Rethinking Data Augmentation for Robust Visual Question Answering

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

Abstract. Data Augmentation (DA) — generating extra training samples beyond the original training set — has been widely-used in today's unbiased VQA models to mitigate language biases. Current mainstream DA strategies are synthetic-based methods, which synthesize new samples by either editing some visual regions/words, or re-generating them from scratch. However, these synthetic samples are always unnatural and error-prone. To avoid this issue, a recent DA work composes new augmented samples by randomly pairing pristine images and other human-written questions. Unfortunately, to guarantee augmented samples have reasonable ground-truth answers, they manually design a set of heuristic rules for several question types, which extremely limits its generalization abilities. To this end, we propose a new **K**nowledge **D**istillation based **D**ata **Aug**mentation for VQA, dubbed **KDDAug**. Specifically, we first relax the requirements of reasonable image-question pairs, which can be easily applied to any question type. Then, we design a knowledge distillation (KD) based answer assignment to generate pseudo answers for all composed image-question pairs, which are robust to both *in-domain* and *out-of-distribution* settings. Since KDDAug is a model-agnostic DA strategy, it can be seamlessly incorporated into any VQA architecture. Extensive ablation studies on multiple backbones and benchmarks have demonstrated the effectiveness and generalization abilities of KDDAug.

Keywords: VQA, Data Augmentation, Knowledge Distillation

1 Introduction

Visual Question Answering (**VQA**), *i.e.*, answering any natural language questions about the given visual content, is regarded as the holy grail of a human-like vision system [21]. Due to its multi-modal nature, VQA has raised unprecedented attention from both CV and NLP communities, and hundreds of VQA models have been developed in recent years. Although current VQA models can achieve really "decent" performance on standard benchmarks, numerous studies have revealed that today's models tend to over-rely on the superficial linguistic correlations between the questions and answers rather than multi-modal reasoning

*Long Chen and Yuhang Zheng are co-first authors with equal contributions.
†Corresponding author. Codes: https://github.com/ItemZheng/KDDAug.

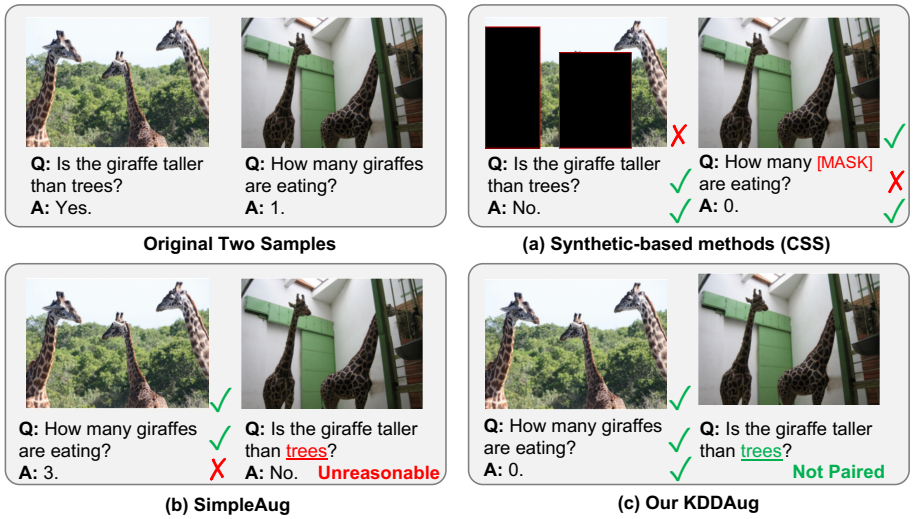

Fig. 1. Comparisons between different DA methods for VQA. (a) **Synthetic-based methods**: Take CSS [16] as an example, it masks some regions or words in the original samples. (b) **SimpleAug** [31]: It pairs images with some specific types of questions, and obtains pseudo labels by predefined heuristic rules. (c) **KDDAug**: It pairs image with any types of questions, and use a KD-based model to predict pseudo answers. Question "`is ...`" is not a reasonable question for the right image as it contains "`trees`".

