CrossVQA: Scalably Generating Benchmarks For Systematically Testing VQA Generalization

Arjun R. Akula[1]*, **Soravit Changpinyo**[3], **Boqing Gong**[3], **Piyush Sharma**[3], **Song-Chun Zhu**[2,4,5], **Radu Soricut**[3]
[1]UCLA Center for Vision, Cognition, Learning, and Autonomy,
[2]Beijing Institute for General Artificial Intelligence (BIGAI),
[3]Google Research, [4]Tsinghua University, [5]Peking University
aakula@ucla.edu, {schangpi,bgong,piyushsharma}@google.com,
s.c.zhu@pku.edu.cn, rsoricut@google.com

Abstract

One challenge in evaluating visual question answering (VQA) models in the cross-dataset adaptation setting is that the distribution shifts are multi-modal, making it difficult to identify if it is the shifts in visual or language features that play a key role. In this paper, we propose a semi-automatic framework for generating disentangled shifts by introducing a controllable visual question-answer generation (VQAG) module that is capable of generating highly-relevant and diverse question-answer pairs with the desired dataset style. We use it to create CrossVQA, a collection of test splits for assessing VQA generalization based on the VQA2, VizWiz, and Open Images datasets. We provide an analysis of our generated datasets and demonstrate its utility by using them to evaluate several state-of-the-art VQA systems. One important finding is that the visual shifts in cross-dataset VQA matter more than the language shifts. More broadly, we present a scalable framework for systematically evaluating the machine with little human intervention.

1 Introduction

Multiple datasets have been proposed to measure the progress on visual question answering (VQA) (Antol et al., 2015; Zhu et al., 2016b; Goyal et al., 2017; Gurari et al., 2018; Hudson and Manning, 2019; Yang et al., 2016; Tu et al., 2014; Qi et al., 2015; Liu et al., 2016). However, these datasets often possess biases introduced in the data collection process and by the human annotators. It has been shown that existing VQA models leverage these spurious biases and take shortcuts (Goyal et al., 2017; Agrawal et al., 2018; Chao et al., 2018a; Akula et al., 2020a). As a result, the performance of those models on a specific VQA dataset can only serve as a rough proxy for the true learning of the VQA *task* (Bras et al., 2020).

Test sets	QA_{vqa2}	QA_{vzwz}
I_{vqa2}	✓	✗
I_{vzwz}	✗	✓
I_{oid}	✗	✗

(a)

Test sets	QA_{vqa2}	QA_{vzwz}
I_{vqa2}	✓	✓
I_{vzwz}	✓	✓
I_{oid}	✓	✓

(b)

Table 1: (a) Existing VQA test sets; (b) CrossVQA (disentangled) test sets generated by our VQAG model.

One common remedy to this is to go beyond in-domain evaluation, in which the test set exhibits some form of "distribution shifts" from the training set (Agrawal et al., 2018; Chao et al., 2018b). The key idea is that a generalizable VQA model should be able to extrapolate, for example, from one dataset to another. One challenge that is quite unique to VQA in this setting is that the distribution shift is *multi-modal*. When one dataset unsatisfactorily transfers to another, it is difficult to identify how much of this is due to vision or language distribution mismatches. To complicate things even more, the frequency of objects occurring in natural images follows a long-tail distribution (Salakhutdinov et al., 2011; Zhu et al., 2014, 2016a). Lack of sufficient instances of minority classes in the test sets further complicates the estimation of generalization capabilities from one dataset to another.

A possible solution to address this issue is to use an iterative, human-in-the-loop approach for dataset collection where human annotators carefully devise new test samples by incorporating visual and language distribution shifts (Nie et al., 2020; Bartolo et al., 2020; Kaushik et al., 2020; Gardner et al., 2020). However, this approach is not scalable and training the human annotators, be they seasoned AI experts or non-experts, would incur huge annotation time and cost.

In this work, we propose to make the process of creating distribution shifts more systematic and automatic. Inspired by recent work on dynamic benchmarks that co-evolve with strong

*Work done in part while AA was an intern at Google.

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 2148–2166
November 7–11, 2021.

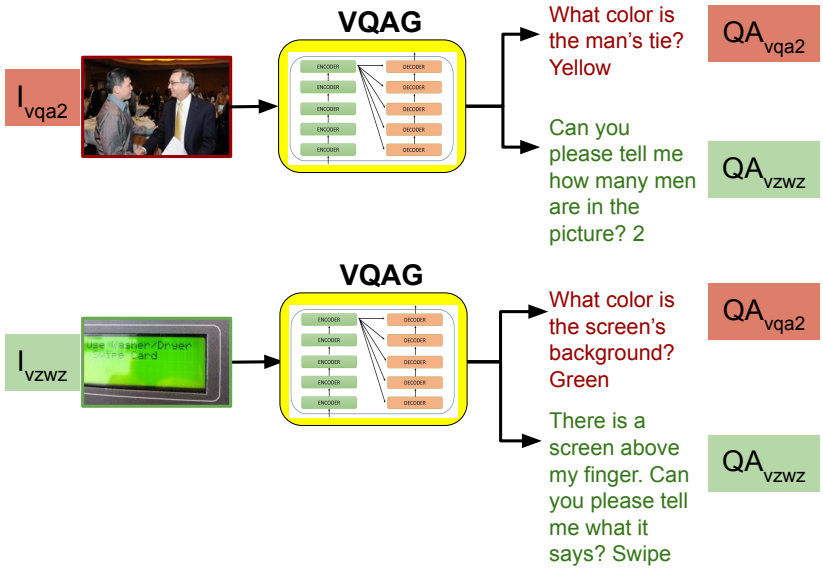

Figure 1: Existing works on VQA domain adaptation between source and target datasets (e.g. VQA2. and VizWiz) can only compare the model's performance on the entangled test splits $\langle I_{vqa2}, QA_{vqa2}\rangle$ and $\langle I_{vzwz}, QA_{vzwz}\rangle$. In this work, we propose a VQAG module to generate novel and scalable VQA test sets, called **CrossVQA**, consisting of additional test sets $\langle I_{vqa2}, QA_{vzwz}\rangle$ and $\langle I_{vzwz}, QA_{vqa2}\rangle$ where visual and language features are disentangled.

models (Zellers et al., 2019b), we propose to bring in visual question-answer generation (VQAG) module in the evaluation process. More specifically, we first build a strong, controllable VQAG engine that is capable of creating particular dataset-style question-answer pairs. Then, we use it to generate novel $\langle image, question, answer\rangle$ test splits, while controlling distribution shifts in vision and language features. This is summarized in Table 1 and exemplified with the VQA2 and VizWiz datasets in Figure 1. Collectively, we refer to the resulting VQA test sets as CrossVQA.

There are at least two advantages in using a VQAG model to construct our CrossVQA test sets: (1) We can evaluate the adaptation skills of VQA models on non-VQA datasets such as Open Images (OID) (Kuznetsova et al., 2018), which contains various image annotations but no question/answer pairs, i.e. $\langle I_{oid}, Q_{vqa2}\rangle$ and $\langle I_{oid}, Q_{vzwz}\rangle$ (see Table 1); (2) Collecting human-annotated test sets is resource-intensive and scales poorly, while the VQAG approach can be massively scaled and applied in a never-ending learning scenario for generating dynamic benchmarks (Nie et al., 2020).

We conduct extensive experiments to evaluate the utility of our proposed framework. First, we validate that our VQAG module is capable of generating relevant questions and correct answers with the desired distribution shifts, which we achieve through a combination of transformer-based architectures, vision-and-language pre-training, and multiple types of control signals. We also find that, when evaluated against state-of-the-art generative models for visual question generation, our VQAG substantially outperforms them in terms of accuracy, diversity, and novelty.

Additionally, we perform analysis and human evaluation of our CrossVQA test sets that are built on VQA2, VizWiz, and Open Images datasets. We show that they are effective at finding and quantifying weaknesses of cross-dataset generalization abilities in the state-of-the-art VQA models. For instance, our experimental results show that VQA models drop up to 40% in absolute accuracy if there is a mismatch in image distribution. On the other hand, VQA models are found to be relatively less sensitive to a mismatch in language distribution.

Finally, inspired by the success of contrastive learning and multi-task learning techniques in improving generalization and robustness of multimodal tasks (Akula et al., 2020a), we investigate whether these techniques improve the performance of VQA models on our CrossVQA test sets. Interestingly, we find that contrastive losses and multi-task regularization do not lead to significant generalization gains on CrossVQA.

