustness gains. Specific models used for G, S, and B/B' are discussed in Section 4.1. The evaluation protocol is discussed further in Section 4.3. A simpler illustration showing the exact components used in the study are shown in Figure 4.2. Off-the-shelf models and datasets

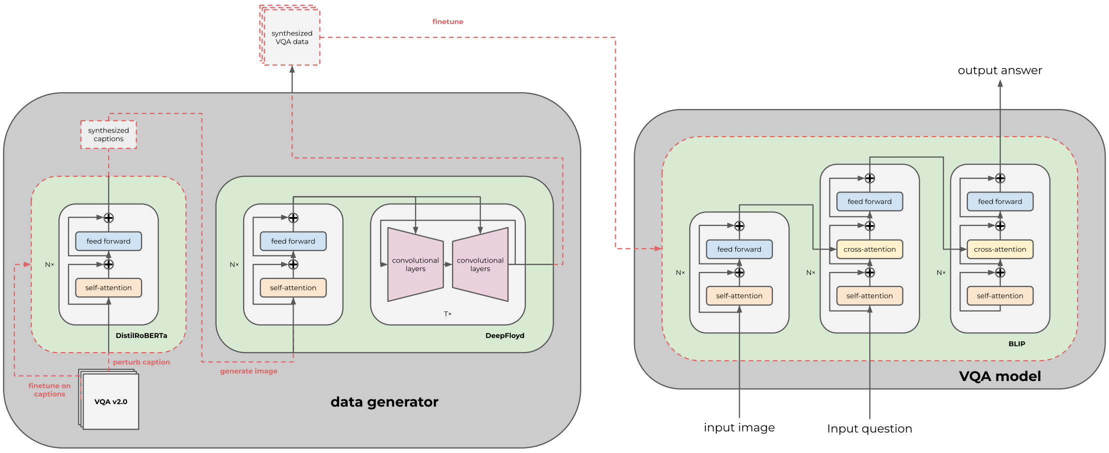


Figure 4.2: Main components of study.

are shown in green. The study's contributions in terms of proposed pipeline and artifacts such as resulting finetuned models and synthetic data are shown outlined in dashed red.

4.1 Experimental details

4.1.1 Base distribution

All experiments use VQA v2.0 [14] as the base distribution. VQA v2.0 sources real images from MS COCO [24], itself a collection of photos containing common everyday objects in common contexts scraped from Flickr, as well as synthetic images using clip art. Each photo is then given multiple relevant questions and answers through human annotation. With 265016 images and at least three questions per image, the final dataset contains 4437570 datapoints for training and 2143540 datapoints for validation.

4.1.2 Text modelling

For G, text modelling of these captions is performed by finetuning a DistilRoBERTa [32], a Transformer-based text encoder. Text is passed as a sequence of tokens, with each token mapped to a learned vector from a vocabulary. Each vector is then updated as a weighted sum of the other vectors in the sequence. DistilRoBERTa is a shrunken version on RoBERTa [25], which is trained with a masked language modelling objective, where a subset of tokens in a sentence are obscured and must be recovered using the remaining tokens as context. As such, given a sentence with a masked word, DistilRoBERTa is able to provide an unbounded score for each word in the vocabulary with higher values indicating a higher likelihood that the word is a proper substitute for the masked word.

The DistilRoBERTa used in this study is finetuned at 16-bit precision on 10000 image captions for 3 epochs with batch size 64, learning rate 8e-5, warmup ratio 0.6, weight decay 0.01, and AdamW optimization with β_1 = 0.9, β_2 = 0.999, and ϵ = 1e-8.

4.1.3 Image generation

Image generation is performed with a DeepFloyd-M as S, an open implementation of Imagen [31]. DeepFloyd leverages diffusion models to generate images, where an initial image of noise is iteratively denoised with the goal of producing an image that matches a provided text description. Specifically, DeepFloyd is a cascade of diffusion models, where one diffusion model is used to generate a 64 × 64 image, a second diffusion model is conditioned off the previous image to produce a 256 × 256 images, and a third diffusion model is conditioned off the previous image to produce a 1024 × 1024 image. For computational efficiency, image generation in this study is performed using only the first diffusion model at 16-bit precision for 27 denoising steps, with the text encoder used in 8-bit precision.

To generate the study's OOD images, 20000 captions are first generated with the method described in Section 3.2. These captions are then filtered down to 5664 captions through a simple filtering heuristic with multimodal encoder CLIP [29], where each image and text are projected to an n-dimensional semantic vector space where similar vector imply similar semantics (e.g. the encoded vector for "A photo of a cat." would be closer to that of an image of a cat than an image of a dog). The vector difference between the generated caption embedding and the the embedding of the concatenation of the question-answer pair are added to the original image embedding, with the resulting embedding compared with original image embedding via cosine similarity; captions that fail to produce a score of at least 0.2 are ignored, as these captions may contradict the question-answer pair. 3624 of these captions are used for training, with the rest reserved as validation and test splits for potential future studies.

