Toward Unsupervised Realistic Visual Question Answering

Yuwei Zhang* Chih-Hui Ho* Nuno Vasconcelos
Department of Electrical and Computer Engineering
University of California, San Diego
{yuz163,chh279, nvasconcelos}@ucsd.edu

‘

A. Real World VQA System

As discussed in the main paper, despite that most recent VQA models have superior performance on AQs, they fail to detect UQs and underperform the proposed methods in terms of AUAF and FF95. To further evaluate these recent classifiers deployed in the real-world VQA system, we investigate the robustness of BLIP [18] under RVQA setting, using their online demo. We use their provided example image and GQA image as visual inputs and ask several unanswerable questions. As shown in Fig. A1, when the user enters a question with objects that do not appear in the image, the model cannot reject or provide further instruction to the user. This shows that models optimizing for better AQ performance does not address the problem of RVQA, which hinders the application of real-world VQA system.

B. Training with Hard Pseudo UQs

Additional details for training the VQA classifiers on only hard pseudo UQs are provided. The hard pseudo UQs are the UQ pairs with higher CLIP similarity scores. We use CLIP to rank the questions for each image according to the similarity and select only top-$1,000$ questions to construct image-question pairs. As shown in Table H1, we observe that model that only trains on hard pseudo UQs performs similarly to our best model on CLIP-Hard and PT-Hard. However, the performance degrades significantly in terms of AUAF by around 7 and 10 points on CLIP-Easy and PT-Easy, respectively. This highlights the need for a dataset with broader coverage of UQ difficulty and indicates that overfitting VQA models on hard pseudo UQs will not address the problem of RVQA in general.

C. Training and Evaluation Details

In this section, the training and evaluation details of the experiments are discussed. The training is conducted using PyTorch [23] for all the experiments. For evaluating different OOD methods, we adopt the VQA classifier of BUTD [1] from https://github.com/siddk/vqa-outliers, LXMERT [25] from https://github.com/airsplay/lxmert and Uniter [4] from https://github.com/ChenRocks/UNITER and https://github.com/YIKUAN8/Transformers-VQA. Both LXMERT and Uniter are initialized from pretrained weights. For BUTD/LXMERT/Uniter, we used the optimizer of Adamax/Adam/Adam, respectively. The learning rate for BUTD/LXMERT/Uniter is set as $2e-3/1e-5/1e-5$, respectively. For RoI Mixup, we select β as $0.7/5/3$ for BUTD/LXMERT/Uniter. Since VQA models use the BCE loss, methods adapted from OOD literature are based on the implementation of this multi-label OOD github. We also use CLIP from huggingface and POS tagger from Spacy to process the text.

For the comparison of 9 different VQA classifiers [9, 25, 1, 4, 29, 20, 13, 18, 14], we further adopt the pretrained checkpoint on GQA of SwapMix [9], Oscar [20], VinVL [29] and MDETR [13] from their official github links. We also finetune Vilt on GQA following the procedure in [13, 20], because Vilt [14] only released the checkpoint from its pretraining stage and does not have checkpoint finetuned on GQA. For BLIP [18], we directly downloaded its checkpoint from their github link, which is trained on Visual Genome [15] and VQA2.0 [8] dataset. Due to the computation constraint of our GPU cluster, we are not able to finetune BLIP on GQA. However, since GQA is also built on Visual Genome, we measure the GQA performance of BLIP without fine-tuning its checkpoint. Note that BLIP supports open-ended VQA, so we follow its VQA setting and use its decoder to rank the GQA candidate answers (rank 1 is selected as prediction). The comparison between different VQA classifiers uses the maximum probability (MSP) as UQ/AQ criterion.

*The first two authors contributed equally to this work.

