

Supplementary- Towards Causal VQA: Revealing and Reducing Spurious Correlations by Invariant and Covariant Semantic Editing

Vedika Agarwal^{1,3*} Rakshith Shetty¹ Mario Fritz²

¹Max Planck Institute for Informatics ² CISPA Helmholtz Center for Information Security ³TomTom
Saarland Informatics Campus Saarland Informatics Campus

vedika.agarwal@tomtom.com

rakshith.shetty@mpi-inf.mpg.de

fritz@cispa.saarland

We structure the supplementary material as follows: In Section A, we discuss further details about the human validation study. In Section B, we describe the VQA models used by us and mention the various hyperparameters used in section B.1. Following which are more visualizations showing predictions of the three VQA models on the original and the synthetic images from our proposed IV-VQA and CV-VQA datasets in section B.2. Also included in section B.2 are an analysis showing how the area of the removed object influences the flip rate and some attention maps for SAAA model on IV-VQA dataset. Section C includes accuracy-flipping numbers for all the models finetuned using real vs real+edit IQAs for different question types for both IV-VQA and CV-VQA datasets along with visualizations. Finally in Section D, we discuss a possible direction to introduce causality into VQA.

A. Synthetic Dataset for Variances and Invariances in VQA

A.1. Human Validation

In order to make sure that our consistency analysis holds and flipping is not due to errors in synthetic dataset, we collect all those IQAs for which labels flip (positively or negatively) for any of the three models (27621 IQAs, 25% of IV-VQAtest set). Of this 25%, we randomly sample 100 IQA from each of the 65 question categories [3] if possible. this results in a total of 4960 edited IQA. Flipping of answers is bad and this number becomes our foundation for the robustness comparisons we make, so it was important for us to get this number validated. For each IQA, the annotator is asked to say if the answer shown is correct for the given image and question (yes/no/ambiguous). We get these numbers validated by three humans and report the results in Table 1. The study reveals that our edited IQA holds 91.3% times according to all three humans. Additionally for

	Yes(%)	No(%)	Ambiguous(%)
User1	97.58	0.89	1.53
User2	96.47	1.15	2.38
User3	94.94	2.5	2.56
User1 \cap User2 \cap User3	91.31	0.04	0.04
User1 \cup User2 \cup User3	99.6	3.87	5.68

Table 1: Human Validation of the edited set: If the given answer is valid for the Image-Question pair.

3.97% IQAs: atleast one of them found it false and 5.68% IQAs- seem to be ambiguous by atleast one of them.

B. Experiments: Consistency analysis

B.1. Models Training

We select three models for our comparison. The first one is a basic CNN+LSTM model, where we use ResNet152 [4] pre-trained on Image-Net [2], to embed the images. The question features are obtained by feeding the tokenized and encoded input question embeddings into the LSTM. The features are then concatenated and fed to the classifier to infer the answer. Secondly, we use an attention model- Show, Ask, Attend and Answer (SAAA) described by Kazemi *et al.* in [8]. The aim of the attention models is the identify and use local image features with different weights. After processing images using ResNet and feeding tokenized questions to LSTM, the concatenated image features and the final state of LSTMs are used to compute multiple attention distributions over image features. Lastly, we use a compositional model, SNMN [7]. The model consists of three different components: layout controller to decompose the question into a sequence of sub-tasks, set of neural modules to perform the sub-tasks and a differentiable memory stack.

For training these models, we use the codes available online with the specified hyperparamters. For SNMN, we use

*Work done at Max Planck Institute for Informatics, Saarland Informatics Campus.

official code available to train the model, [6]. For training SAAA, we use the code available online, [9]. We modified the available SNMN code in order to get CNN+LSTM model- we just removed the attention layers from the network. As we use the validation split for consistency evaluation and testing, we cannot let the models train on it. We keep aside the validation set for testing, and only the training split is used to train the models. All these models use standard Cross Entropy Loss and follow the standard VQA practices. We follow all respective pre-processing and training procedure given on the github sites (SNMN: link [6], SAAA link: [6] for pre-processing IQAs and for training. For SAAA and CNN+LSTM: ResNet152, conv layer-4 is used to extract $14 * 14$ features for image whereas SNMN uses ResNet 152, layer-5 resulting into $7 * 7$ features. The learning rate used to train each model is e^{-3} , batch size for learning is set to 128 for all 3 models.