(*a.k.a.*, **language biases**) [4,3,54,28,24]. For example, blindly answering "2" for all counting questions or "tennis" for all sport-related questions can still get a satisfactory performance. To mitigate these bias issues and realize robust VQA, a recent surge of VQA work [16,18,32,2,22,9,46,30,29,23,8,7,51] resort to different data augmentation techniques (*i.e.*, generating extra training samples beyond original training set), and achieve good performance on both the in-domain (ID) (*e.g.*, VQA v2 [24]) and out-of-distribution (OOD) datasets (*e.g.*, VQA-CP [4]).

Currently, mainstream Data Augmentation (DA) strategies for robust VQA are **synthetic-based** methods. As shown in Fig. 1(a), from modality viewpoint, these synthetic-based DA methods can be categorized into two groups: 1) *Visual-manipulated*: They usually edit some visual regions in original images [16,18,32], or re-generate counterfactual/adversarial images with generative networks [2,22] or adversarial attacks [46]. 2) *Textual-generated*: They edit some words in original questions [16,18,32] or re-generate the sentence from scratch with back translation [46,30] or visual question generation (VQG) methods [51]. Although these synthetic-based methods have dominated the performance on OOD benchmarks (*e.g.*, VQA-CP), and significantly improved VQA models' interpretability and consistency, there are several inherent weaknesses: 1) Photo-realistic image generation or accurate sentence generation themselves are still open problems, *e.g.*, a significant portion of the generated questions have grammatical errors [46]. 2) They always need extra human annotations to assign reasonable answers [22].

To avoid these unnatural and synthetic training samples, a recent work SimpleAug [31] starts to compose new training samples by randomly pairing images

and questions. As the example shown in Fig. 1(b), they pair the left image and human-written questions from the right image (*i.e.*, "`How many giraffes are eating?`") into a new image-question (VQ) pair. To obtain "reasonable" pseudo ground-truth answers for these new VQ pairs, they manually design a set of heuristic rules for several specific question types, including "`Yes/No`", "`Color`", "`Number`", and "`What`" type questions. Although SimpleAug avoids the challenging image/sentence generation procedures, there are still several drawbacks: 1) These predefined rules for answer assignment are fallible (*e.g.*, the human-check accuracy for "`Yes/No`" type questions is only 52.20%, slightly higher than a random guess.) 2) Due to the limitations of these rules, it only covers several specific types of answers, and it is difficult to extend to other question types[1]. 3) It still relies on some human annotations (*e.g.*, object annotations in COCO).

In this paper, we propose a **K**nowledge **D**istillation based **D**ata **Aug**mentation (**KDDAug**) strategy for robust VQA, which can avoid all the mentioned weaknesses in existing DA methods. Specifically, we first relax the requirement for reasonable VQ pairs by only considering the object categories in the images and nouns in the questions. As illustrated in Fig. 1(c), question "`how many giraffes are eating`" is a reasonable question for the left image which contains "`giraffe`" objects. To avoid extra human annotations, we only utilize an off-the-shelf object detector to detect objects[2]. After obtaining all the reasonable VQ pairs, we design a multi-teacher knowledge distillation (KD) based answer assignment to generate corresponding pseudo ground-truth answers. We first pretrain two teacher models (ID and OOD teacher) with the original training set, and then utilize these teacher models to predict a "soft" answer distribution for each VQ pair. Last, we combine the predicted distributions (*i.e.*, knowledge) from both teachers, and treat them as the pseudo answers for augmented VQ pairs. Benefiting from our designs, KDDAug achieves the best trade-off results on both ID and OOD settings with even fewer samples (Fig. 2)[3].