4.2 Data Description

Examples of data generated with this method can be seen in Figure 4.3.

This study performs a brief exploratory analysis on the distribution generated by the proposed method.

Figure 4.4 shows the top 25 words by relative usage, where in-distribution corresponds to the training captions and out-of-distribution corresponds to the generated data. For this particular experiment, a word is an item in Dis-

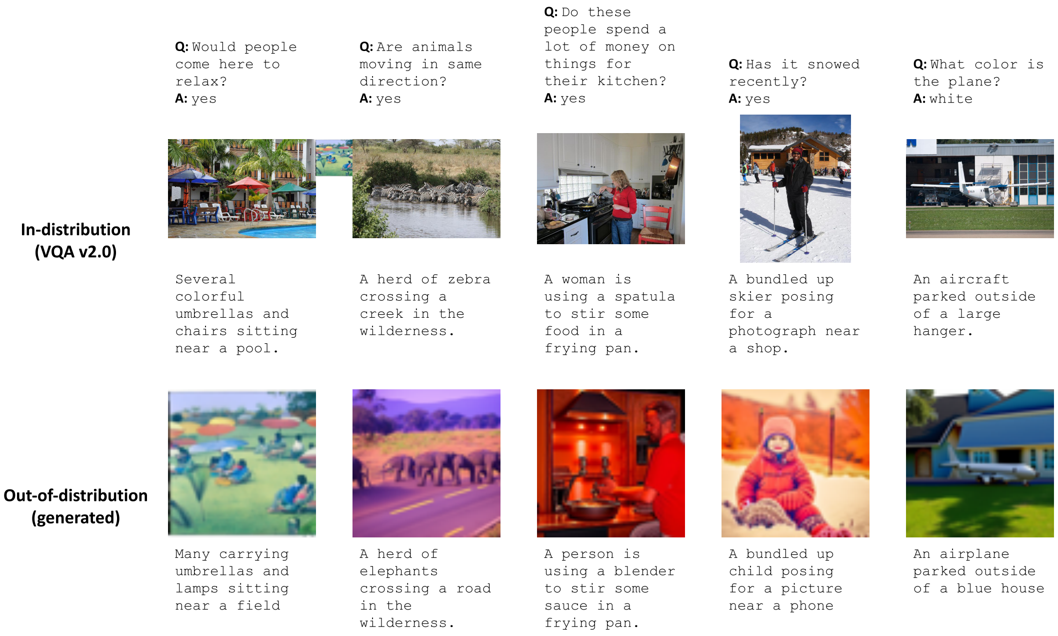
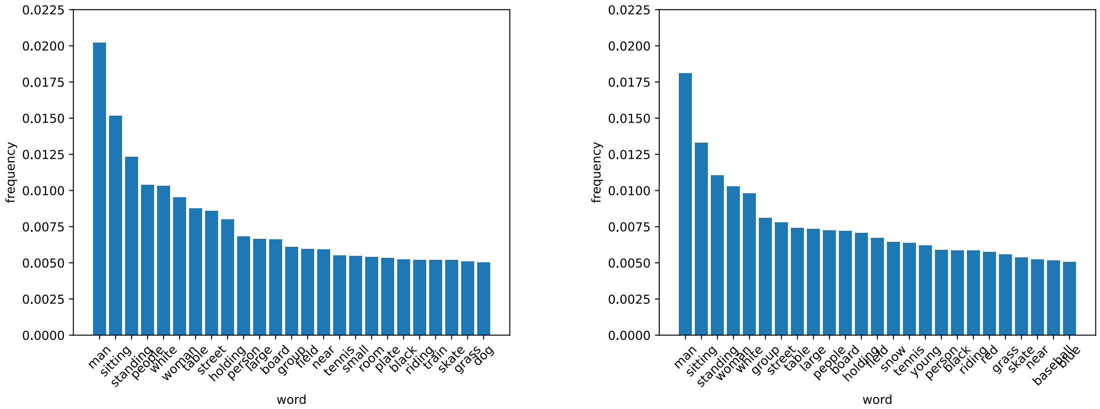

Q: Would people come here to relax?
A: yes
Q: Are animals moving in same direction?
A: yes
Q: Do these people spend a lot of money on things for their kitchen?
A: yes
Q: Has it snowed recently?
A: yes
Q: What color is the plane?
A: white
In-distribution (VQA v2.0)
Several colorful umbrellas and chairs sitting near a pool.
A herd of zebra crossing a creek in the wilderness.
A woman is using a spatula to stir some food in a frying pan.
A bundled up skier posing for a photograph near a shop.
An aircraft parked outside of a large hanger.
Out-of-distribution (generated)
Many carrying umbrellas and lamps sitting near a field
A herd of elephants crossing a road in the wilderness.
A person is using a blender to stir some sauce in a frying pan.
A bundled up child posing for a picture near a phone
An airplane parked outside of a blue house

Figure 4.3: Sample generations with study's proposed method.

tilRoBERTa's vocabulary after lowercasing and whitespace removal. Stopwords as provided by Spacy and punctuation marks are also excluded from this analysis. While there exists overlap between the top 25 words of the original and generated captions, with 20 words existing in the top 25 of both distributions, a stronger gap between in-distribution and out-of-distribution data is revealed through an analysis of the generated images.