B.2. Visualizations

Figures 3, 4 show the predictions of 3 models on original and edited IQA from IV-VQAdataset. We expect the models to make consistent predictions across original and edited images. However we see that this isn't the case.

Figure 5 shows the predictions on original and edited IQA from CV-VQAdataset. Here we expect the models to maintain n/n-1 consistency. Counting is a hard problem for VQA models and enforcing consistency seems to break these models completely.

Area of the object removed vs flip. To study the correlation of area of the object removed on different types of flips, we plot the flip rate for different area ranges for objects being removed in Figure 1. According to our analysis, there is no large dependence between removed object area and the flip rate. For example, for objects of size 0-1% of the image area, pos→neg flip rate was about 7% for CNN-LSTM and for objects of size 9-10%, the flip rate only marginally higher at 8.7%.

Heatmaps for inspection. SAAA[8] has attention mechanisms incorporated in its architecture. The model uses concatenated image features and final state of LSTMs to compute multiple attention distributions over image features. One would expect these attention maps to provide a clue as to where the model is looking in order to explain the flipping behaviour under editing. To see if this is true we visualize the attention maps for SAAA on original/edited examples in Figure 2. On the bottom of every image, we visualize the corresponding attention distributions produced by the model. As we can see from the figure, the heatmaps are not conclusive. They are diffuse and does not clearly show one object where the model pays attention to.

	C+L (%)	SAAA (%)	SNMN (%)
Accuracy orig	60.21	70.26	66.04
Predictions flipped	17.15	7.53	6.38
neg→pos	6.79	2.71	2.54
pos→neg	7.34	3.42	2.84
neg→neg	3.02	1.39	1.01

Table 2: Accuracy-flipping on VQA-IR edit test split with zero overlap.

C. Robustification by Data Augmentation

C.1. Models Performance

For our fine-tuning experiments, we use a strict subset of IV-VQA with an overlap score of zero. As promised in the paper, Table 2 shows the accuracy-flipping analysis for all the models on this strict subset. As we see, the numbers are comparable to the model's performance on the overall set.

C.2. InVariant VQA Augmentation

In Table 3, we show the accuracy-flipping analysis for all the specialized models. Figure 6 shows some examples where using additional synthetic data makes models more consistent. In the paper, we show a compact representation of the table by plotting the reduction in flips/improvement in accuracy for models finetuned using real+edit data relative to the models finetuned only using the real data. For instance, for question-type 'is this a' for CNN+LSTM: we see there is $(12.72-9.77)/12.72$ which is about 23% reduction in flips. These numbers show that using synthetic data always leads to a reduction in flips and in some cases- also results in improved accuracy on the original VQA set. Figure 7 shows some of the examples where using the synthetic data makes the models n/n-1 consistent and accurate as well.

C.3. CoVariant VQA Augmentation

Table 4 shows the accuracy/flips for models fine-tuned using real/ real+synthetic IQAs. In the paper we compress the information in the form of plot as we do in the case of IV-VQAugmentation.

D. Outlook on building causal VQA models

In recent works [1, 5], image classifiers are taught to rely on causal features by imposing regularization across data from different environments/ identities. A requirement for this is to have pairs of data points where the only change is in non-causal features, so one can regularize the network response to these. In our work, we explicitly create such data for the VQA task. We believe, future work can exploit this data for imposing consistency losses across original/edited IQA triplets while training or providing part of the causal structure as part of the supervision.

