Rethinking Data Augmentation for Robust Visual Question Answering ****** Supplementary Manuscript ******

Long Chen[1⋆], Yuhang Zheng[2*], and Jun Xiao[2⋆⋆]

[1]Columbia University [2]Zhejiang University
zjuchenlong.com, itemzhang@zju.edu.cn, junx@cs.zju.edu.cn
https://github.com/ItemZheng/KDDAug

This supplementary manuscript is organized as follows:

1. In Section A, we introduce more details about experiments, including experimental settings, training process, KDDAug settings, and paraphrasing steps mentioned in Section 4.1 (*cf.*, Experimental Settings and Implementation Details).
2. In Section B, we describe more details about the $\text{CLIP}_{\text{rank}}$ used in Section 4.4 (*cf.*, ablation studies Q1 & Q2).
3. In Section C, we add additional experimental results to demonstrate the effects of augmentation diversity.
4. In Section D, we demonstrate more visualization results, including comparisons of the generated pseudo ground-truth answers between KDDAug and SimpleAug [5], and more diversity augmented samples in $D_{\text{aug}}^{\text{extra}}$.

A More Details about Experiments

A.1 Details about Experimental Settings

Datasets. We evaluated the proposed KDDAug on two datasets: the ID benchmark **VQA v2** [4] and OOD benchmark **VQA-CP v2** [1]. VQA v2 is a "balanced" VQA dataset, where each question has complementary images with opposite answers. Although VQA v2 has reduced language biases to some extent, the statistical biases from questions still can be leveraged [1]. To disentangle the biases and clearly monitor the progress of VQA, VQA-CP re-organizes VQA v2, and deliberately keeps different QA distributions in the training and test sets.
Evaluation Metrics. For model accuracies, we followed standard VQA evaluation metric [2], and reported accuracy on three different categories separately: Yes/No (`Y/N`), number counting (`Num`), and other (`Other`) categories. For the ID evaluation, we reported the results on the VQA v2 val set. For the OOD evaluation, we reported the results on the VQA-CP v2 test set. Meanwhile, we followed [8] and used Harmonic Mean (**HM**) of the accuracies on both two datasets (VQA v2 val & VQA-CP test) to evaluate the trade-off between ID and OOD evaluations.

⋆ Long Chen and Yuhang Zheng are co-first authors with equal contribution.
⋆⋆ Jun Xiao is the corresponding author.

A.2 Details about Training Process

To effectively train VQA models with both original and new augmented samples, we first pre-trained VQA models with only original samples following their respective settings. Then, we fine-tuned pre-trained VQA models[1] with augmented samples for 5 epochs. The batch size was set to 512. We used the Adamax [6] as the optimizer and the random seed was set to 0.

A.3 Details about KDDAug Settings

For the object detector, we used the Faster R-CNN [10] pre-trained on VG [7] to detect 36 objects and attributes (*e.g.*, color) for every image. To keep highly-confident predictions, we set the score thresholds for object and attribute to 0.8 and 0.4. *It is worth noting that we only used detection results from the Faster R-CNN without relying on other extra human annotations.*

A.4 Details about Paraphrasing

In this section, we introduce more details about the paraphrasing in Section 4.1. Paraphrasing is a supplementary data augmentation trick proposed by SimpleAug [5] which composes new VQ pairs by searching similar questions. Specifically, for each original sample (I_i, Q_i, a_i), if question Q_j is similar to Q_i, they construct a new augmented training sample (I_i, Q_j, a_i). By "similar", we mean that the cosine similarity between question BERT embeddings[2] [3] is large than 0.95. For each original sample, we choose all top-3 similar questions for composing new samples according to cosine similarity scores.

B Details about $\text{CLIP}_{\text{rank}}$

$\text{CLIP}_{\text{rank}}$ aims to rank the quality of all SimpleAug assigned answers, *i.e.*, we used it to rank the similarity between each image and the augmented question-answer (QA) pair. We firstly generated a prompt for each augmented QA pair, and utilized a pretrained CLIP [9] to calculate the similarity score between the prompt and the image. Specifically, we designed different strategies to generate prompts for different question types. For "`Number`" and "`Color`" questions, we generated prompts by removing the question type category prefix and inserting the answer in front of the noun[3]. For example, if the question is "`how many umbrellas are there`", and its pseudo answer is "`2`", then the prompt is "`2 umbrellas are there`". For the other questions, we generated prompts by simply replacing question type category prefix with the answer. For example, if

[1] We only fine-tune the basic VQA backbone (UpDn) in the fine-tuning stage, *i.e.*, for ensemble-based models, we removed the auxiliary question-only branches.

[2] The BERT is pre-trained on BookCorpus [11] and English Wikipedia. We get the pre-trained BERT model from https://github.com/google-research/bert.

[3] For "`Number`" and "`Color`" questions, there is an only single noun in the questions.