Datapoints with identical question-answer pairs are identified and encoded via the same CLIP described in Section 4.1.3. After reducing to 8 dimensions via PCA for numerical stability, the Mahalanobis distance of datapoints in these sample question-answer clusters compared to the mean of the in-distribution datapoints are seen in Table 4.1. For these sample question-answer categories, the average image generation is empirically shown to lie away from its in-distribution counterparts in semantic space. For a visual representation of this distance, the vectors can instead be re-



(a) VQA v2.0 (in-distribution)

(b) generated data (out-of-distribution)

Figure 4.4: Top 25 words by frequency.

duced to two-dimensional space via PCA and plotted, as shown in in Figure 4.5. Blue points represent the original distribution while orange points represent the study's generated data.

Lastly, the generated data is evaluated for potential factual incorrectness. Text-based image generation is still prone to producing images inconsistent with the provided prompt. The CLIP embeddings of the generated images are compared with the CLIP embeddings of the concatenation of their corresponding question and answer, similar to the filtering step discussed in Section 4.1.3. On average, the cosine similarity of an image generated with the study's method with its question and answer is 0.18, below the minimum 0.2 or 0.3 threshold that most studies consider as an indicator of data cleanliness. Despite this relatively low cosine similarity, the results in Section 4.4 empirically show the potential for the study's method is concretely improve OOD generalization in the VQA task setting.

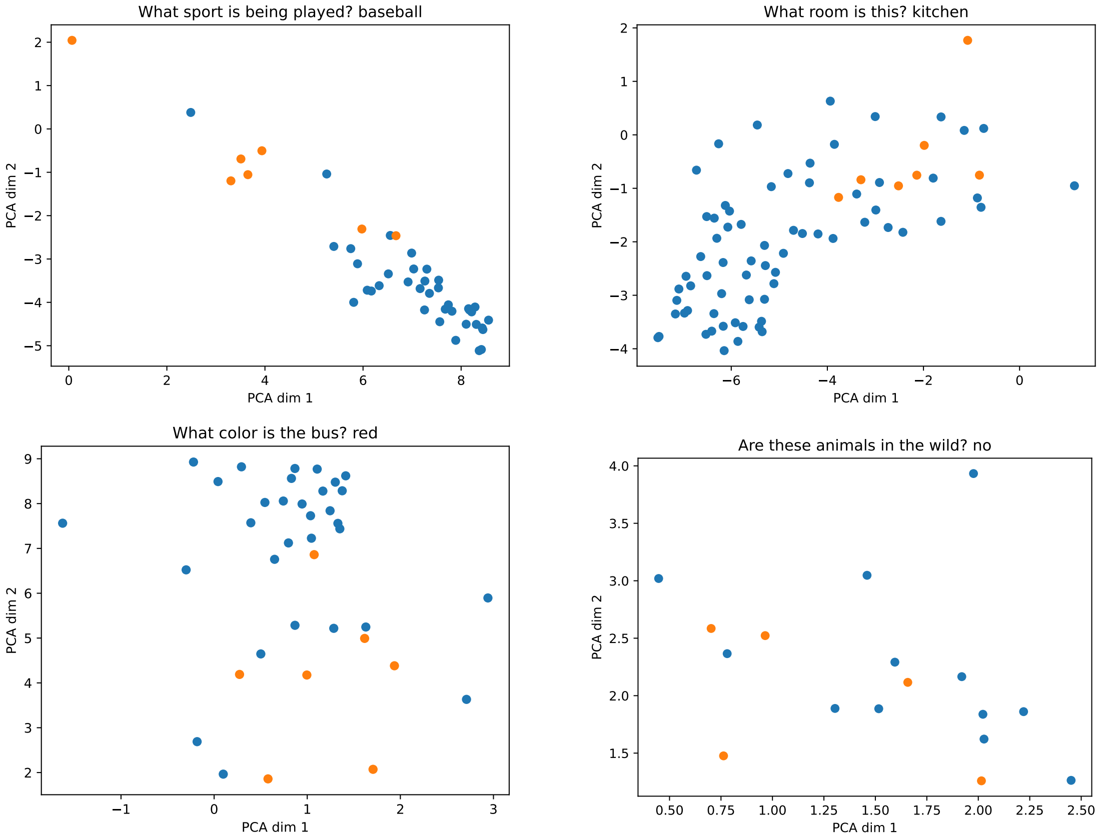


Figure 4.5: PCA-reduced plots of datapoints in CLIP space.