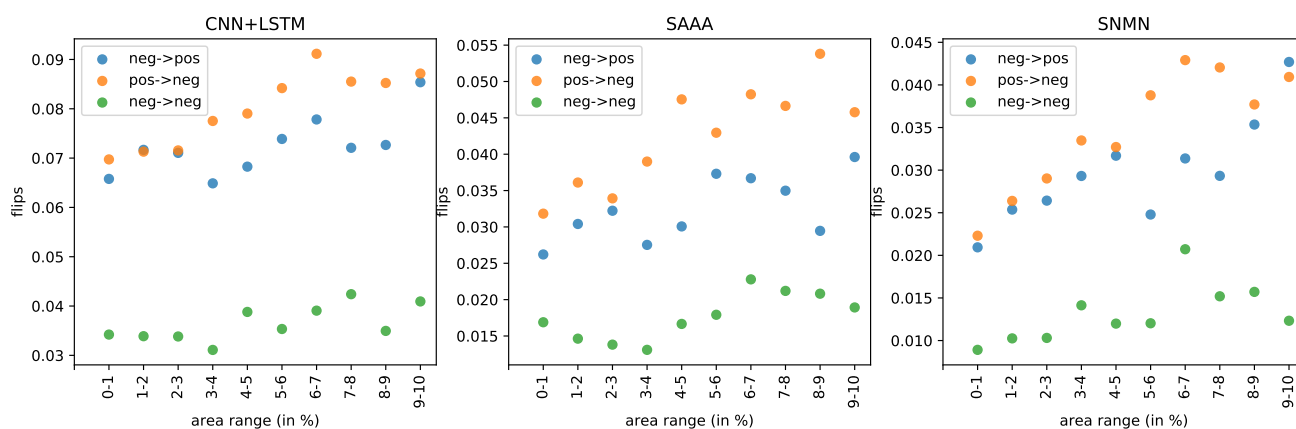


Figure 1: Flip rate vs. area of the object.