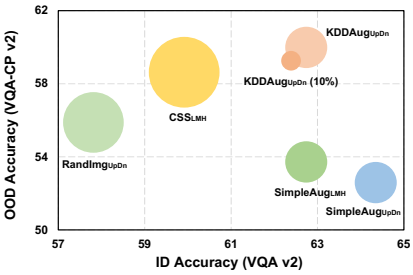

Fig. 2. Performance of SOTA DA methods. Circle sizes are in proportion to the number of their augmented samples[3].

Extensive ablation studies have demonstrated the effectiveness of KDDAug. KDDAug can be seamlessly incorporated into any VQA architecture, and consistently boost their performance. Particularly, by building on top of some SOTA

[1] VQA datasets typically have much more question types (*e.g.*, 65 for VQA v2 [24]).

[2] Since all the compared state-of-the-art VQA models follow UpDn [5] and use VG [33] pretrained detector to extract visual features, we don't use extra annotations.

[3] CSS & RandImg dynamically generate different samples in each epoch, their sizes are difficult to determine. Here, their sizes are for illustration (larger than SimpleAug).

debiasing methods (*e.g.*, LMH [19], RUBi [11], and CSS [16]), KDDAug consistently boost their performance on both ID and OOD benchmarks.

In summary, we make three main contributions in this paper:

1. We systematically analyze existing DA strategies for robust VQA, and propose a new KDDAug that can avoid all the weaknesses of existing solutions.
2. We use multi-teacher KD to generate pseudo answers, which not only avoids human annotations, but also is more robust to both ID and OOD settings.
3. KDDAug is a model-agnostic DA method, which empirically boosts multiple different VQA architectures to achieve state-of-the-art performance.

2 Related Work

Language Biases in VQA. In order to overcome the language biases issues in VQA, many debiasing methods have been proposed recently. Specifically, existing methods can be roughly divided into two groups: 1) Ensemble-based debiasing methods. These methods always design an auxiliary branch to explicit model and exclude the language biases [45,25,11,19,38,40,26,35,52]. 2) Model-agnostic debiasing methods. These methods mainly include balancing datasets [24,54], data augmentation by generating augmented training samples [22,16,18,31,1,30], and designing extra training objectives [22,56,36]. Almost existing debiasing methods significantly improve their OOD performance, but with the cost of ID performance drops. In this paper, we deeply analyze existing DA methods, and propose a new DA strategy to achieve a decent trade-off between ID & OOD performance.

Data Augmentation in VQA. In addition to the mainstream synthetic-based methods, there are other DA methods: some existing methods generate negative samples by randomly selecting images or questions [48,56], or compose reasonable image-question (VQ) pairs as new positive training samples [31]. For these generated VQ pairs, they utilize manually pre-defined rules to obtain answers, which are designed for some specific question types. However, these DA methods almost either suffer a severe ID performance drop [16,48,18,32] or their answer assignment mechanisms rely on human annotations and lack generality [29,22,23,31,7]. Instead, our KDDAug overcomes all these weaknesses.

Knowledge Distillation. KD is a method that helps the training process of a smaller student network under the supervision of a larger teacher network [49]. The idea of KD has been applied to numerous vision tasks, *e.g.*, object detection [50,12] or visual-language tasks [55,42,34]. Recently, Niu *et.al.* [41] began to study KD for VQA and propose a KD-based method to generate "soft" labels in training. Inspired by them, we propose to use a multi-teacher KD to generate robust pseudo ground-truth labels for all new composed VQ pairs.

3 KDDAug: A New DA Framework for VQA

Following same conventions of existing VQA works, VQA task is typically formulated as a multi-class classification problem. Given a dataset $\mathcal{D}_{\text{orig}} = \{I_i, Q_i, a_i\}_i^N$

consisting of triplets of images $I_i \in \mathcal{I}$, questions $Q_i \in \mathcal{Q}$ and ground-truth answers $a_i \in \mathcal{A}$, VQA model learns a multimodal mapping: $\mathcal{I} \times \mathcal{Q} \rightarrow [0,1]^{|\mathcal{A}|}$, which produces an answer distribution given an image-question (VQ) pair.