4.3 Evaluation Protocol

To evaluate robustness against OOD images, this study adopts the evaluation protocol proposed by Agrawal et al. [3]. This study first evaluates the performance of a VQA model, B, on the test set of its training distribution, followed by the test sets of separate distributions to observe the drop in accuracy resulting from the distribution shift. With these results as a baseline, the model is then finetuned on the study's generated data to create B', which is evaluated on the same out-of-distribution validation or test sets. Higher accuracy implies increased robustness while converse implies worsened robustness. Accuracy is measured with the protocols corresponding to the respective benchmark. Specific implementation details are as follows.

4.3.1 Model and finetuning

For B and subsequently B', this study selects BLIP [23] as the model. BLIP is a multimodal mixture of encoder-decoders (MED) comprised of an image encoder, text encoder, and text decoder, all of which are Transformer-based. The model is pretrained on a large corpus of web-scraped images-with-caption pairs optimized with three loss functions/objectives: 1) a contrastive loss between matching/mis-matching image-text pairs where the encoders act independently without cross-attention; 2) a binary classification task to identify matching image-text pairs where the text encoder interacts with the vision encoder via cross-attention (image-grounded text encoding); and 3) a causal language modelling task to generate captions given an image where the text decoder interacts with the vision encoder via cross-attention (image-grounded text decoding). The pretraining data is augmented and filtered via a bootstrapping method, where an identically configured MED is trained with the same objectives but on a human annotated image-text pair dataset, which is then used to adjust the pretraining corpus by adding new captions to the images as well as filter out noisy image-text pairs.

To use BLIP for VQA, the image is first encoded with the image encoder, then the question is encoded by the text-encoder in an image-grounded manner, and finally the answer is generated by the text-decoder via cross-attention with the encoding of the question. A model card [27] for the study's finetuned model can be found in Appendix A.

4.3.2 Evaluation data

Two datasets are used for the robustness evaluation protocol: 1) the Art QUestion Answering (AQUA) [9] dataset, and 2) VizWiz dataset [15].

AQUA is comprised of image-question-answer triples where the images are fine-art paintings. With respect to VQA v2.0, AQUA represents a distribution shift by focusing primarily on artistic renderings and paintings rather than photos of the real world. The study focuses on the subset of image-question-answer triples that can be answered without historical context and art knowledge. These are generated using two methods: 1) the use of an object detector to identify objects in the images, which are passed to a model which generates a question conditioned on an image and an answer, in this case an identified object; and 2) the generation of an image caption using off-the-shelf tools which is then converted into a question and answer pair using rules-based methods. The resulting image-question-answer triples are filtered by grammatical and factual correctness using Amazon Mechanical Turk (AMT) workers. This study evaluates on the test split of this subset, amounting to 1270 datapoints. Examples from this dataset are shown in Figure B.1.

VizWiz is comprised of photos taken by the visually impaired, with each photo accompanied with a question asked by the photo's originator as well as corresponding answers. With respect to VQA v2.0, VizWiz represents a distribution shift in image quality with lighting, camera focus, and framing differing from the conditions normally found in photos taken by the visually unimpaired. Image-question pairs are sourced from visually impaired participants, with manual filtering performed by AMT workers and a specialized committee to remove data violating individuals' privacy. AMT workers are also used to provide answers to the image-question pairs. As answers are not available for the test set are not publicly available, this study utilizes the validation set for evaluation, specifically the subset of data where the questions are indicated by annotators to be properly answerable, resulting in 2934 datapoints. Examples from this dataset are shown in Figure B.2.

4.4 Results

Table 4.2 shows the performance gains produced by finetuning on the study's synthetic OOD data. The performance of BLIP on VQA v2.0, the original distribution used for training, is 77.54%. In concurrence with prior results [3], because AQUA, and VizWiz carry underlying distribution shifts, BLIP does not maintain this same level of performance across these benchmarks despite all being VQA tasks, dropping in accuracy to 27.72% and 19.54% on AQUA and VizWiz respectively.

With finetuning on less than 4000 synthetically generated datapoints and relying only on large pretrained models and the base distribution, these performance drops across distributions shifts are reduced. After finetuning, accuracy is raised to 31.50% for AQUA, and to 20.37% for VizWiz. This entails accuracy drop reductions of 7.59% and 1.43% respectively. While the performance gap is not completely closed, the presence of positive accuracy drop reduction shows the potential for using large pretrained models for synthetic OOD data generation for training purposes.

Table 4.1: Mahalanobis distance of sample generations from in-distribution counterparts.