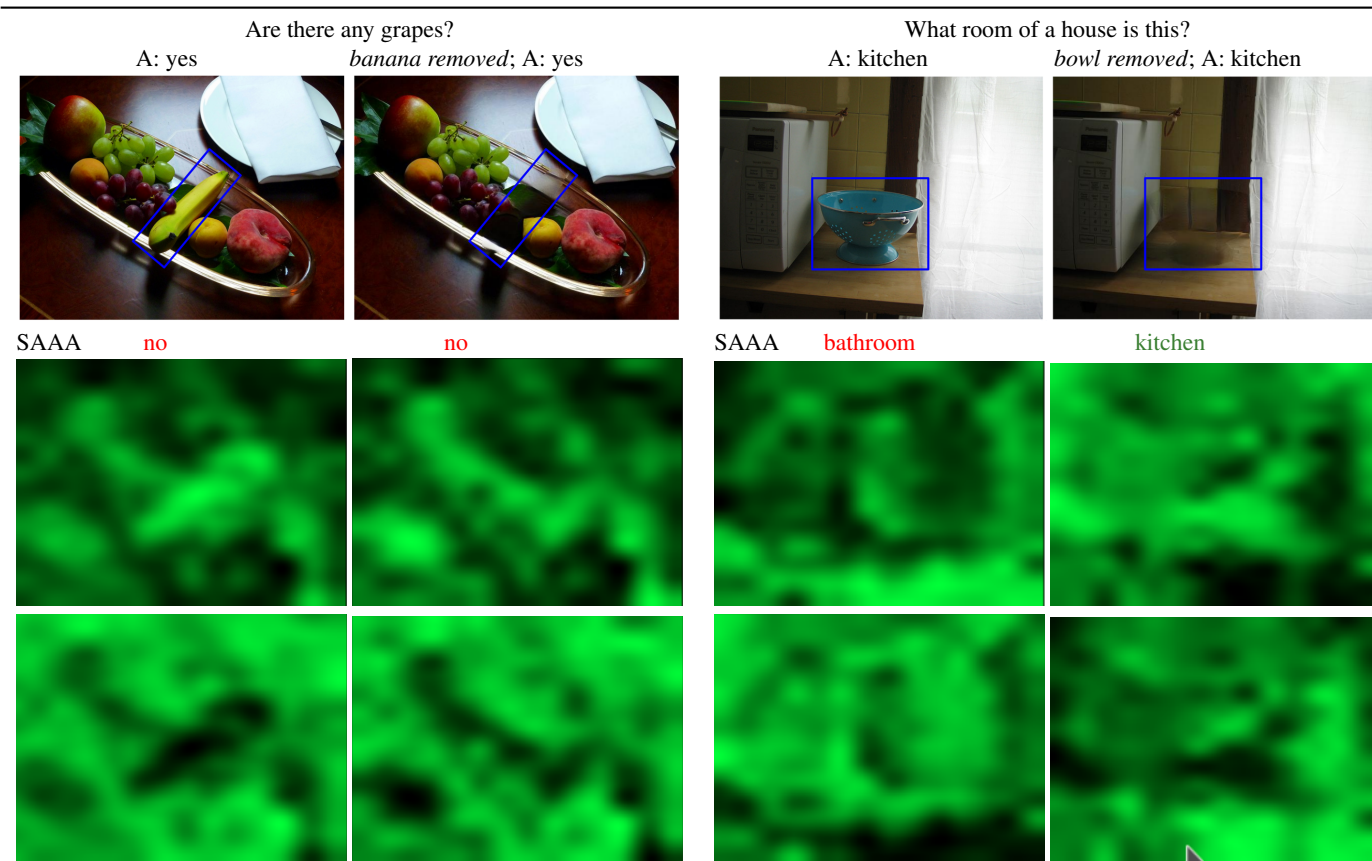


Figure 2: Shown above are the attention maps for SAAA on original and edited images from our synthetic dataset IV-VQA. The attention maps are diffuse and does not clearly show one object where the model pays attention to.

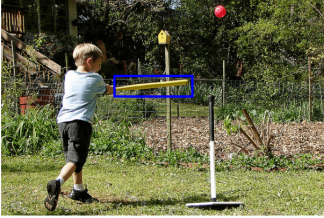
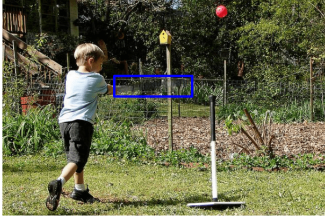




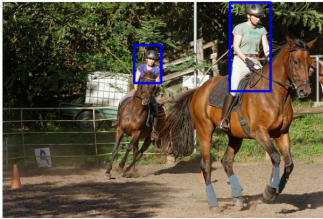
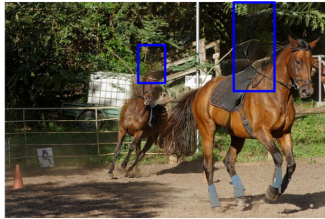


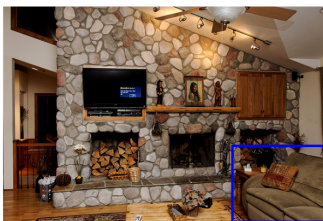
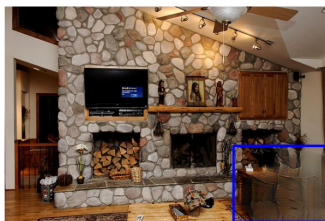


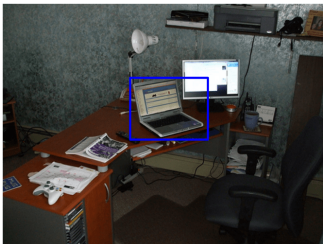
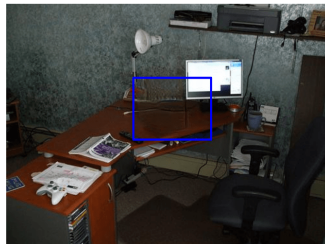
Q: What color is the bird house? A: yellow		Q: What color is the sauce? A: red	
<i>baseball bat removed; A: yellow</i>		<i>cup removed; A: red</i>	
			