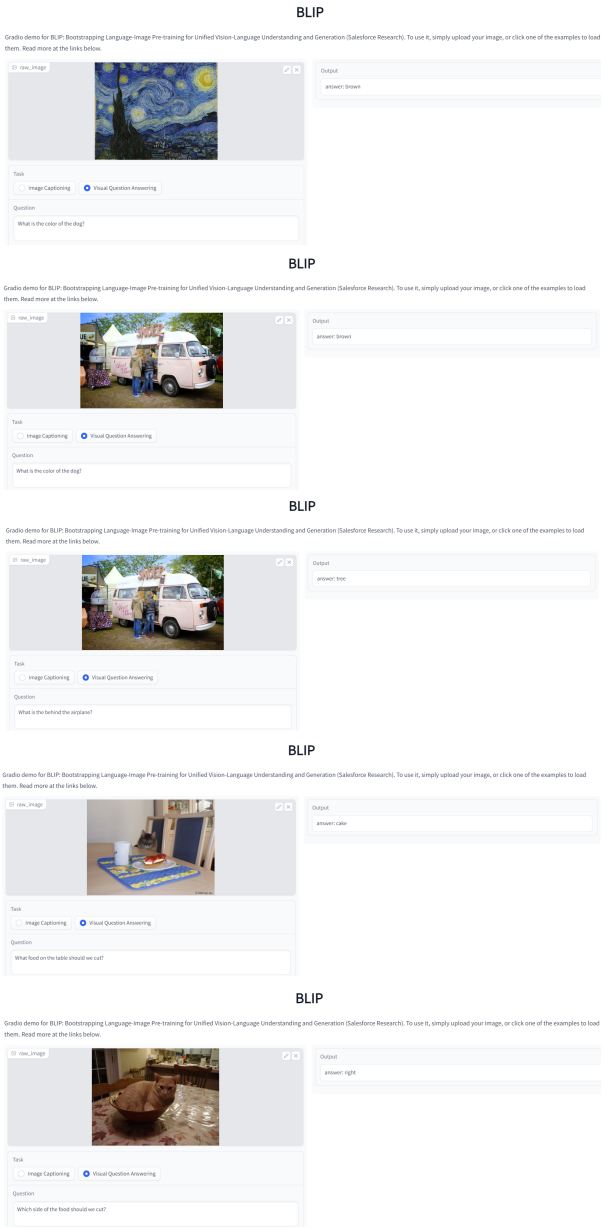


Figure A1: Illustration of the RVQA problem in real-world VQA system using the BLIP [18] demo website. The top image is provided on BLIP's demo website and the rest are GQA images.

D. Additional RGQA Details

Dataset Annotation. As mentioned in Sec. 3.1, the candidate UQs are passed to the annotators. The annotators are asked to read the instruction with few AQs (i.e. Valid) and UQs (i.e. Invalid) examples, as shown in Fig. D1(a). After reading the instruction, the annotator is given the tasks, where 2 questions in random order and an image are given. One of the questions requires annotation, while the other question is the "filter question". The filter question is used to ensure that the annotator fully comprehends the task and is paying attention during the process. Some examples of the task are shown in Fig. D1(b-e). Take Fig. D1(b) for example. 2 questions are presented to the annotator and the first question "Is there a tv stand?" is the filter question. The annotator is expected to answer "valid" for the filter question since it is an answerable question with answer "No". Fig. D1(d) is another example, where the filter question of "What color is the hills above the cat?" is shown on the second question. Obviously, the annotator is expected to answer "invalid", because there is no hill above the cat in the image.

More specifically, the filter question is guaranteed to be either answerable or unanswerable. To create filter questions automatically, we extract all the object names from the annotated scene graph in GQA [11] and curate a set of object names. For the candidate set of answerable filter questions, the template of "Are/Is there a ⟨obj⟩?" is used, where ⟨obj⟩ is a randomly selected object from the object set. Furthermore, the candidate set of answerable filter questions is augmented with ""Is this indoor or outdoor?", "Is this a color image?" and "What place is this ?" to increase the diversity of answerable filter questions.