	Mahalanobis distance ↑	
question-answer pair	in-distribution image	out-of-distribution image

`Q: What sport is being played?`
`A: baseball`
	2.55	30.53

`Q: What room is this?`
`A: kitchen`
	2.77	23.26

`Q: What color is the bus?`
`A: red`
	2.73	26.66

`Q: Are these animals in the wild?`
`A: no`
	2.57	27.66

Table 4.2: Accuracy across different test sets.

	AQUA	VizWiz
baseline	27.72%	19.54%
finetuned	**31.50%**	**20.37%**

CHAPTER V

CONCLUSION

This study shows pilot work for utilizing pretrained foundation models to synthetically intervene on a VQA data distribution. It proposes a novel sampling method to extract low probability token replacements from image captions via a masked language model. By leveraging a pretrained text-to-image model, these generated captions can then be used to synthesize new datapoints to finetune a pretrained VQA model. Finetuning with this generated data is empirically shown to improve robustness in VQA accuracy across multiple investigated distribution shifts, with accuracy drop reductions of 7.59% on art-based distribution shift dataset AQUA and 1.43% on visual impairment-based distribution shift dataset VizWiz. This demonstrates the potential for improving model robustness through foundation models without additional data collection.

Several future directions exist for this work. Firstly, future studies can examine the effect of experimenting with larger but more computationally expensive generative models on the quality of generated outputs. Secondly, studies can investigate an alternative approach to simulating interventions by leveraging an image-to-image translation method to intervene on individual datapoints. Lastly, future studies can experiment with generative text modelling as opposed to masked language modeling to understand whether data quality can be improved if generation were not reliant on a single uniform threshold.

REFERENCES

[1] Agarwal, V., Shetty, R. and Fritz, M. [2020], Towards causal vqa: Revealing and reducing spurious correlations by invariant and covariant semantic editing, *in* 'Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition', pp. 9690–9698.

[2] Agrawal, A., Batra, D., Parikh, D. and Kembhavi, A. [2018], Don't just assume; look and answer: Overcoming priors for visual question answering, *in* 'Proceedings of the IEEE conference on computer vision and pattern recognition', pp. 4971–4980.

[3] Agrawal, A., Kajić, I., Bugliarello, E., Davoodi, E., Gergely, A., Blunsom, P. and Nematzadeh, A. [2022], 'Rethinking evaluation practices in visual question answering: A case study on out-of-distribution generalization', *arXiv preprint arXiv:2205.12191* .

[4] Agrawal, A., Teney, D. and Nematzadeh, A. [2022], Vision-language pretraining: Current trends and the future, *in* 'Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts', pp. 38–43.

[5] Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M. et al. [2022], 'Flamingo: a visual language model for few-shot learning', *arXiv preprint arXiv:2204.14198* .

[6] Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L. and Parikh, D. [2015], Vqa: Visual question answering, *in* 'Proceedings of the IEEE international conference on computer vision', pp. 2425–2433.

[7] Bengio, Y., Deleu, T., Rahaman, N., Ke, R., Lachapelle, S., Bilaniuk, O., Goyal, A. and Pal, C. [2019], 'A meta-transfer objective for learning to disentangle causal mechanisms', *arXiv preprint arXiv:1901.10912* .

[8] Chen, L., Yan, X., Xiao, J., Zhang, H., Pu, S. and Zhuang, Y. [2020], Counterfactual samples synthesizing for robust visual question answering, *in* 'Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition', pp. 10800–10809.

[9] Garcia, N., Ye, C., Liu, Z., Hu, Q., Otani, M., Chu, C., Nakashima, Y. and Mitamura, T. [2020], A dataset and baselines for visual question answering on art, *in* 'European Conference on Computer Vision', Springer, pp. 92–108.

[10] Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M. and Wichmann, F. A. [2020], 'Shortcut learning in deep neural networks', *Nature Machine Intelligence* **2**(11), 665–673.

[11] Gokhale, T., Banerjee, P., Baral, C. and Yang, Y. [2020], 'Mutant: A training paradigm for out-of-distribution generalization in visual question answering', *arXiv preprint arXiv:2009.08566* .

[12] Goyal, A., Didolkar, A., Lamb, A., Badola, K., Ke, N. R., Rahaman, N., Binas, J., Blundell, C., Mozer, M. and Bengio, Y. [2021], 'Coordination among neural modules through a shared global workspace', *arXiv preprint arXiv:2103.01197* .

[13] Goyal, A., Lamb, A., Hoffmann, J., Sodhani, S., Levine, S., Bengio, Y. and Schölkopf, B. [2019], 'Recurrent independent mechanisms', *arXiv preprint arXiv:1909.10893* .

[14] Goyal, Y., Khot, T., Summers-Stay, D., Batra, D. and Parikh, D. [2017], Making the v in vqa matter: Elevating the role of image understanding in visual question answering, *in* 'Proceedings of the IEEE conference on computer vision and pattern recognition', pp. 6904–6913.