To reduce the language biases, a surge of data augmentation (DA) methods have been proposed for VQA. Specifically, given the original training set $\mathcal{D}_{\text{orig}}$, DA methods generate an augmented training set $\mathcal{D}_{\text{aug}}$ automatically. Then, they can train any VQA architectures with both two training sets ($\mathcal{D}_{\text{orig}} \cup \mathcal{D}_{\text{aug}}$).

In this section, we first compare proposed KDDAug with existing DA methods in Sec. 3.1. Then, we introduce details of KDDAug, including image-question pair composition in Sec. 3.2, and KD-based answer assignment in Sec. 3.3.

3.1 KDDAug vs. Existing DA Pipelines

Synthetic-based Methods. For each human-labeled sample $(I_i, Q_i, a_i) \in \mathcal{D}_{\text{orig}}$, the synthetic-based methods (*e.g.*, CSS [16]) always synthesize one corresponding augmented sample by either editing the image I_i or question Q_i, denoted as $\hat{I}_i$ or $\hat{Q}_i$. Then, original image I_i and its synthesized question $\hat{Q}_i$ compose a new VQ pair $(I_i, \hat{Q}_i)$ (similar for VQ pair $(\hat{I}_i, Q_i)$ and $(\hat{I}_i, \hat{Q}_i)$). Lastly, different answer assignment mechanisms are designed to generate pseudo answers for these augmented VQ pairs, and these samples constitute the augmented set $\mathcal{D}_{\text{aug}}$. As discussed in Sec. 1, these new synthetic VQ pairs are unnatural and error-prone.
SimpleAug [31]. Unlike synthetic-based methods, SimpleAug tries to compose new VQ pairs by randomly sampling an image $I_i \in \mathcal{I}$ and other possible question $Q_j \in \mathcal{Q}$, *i.e.*, (I_i, Q_j). This simple strategy can make sure both image I_i and question Q_j are always pristine. Not surprisingly, there is no free lunch — this arbitrary composition strategy significantly increases the difficulty of pseudo answers assignment. To this end, SimpleAug proposes a set of heuristic rules for only four types of questions ("`Yes/No`", "`Number`", "`What`", and "`Color`"), which limits its diversity and generalization ability.
Proposed KDDAug. To solve all the weaknesses in existing DA methods (*i.e.*, both synthetic-based methods and SimpleAug), we take two steps to generate the augmented set $\mathcal{D}_{\text{aug}}$: 1) We randomly compose image $I_i \in \mathcal{I}$ and all reasonable questions $Q_j \in \mathcal{Q}$ without limiting question types. 2) We utilize a knowledge distillation (KD) based answer assignment to automatically generate "soft" pseudo answers for each VQ pair. Next, we detailed introduce these two steps.

3.2 Image-Question Pair Composition

To extremely increase the diversity of the new augmented training set $\mathcal{D}_{\text{aug}}$, we relax the requirements for *reasonable* VQ pairs by only considering the object categories in the images and nouns in the questions. By "reasonable", we mean that: 1) The question is suitable for the image content. 2) There are some ground-truth answers for this VQ pair. For example, as shown in Fig. 3(a), the question "`Why is the suitcase in the trunk?`" is not a reasonable question for the image, because "`truck`" does not appear in the image. Therefore, we treat question Q_j as a reasonable question for image I_i as long as all the meaningful

Fig. 3. Example of three randomly composed VQ pairs. (a) **Unreasonable pair**: The question contains noun "`trunk`" which are not in the image. (b) **Reasonable pair**: All nouns in the question ("`girl`" and "`sock`") are in the image. (c) **Reasonable pair**: Although the question contains "`hat`" which are not in the image, it is still reasonable.

nouns in Q_j appear in I_i. For example in Fig. 3(b), questions only containing "`girl`" and "`sock`" are all reasonable questions for this image[4].