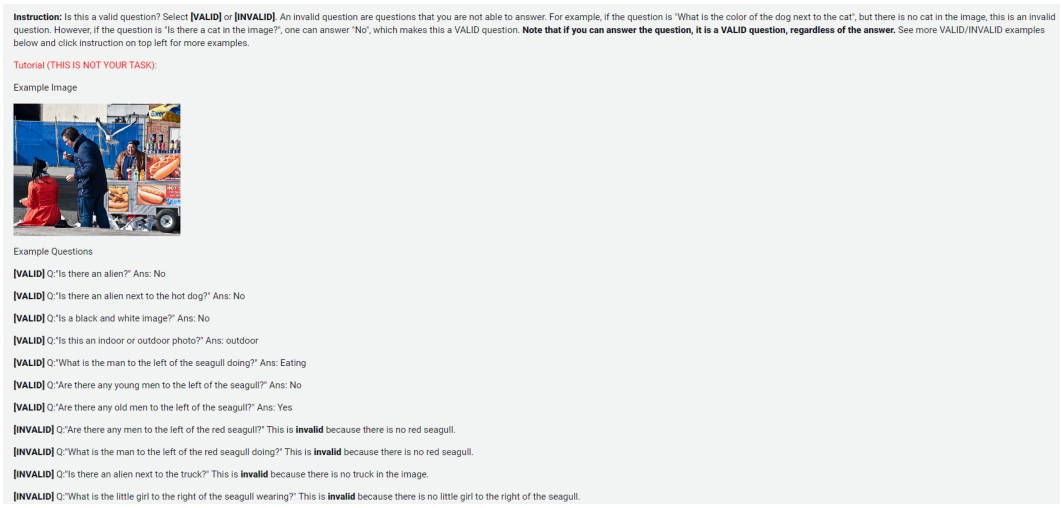
For the candidate set of unanswerable questions, we adopt the template of "What color is the ⟨obj0⟩ ⟨rel⟩ the ⟨obj1⟩?", where ⟨obj0⟩ and ⟨obj1⟩ are 2 randomly selected objects from the object set and the ⟨rel⟩ is a randomly selected relation from a set of predefined relations (e.g. next to, around, under, on and above).

AQ vs UQ Ratio. As mentioned in the main paper, each UQ is paired with an AQ. However, this could result in duplicated AQs, because the number of UQs could be larger than that of AQs for some images. As a result, the duplicated AQs are removed from the proposed dataset, which explains the reason that the proportion of UQ in the main paper is around 52%.

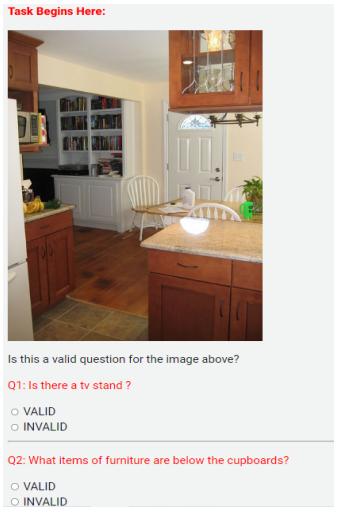
AQ vs UQ Question Structure We further analyze the difference between AQ and UQ from its question structure. This is done by plotting the distribution of questions by the first three words, as shown in Fig. D2, While the three most popular words ("Are", "Who" and "Which") in AQs and UQs have minor difference in their order and proportions, there are no major differences between the question structure of AQ (Fig. D2(a)) and UQ (Fig. D2(b)). This indicates that AQs/UQs cannot be easily separated by word frequency and distribution.

Conflicting Candidate UQs Removal: Conflicting Candidate UQs like "What color are the black shoes?" are filtered using predefined rules. For example, for a question asking about color, a program checks whether the answer (i.e. black) is in the text.

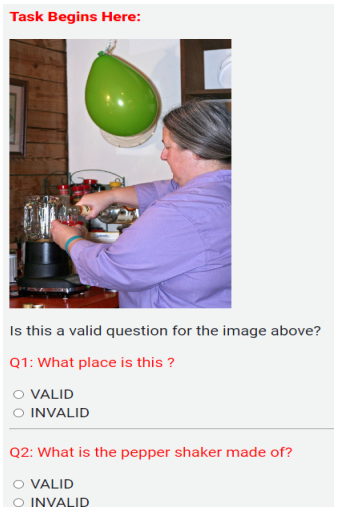
CLIP Bias: We notice that CLIP might have a bias when producing UQs (e.g. confuses attributes of multiple objects in the same image [26]). We prevent these biases by introducing PT-based UQs and using human annotators to con-



(a) Instruction to the annotators.