[15] Gurari, D., Li, Q., Stangl, A. J., Guo, A., Lin, C., Grauman, K., Luo, J. and Bigham, J. P. [2018], Vizwiz grand challenge: Answering visual questions from blind people, *in* 'Proceedings of the IEEE conference on computer vision and pattern recognition', pp. 3608–3617.

[16] Han, X., Wang, S., Su, C., Huang, Q. and Tian, Q. [2021], Greedy gradient ensemble for robust visual question answering, *in* 'Proceedings of the IEEE/CVF International Conference on Computer Vision', pp. 1584–1593.

[17] Hudson, D. A. and Manning, C. D. [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', pp. 6700–6709.

[18] Ilse, M., Tomczak, J. M. and Forré, P. [2021], Selecting data augmentation for simulating interventions, *in* 'International Conference on Machine Learning', PMLR, pp. 4555–4562.

[19] J., P. [2018], *Causality: Models, Reasoning, and Inference 2ed., 2013 printing*.

[20] Kojima, T., Gu, S. S., Reid, M., Matsuo, Y. and Iwasawa, Y. [2022], 'Large language models are zero-shot reasoners', *arXiv preprint arXiv:2205.11916* .

[21] Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.-J., Shamma, D. A. et al. [2017], 'Visual genome: Connecting language and vision using crowdsourced dense image annotations', *International journal of computer vision* **123**(1), 32–73.

[22] Lamb, A., He, D., Goyal, A., Ke, G., Liao, C.-F., Ravanelli, M. and Bengio, Y. [2021], 'Transformers with competitive ensembles of independent mechanisms', *arXiv preprint arXiv:2103.00336* .

[23] Li, J., Li, D., Xiong, C. and Hoi, S. [2022], Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, *in* 'International Conference on Machine Learning', PMLR, pp. 12888–12900.

[24] Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P. and Zitnick, C. L. [2014], Microsoft coco: Common objects in context, *in* 'European conference on computer vision', Springer, pp. 740–755.

[25] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. and Stoyanov, V. [2019], 'Roberta: A robustly optimized bert pretraining approach', *arXiv preprint arXiv:1907.11692* .

[26] Madan, K., Ke, N. R., Goyal, A., Schölkopf, B. and Bengio, Y. [2021], 'Fast and slow learning of recurrent independent mechanisms', *arXiv preprint arXiv:2105.08710* .

[27] Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D. and Gebru, T. [2019], Model cards for model reporting, *in* 'Proceedings of the conference on fairness, accountability, and transparency', pp. 220–229.

[28] Mittal, S., Lamb, A., Goyal, A., Voleti, V., Shanahan, M., Lajoie, G., Mozer, M. and Bengio, Y. [2020], Learning to combine top-down and bottom-up signals in recurrent neural networks with attention over modules, *in* 'International Conference on Machine Learning', PMLR, pp. 6972–6986.

[29] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J. et al. [2021], Learning transferable visual models from natural language supervision, *in* 'International Conference on Machine Learning', PMLR, pp. 8748–8763.

[30] Ramakrishnan, S., Agrawal, A. and Lee, S. [2018], 'Overcoming language priors in visual question answering with adversarial regularization', *Advances in Neural Information Processing Systems* **31**.

[31] Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E. L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T. et al. [2022], 'Photorealistic text-to-image diffusion models with deep language understanding', *Advances in Neural Information Processing Systems* **35**, 36479–36494.

[32] Sanh, V., Debut, L., Chaumond, J. and Wolf, T. [2019], 'Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter', *ArXiv* **abs/1910.01108**.

[33] Schölkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A. and Bengio, Y. [2021], 'Toward causal representation learning', *Proceedings of the IEEE* **109**(5), 612–634.

[34] Shah, M., Chen, X., Rohrbach, M. and Parikh, D. [2019], Cycle-consistency for robust visual question answering, *in* 'Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition', pp. 6649–6658.

[35] Singh, A., Hu, R., Goswami, V., Couairon, G., Galuba, W., Rohrbach, M. and Kiela, D. [2022], Flava: A foundational language and vision alignment model, *in* 'Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition', pp. 15638–15650.

[36] Teney, D., Abbasnedjad, E. and Hengel, A. v. d. [2020], Learning what makes a difference from counterfactual examples and gradient supervi-

sion, *in* 'European Conference on Computer Vision', Springer, pp. 580–599.

[37] Teney, D., Abbasnejad, E. and van den Hengel, A. [2021], Unshuffling data for improved generalization in visual question answering, *in* 'Proceedings of the IEEE/CVF International Conference on Computer Vision', pp. 1417–1427.

[38] Von Kügelgen, J., Sharma, Y., Gresele, L., Brendel, W., Schölkopf, B., Besserve, M. and Locatello, F. [2021], 'Self-supervised learning with data augmentations provably isolates content from style', *Advances in neural information processing systems* **34**, 16451–16467.