Thus, similar to other DA methods [31,16,18], we first extract these meaningful nouns from all questions $\mathcal{Q}$. We utilize the spaCy POS tagger [27] to extract all nouns and unify their singular and plural forms[5]. We ignore the nouns such as "picture" or "photo". We remove all the questions without any meaningful nouns (the proportion is small, *e.g.*, $\approx 9\%$ in VQA-CP v2). For all images $\mathcal{I}$, we leverage an off-the-shelf object detector to detect all proposals in each image, and predict their object categories. Lastly, we compose all possible reasonable VQ pairs by traversing all the questions and images in the original training set.

Since the number of "`Yes/No`" questions is quite large (*e.g.*, $\approx 42\%$ in VQA-CP v2), to prevent creating too many "`Yes/No`" samples, we group all "`Yes/No`" questions with the same set of nouns into one group, and randomly select three questions per group for new sample compositions.

CLIP-based Filtering. One potential weakness for our VQ pair composition strategy is the excessive training samples, which may increase the training times. To achieve a good trade-off between efficiency and effectiveness, we can utilize a pretrained visual-language model CLIP [43] to filter out less-efficient augmented samples. By "less-efficient", we mean that the improvements provided by these training samples are marginal. This filtering design is based on the observation that people tend to ask questions about salient objects in the image, *i.e.*, the nouns mentioned in questions should appear prominently in the image. Specifically, we firstly use the template "`a photo of <NOUN>`" to generate prompts for each meaningful noun in the question and utilize CLIP to calculate the similarity score between the image and all corresponding prompts. Then, we use the average similarity score over all meaningful nouns to get the relevance score

[4]Some questions containing extra nouns may also be reasonable questions, especially for "`Yes/No`" questions (*cf.* example in Fig. 3(c)). However, almost all VQ pairs that meet this more strict requirement are always reasonable.

[5]We tried to use WordNet to map between nouns' synsets, lemmas or hypernyms (*e.g.*, "dog" and "animal"). But empirically, the VQA performance is quite similar.

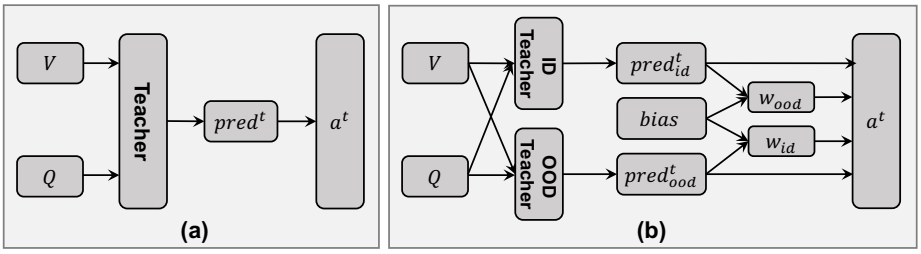

Fig. 4. Pipelines of single-teacher KD (a) and multi-teacher KD (b) answer assignment.

between the question and the image. We sort all composed VQ pairs according to the relevance score and only the $\alpha\%$ samples with the top highest relevance scores are reserved (See Table 5 for more details about influence of different α). **Advantages.** Compared with existing VQ pair composition strategies, ours has several advantages: 1) Our definition for "reasonable" simplifies the composition step and further improves the diversity of the VQ pairs. 2) Our strategy gets rid of human annotations, which means it can be easily extended to other datasets.