(b) Example 1

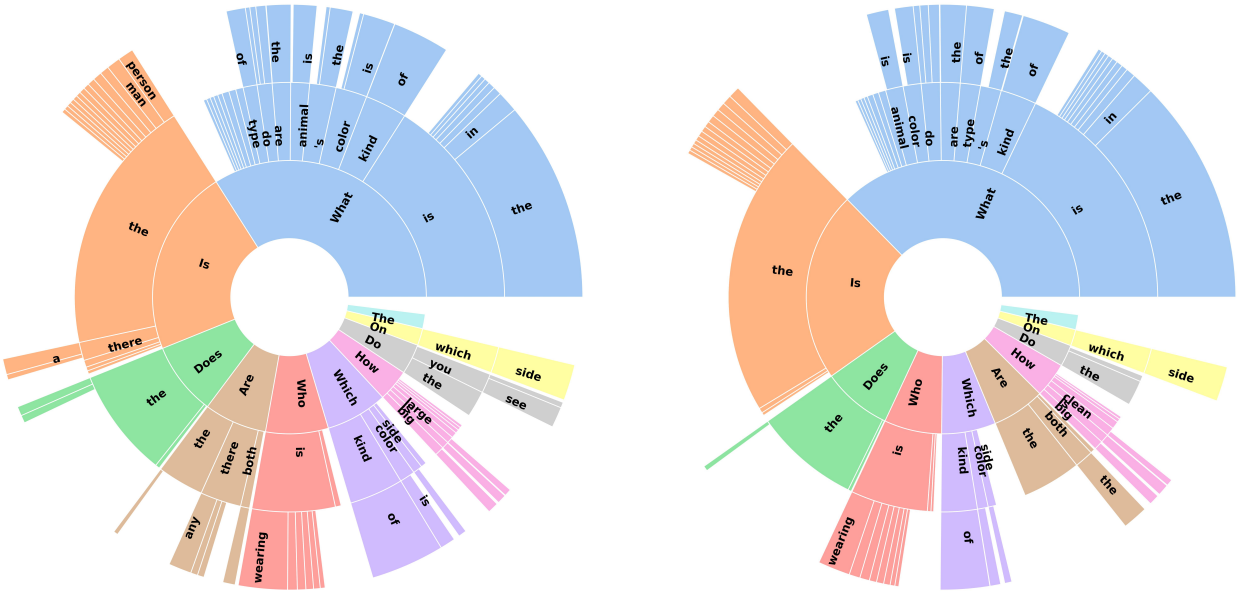


(c) Example 2

(d) Example 3

(e) Example 4

Figure D1: The annotator is asked to read the instruction in (a). (b-e) are the task that are assigned to the annotator. See text in Sec. D for more details.



(a) Answerable question (AQ)

(b) Unanswerable question (UQ)

Figure D2: Distribution of questions by first three words for all subsets in RGQA. The white regions are marginal probabilities for those less populated words.

firm the validity of UQs. Obviously, there could be other biases, e.g. a preponderance of certain types of objects in the dataset. The characterization of these is a project in itself and left for future work.

E. Additional Related Work

Visual language pretraining (VLP) [29, 14, 20, 30, 17, 19, 22, 21, 6, 4, 25, 18] has been a dominated manner to learn generalizable multi-modal feature for visual language task. During the pre-training stage, the models are usually trained on self-supervised tasks: (a) masked language modeling, (b) masked image modeling, and (c) image-text matching. For (a) and (b), the model predicts the masked words and masked patches using the rest of the unmasked text and image. For (c), the image and text are randomly paired and the model is asked whether the pair are matched. The universal feature from the pre-training stage are shown to be applicable to various downstream tasks, including image captioning [14, 30, 29], visual grounding [4, 6, 10, 13] and VQA [22, 25, 21]. For more detailed related work, please refer to recent surveys [5, 24].

Despite that most of the recent VQA models [22, 25, 21, 30, 29, 20, 19, 21, 6, 4] are fine-tuned after VLP, there is no evidence that these model are robust toward UQ. Our experiments analyze the SoTA VQA models [25, 4, 20, 29, 18] and found their vulnerability to the UQs. This is surprising since the proxy task of image-text matching is optimized during the pretraining stage. In this work, we proposed a new training scheme specifically tailored for the RVQA task, which does not require any annotated UQs during training and is robust to UQ during inference.

Mixup Inspired by the mixup data augmentation [28, 27, 3], we proposed RoI mixup to encourage the RVQA model to be aware of the fine-grain mismatch between image and text. While the proposed RoI mixup is similar to [9] and [2], the goal is entirely different.[9] mitigates the reliance of the VQA model on irrelevant background, while [2] mixes object features of the same classes from different domains, to achieve domain invariance between synthetic and real VQA datasets. Our aim is to detect UQs, which these methods cannot. For example, although [9] leverages scene graph annotation of objects "relevant" to the text, to swap object features, it fails to detect UQs, as shown in the experiment section. No additional annotation is needed for the proposed RoI Mixup.