[39] Wang, Z., Yu, J., Yu, A. W., Dai, Z., Tsvetkov, Y. and Cao, Y. [2021], 'Simvlm: Simple visual language model pretraining with weak supervision', *arXiv preprint arXiv:2108.10904* .

[40] Wu, J. and Mooney, R. [2019], 'Self-critical reasoning for robust visual question answering', *Advances in Neural Information Processing Systems* **32**.

[41] Yu, J., Wang, Z., Vasudevan, V., Yeung, L., Seyedhosseini, M. and Wu, Y. [2022], 'Coca: Contrastive captioners are image-text foundation models', *arXiv preprint arXiv:2205.01917* .

[42] Yuan, L., Chen, D., Chen, Y.-L., Codella, N., Dai, X., Gao, J., Hu, H., Huang, X., Li, B., Li, C. et al. [2021], 'Florence: A new foundation model for computer vision', *arXiv preprint arXiv:2111.11432* .

APPENDIX A
MODEL CARD

Table A.1 shows the model card for the finetuned VQA model produced by this study. Because it is a finetuned version of the original BLIP model, most details in this model card are focused on information relevant to the particular finetuned version.

Table A.1: Finetuned BLIP Model Card.

Model Details	
Organization developing model	Salesforce (base model); ALIVE (finetuning)
Model date	July 27, 2023 (finetuning)
Model version	v1
License	BSD 3-Clause
Intended Use	
Primary intended use	visual question answering (VQA) research, robust VQA
Primary intended users	researchers in computer vision, natural language processing, foundation modals, multimodal learning
Out-of-scope use cases	Deployment in practical applications outside research settings
Metrics	
Model performance measures	accuracy on out-of-distribution VQA datasets following said datasets' prescribed evaluation protocols

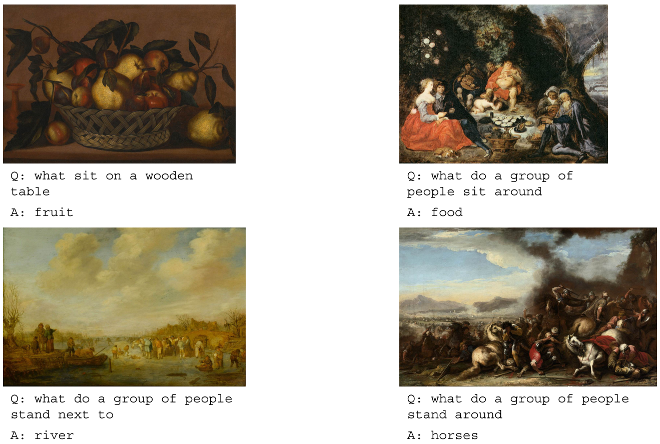
Decision thresholds	Not applicable
Approaches to uncertainty and variability	To conserve computational and energy costs, only one finetuned model was produced.
Evaluation Data	
Datasets	AQUA [9]; VizWiz [15]; VQA v2.0 [14]
Motivation	Following Agrawal et al. [3], we also use evaluation splits from datasets other than the original training dataset in order to probe robustness to distribution shifts.
Preprocessing	None
Training Data	
Datasets	Synthetically generated OOD VQA data described in Section 3
Motivation	We finetune on synthetically generated OOD data in order to examine the possible improvements in robustness across distribution shifts.
Preprocessing	0.20 CLIP-filtering used in training data construction as described in Section 4.1
Ethical Considerations	
Data	This model has been trained on data generated with diffusion model-based methods, which have been shown to be capable of memorizing exact instances of training data.

<table>
<tr><td><b>Human life</b></td><td>This model is not intended to infer anything about humans other than what can be understood visually (e.g. ”Is this person smiling?”, ”What is the person holding?”).</td></tr>
<tr><td><b>Risks & harms</b></td><td>As this model has not been evaluated for bias, toxicity, etc., this model could potentially pose a risk if deployed if its outputs were to be used in decision-making scenarios regarding people.</td></tr>
<tr><td><b>Use cases</b></td><td>This model is intended for research purposes only. It has not been evaluated for toxicity, bias, etc. and should not be deployed outside of a research setting.</td></tr>
<tr><th colspan="2">Caveats and Recommendations</th></tr>
<tr><td colspan="2">This model has not been evaluated for bias, toxicity, etc. and is recommended to strictly be used only within a research setting unless further analysis on safety has been conducted.</td></tr>
</table>

APPENDIX B

EVALUATION DATA

Because each VizWiz question is annotated with ten answers, Figure B.2 shows only the most common answer for demonstration purposes.