In summary, our key contributions are three-fold. First, we introduce the CrossVQA benchmark for

systematically assessing the generalization skills of VQA models, and provide analysis and experiments to support its utility. Second, we describe a scalable data collection and benchmarking framework for semi-automatically constructing the proposed benchmarks using a strong and controllable visual question-answer generation (VQAG) module. Finally, we empirically demonstrate the superiority of our VQAG module by achieving new state-of-the-art results in visual question generation.

2 Related Work

Cross-Dataset Distribution Shifts. There is a large body of work analyzing the generalization skills of neural networks from a labeled source domain to a target domain where there is no or limited labeled data (Ganin and Lempitsky, 2015; Gong et al., 2012; Guo and Xiao, 2012; Tzeng et al., 2015; Akula et al., 2020b). However, these works focus either on language modeling or visual recognition tasks. Here, we investigate adaptation skills using the multi-modal VQA task, for which distribution mismatches can occur in both language and visual features.

There are a few works that study systematic compositional skills in multi-modal tasks. For example, Lampert et al. (2009) study the use of attributes in transferring information between object classes. Jabri et al. (2016) explore several variants of the VQA task and show that VQA models struggle with transferring knowledge across datasets. Agrawal et al. (2018) study the extent to which a model is visually grounded, by evaluating its ability to generalize to a different answer distribution for each question type. Chao et al. (2018b) investigate the issue of cross-dataset generalization, using a specific setting where the source domain contains a large amount of training data and the target domain contains insufficient data to train a VQA system from scratch. Unlike these works, our work performs a more fine-grained analysis by disentangling the distribution mismatches in language and vision, achieved by generating out-of-distribution shifts using a learned VQAG module.

Visual Question Generation (VQG). The goal of VQG is to generate natural questions for an image. This task has drawn much attention due to its ability to test a model's understanding of natural language in the context of visual grounding and its application in downstream tasks such as image retrieval and question answering (Antol et al., 2015; Zhu et al., 2016b; Akula, 2015; Palakurthi et al., 2015).

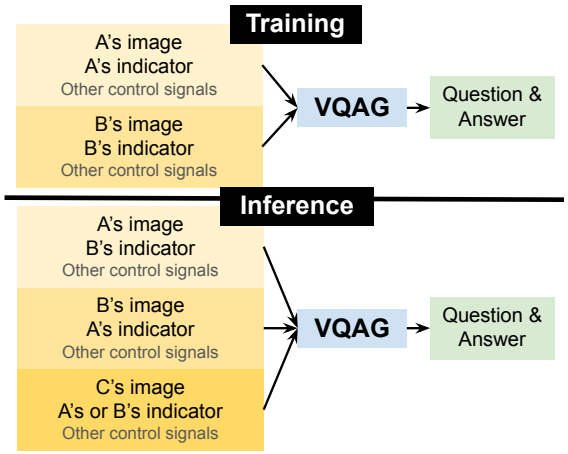

Figure 2: **Overview of CrossVQA**. We train a controllable visual question-answer generation (VQAG) engine and use the dataset indicators and control signals to generate the desired cross-dataset shifts.

While the task of generating question automatically is well studied in the language domain, it has been under-explored for image-related natural questions (Mostafazadeh et al., 2016). Prior works explored VQG using autoencoder-based architectures (Jain et al., 2017; Yang et al., 2018; Alberti et al., 2019; Krishna et al., 2019). Jain et al. (2017) employ a variational autoencoder paradigm where they first learn to embed a given question and image into a low dimensional latent space. The latent codes are subsequently mapped to a high-dimensional representation using RNNs during inference to generate the question. Krishna et al. (2019) model question generation as a process that maximizes mutual information between the image and the expected answer's category. They incorporate fine-grained answer type as the guidance to generate goal-driven questions. Xu et al. (2020) propose an answer-centric approach where they model the complex relationship between an answer and its relevant image regions. Unlike these works, our approach uses a simple encoder-decoder framework, but we enhance it using a transformer-based architecture, vision-and-language pre-training, and various control signals, which together lead to a stronger VQG model. Furthermore, our work not only improves the VQG performance, but also takes a step further by exploring *using* VQG in the context of VQA evaluation.

3 Approach

3.1 Overview

Figure 2 overviews our approach to systematically generating cross-dataset distribution shifts. During training, we train a visual question-and-answer generation (VQAG) engine using multiple sources of VQA data (denoted by A and B). This VQAG module uses a dataset indicator to learn and generate question-answer pairs of a particular dataset's style.

During inference, we apply the trained VQAG model to multiple image sources (denoted by A, B, and C), while varying the dataset indicator. For example, we turn on the dataset B indicator for the images of A, which generates B-style questions/answers for the images in A. Furthermore, VQAG can also be applied to images from a different dataset C, for which no VQA annotations are available, yet we can still control the style of annotations generated. In the post-processing step, the resulting VQA datasets are validated by human annotators.

We first provide more details on our VQAG engine (Sec. 3.2) and then describe how it is used to generate CrossVQA benchmarks (Sec. 3.3).

3.2 Visual Question-Answer Generation

We start from a transformer-based encoder-decoder model that learns to generate question-answer pairs from images. We then enhance this model in two ways. First, we perform image-text pre-training using a recently introduced Conceptual 12M (CC12M) dataset (Changpinyo et al., 2021). Second, we experiment with multiple control signals. As we will show in our experimental results, these signals help improve the accuracy and the diversity of the generated outputs when applied to diverse sources of images.

Base VQAG Model and Input-Output Format. We adopt a transformer-based encoder-decoder framework (Vaswani et al., 2017) for image-to-text generation as our base model, following recent work on large-scale image captioning (Sharma et al., 2018; Changpinyo et al., 2019). In particular, we represent each input image as a sequence of feature vectors, and the model learns to produce relevant questions and their corresponding correct answers.

Each input image is represented by multiple types of visual features (Changpinyo et al., 2019), which we briefly describe here (see Appendix D for more details):
(i) a global feature vector extracted by Graph-RISE (Juan et al., 2019), a ResNet-101 (He et al., 2016) trained for image classification at ultrafine granularity levels;
(ii) 16 regional feature vectors, obtained from Graph-RISE featurization of top-16 proposals of a Faster RCNN (Ren et al., 2015) object detector trained on Visual Genome (Krishna et al., 2017);
(iii) top semantic object label vectors, where labels (e.g. "river", "man", "football") are produced by the Google's Vision API[1].

Our target is a question-answer pair in the format $q\ \langle sep \rangle\ a$, where q is the question tokens, a is the answer tokens, and $\langle sep \rangle$ is the chosen delimiter. Furthermore, since a is not limited to a single answer (Bhattacharya et al., 2019), a is represented as $a_1 \langle dsep \rangle a_2 \langle dsep \rangle \ldots \langle dsep \rangle a_K$, where $a_1, a_2, \ldots, a_k$ are possible answers for q. We use beam search to generate the target question and answer(s) during the decoding stage.

Next we incorporate two enhancements into this base model to (a) maximize the relevance between image, question and expected answer in the generated test sets; (b) improve generalization capability of the model to out-of-domain images; and (c) increase the diversity and novelty of the questions.

Enhancement 1: Image-To-Text Pre-Training. We pre-train our base VQAG model on Conceptual 12M (Changpinyo et al., 2021), a large-scale dataset specifically designed for vision-and-language pre-training. It consists of 12.4 million image–Alt-text pairs harvested from the Web. We use the standard image captioning objective for pre-training (Changpinyo et al., 2021). Despite this task mismatch (i.e., image captioning vs. visual question/answer generation), we observe the utility of pre-training in addressing the long-tail distribution of objects (see Sec. 4.2)

Enhancement 2: Dataset-Agnostic Control Signals. In addition to the image features, we also condition our model on up to three control knobs more directly related to visual question generation and answering. In particular, we explore three main types of dataset-agnostic control signals, summarized in Table 2: the expected first two words of the question (i.e. question prefix), the expected answer category, and the expected answer(s). See

[1] https://cloud.google.com/vision/docs/labels

Table 2: Dataset-agnostic control signals and examples.

Notation	Description
P	Question prefix
C	Answer category
A	Most common answer
$\tilde{A}$	All answers

Examples
Question 1: Is the screen's background blue?
P : *Is the*, **C** : *Color*, $\tilde{\mathbf{A}}$: *yes <dsep> true <dsep> blue screen <dsep> yes,* ***A*** *: yes*
Question 2: How many men are in the picture?
P : *How many*, **C** : *Counting*, $\tilde{\mathbf{A}}$: *2 <dsep> 2 <dsep> 3 <dsep> 5,* ***A****: 2*

Appendix A for further discussion.