3.3 KD-based Answer Assignment

Given the original training set $\mathcal{D}_{\text{orig}}$ and all new composed reasonable VQ pairs ($\{I_i, Q_j\}$), we use a KD-based answer assignment to generate pseudo answers.
Single-Teacher KD for Answer Assignment. Inspired by existing KD work for pseudo labeling [44,41], we begin to shift our gaze from manual rules to KD. A straightforward KD-based strategy is training a *teacher* VQA model with the original training set $\mathcal{D}_{\text{orig}}$, and then using the pre-trained teacher model to predict answer distributions $pred^t$ for each composed VQ pair. And the $pred^t$ is treated as the pseudo answer for this VQ pair (*cf.* Fig. 4(a)). Obviously, the quality of assigned answers is determined by the performance of the teacher model, and the biases learned by the teacher model are also included in $pred^t$.
Multi-Teacher KD for Answer Assignment. To generate more accurate and robust pseudo answers, an extension is merging knowledge from multiple teachers. Given N pre-trained teacher models, and each trained model can predict an answer distribution $\{pred_i^t\}$, and the pseudo answer is:

$$a^t = \sum_{i=1}^{N} w_i * pred_i^t, \quad w.r.t \quad \sum_{i=1}^{N} w_i = 1, \tag{1}$$

where w_i denotes the weight of i-th teacher model.

To make pseudo ground-truth answers more informative and robust to both ID and OOD benchmarks, in KDDAug, we adopt two expert teacher models: *ID teacher* and *OOD teacher*. Among them, ID and OOD teachers can effectively extract ID and OOD knowledge, respectively. Given the predicted answer distributions of ID teacher and OOD teacher (denoted as $pred_{id}^t$ and $pred_{ood}^t$), we then need to calculate the weights for the two teachers. Following Niu *et.al.* [41], to obtain unbiased labels, we also assign a smaller weight to a more biased teacher. Since we lack any human-annotated ground-truth answers for these composed

VQ pairs, we directly measure the bias degree of each teacher by calculating the cross-entropy (`XE`) loss between question-type bias (*bias*) and their predictions:

$$c_{id} = \frac{1}{\texttt{XE}(bias, pred^t_{id})}, \quad c_{ood} = \frac{1}{\texttt{XE}(bias, pred^t_{ood})}, \tag{2}$$

where *bias* is the statistical answer distribution of each question type, which is calculated from the original training set $\mathcal{D}_{\text{orig}}$. Obviously, if the prediction is more closer to *bias*, the teacher is more like to be biased. Then, we obtain:

$$w_{id} = c_{ood}/(c_{id} + c_{ood}), \quad w_{ood} = c_{id}/(c_{id} + c_{ood}). \tag{3}$$

Lastly, we obtain pseudo answers a^t by Eq. (1) for each VQ pair (*cf.* Fig. 4(b)).

Benefiting from Rule-based Initial Answers. Based on Eq. (2), our answer assignment can be easily extended to further benefit from high-quality rule-based initial answers (*i.e.*, replacing *bias* with high-quality initial answers). We denote these initial pseudo ground-truth answers as a^{init}. Following SimpleAug [31], we consider three types of questions with a single noun:

1) **"`Color`" questions.** For each paired image, the detector may output some color attributes. We assign the color of the noun in the question as a^{init}.

2) **"`Number`" questions.** For each paired image, we assign the count of detected objects which are same as the noun in the question as a^{init}.

3) **"`What`" questions.** For each original sample (I_i, Q_i, a_i), and new paired image I_j for Q_i, if a_i is in I_j's object labels, we assign a_i as the initial answer for VQ pair (I_j, Q_i). For example, if original sample is "`What is near the fork? Knife.`" If paired image contains "`knife`", we assign "`knife`" as a^{init}.

We refer the readers to SimpleAug paper [31] for more details. After obtaining a^{init}, we can replace *bias* to a^{init} in these question types. Then, Eq. (2) becomes:

$$c_{id} = \frac{1}{\texttt{XE}(a^{init}, pred^t_{id})}, \quad c_{ood} = \frac{1}{\texttt{XE}(a^{init}, pred^t_{ood})}. \tag{4}$$

Considering that the initial answer a^{init} is more accurate than the $pred^t_{id}$ *e.g.*, ID teacher's ID performance is only 63.01% (*cf.* Table 6), we follow [41] and use a^{init} as ID knowledge, *i.e.*, $a^t = w_{id} * a^{init} + w_{ood} * pred^t_{ood}$.