Q: what sit on a wooden table

A: fruit

Q: what do a group of people sit around

A: food

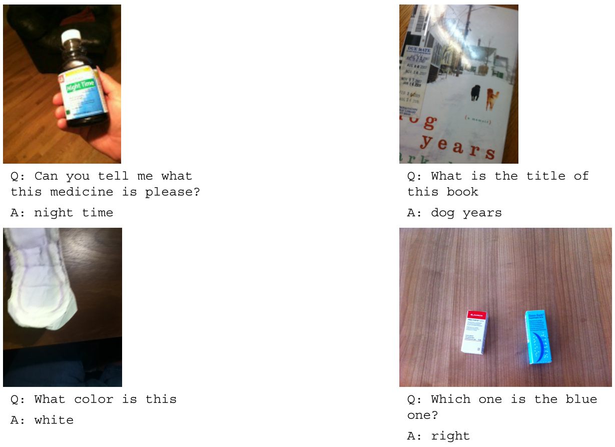
Q: what do a group of people stand next to

A: river

Q: what do a group of people stand around

A: horses

Figure B.1: Examples of data from AQUA used in this study.



Q: Can you tell me what this medicine is please?

A: night time

Q: What is the title of this book

A: dog years

Q: What color is this

A: white

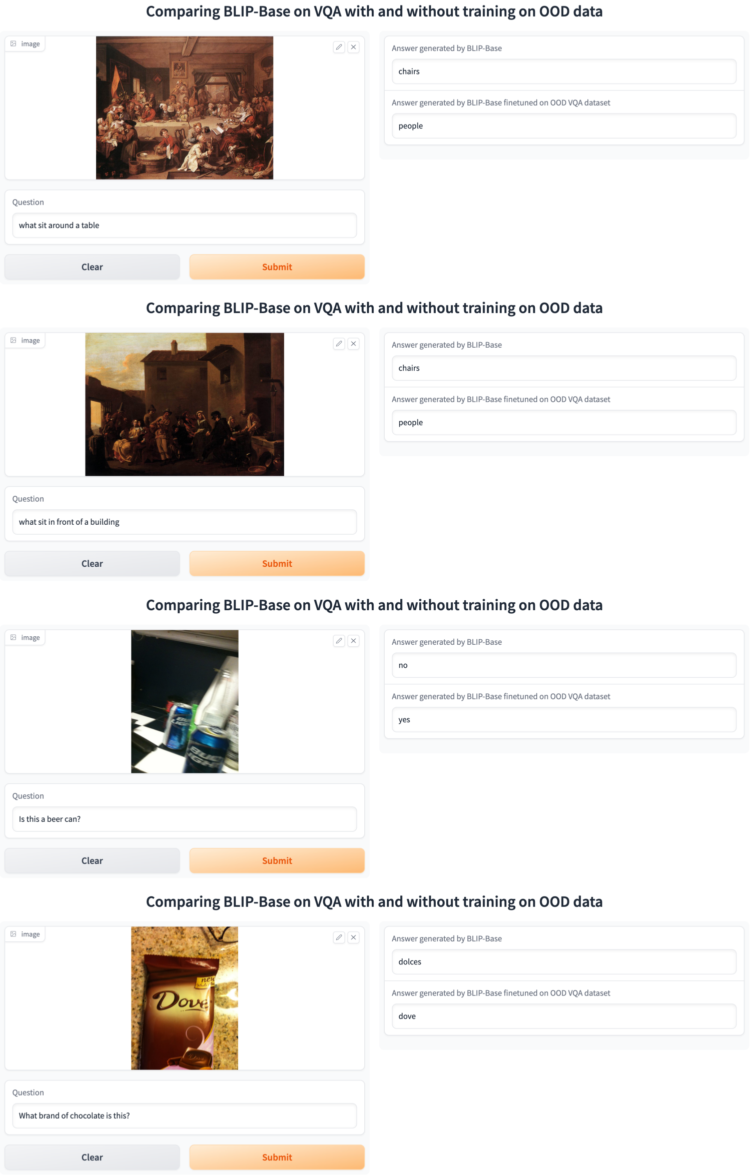
Q: Which one is the blue one?

A: right

Figure B.2: Examples of data from VizWiz used in this study.

APPENDIX C
SAMPLE EVALUATIONS

Figure C.1 shows example VQA datapoints where answers differ between the finetuned and baseline models.



Comparing BLIP-Base on VQA with and without training on OOD data

Question: what sit around a table

Answer generated by BLIP-Base: chairs

Answer generated by BLIP-Base finetuned on OOD VQA dataset: people

Comparing BLIP-Base on VQA with and without training on OOD data

Question: what sit in front of a building

Answer generated by BLIP-Base: chairs

Answer generated by BLIP-Base finetuned on OOD VQA dataset: people

Comparing BLIP-Base on VQA with and without training on OOD data

Question: Is this a beer can?

Answer generated by BLIP-Base: no

Answer generated by BLIP-Base finetuned on OOD VQA dataset: yes

Comparing BLIP-Base on VQA with and without training on OOD data

Question: What brand of chocolate is this?

Answer generated by BLIP-Base: dolces

Answer generated by BLIP-Base finetuned on OOD VQA dataset: dove

Figure C.1: Sample inference with baseline and finetuned models.