To condition the VQAG model on these control signals, the embeddings for the control signals are fed to the encoder together with the image embeddings. The visual and language features from the image embeddings and the control signals are allowed to attend to all other features through the self-attention mechanism.

Dataset indicator as additional control signal. As the main focus of this paper is cross-dataset shifts, we consider the dataset indicator control signal as an additional input. This signal helps inform the model of the desired domain or style of visual questions. Similar to dataset-agnostic control signals above, the one-hot embedding for the dataset indicator is concatenated to the image and other control signal embeddings and fed to the encoder.

3.3 Generating CrossVQA Benchmarks

We now describe how to use the enhanced VQAG model together with the dataset indicator described in previous section for generating CrossVQA benchmarks.

Datasets. We consider two VQA datasets: VQA2 (Goyal et al., 2017) and VizWiz (Gurari et al., 2018). The two datasets are drastically different visually and textually. VQA2 is built on top of high-quality COCO images (Lin et al., 2014) with visual questions intended to fool "smart robot" but not humans. VizWiz, on the other hand, is collected in-the-wild from the visually-impaired users, often with lower image quality and more conversational and simpler questions intended to be useful if answered correctly.

Additionally, we consider the images from Open Images (OID) (Kuznetsova et al., 2018), which is known to have more diverse objects than COCO (Agrawal et al., 2019).

3.3.1 Training

We mix the training splits of VQA2 and VizWiz and use that for training our VQAG. We experiment with pre-training and different combinations of dataset-agnostic control signals (Sec. 4). We leverage ground-truth control signals in the training set whenever available; question prefixes and answers are available for both datasets, while the answer categories are available on a subset of VQA2, as provided by (Krishna et al., 2019).

3.3.2 Inference

Creating Disentangled Shifts. By varying the dataset-indicator control knob of our best-performing VQAG models, we generate our desired disentangled shifts. More specifically, denote by $\langle I_A, QA_B \rangle$ a dataset with A-style images and B-style questions. We generate the following four VQA splits: VQA2-style question-answer pairs on a subset of VizWiz validation images $\langle I_{vzwz}, QA_{vqa2} \rangle$, VizWiz-style question-answer pairs on a subset of VQA2 validation images $\langle I_{vqa2}, QA_{vzwz} \rangle$, and additionally both VQA2-style and VizWiz-style pairs on a subset of OID validation images $\langle I_{oid}, QA_{vqa2} \rangle$ and $\langle I_{oid}, QA_{vzwz} \rangle$. In addition, we also generate $\langle I_{vqa2}, QA_{vqa2} \rangle$ and $\langle I_{vzwz}, QA_{vzwz} \rangle$ as a sanity check to verify if our model learns to understand the styles of VQA2 and VizWiz.

Dataset-agnostic control signals. There are no ground-truth control signals for the images during inference. Thus, we train an image tagger with the multi-label sigmoid cross entropy loss to predict top-k most relevant first two words (i.e. question prefix), answer categories, and answers from the input image and the target dataset indicator This is more flexible than the approach used in (Krishna et al., 2019) where all the pre-annotated answer categories are used during inference for all images.

3.3.3 Postprocessing

We further clean CrossVQA by using the human annotators to assess question relevance and answer correctness (Sec. 4.2).

4 Experiments

In this section, we first evaluate the performance of our VQAG model against existing state-of-the-art baselines (Krishna et al., 2019; Wang et al.,

2017; Jain et al., 2017). We then demonstrate the importance of conditioning our VQAG model on the proposed control signals by performing several ablation studies. Next, we present CrossVQA examples and several statistics based on the generated data. We finally show that CrossVQA is effective at identifying the limitations of state-of-the-art VQA models, and examine the extent to which existing adaptation techniques help in improving performance of VQA models as measured by CrossVQA.

4.1 In-Domain Evaluation of VQAG

We first benchmark the in-domain performance of our VQAG model by training and testing on VQA2 (Goyal et al., 2017) against existing models for visual question generation (VQG). Note that, unlike those models which focus on generating only questions, our model also generates answers; we discard the generated answers when evaluating the generated questions against existing work.

Metrics. We consider two sets of evaluation metrics. The first set of metrics measure **question relevance**. It consists of multiple automatic text similarity metrics widely used for image captioning and VQG: BLEU (Papineni et al., 2002), ROUGE-L (Lin, 2004), METEOR (Banerjee and Lavie, 2005), SPICE (Anderson et al., 2016) and CIDEr (Vedantam et al., 2015). The second set of metrics measure the **diversity and novelty** of questions and answers (Vijayakumar et al., 2016; Jain et al., 2017): (i) generative strength: the percentage of unique generated questions normalized by the number of unique ground truth questions, (ii) inventiveness: the percentage of unique generated questions that are unseen during training, (iii) oracle CIDEr: the maximum value of the CIDEr over a list of all references. Note that, although not considered by previous work, both generative strength and inventiveness for questions (QS and QI, respectively) can be extended to measure the diversity and novelty of generated answers as well (AS and AI, respectively).

Notation. We use X2Y to denote the model with X as input and Y as output. We use I, Q, A, C to refer to image, question, answer, and answer category, respectively. Furthermore, we use Ã to refer to multiple answers and P to question prefix. See Table 2 for examples of our control signals.

Baselines. We compare the performance of our VQAG model against the following baselines: **IA2Q** (Wang et al., 2017), a non-variational model that takes an image and answer as input and generates a question; **V-IA2Q** (Wang et al., 2017), a variational-autoencoder based approach that embeds the input image and question to a latent space before generating a question; **IC2Q** and **V-IC2Q**, extensions to the IA2Q and V-IA2Q models, respectively, where the models are conditioned on answer categories (Krishna et al., 2019) instead of ground-truth answers; **MI-IA2Q** (Krishna et al., 2019) and **MI-IC2Q**, also variational models posing the question generation as a process that maximizes mutual information between the image, the expected answer and the answer category.

Results. Results are reported in Table 3. Our models (IÃC2QÃ, IÃP2QÃ) significantly outperform all the baselines on standard automatic metrics by large margins, especially improving the BLEU-4, METEOR and CIDEr scores by +29.5%, +23.17% and +0.62, respectively, compared to the current state-of-the-art methods MI-IC2Q and MI-IA2Q. In addition, our best model IÃP2QÃ outperforms the state-of-the-art MI-IC2Q by +7.06% in QS, suggesting that we generate a diverse pool of questions. Moreover, for question inventiveness, a +30.39% QI improvement paired with a high oracle CIDEr score indicates that our model also generates novel and appropriate questions by using new combinations of objects and question patterns. We also find a +20% improvements in AS and AI with the enhancements discussed in Sec. 3.2.

4.2 Analysis of Generated Data

Now that we establish the superiority of our VQAG engine to existing approaches, we analyze the outputs of our best model (IÃP2QÃ with pre-training) when used to generate CrossVQA benchmarks (Sec. 3.3.2).

Statistics and Examples of CrossVQA. Table 4 presents basic statistics of the six CrossVQA test splits generated by our VQAG model. Figure 3 provides examples.

Human Evaluation. We first conduct a human study to verify **question relevance and answer correctness** of 3000 samples from the generated splits. More concretely, we present each <image, question, answer> triplet to three crowd workers and ask them to verify if the generated question is relevant to the image. Questions that are annotated as not relevant by at least two workers are discarded. For each of the relevant questions, we also ask the

Model	Pre-train?	B1	B4	M	R	S	C	QS	QI	AS (0−100)	AI (0−100)	OC (0−10)
IC2Q (Wang et al., 2017)	✗	30.42	4.44	9.42	-	-	0.27	11.37	34.76	-	-	-
V-IC2Q (Jain et al., 2017)	✗	35.40	10.78	13.35	-	-	0.42	12.97	38.32	-	-	-
MI-IC2Q (Krishna et al., 2019)	✗	47.40	14.49	18.35	40.27	-	0.86	26.06	52.11	-	-	-
Ours (IC2QA)	✗	55.77	27.54	22.18	49.60	21.80	0.98	27.00	53.90	2.80	11.15	2.78
Ours (IC2QA)	✓	61.34	32.01	29.09	52.18	26.03	1.15	27.94	57.00	3.79	15.00	3.12
IA2Q (Wang et al., 2017)	✗	32.43	6.23	11.21	-	-	0.36	-	-	-	-	-
V-IA2Q (Jain et al., 2017)	✗	36.91	6.25	12.39	-	-	0.36	-	-	-	-	-
MI-IA2Q (Krishna et al., 2019)	✗	48.09	15.17	18.78	49.10	-	0.92	-	-	-	-	-
Ours (IA2QA)	✗	57.12	29.00	24.16	51.13	23.69	1.02	27.20	54.09	2.90	11.20	3.02
Ours (IA2QA)	✓	63.00	34.82	30.05	55.00	27.18	1.18	28.90	58.11	3.89	16.01	3.18
Ours (IÃ2QÃ)	✓	66.02	37.15	32.00	58.16	30.62	1.20	29.10	61.09	4.96	18.89	4.56
Ours (IÃC2QÃ)	✓	75.34	42.09	**41.52**	**69.41**	38.60	1.40	33.00	80.50	22.09	39.80	4.98
Ours (IÃP2QÃ)	✓	**79.52**	**44.74**	41.01	68.20	**39.87**	**1.54**	**33.12**	**82.50**	**23.50**	**39.86**	**5.74**

Table 3: Performance of our VQAG model against the baselines using the metrics BLEU-1 (B1), BLEU-4 (B4), METEOR (M), ROUGE-L (R), SPICE (S), CIDEr (C), Question generative strength (QS) and inventiveness (QI), answer generative strength (AS) and inventiveness (AI), and oracle cider (OC). "Pre-train?" refers to whether or not we pre-train our VQAG on Conceptual 12M (Changpinyo et al., 2021).