VQA-OOD While there are certain similarities between OOD and realistic VQA, they are different. [7, 16, 12] address OOD where the distributions of training and test set are different. However, *there is no answer* for UQs, so [7, 16, 12] are not applicable to the proposed RVQA task.

F. GQA test set performance

The goal of realistic VQA is to detect the UQs without sacrificing the VQA performance on AQs. To ensure that the proposed training strategy does not degrade the VQA performance, we report the accuracy of the GQA test set. As shown in Table F1, the original LXMERT [25] achieves 77.8 accuracy, while the proposed RP and Mix have similar performance. We also report the accuracy for binary (e.g. yes/no) questions and open questions.

Table F1: GQA test set performance without rejections. The backbone is LXMERT.

Methods	Accuracy	Binary	Open
Original	77.8	45.0	60.3
RP	76.3	45.6	60.0
Mix	77.1	45.8	60.5
Ens	77.1	46.2	60.7

G. Additional examples from RGQA

We show more examples from RGQA in Figure H1, where (a-d), (e-h), (i-l) and (m-p) are CLIP-Hard, PT-Hard, CLIP-Easy, and PT-Easy, respectively.

H. Social Impact and Future Work

Social Impact: In this work, we study the problem of realistic VQA and proposed a new dataset containing UQs to evaluate the existing VQA classifiers. Note that the goal is not to falsify the VQA system, but to evaluate and improve the robustness of the existing system to the UQ. We plan to expand the proposed dataset with more diverse types of UQ and include more annotations, such as the category of UQs. We hope this dataset will encourage more research on the realistic VQA.

Future Work: While we have evaluated the open-ended VQA model (See BLIP experiment) on the proposed dataset, we would like to further explore open-ended VQA in the future, especially by allowing the model to explain why a question is a UQ. However, this might require further annotations. Hence, we see this as a direction for future work. Another direction is to compare the distribution between the proposed RGQA dataset and the unanswerable question collected from the real-world system.

CLIP-hard: Are the cabinets below the stove wooden and open?

(a)

CLIP-hard: Is the black bag to the left or to the right of the bed?

(b)

CLIP-hard: Do the snowpants look black and long?

(c)

CLIP-hard: What is around the open window?

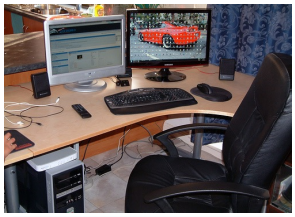
(d)

PT-hard: What is the surfing person in front of?

(e)

PT-hard: Does the rolled meat on the stacked plate look roasted?

(f)



PT-hard: Which kind of small device is pink?

(g)

PT-hard: What is the item of furniture to the left of the light white couch?

(h)

CLIP-easy: Is the color of the keyboard the same as the color of the plant?

(i)

CLIP-easy: On which side is the doll?

(j)

CLIP-easy: Are the baseball mitt and the belt the same color?

(k)

CLIP-easy: Which kind of vegetable is on top of the cutting board?

(l)

PT-easy: On which side of the photo is the huge man?

(m)

PT-easy: Is the mirror in front of the cap clean and metallic?

(n)

PT-easy: What kind of containers does the map lie on top of?

(o)

PT-easy: What is parked near the piano the nearest traffic light is across from?

(p)

Figure H1: More examples from RGQA dataset across 4 different subsets.

Table H1: Comparison between different RVQA approaches. Larger AUAF and smaller FPR@95 is better. Cells with light cyan background denote training with pseudo UQs.