Models	**Extra**	VQA-CP v2	VQA v2	**HM**
KDDAug‡		53.03	61.59	59.99
+Initial Answers		59.02	61.07	60.03
KDDAug$^{+\ddagger}$	✓	53.76	62.58	57.84
+Initial Answers	✓	**60.03**	**62.24**	**61.12**

Table 8. Accuracies (%) on VQA-CP v2 and VQA v2.‡ denotes without paraphrasing samples. **"Extra"** denotes using D_{aug}^{extra}.

the question is "`what food is that`", and its pseudo answer is "`donut`", then its prompt is "`donut is that`". Based on their respective similarity scores, we ranked all the augmented samples.

C Additional experimental results

In this section, we add additional experimental results to demonstrate the effects of augmentation diversity. As shown in Table 8, diversity indeed helps model performance (both w/ and w/o initial answers.). However, when using the extra paraphrasing [5], the improvement gains brought by diversity in the smaller size KDDAug$^{+\ddagger}$ is overwhelmed. Moreover, we only use KDDAug for SOTA comparison rather than KDDAug^{+} for two reasons: 1) The sample size of whole $D_{\text{aug}}^{\text{extra}}$ is enormous. Thus, it is infeasible to directly train models with whole $D_{\text{aug}}^{\text{extra}}$ (KDDAug^{+}). 2) For efficiency and fair comparison with prior work SimpleAug [5], we controlled the number of samples to be the same as SimpleAug (*i.e.*, KDDAug$^{+\ddagger}$ in Table 7 and 8).

D More Visualization Results

D.1 KDDAug vs. SimpleAug

To further compare KDDAug and SimpleAug, we show some augmented samples and their answers assigned by KDDAug and SimpleAug in Fig. 6. Take the second question "`How many signs are yellow?`" as an example, SimpleAug directly uses the count of signs appearing in the image as the answer, *e.g.*, "`1`". In contrast, our KDDAug takes "`0`" as the answer, which demonstrates the robustness of KDDAug assigned answers. Meanwhile, for some questions with multiple possible answers (*e.g.*, the question "`what else is on the desk`" for the third sample), our "soft" version ground-truth answer is inherently more accurate and better for VQA model training.

D.2 Augmented Samples in $D_{\text{aug}}^{\text{extra}}$

As shown in Fig. 7, we show some augmented samples in $D_{\text{aug}}^{\text{extra}}$. All these samples can't be generated by SimpleAug due to the limitations of its image-question pair

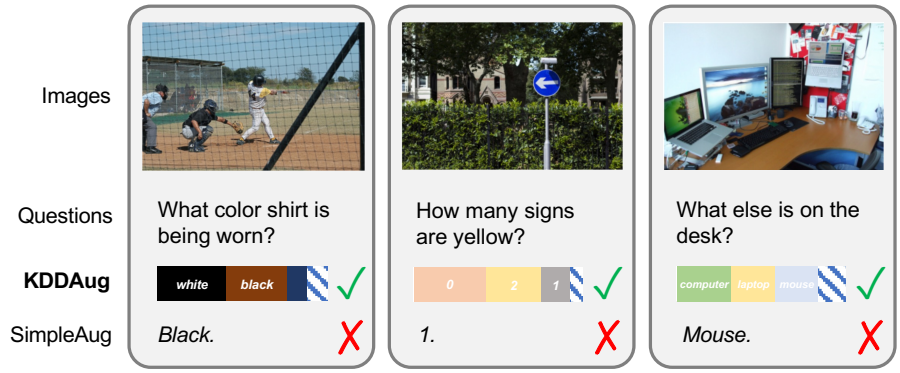

Fig. 6. Visualization results of some augmented samples and their pseudo ground-truth answers assigned by our KDDAug and SimpleAug.

Fig. 7. Visualization results of some augmented samples in D_{aug}^{extra}. "Type" denotes the question type, which are all excluded in original SimpleAug.

composition strategy. Take the third question "`Are the zebras' tails up or down?`" as an example, SimpleAug can't generate it since it doesn't belong to "`Color`", "`What`", "`Number`" or "`Yes/No`" questions. In contrast, our KDDAug can generate this augmented sample and assign a reasonable answer "`down`" for it, which demonstrates the generalization of KDDAug.

References

1. Agrawal, A., Batra, D., Parikh, D., Kembhavi, A.: Don't just assume; look and answer: Overcoming priors for visual question answering. In: CVPR (2018) 1
2. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C.L., Parikh, D.: Vqa: Visual question answering. In: ICCV. pp. 2425–2433 (2015) 1
3. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: NAACL. pp. 4171–4186 (2019) 2

4. Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., Parikh, D.: Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In: CVPR. pp. 6904–6913 (2017) 1
5. Kil, J., Zhang, C., Xuan, D., Chao, W.L.: Discovering the unknown knowns: Turning implicit knowledge in the dataset into explicit training examples for visual question answering. In: EMNLP (2021) 1, 2, 3
6. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: ICLR (2015) 2
7. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., et al.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. In: IJCV. pp. 32–73 (2017) 2
8. Niu, Y., Tang, K., Zhang, H., Lu, Z., Hua, X.S., Wen, J.R.: Counterfactual vqa: A cause-effect look at language bias. In: CVPR (2021) 1
9. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: ICML. pp. 8748–8763 (2021) 2
10. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: NeurIPS. pp. 91–99 (2015) 2
11. Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., Fidler, S.: Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In: ICCV. pp. 19–27 (2015) 2