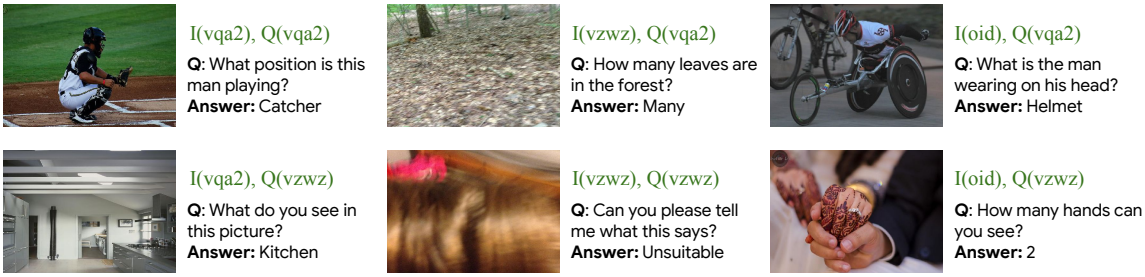

Figure 3: Qualitative examples of questions and answers in our CrossVQA dataset.

Test Set	#Images	#Questions	Question Vocab	Unique Answers
$\langle I_{vqa2}, QA_{vqa2} \rangle$	3000	8418	976	464
$\langle I_{vqa2}, QA_{vzwz} \rangle$	3000	8986	927	389
$\langle I_{vzwz}, QA_{vqa2} \rangle$	3000	8438	872	440
$\langle I_{vzwz}, QA_{vzwz} \rangle$	3000	3014	1004	325
$\langle I_{oid}, QA_{vqa2} \rangle$	3000	8986	963	332
$\langle I_{oid}, QA_{vzwz} \rangle$	3000	8986	982	427

Table 4: Statistics of CrossVQA before human validation.

Test Set	QR	AC	Categories with AC $< 30\%$
$\langle I_{vqa2}, QA_{vqa2} \rangle$	97.8	69.8	*count, time*
$\langle I_{vqa2}, QA_{vzwz} \rangle$	96.0	74.8	*count, time, spatial*
$\langle I_{vzwz}, QA_{vqa2} \rangle$	69.8	52.07	*time, food, spatial*
$\langle I_{vzwz}, QA_{vzwz} \rangle$	82.2	61.2	*food, spatial, attribute*
$\langle I_{oid}, QA_{vqa2} \rangle$	77.4	51.6	*count, time, attribute*
$\langle I_{oid}, QA_{vzwz} \rangle$	81.4	63.7	*count, time, spatial*

Table 5: Human Evaluation: question relevance (QR) and answer correctness (AC).

workers to verify if the generated answer is correct, and, if incorrect, ask them to write a correct answer (See Appendix C).

As shown in Table 5, workers annotate a large portion of the generated questions by our VQAG model as relevant (QR percentages between 77.4% and 97.8%), showcasing the effectiveness of the proposed VQAG model. Answer correctness is found to be relatively lower (AC percentages between 51.6% and 74.8%), a result that indicates that CrossVQA is a challenging new benchmark for visual question answering. We find that the questions belonging to *count*, *time*, *spatial*, *food* and *attribute* categories are relatively more difficult for our model to generate correct answers.

Further Analysis. We first assess the controllability ability of our VQAG model in the generation of VQA2-style or VizWiz-style questions. In Table 6, we use the Jensen-Shannon (JSD) divergence between the unigrams and bigrams distributions of questions between each data pair to measure their "style" distance. Regardless of the image sources, the generated VQA2-style (VizWiz-style) questions are much more similar to VQA2 (VizWiz) than the original VizWiz (VQA2) questions are.

We then focus on the generated ques-

Q_A from	Q_B from	JSD unigram	JSD bigram
VQA2	VizWiz	0.57	0.59
$\langle I_{vqa2}, QA_{vqa2} \rangle$	VQA2	0.06	0.07
$\langle I_{vqa2}, QA_{vzwz} \rangle$	VizWiz	0.09	0.08
$\langle I_{vzwz}, QA_{vqa2} \rangle$	VQA2	0.11	0.09
$\langle I_{vzwz}, QA_{vzwz} \rangle$	VizWiz	0.06	0.07

Table 6: Comparison of question distribution of source and the generated datasets measured using the Jensen-Shannon (JSD) divergence

Figure 4: Pre-training improves the ability of the VQAG model to generate questions and answers about long-tail concepts (images in the figure are from OID).

tions/answers on OID and assess the benefits of pre-training and control signals on out-of-domain images. Figure 4 shows a qualitative comparison of questions generated without (red) and with pre-training (green). We observe that the pre-trained model generates more accurate and informative questions (e.g., *fire hydrant* vs. *fire extinguisher*, *fish* vs. *shark*). In Figure 5, the sunburst plots (shown at the top) of the first three words of the questions exhibit much higher diversity with control signals. Further, in Figure 6, the distribution of answer categories demonstrate that control signals increase the entropy of answer category distribution, helping the heavy tail ones.

4.3 Cross-Dataset VQA Experiments

Performance of Existing VQA Systems on Human-Validated CrossVQA. On the 2100 human-validated CrossVQA relevant questions, we evaluate the VQA adaptation performance of the

Model	vqa2,vqa2	vqa2,vzwz	vzwz,vqa2	oid,vqa2
LXMERT	60.1	50.5	25.0	38.6
VisualBERT	58.1	55.1	21.4	43.6
ViLBERT(VB)	62.5	57.8	26.6	44.8
VB+Sum-H	62.8	57.8	26.9	43.9
VB+Max-H	64.1	58.0	26.9	42.8
VB+GQA	65.3	57.8	25.7	40.4
VB+RER	63.0	58.1	27.2	44.0
VB+VCR	61.0	54.3	24.1	39.6

Table 7: Performance on human-validated CrossVQA test sets with VQA2 images or VQA2-style questions for (i) the state-of-the-art models (top three rows) and (ii) ViLBERT (VB) with contrastive (Sum-H, Max-H) and multi-task (GQA, RER, VCR) losses.

state-of-the-art VQA models: ViLBERT (VB) (Lu et al., 2019a), LXMERT (Tan and Bansal, 2019), and VisualBERT (Li et al., 2019), all trained on VQA2. In Table 7 (top three rows), we find that ViLBERT outperforms other baselines on CrossVQA splits with VQA2 images or VQA-style questions, so we provide a detailed analysis of ViLBERT.

Figure 7 compares the CrossVQA performance of (a) ViLBERT trained on the VQA2 dataset, and (b) ViLBERT trained on VQA2 and fine-tuned on VizWiz. We find that both VQA models show accuracy drops on all six splits, compared to the SOTA accuracy 71.0% on VQA2 test set and 54.7% on VizWiz test set (left-most column). This indicates that the questions in CrossVQA are harder for SOTA models to get right. Moreover, the model trained on VQA2 drops by up to 40% on VizWiz and OID input images, a rather unexpected (and never-before quantified) result. Similarly, the model trained on VizWiz underperforms on splits with VQA and OID images by similarly large margins. This suggests that the VQA models struggle to generalize when there is a mismatch in image distribution. In contrast, the drop in accuracy is relatively low for mismatches in language distribution, indicating that these models are relatively less robust to visual features compared to language features. We believe that the rich object-level features and interactions available in the visual space could be causing the models to overfit to training image distribution and therefore the models struggle to generalize to new image distribution.