CNN+LSTM	yellow	CNN_LSTM	red
SAAA	white	SAAA	orange
SNMN	yellow	SNMN	orange
Q: What color is the toilet seat? A: white		Q: What color is cone? A: orange	
<i>sink removed; A: white</i>		<i>person removed; A: orange</i>	
			
CNN_LSTM	green	CNN_LSTM	brown
SAAA	green	SAAA	orange
SNMN	green	SNMN	white
Is this a kite? A: yes		Q: Is this a museum? A: no	
<i>backpack removed; A: yes</i>		<i>couch removed; A: no</i>	
			
CNN_LSTM	yes	CNN_LSTM	no
SAAA	no	SAAA	yes
SNMN	yes	SNMN	no
How many bowls of food are there? A: 2		How many desk lamps are there? A: 1	
<i>bottle removed; A: 2</i>		<i>laptop removed; A: 1</i>	
			
CNN+LSTM	3	CNN+LSTM	2
SAAA	3	SAAA	0
SNMN	2	SNMN	1

Figure 3: Models tend to look at different objects while predicting the answers. Shown above are the models' predictions on original and edited images from our synthetic dataset IV-VQA.

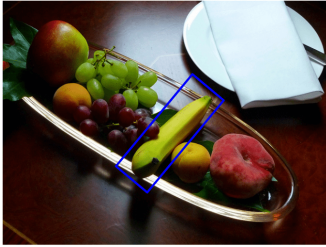
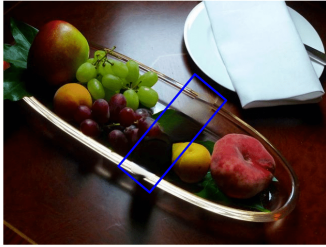


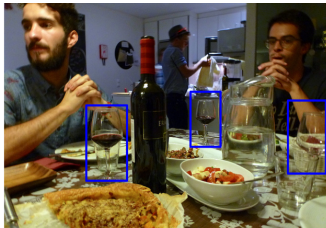
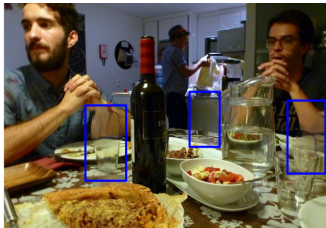

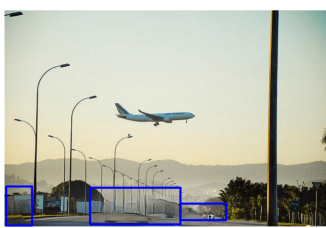





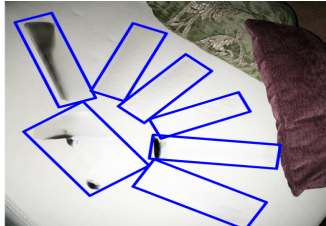


Are there any grapes?		Is there a trash can?	
A: yes	<i>banana removed; A: yes</i>	A: yes	<i>toilet removed; A: yes</i>
			
CNN+LSTM	yes	CNN+LSTM	no
SAAA	no	SAAA	yes
SNMN	no	SNMN	yes
What is the liquid in the pitcher?		Is there an airport nearby? [car]	
A: water	<i>wine glass removed; A: water</i>	A: yes	<i>car removed; A: yes</i>
			
CNN+LSTM	wine	CNN+LSTM	yes
SAAA	wine	SAAA	yes
SNMN	wine	SNMN	yes
What is in the sky?		What is room to the right called?	
A: nothing	<i>airplane removed; A: nothing</i>	A: kitchen	<i>toilet removed; A: kitchen</i>
			
CNN+LSTM	plane	CNN+LSTM	bathroom
SAAA	plane	SAAA	bathroom
SNMN	plane	SNMN	bathroom
What is the purple thing?		What are the kids doing?	
A: pillow	<i>remote removed; A: pillow</i>	A: petting horse	<i>bench removed; A: petting horse</i>
			