RVQA Approaches	CLIP-Easy			CLIP-Hard			PT-Easy			PT-Hard			Avg.
	AUAF	FF95↓	FACC	AUAF	FF95↓	FACC	AUAF	FF95↓	FACC	AUAF	FF95↓	FACC	AUAF
BUTD [1]													
FRCNN	33.58	93.28	53.50	30.73	93.94	53.08	31.43	93.77	53.02	26.94	94.65	51.31	30.67
MSP	38.45	64.75	53.50	36.13	79.14	53.08	37.83	66.05	53.02	33.60	83.11	51.31	36.50
ODIN	38.47	64.66	53.53	36.14	79.19	53.11	37.80	66.14	52.97	33.60	83.41	51.33	36.50
Maha	30.05	80.66	48.76	25.75	92.16	48.42	25.34	94.90	47.70	23.93	95.43	46.39	26.26
Energy	38.47	64.14	53.50	36.19	79.42	53.08	37.77	66.12	53.02	33.67	82.99	51.31	36.52
Q-C	53.04	3.48	53.50	36.20	69.25	53.08	47.14	42.18	53.02	29.06	85.65	51.31	41.36
Resample	40.25	65.23	56.20	37.73	79.64	55.45	39.54	66.43	55.41	34.78	83.73	53.79	38.07
RP(w/ hard UQ)	43.74	66.33	56.04	43.27	70.38	55.40	37.62	81.98	55.21	36.17	84.97	53.81	40.2
RP(Ours)	56.31	1.82	56.64	44.09	56.57	55.66	50.51	27.41	55.03	37.18	80.38	53.88	47.02
Mix(Ours)	56.85	1.65	57.17	44.70	58.84	56.59	51.27	29.28	55.99	37.59	83.41	**55.24**	47.60
Ens(Ours)	**57.25**	**1.31**	**57.50**	**45.46**	**56.04**	**56.90**	**51.95**	**24.69**	**56.02**	**38.46**	**80.08**	54.85	**48.28**
UNITER [4]													
FRCNN	35.81	93.28	57.08	33.09	93.93	57.10	33.67	93.77	56.82	28.82	94.68	55.08	32.84
MSP	40.03	73.15	57.08	39.42	80.48	57.10	41.45	61.76	56.82	35.17	83.52	55.08	39.01
ODIN	40.04	73.22	57.12	39.43	80.48	57.15	41.45	61.83	56.85	35.16	83.54	55.06	39.02
Maha	37.52	67.07	55.38	33.74	81.09	54.88	35.87	63.98	54.68	31.68	85.78	52.80	34.70
Energy	40.10	71.45	57.08	39.42	79.78	57.10	41.41	61.31	56.82	35.19	83.63	55.08	39.03
Q-C	56.61	3.53	57.08	38.67	69.56	57.10	50.12	45.64	56.82	30.93	86.18	55.08	44.08
Resample	58.66	0.755	58.85	48.08	47.10	57.60	53.65	22.42	57.48	39.84	73.46	55.33	50.05
RP(w/ hard UQ)	44.92	70.71	59.02	47.14	59.81	57.91	41.89	70.89	58.36	37.92	80.19	55.70	42.96
RP(Ours)	58.35	0.615	58.49	48.37	47.08	57.69	54.42	20.43	57.83	40.27	73.20	55.44	50.35
Mix(Ours)	59.08	0.615	59.37	49.00	47.00	58.06	54.63	21.44	58.08	41.50	73.29	56.68	51.05
Ens(Ours)	**59.62**	**0.58**	**59.82**	**49.65**	**46.71**	**58.84**	**55.79**	**20.08**	**59.11**	**42.14**	**72.71**	**57.17**	**51.8**
LXMERT [25]													
FRCNN	38.43	93.21	60.87	35.22	93.88	60.49	35.73	93.72	59.94	31.00	94.62	58.76	35.09
MSP	42.39	76.25	60.87	42.60	78.92	60.49	47.30	61.79	59.94	38.12	85.14	58.76	42.60
ODIN	42.41	76.43	60.92	42.59	78.96	60.46	47.33	61.97	59.97	38.12	84.78	58.73	42.61
Maha	57.68	9.79	58.98	44.96	61.09	58.16	49.44	44.43	57.27	39.25	75.25	56.29	47.83
Energy	38.76	76.88	60.87	42.11	78.85	60.49	47.00	61.84	59.94	37.90	85.53	58.76	41.44
Q-C	60.39	3.42	60.87	41.31	68.72	60.49	53.11	44.50	59.94	33.18	85.65	58.76	46.99
Resample	60.47	0.58	60.66	50.80	46.49	60.37	55.74	25.30	59.84	42.18	76.78	58.27	52.29
RP(w/ hard UQ)	53.60	40.44	60.15	51.39	47.80	59.40	46.95	57.51	58.74	42.96	**68.56**	57.17	48.72
RP(Ours)	60.51	0.527	60.66	51.49	45.02	60.69	56.08	23.18	59.74	42.53	75.78	58.37	52.65
Mix(Ours)	60.79	**0.298**	61.03	51.91	43.43	60.67	56.83	22.58	60.40	43.56	73.02	58.64	53.27
Ens(Ours)	**61.03**	0.351	**61.19**	**52.42**	**42.84**	**61.19**	**56.90**	**22.40**	**60.47**	**43.75**	73.01	**58.83**	**53.52**