Adaptation Techniques with Auxiliary Losses. We also examine if the contrastive and multi-task (MTL) losses (Akula et al., 2020a) improve the adaptation performance of ViLBERT on CrossVQA in Table 7. In contrastive leaning, neg-

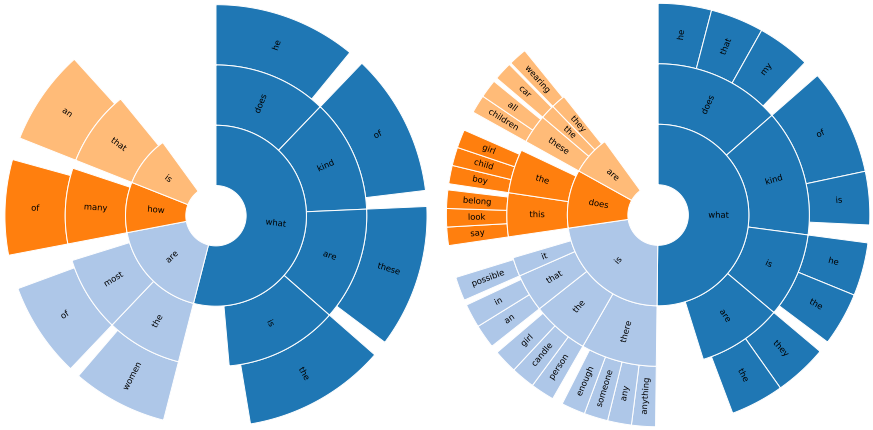

Figure 5: Distribution of the first three words for questions generated without (left) and with (right) control signals (on OID).

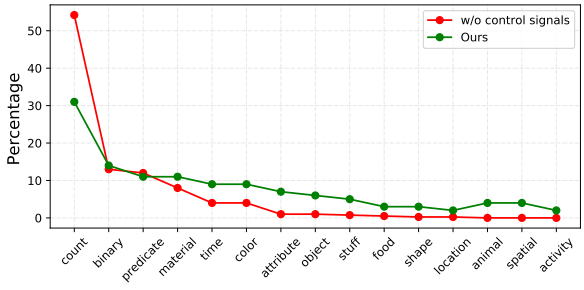

Figure 6: Distribution of answer categories generated without (red) and with (green) control signals (on OID).

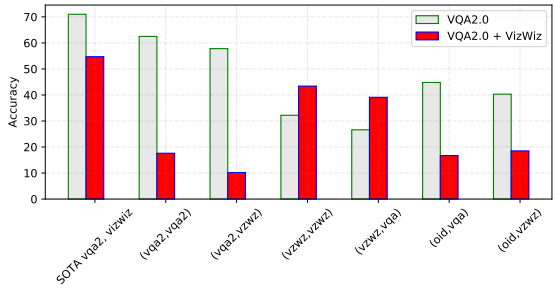

Figure 7: The CrossVQA performance of ViLBERT, trained on VQA2 only (VQA2.0) or trained on VQA2 and then fined-tuned on VizWiz (VQA2.0 + VizWiz). Left-most column indicates the reference state-of-the-art performance.

ative examples that are close to the current example are mined, and used to learn to jointly minimize the loss on the current (positive) example and maximize the loss on the (hard) negative examples. Two versions of contrastive losses are considered: Sum of Hinges (Sum-H), taking a sum over all negative samples, and Max of Hinges (Max-H), which only considers the loss on hardest negative sample by applying the max operation. For MTL, the following auxiliary tasks are used: GQA (Hudson and Manning, 2019), visual common sense reasoning (VCR) (Zellers et al., 2019a), and referring expression recognition with RefCOCOg (RER) (Mao et al., 2016). The last five rows of Table 7 show the performance of ViLBERT (VB) using these contrastive and MTL losses. Although the losses slightly improve the accuracy on in-domain CrossVQA split $\langle I_{vqa2}, QA_{vqa2} \rangle$, they fail to improve generalization on cross-domain splits $\langle I_{vqa2}, QA_{vzwz} \rangle$, $\langle I_{vzwz}, QA_{vqa2} \rangle$ and $\langle I_{oid}, QA_{vqa2} \rangle$, suggesting that there is ample room for improvement (see Appendix E).

5 Conclusion

We present a step toward scalable and systematic evaluation of VQA systems. Key to our approach is an accurate and controllable VQAG module that is capable of generating disentangled distribution shifts. We generate CrossVQA benchmarks, a collection of test splits based on VQA2, VizWiz, and Open Images datasets. We validate their utility by showing that existing VQA models struggle to perform well in this evaluation scenario and identifying the image distribution mismatch as the main factor.

Acknowledgments. We would like to thank Prof. Joyce Chai, Prof. Siva Reddy, Spandana Gella for helpful discussions, Sebastian Goodman and Nan Ding for their feedback on the code, Ashish Thapliyal on his feedback on an earlier version of the draft, Keze Wang for his help with technical issues, and Google data team for their help with human annotations. We are grateful to the anonymous reviewers for their useful feedback.

References

Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. 2018. Don't just assume; look and answer: Overcoming priors for visual question answering. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 4971–4980. IEEE Computer Society.

Harsh Agrawal, Peter Anderson, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. nocaps: novel object captioning at scale. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 8947–8956. IEEE.

Arjun Akula, Spandana Gella, Yaser Al-Onaizan, Song-Chun Zhu, and Siva Reddy. 2020a. Words aren't enough, their order matters: On the robustness of grounding visual referring expressions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6555–6565, Online. Association for Computational Linguistics.

Arjun R Akula. 2015. A novel approach towards building a generic, portable and contextual nlidb system. *International Institute of Information Technology Hyderabad*.

Arjun R. Akula, Shuai Wang, and Song-Chun Zhu. 2020b. Cocox: Generating conceptual and counterfactual explanations via fault-lines. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 2594–2601. AAAI Press.

Chris Alberti, Jeffrey Ling, Michael Collins, and David Reitter. 2019. Fusion of detected objects in text for visual question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2131–2140, Hong Kong, China. Association for Computational Linguistics.

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. Spice: Semantic propositional image caption evaluation. In *European Conference on Computer Vision*, pages 382–398. Springer.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: visual question answering. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 2425–2433. IEEE Computer Society.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Max Bartolo, Alastair Roberts, Johannes Welbl, Sebastian Riedel, and Pontus Stenetorp. 2020. Beat the AI: Investigating adversarial human annotation for reading comprehension. *Transactions of the Association for Computational Linguistics*, 8:662–678.

Nilavra Bhattacharya, Qing Li, and Danna Gurari. 2019. Why does a visual question have different answers? In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 4270–4279. IEEE.

Ronan Le Bras, Swabha Swayamdipta, Chandra Bhagavatula, Rowan Zellers, Matthew E. Peters, Ashish Sabharwal, and Yejin Choi. 2020. Adversarial filters of dataset biases. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 1078–1088. PMLR.

Soravit Changpinyo, Bo Pang, Piyush Sharma, and Radu Soricut. 2019. Decoupled box proposal and featurization with ultrafine-grained semantic labels improve image captioning and visual question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1468–1474, Hong Kong, China. Association for Computational Linguistics.

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. 2021. Conceptual 12M: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *CVPR*.

Wei-Lun Chao, Hexiang Hu, and Fei Sha. 2018a. Being negative but constructively: Lessons learnt from creating better visual question answering datasets. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 431–441, New Orleans, Louisiana. Association for Computational Linguistics.

Wei-Lun Chao, Hexiang Hu, and Fei Sha. 2018b. Cross-dataset adaptation for visual question answering. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 5716–5725. IEEE Computer Society.

Volkan Cirik, Louis-Philippe Morency, and Taylor Berg-Kirkpatrick. 2018. Visual referring expression

recognition: What do systems actually learn? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 781–787, New Orleans, Louisiana. Association for Computational Linguistics.

Yaroslav Ganin and Victor S. Lempitsky. 2015. Unsupervised domain adaptation by backpropagation. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 1180–1189. JMLR.org.

Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, and Ben Zhou. 2020. Evaluating models' local decision boundaries via contrast sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1307–1323, Online. Association for Computational Linguistics.

Boqing Gong, Yuan Shi, Fei Sha, and Kristen Grauman. 2012. Geodesic flow kernel for unsupervised domain adaptation. In *2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, June 16-21, 2012*, pages 2066–2073. IEEE Computer Society.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 6325–6334. IEEE Computer Society.