CNN+LSTM	scissors	CNN+LSTM	racing
SAAA	remote	SAAA	standing
SNMN	remote	SNMN	playing

Figure 4: Existing VQA models are brittle to semantic variations in the images. Shown above are examples showing different sorts of flips for IV-VQA

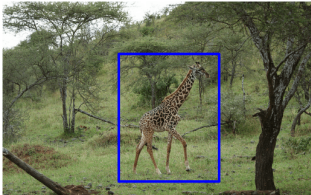
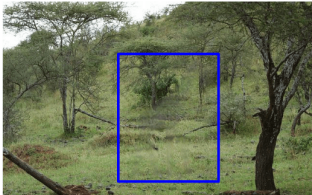





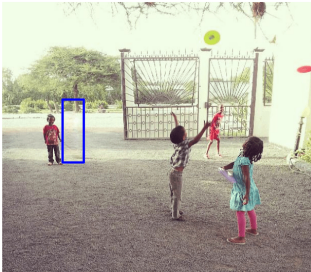






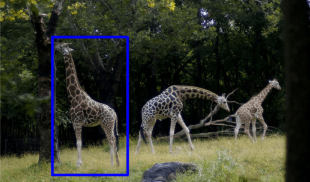
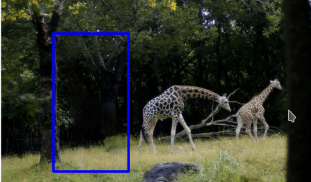
How many animals?		How many planes are in the air?	
A: 1	<i>giraffe removed; A: 0</i>	A: 1	<i>plane removed; A: 0</i>
			
CNN_LSTM	1	CNN_LSTM	1
SAAA	1	SAAA	1
SNMN	1	SNMN	1
How many clocks are there?		How many children are there?	
A: 2	<i>clock removed; A: 1</i>	A: 5	<i>child removed; A: 4</i>
			
CNN+LSTM	2	CNN+LSTM	2
SAAA	2	SAAA	2
SNMN	2	SNMN	2
How many horses are in the picture?		How many zebras?	
A: 1	<i>horse removed; A: 0</i>	A: 2	<i>zebra removed; A: 1</i>
			
CNN_LSTM	2	CNN_LSTM	3
SAAA	2	SAAA	3
SNMN	1	SNMN	2
How many people are in the image?		How many giraffes are here?	
A: 1	<i>person removed; A: 0</i>	A: 3	<i>giraffe removed; A: 2</i>
			
CNN_LSTM	1	CNN_LSTM	2
SAAA	1	SAAA	2
SNMN	1	SNMN	3

Figure 5: Shown above are models' predictions on original and edited images from CV-VQA.

	CL (%)	SAAA (%)	SNMN (%)
what color is the			
Acc orig	65.48 → 65.06	82.12 → 83.75	78.78 → 80.1
Pred flipped	11.79 → 10.89	7.25 → 6.27	7.41 → 7.21
is there a			
Acc orig	61.96 → 63.61	69.44 → 69.36	71.32 → 72.26
Pred flipped	13.3 → 10.81	8.75 → 7.51	8.83 → 7.79
is this a			
Acc orig	64.99 → 64.87	74.33 → 72.84	76.54 → 76.79
Pred flipped	12.72 → 9.77	6.09 → 5.14	6.96 → 6.52
how many			
Acc orig	43.24 → 43.2	50.38 → 50.12	49.71 → 50.56
Pred flipped	21 → 20.1	13.35 → 11.04	14.04 → 13.35
counting			
Acc orig	42.87 → 43.58	51.05 → 49.94	50.22 → 50.26
Pred flipped	21.08 → 19.06	12.81 → 12.60	14.76 → 12.96

Table 3: IV-VQAAugmentation: numbers on the left side of the arrow denote the accuracy/flipping for models finetuned using just real data whereas numbers on the right side show the performance of models when finetuned with real+synthetic data