References

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6077–6086, 2018. 1, 7

[2] Paola Cascante-Bonilla, Hui Wu, Letao Wang, Rogerio Feris, and Vicente Ordonez. Sim vqa: Exploring simulated environments for visual question answering. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5046–5056, 2022. 4

[3] Jie-Neng Chen, Shuyang Sun, Ju He, Philip Torr, Alan Yuille, and Song Bai. Transmix: Attend to mix for vision transformers. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022. 4

[4] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *ECCV*, 2020. 1, 4, 7

[5] Yifan Du, Zikang Liu, Junyi Li, and Wayne Xin Zhao. A survey of vision-language pre-trained models. In *IJCAI*, 2022. 4

[6] Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, and Jingjing Liu. Large-scale adversarial training for vision-and-language representation learning. In *NeurIPS*, 2020. 4

[7] Tejas Gokhale, Pratyay Banerjee, Chitta Baral, and Yezhou Yang. MUTANT: A training paradigm for out-of-distribution generalization in visual question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 878–892, Online, Nov. 2020. Association for Computational Linguistics. 4

[8] 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. *International Journal of Computer Vision*, 127:398–414, 2017. 1

[9] Vipul Gupta, Zhuowan Li, Adam Kortylewski, Chenyu Zhang, Yingwei Li, and Alan Loddon Yuille. Swapmix: Diagnosing and regularizing the over-reliance on visual context in visual question answering. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5068–5078, 2022. 1, 4

[10] Chih-Hui Ho, Srikar Appalaraju, Bhavan Jasani, R Manmatha, and Nuno Vasconcelos. Yoro-lightweight end to end visual grounding. In *ECCV 2022 Workshop on International Challenge on Compositional and Multimodal Perception*, 2022. 4

[11] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[12] Jingjing Jiang, Ziyi Liu, Yifan Liu, Zhixiong Nan, and Nanning Zheng. X-ggm: Graph generative modeling for out-of-distribution generalization in visual question answering. In *Proceedings of the 29th ACM International Conference on Multimedia*, MM '21, page 199–208, New York, NY, USA, 2021. Association for Computing Machinery. 4

[13] Aishwarya Kamath, Mannat Singh, Yann LeCun, Ishan Misra, Gabriel Synnaeve, and Nicolas Carion. Mdetr - modulated detection for end-to-end multi-modal understanding. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1760–1770, 2021. 1, 4

[14] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *ICML*, 2021. 1, 4

[15] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalanditis, Li-Jia Li, David A Shamma, Michael Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. 2016. 1

[16] Gouthaman KV and Anurag Mittal. On the role of question encoder sequence model in robust visual question answering. *Pattern Recogn.*, 131(C), nov 2022. 4

[17] Gen Li, Nan Duan, Yuejian Fang, Daxin Jiang, and Ming Zhou. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *AAAI*, 2020. 4

[18] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022. 1, 2, 4

[19] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *ArXiv*, abs/1908.03557, 2019. 4

[20] Xiujun Li, Xi Yin, Chunyuan Li, Xiaowei Hu, Pengchuan Zhang, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *ECCV*, 2020. 1, 4

[21] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *NeurIPS*, 2019. 4

[22] Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 12-in-1: Multi-task vision and language representation learning. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10434–10443, 2020. 4

[23] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017. 1

[24] Ludan Ruan and Qin Jin. Survey: Transformer based video-language pre-training. *AI Open*, 3:1–13, 2022. 4

[25] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, 2019. 1, 4, 7

[26] Yutaro Yamada, Yingtian Tang, and Ilker Yildirim. When are lemons purple? the concept association bias of clip. *ArXiv*, abs/2212.12043, 2022. 2

[27] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable

features. In *International Conference on Computer Vision (ICCV)*, 2019. 4

[28] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *International Conference on Learning Representations*, 2018. 4

[29] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Making visual representations matter in vision-language models. *CVPR 2021*, 2021. 1, 4

[30] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason J. Corso, and Jianfeng Gao. Unified vision-language pre-training for image captioning and vqa. *arXiv preprint arXiv:1909.11059*, 2019. 4