Yuhong Guo and Min Xiao. 2012. Cross language text classification via subspace co-regularized multi-view learning. In *Proceedings of the 29th International Conference on Machine Learning, ICML 2012, Edinburgh, Scotland, UK, June 26 - July 1, 2012*. icml.cc / Omnipress.

Danna Gurari, Qing Li, Abigale J. Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P. Bigham. 2018. Vizwiz grand challenge: Answering visual questions from blind people. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 3608–3617. IEEE Computer Society.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society.

Drew A. Hudson and Christopher D. Manning. 2019. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 6700–6709. Computer Vision Foundation / IEEE.

Allan Jabri, Armand Joulin, and Laurens Van Der Maaten. 2016. Revisiting visual question answering baselines. In *Proceedings of ECCV*.

Unnat Jain, Ziyu Zhang, and Alexander G. Schwing. 2017. Creativity: Generating diverse questions using variational autoencoders. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 5415–5424. IEEE Computer Society.

Da-Cheng Juan, Chun-Ta Lu, Zhen Li, Futang Peng, Aleksei Timofeev, Yi-Ting Chen, Yaxi Gao, Tom Duerig, Andrew Tomkins, and Sujith Ravi. 2019. Graph-rise: Graph-regularized image semantic embedding. *ArXiv preprint*, abs/1902.10814.

Divyansh Kaushik, Eduard H. Hovy, and Zachary Chase Lipton. 2020. Learning the difference that makes A difference with counterfactually-augmented data. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. 2019. Information maximizing visual question generation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 2008–2018. Computer Vision Foundation / IEEE.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. 2017. Visual Genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73.

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Tom Duerig, et al. 2018. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *ArXiv preprint*, abs/1811.00982.

Christoph H. Lampert, Hannes Nickisch, and Stefan Harmeling. 2009. Learning to detect unseen object classes by between-class attribute transfer. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, pages 951–958. IEEE Computer Society.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. *ArXiv preprint*, abs/1908.03557.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: common objects in context. In *ECCV*.

Changsong Liu, Shaohua Yang, Sari Saba-Sadiya, Nishant Shukla, Yunzhong He, Song-Chun Zhu, and Joyce Chai. 2016. Jointly learning grounded task structures from language instruction and visual demonstration. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1482–1492, Austin, Texas. Association for Computational Linguistics.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019a. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 13–23.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019b. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 13–23.

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L. Yuille, and Kevin Murphy. 2016. Generation and comprehension of unambiguous object descriptions. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 11–20. IEEE Computer Society.

Nasrin Mostafazadeh, Ishan Misra, Jacob Devlin, Margaret Mitchell, Xiaodong He, and Lucy Vanderwende. 2016. Generating natural questions about an image. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1802–1813, Berlin, Germany. Association for Computational Linguistics.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901, Online. Association for Computational Linguistics.

Ashish Palakurthi, Ruthu S M, Arjun Akula, and Radhika Mamidi. 2015. Classification of attributes in a natural language query into different SQL clauses. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 497–506, Hissar, Bulgaria. INCOMA Ltd. Shoumen, BULGARIA.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari. 2019. Connecting vision and language with localized narratives. *ArXiv preprint*, abs/1912.03098.

Hang Qi, Tianfu Wu, Mun-Wai Lee, and Song-Chun Zhu. 2015. A restricted visual turing test for deep scene and event understanding. *ArXiv preprint*, abs/1512.01715.

Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. 2015. Faster R-CNN: towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 91–99.

Ruslan Salakhutdinov, Antonio Torralba, and Joshua B. Tenenbaum. 2011. Learning to share visual appearance for multiclass object detection. In *The 24th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2011, Colorado Springs, CO, USA, 20-25 June 2011*, pages 1481–1488. IEEE Computer Society.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, Melbourne, Australia. Association for Computational Linguistics.

Hao Tan and Mohit Bansal. 2019. LXMERT: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on*

Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 5100–5111, Hong Kong, China. Association for Computational Linguistics.

Ashish V. Thapliyal and Radu Soricut. 2020. Cross-modal Language Generation using Pivot Stabilization for Web-scale Language Coverage. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 160–170, Online. Association for Computational Linguistics.

Kewei Tu, Meng Meng, Mun Wai Lee, Tae Eun Choe, and Song-Chun Zhu. 2014. Joint video and text parsing for understanding events and answering queries. *IEEE MultiMedia*, 21(2):42–70.

Eric Tzeng, Judy Hoffman, Trevor Darrell, and Kate Saenko. 2015. Simultaneous deep transfer across domains and tasks. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 4068–4076. IEEE Computer Society.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 4566–4575. IEEE Computer Society.

Ashwin K Vijayakumar, Michael Cogswell, Ramprasath R Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. 2016. Diverse beam search: Decoding diverse solutions from neural sequence models. *ArXiv preprint*, abs/1610.02424.

Tong Wang, Xingdi Yuan, and Adam Trischler. 2017. A joint model for question answering and question generation. *ArXiv preprint*, abs/1706.01450.

Xing Xu, Tan Wang, Yang Yang, Alan Hanjalic, and Heng Tao Shen. 2020. Radial graph convolutional network for visual question generation. *IEEE Transactions on Neural Networks and Learning Systems*.

Jianwei Yang, Jiasen Lu, Stefan Lee, Dhruv Batra, and Devi Parikh. 2018. Visual curiosity: Learning to ask questions to learn visual recognition. *ArXiv preprint*, abs/1810.00912.

Shaohua Yang, Qiaozi Gao, Changsong Liu, Caiming Xiong, Song-Chun Zhu, and Joyce Y. Chai. 2016. Grounded semantic role labeling. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 149–159, San Diego, California. Association for Computational Linguistics.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019a. From recognition to cognition: Visual commonsense reasoning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 6720–6731. Computer Vision Foundation / IEEE.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019b. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.

Xiangxin Zhu, Dragomir Anguelov, and Deva Ramanan. 2014. Capturing long-tail distributions of object subcategories. In *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*, pages 915–922. IEEE Computer Society.

Xiangxin Zhu, Carl Vondrick, Charless C Fowlkes, and Deva Ramanan. 2016a. Do we need more training data? *International Journal of Computer Vision*, 119(1):76–92.

Yuke Zhu, Oliver Groth, Michael S. Bernstein, and Li Fei-Fei. 2016b. Visual7w: Grounded question answering in images. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 4995–5004. IEEE Computer Society.

A Appendix

In this supplementary material, we begin by providing more details on our VQAG implementation. We then provide additional results and detailed analysis comparing the diversity, novelty of questions generated using our VQAG model and baselines. Next, we present our experiment interfaces used for conducting human studies and show additional results. We then present details of contrastive learning and multi-task learning models used in our adaptation analysis. Finally, we present more statistics from CrossVQA.

B Implementation Details

The models are optimized with Adam (Kingma and Ba, 2015) with an initial learning rate of 0.000032. We use a linear decay learning rate schedule with warm up and employ early stopping based on validation set accuracy. If not pre-trained, we train our VQAG model for a maximum of 2M iterations. With pre-trained initialization, we train our VQAG model for a maximum of $500,000$ iterations. Both the encoder and decoder layers of transformer have 6 layers each with 8 heads for multiheaded attention. The vocabulary embedding size is 512, and the hidden embedding size is 1024. We train our models with a global batch size of 4096 over Google Cloud 32-core TPUs[2]. The average training time for pre-training on conceptual captions dataset is 52 hours, and training on VQA2.0 and VizWiz takes up to 21 hours.

We condition our VQAG model using the expected answer categories ($\tilde{A}$) of the output answer as one of the control signals, in order to maximize the relevance between image, question and expected answer in the generated test sets. These answer categories can be objects, attributes, colors, materials, time, etc. Specifically we use 16 categories (similar to (Krishna et al., 2019)), covering more than 80 objects, 40 attributes, 17 colors, and 8 materials. Table 8 presents the list of all the 16 categories and provides examples of answers for each of the categories.

The decoder generates the question and the answer(s) separated by delimiters, for example, question $\langle sep \rangle$ answer1 $\langle dsep \rangle$ answer2. We use beam search (width = 5, alpha = 0.6) to generate the target question and answer(s) during decoding.

