model won the VQA Challenge in 2017 and achieves 66.25% accuracy on VQA v2.0 test-dev.

Pythia[41]3 extends the BUTD model by incorporating co-attention [27] between question and image regions. Pythia uses features extracted from Detectron [8] pretrained on Visual Genome. An ensemble of Pythia models won the 2018 VQA Challenge using extra training data from Visual Genome [21] and using Resnet[11] features. In this study, we use Pythia models which do not use Resnet features.

Footnote 3: https://github.com/facebookresearch/pythia

Bilinear Attention Networks (BAN) [19]4 combines the idea of bilinear models and co-attention [27] between image regions and words in questions in a residual setting. Similar to [3], it uses Faster-RCNN [33] pretrained on Visual Genome [21] to extract image features. In all our experiments, for a fair comparison, we use BAN models which do not use additional training data from Visual Genome. BAN achieves the current state-of-the-art single-model accuracy of 69.64 % on VQA v2.0 test-dev without using additional training data from Visual Genome.

Footnote 4: https://github.com/jnhwkim/ban-vqa

Implementation Details For all models trained with our cycle-consistent framework, we use the values T_{sim} =0.9, λ_G =1.0, λ_C =0.5 and A_{iter} =5500. When reporting results on the validation split and VQA-Rephrasings we train on the training split and when reporting results on the test split we train on both training and validation splits of VQA v2.0. Note that we *never* explicitly train on the collected VQA-Rephrasings dataset and use it purely for evaluation purposes. We use publicly available implementations of each backbone VQA model.

We measure the robustness of each of these models on our proposed VQA-Rephrasings dataset using the consensus score (Eq. 2). Table 1 shows the consensus scores at different values of k for several VQA models. We see that all models suffer significantly when measured for consistency across rephrasings. For e.g., the performance of Pythia (winner of 2018 VQA challenge) is reduced to a consensus score of 39.49% at k=4. Similar trends are observed for MUTAN, BAN and BUTD. The drop increases with increasing k, the number of rephrasings used to measure consistency. Models like BUTD, BAN and Pythia which use word-level encodings of the question suffer significant drops. It is interesting to note that even MUTAN which uses skip-thought based sentence encoding [20] suffers a drop when checked for consistency across rephrasings. We observe that BAN + CC model trained with our proposed cycle-consistent training framework outperforms its counterpart BAN and all other models at all values of k.

Fig 4 qualitatively compares the textual and visual attention (over image regions) over 4 rephrasings of a question. The top row shows attention and predictions from a Pythia model, while the bottom row shows attention and predictions from the same Pythia model, but trained using our framework. Our model attends at relevant image regions

Model	CS(k)			VQA Accuracy		
	k=1	k=2	k=3	k=4	ORI	REP
MUTAN [5]	56.68	43.63	38.94	32.76	59.08	46.87
BUTD [3]	60.55	46.96	40.54	34.47	61.51	51.22
BUTD + CC	61.66	50.79	44.68	42.55	62.44	52.58
Pythia [41]	63.43	52.03	45.94	39.49	64.08	54.20
Pythia + CC	64.36	55.45	50.92	44.30	64.52	55.65
BAN [19]	64.88	53.08	47.45	39.87	64.97	55.87
BAN + CC	65.77	56.94	51.76	48.18	65.87	56.59

Table 1: Consensus performance on VQA-Rephrasings dataset. CS(k) as defined in Eq. 2 is consensus score which is non-zero only if at least k rephrasings are answered correctly, zero otherwise; averaged across all group of questions. ORI represent a split of questions from VQA-Rephrasings which are original questions from VQA v2.0 and their corresponding rephrasings are represented by the split REP. Models trained with our cycle-consistent (CC) framework consistently outperform their baseline counterparts at all values of k.

Model	val	test-dev	
MUTAN [5]	61.04	63.20	
BUTD [3]	65.05	66.25	
+ Q-consistency	65.38	66.83	
+ A-consistency	60.84	62.18	
+ Gating	65.53	67.55	
Pythia [41]	65.78	68.43	
+ Q-consistency	65.39	68.58	
+ A-consistency	62.08	63.77	
+ Gating	66.03	68.88	
BAN [19]	66.04	69.64	
+ Q-consistency	66.27	69.69	
+ A-consistency	64.96	66.31	
+ Gating	66.77	69.87	

Table 2: VQA Performance and ablation studies on VQA v2.0 validation and test-dev splits. Each row in blocks represents a component of our cycle-consistent framework added to the previous row. First row in each block represents the baseline VQA model F. Q-consistency implies addition of a VQG module G to generate rephrasings Q' from the image I and the predicted answer A' with an associated VQG loss $\mathcal{L}_{vag}(Q,Q')$. A-consistency implies passing all the generated questions Q' to the VQA model F and an associated loss $\mathcal{L}_{cycle}(A,A')$. Gating implies the use of gating mechanism to filter undesirable generated questions in Q' and passing the remaining to VQA model F. Models trained with our cycle-consistent (last row in each block) framework consistently outperform baselines.