	CL (%)	SAAA (%)	SNMN (%)
CV-VQA			
Acc orig	43.65 → 42.04	50.87 → 50.24	50.67 → 49.99
Pred flipped	83.84 → 50.74	77.74 → 45.85	73.12 → 44.19
CV-VQA+IV-VQA			
Acc orig	43.65 → 43.94	50.87 → 50.45	50.67 → 50.61
Pred flipped	83.84 → 59.58	77.74 → 52.71	73.12 → 51.91

Table 4: CV-VQAAugmentation: numbers on the left side of the arrow denote the accuracy/flipping for models finetuned using just real data whereas numbers on the right side show the performance of models when finetuned with real+synthetic data











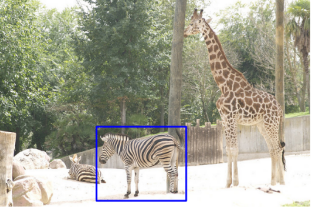

Q: What color is the floor?				
A: green		chair removed; A: green		
				
	real	real+edit	real	real+edit
CL	brown	brown	white	brown
SAAA	green	green	gray	green
SNMN	green	green	green	green
Q: Is this a bookstore?				
A: no		person removed; A: no		
				
	real	real+edit	real	real+edit
CL	yes	no	yes	no
SAAA	no	no	no	no
SNMN	no	no	yes	no
Q: Is there a pier in the picture?				
A: yes		boat removed; A: yes		
				
	real	real+edit	real	real+edit
CL	yes	yes	yes	yes
SAAA	no	yes	no	yes
SNMN	yes	yes	no	no
Q: How many bowls of food are there?				
A: 2		bottle removed; A: 2		
				
	real	real+edit	real	real+edit
CL	2	2	3	2
SAAA	2	2	2	2
SNMN	2	2	1	2

Figure 6: InVariant VQA Augmentation: Some visualizations from fine-tuning experiments using real/real+edit data from IV-VQA. Using real+edit makes models more consistent.

Q: How many planes are in the air?				
A: 1		<i>plane removed; A: 0</i>		
				
	real	real+edit	real	real+edit
CL	1	1	1	0
SAAA	1	1	1	0
SNMN	1	1	1	0

Q: How many zebras are there in the picture?				
A: 2		<i>zebra removed; A: 1</i>		
				
	real	real+edit	real	real+edit
CL	2	2	2	1
SAAA	2	2	2	1
SNMN	2	2	2	1



Q: How many boys are playing Frisbee?				
A: 2		<i>person removed; A: 1</i>		
				
	real	real+edit	real	real+edit
CL	1	2	1	1
SAAA	2	2	2	1
SNMN	1	2	1	1

Figure 7: CoVariant VQA Augmentation: Some visualizations from fine-tuning experiments using real/real+edit data from both CV-VQA and IV-VQA. Using real+edit makes models more consistent and in these examples- also accurate.

References

- [1] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019. 2
- [2] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR*, 2009. 1
- [3] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *CVPR*, 2017. 1
- [4] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 1
- [5] Christina Heinze-Deml and Nicolai Meinshausen. Conditional variance penalties and domain shift robustness. *stat*, 1050:8, 2018. 2
- [6] Ronghang Hu. Official code release for explainable neural computation via stack neural module networks. <https://github.com/ronghanghu/snmn>, 2018. 2
- [7] Ronghang Hu, Jacob Andreas, Trevor Darrell, and Kate Saenko. Explainable neural computation via stack neural module networks. In *ECCV*, 2018. 1
- [8] Vahid Kazemi and Ali Elqursh. Show, ask, attend, and answer: A strong baseline for visual question answering. In *ArXiv*, volume abs/1704.03162, 2017. 1, 2, 9
- [9] Yan Zhang. Re-implementation of show, ask, attend, and answer: A strong baseline for visual question answering [8] in pytorch. <https://github.com/Cyanogenoid/pytorch-vqa>, 2017. 2