[2]https://cloud.google.com/tpu/

Categories	Examples
Count	0, 1, 2, 30, 40, 200, many, lot, very
Binary	yes, no
Predicate	on ground, on plate
Material	wood, plastic, concrete, oak, plaid
Time	afternoon, sunset, morning, spring
Color	white, blue, red, black
Attribute	sunny, male, winter, stripes, open
Object	frisbee, water, grass, skateboard, phone
Stuff	sky
Food	vegetables, tomato, salad, milk, dessert
Shape	rectangle, triangle, oval, round
Other	nothing, english, electricity, united
Location	living room, beach, ocean, mountains
Animal	cat, dog, zebras, person, police
Spatial	right, left, front, downhill, north
Activity	skateboarding, standing, playing wii

Table 8: Answer categories in our VQAG Model

C More Results on Diversity and Novelty

In Section 4 of the main paper, we show that control signals improve the diversity and novelty of the generated questions through the metrics question generative strength (QS) and inventiveness (QI), answer generative strength (AS) and inventiveness (AI). To do this, we trained our VQAG model on VQA2.0 train split and evaluated the model performance on the in-domain VQA2.0 val split. In this section, we additionally show the performance of VQAG model on out-of-domain (o.o.d) splits, namely, VizWiz val split and OID val split. Figure 8 shows the results. As we can see, there is no significant drop in QS and AS on o.o.d splits, suggesting the superior generalization skills of our model. Moreover, increase in QI and AI indicates that model is relatively more creative in inventing new questions and answers on o.o.d splits compared to in-domain splits. Table 9 presents exam-

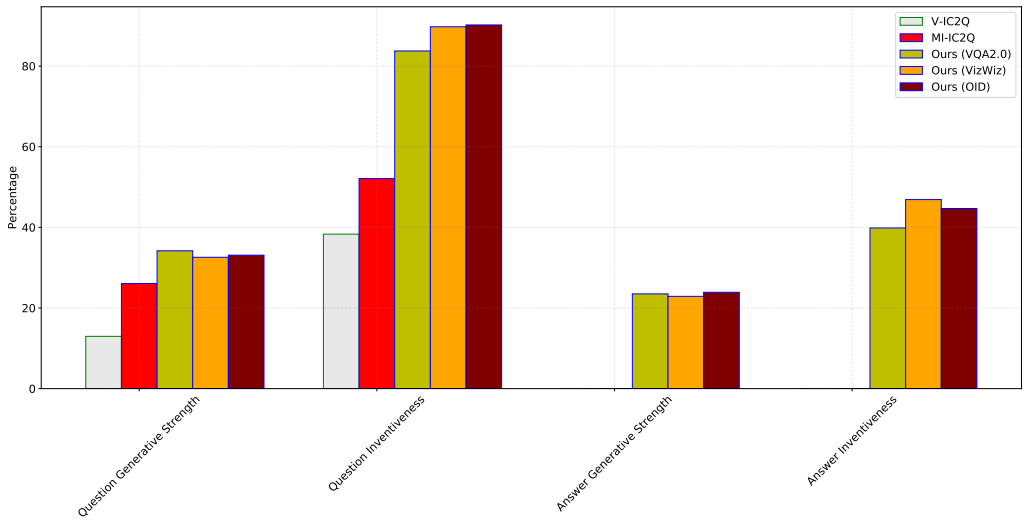

Figure 8: Diversity and Novelty of our VQAG model on out-of-domain splits: VizWiz val split and OID val split.

Examples of Invented Questions	Examples of Invented Answers
Q1: What hand is the man using to write with?	{at least 10 years, above doorway, inside the baggage, behind red car, towards bottom left side, dirt bikes, fishing boats, fork and sharp knife, riding big elephants, right side of road}
Q2: Are most of the lights on or off in the living room?	
Q3: Will this woman be drinking beer?	
Q4: What is the number on the front side of the bike?	
Q5: In this scene how many sheep can be clearly seen?	
Q6: What is the purpose of the number on the yellow board?	
Q7: Which sheep is the older in the picture?	
Q8: Is the fire hydrant old or new?	
Q9: What is the first letter of the word on the blue sign?	
Q10: What is the name of the logo on top of the keyboard?	

Table 9: Examples of unseen questions and answers invented by our VQAG Model

ples of the invented/unseen questions and answers that are not seen by our VQAG model during training. In the next section, we verify the question relevance and answer correctness of these o.o.d questions.

D Additional Human Evaluation Results

We verify question relevance and answer correctness of the samples in CrossVQA splits where the VQAG model is trained on combined train sets of VQA2.0 and VizWiz. In this section, we present additional results on human evaluation of VQAG model that is trained on only VQA2.0 train split. We generate questions and answers for VQA2.0 val split (in-domain) and VizWiz, OID val splits (o.o.d). Figure 9 shows the interface used for conducting this study. Questions that are annotated as not relevant by at least two workers are considered as irrelevant. For each of the relevant questions, we ask the workers to verify if the generated answer is correct, and if incorrect, ask them to write the correct answer. Table 10 present human evaluation results. A significant portion of generated questions are annotated as relevant. Moreover, we do not find significant differences in QR and AC metrics across in-domain and o.o.d samples, confirming that the higher percentage of invented questions on o.o.d splits (in Figure 9) are indeed relevant and not due to random noise. Furthermore, in Table 10, we also show the QA and AC percentages across seen and unseen questions generated by VQAG model. We see higher drop in AC percentage on unseen questions compared to the drop in QR, indicating that unseen questions are relatively harder for the model to generate correct answers.

E More Details on our Base Model

Both the encoder and the decoder contain a stack of L layers, with each layer consisting of a multi-head self-attention layer followed by a feedforward layer. For a given token embedding, the self-attention layer produces a weighted representation of all other tokens in the input. This weighted representation is then combined with the input representation

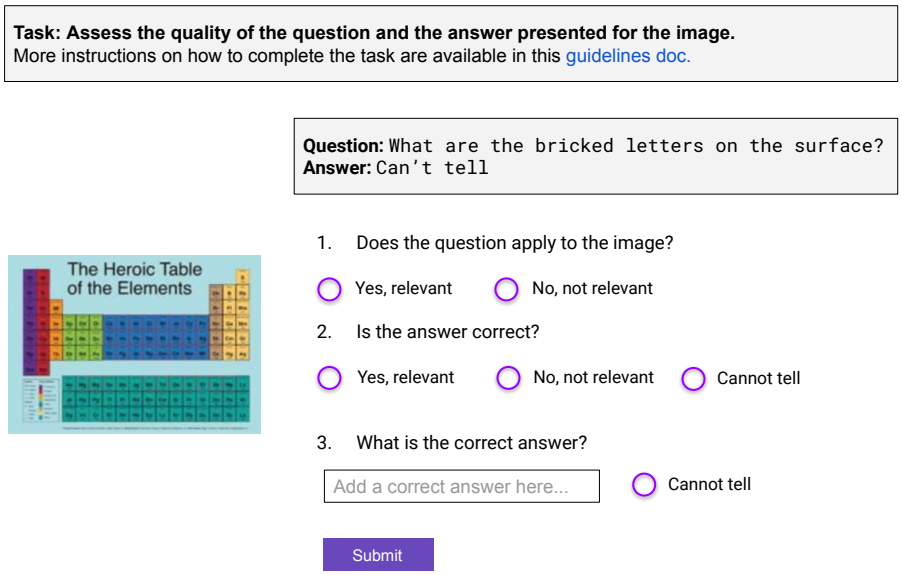

Figure 9: Experiment interface for human evaluation to verify question relevance and answer correctness.

	Seen+Unseen		Seen		Unseen	
	QR	AC	QR	AC	QR	AC
VQA2.0 val split	90.6	61.7	93.2	74.7	84.6	58.8
VizWiz val split	91.2	54.2	92.8	59.7	86.1	48.3
OpenImages val split	88.8	57.0	89.1	60.9	85.7	49.1

Table 10: Comparison of question relevance (QR) and answer correctness (AC) on in-domain val splits (VQA2.0) and out-of-domain splits (VizWiz, OpenImages).

of the given token and it is passed to the next layer.

Specifically, each attention head first calculates the queries Q, keys K and values V as follows:

$$Q = XW_Q, K = XW_K, V = XW_V \quad (1)$$

where X contains all the input features stacked into a matrix, and W_Q, W_K, and W_V are learned projection matrices.

The output of the attention head is then computed as follows:

$$\text{ATTN}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V \quad (2)$$

where d_k, d_v are the dimension of the keys K and values V respectively. Intuitively, with the above attention, the encoder jointly attends to information from different representation subspaces at different positions in the input image.

The point-wise feedforward network (FFN) is applied to each output of the attention layer and it consist of two linear transformations, with a ReLU activation in between,

$$\text{FFN}(x) = max(0, xW_1 + b_1)W_2 + b_2 \quad (3)$$

where W_1, b_1 and W_2, b_2 are the weights and biases of two fully connected layers.

Embedding Regional Image Features We extract image objects and their features using a Faster RCNN (Ren et al., 2015) object detector model, trained on Visual Genome (Krishna et al., 2017). We extract 100 object regions per image. The resulting bounding boxes are considered as *visual tokens*. Similar to the positional encoding in language models (Vaswani et al., 2017), for each visual token, the spatial position of bounding box is also encoded. We use a 5-d vector, $p_{spatial}$, to encode the top-left, bottom-right, and the bounding box area relative to the image, i.e., $p_{spatial} = [\frac{x_{tl}}{W}, \frac{y_{tl}}{H}, \frac{x_{br}}{W}, \frac{x_{br}}{H}, \frac{w \cdot h}{W \cdot H}]$.

Embedding Global Image Features Similar to (Thapliyal and Soricut, 2020; Changpinyo et al., 2019; Pont-Tuset et al., 2019), we also use a global image representation using the Graph-RISE model (Juan et al., 2019), a ResNet-101 model (He et al., 2016) trained for image classification at ultra-

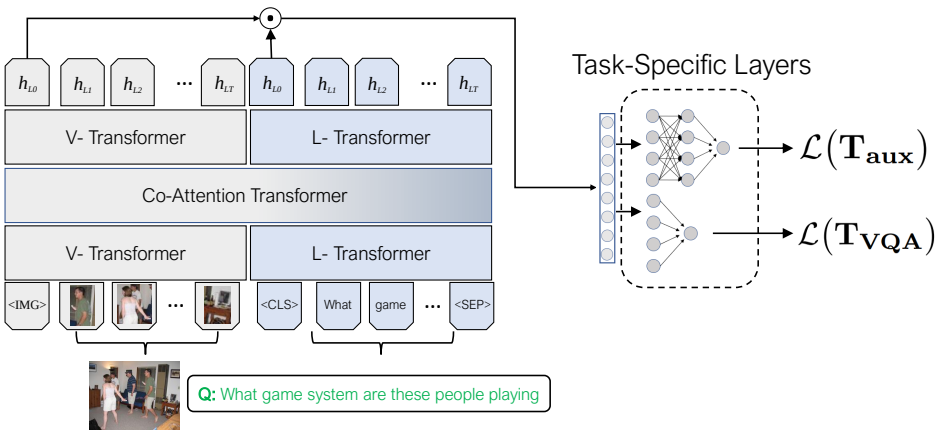

Figure 10: Multi-task learning model for VQA with auxiliary tasks such as GQA, REF, and VCR.

fine granularity levels. These regional and global image features $f_I = (f_r, f_g)$ are fixed during training.

$$f_r = \text{RCNN}(I; \theta_{RCNN})$$
$$f_g = \text{GraphRISE}(I; \theta_{GraphRISE}) \quad (4)$$

F Models for Adaptation Analysis

ViLBERT Training: As discussed in Section 4 of the main paper, we use **ViLBERT** (Lu et al., 2019b) for our adaptation experiments. ViLBERT uses a pretrain-then-transfer learning approach to jointly learn visual and textual representations from large-scale data, and utilizes them to answer VQA questions. Specifically, we consider 8-layer ViLBERT implementation available at the link https://github.com/jiasenlu/vilbert_beta. On VQA train splits, we train the model for a maximum of 25 epochs and use early-stopping based on the validation performance. We use an initial learning rate of $3e^{-5}$ and use a linear decay learning rate schedule with warm up. We train on 8 Tesla V100 GPUs with a total batch size of 512.

Contrastive Learning using ViLBERT: In implementing the contrastive loss functions, we randomly sample negatives from the mini-batch for computational efficiency (similar to (Akula et al., 2020a)). We sampled 64 negatives from each batch for both Sum-H and Max-H losses and fine-tune the margin parameters based on development split.

Multi-Task Learning using ViLBERT: We present our multi-task learning (MTL) architecture in Figure 10. The shared layers of ViLBERT constitute transformer blocks (TRM) and co-attentional transformer layers (Co-TRM) (Lu et al., 2019b). The weights for the task-specific layers are randomly initialized, whereas the shared layers are initialized with weights pre-trained on 3.3 million image-caption pairs from Conceptual Captions dataset (Sharma et al., 2018). We use a binary cross-entropy loss for all the auxiliary tasks GQA (Hudson and Manning, 2019), visual common sense reasoning (VCR) (Zellers et al., 2019a), and referring expression recognition (REF) (Cirik et al., 2018). We considered RefCOCOg (Mao et al., 2016) dataset for REF task. We optimize each task alternatively in mini-batches based on a mixing ratio and employ early-stopping based on the validation performance. In all our contrastive learning and multi-task learning experiments, we use an initial learning rate of 4e-5, and use a linear decay learning rate schedule with warm up. We train on 4 RTX 2080 GPUs with a total batch size of 256.

Transfer Learning using ViLBERT: In addition to the contrastive learning and MTL based adaptation results presented in Section 4 of main paper, we also explore transfer learning (TL) based models. Specifically, we first pre-train ViLBERT on auxiliary tasks, in contrast to joint training in MTL, and then fine-tune it on VQA train split. As shown in Table 11, we did not find any significant improvement in model's performance on CrossVQA.

G More Details on CrossVQA

In addition to the statistics presented in the Section 4 of the main paper, we present additional details of our CrossVQA splits. Figure 12a and Figure 12b show a word cloud plot for the majority

Model	vqa2,vqa2	vqa2,vzwz	vzwz,vqa2	oid,vqa2
VB	62.5	57.8	26.6	44.8
VB+TL(GQA)	59.3	57.9	26.0	42.1
VB+TL(REF)	58.4	54.2	24.1	40.2
VB+TL(VCR)	59.7	56.3	25.0	41.4

Table 11: Adaptation Results on CrossVQA with Transfer Learning

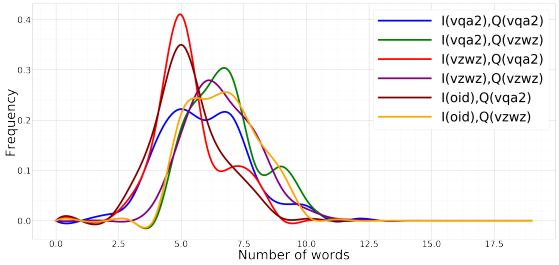

Figure 11: Question length distribution for all the six CrossVQA splits.

questions and answers across all the six splits. A variety of objects and answers can be seen in the plots, suggesting that our splits are diverse. Moreover, the relative frequency of the most frequent spatial relationships across all the six splits in Figure 13 show that CrossVQA comprises of rich and diverse spatial relationships. Figure 11 shows question length distribution of all the six splits. As we expected, we find that splits with VizWiz style questions, i.e. $\langle I_{vqa2}, QA_{vzwz} \rangle$, $\langle I_{vzwz}, QA_{vzwz} \rangle$, and $\langle I_{oid}, QA_{vzwz} \rangle$ contain more words in the question on average than other splits in CrossVQA.

(a) (b)

Figure 12: Wordcloud for (a) questions and (b) answers across all the CrossVQA splits.

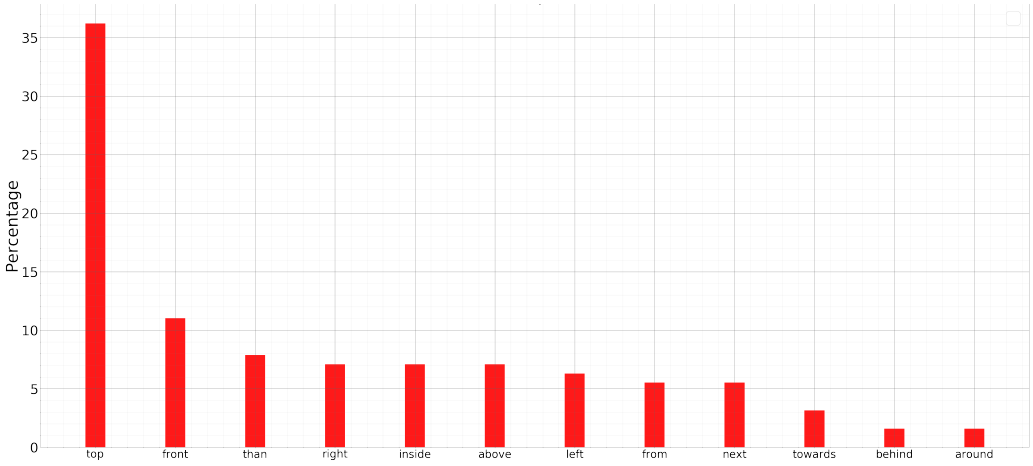

Figure 13: Relative frequency of the most frequent spatial relationships in CrossVQA.