Advantages. Compared with the existing answer assignment mechanism, our solution gets rid of heuristic rules and human annotations. Meanwhile, it can be easily extended to generate better answers with more advanced teacher models. Besides, it is more general, which can theoretically be applied to any VQ pair.

Why KDDAug can work? KDDAug improves performance from two aspects: 1) It composes new samples to increase the diversity of the training set, which implicitly mitigates the biases with more balanced data. 2) It assigns more robust and informative answers for *new* samples.

4 Experiments

4.1 Experimental Settings and Implementation Details

Evaluation Datasets. We evaluated the proposed KDDAug on two datasets: the ID benchmark **VQA v2** [24] and OOD benchmark **VQA-CP v2** [4]. For

model accuracies, we followed the standard VQA evaluation metric [6]. Meanwhile, we followed [40] and used Harmonic Mean (**HM**) to evaluate the trade-off between ID and OOD evaluations. More details are left in the appendix.

VQA Models. Since KDDAug is an architecture-agnostic DA method, we evaluated the effectiveness of KDDAug on multiple different VQA models: UpDn [5], LMH [19], RUBi [11] and CSS [16,18]. Specifically, UpDn is a simple but effective VQA model, which always serves as a backbone for other advanced VQA models. LMH, RUBi, and CSS are SOTA ensemble-based VQA models for debiasing. For each specific VQA baseline, we followed their respective configurations (*e.g.*, hyperparameter settings) and re-implemented them using the official codes.

ID & OOD Teachers. The ID and OOD teachers were from a same LMH-CSS [16] model with different architectures [40,41]. Since LMH-CSS is an ensemble-based debiasing model, we took the whole ensemble model (VQA w/ bias-only model) as ID teacher, and the bare VQA model as OOD teacher (original for debiasing). Benefiting from the different architectures, they can extract ID and OOD knowledge, respectively. We used the official LMH-CSS codes to train ID and OOD teachers simultaneously *on a same dataset*, *e.g.*, for VQA-CP evaluation, both teachers were trained on the VQA-CP training set.

Two Augmented Dataset Versions. Due to the huge amount of all reasonable composed VQ pairs, and to keep fair comparisons with existing DA methods (especially SimpleAug), we constructed two versions of augmented sets: 1) $D_{\text{aug}}^{\text{basic}}$: It only contains the same four types of questions as SimpleAug. Meanwhile, it also contains a same set of extra training samples by paraphrasing[6]. Thus, $D_{\text{aug}}^{\text{basic}}$ can clearly demonstrate the effectiveness of our proposed KD-based answer assignment strategy. 2) $D_{\text{aug}}^{\text{extra}}$: It contains all possible question types. Obviously, all VQ pairs from paraphrasing[6] are a subset of $D_{\text{aug}}^{\text{extra}}$. To decrease the number of augmented samples, we applied the CLIP-based filtering (*cf.* Sec. 3.2) to keep top 10 % samples. In the following experiments, we denote the model trained with $D_{\text{aug}}^{\text{basic}}$ and $D_{\text{aug}}^{\text{extra}}$ as **KDDAug** and **KDDAug**$^{+}$, respectively. Meanwhile, unless otherwise specified, we used the rule-based initial answers for $D_{\text{aug}}^{\text{basic}}$.

Training Details & KDDAug Settings. Details are left in the appendix.

4.2 Architecture Agnostic

Settings. Since KDDAug is a model-agnostic data augmentation method, it can seamlessly incorporated into any VQA architectures. To validate the generalization of KDDAug, we applied it to multiple different VQA models: UpDn [5], LMH [19], RUBi [11] and CSS^{+} [18]. All the results are shown in Table 1.

Results. Compared to these baseline models, KDDAug can consistently improve the performance for all architectures, and push all models' performance to the state-of-the-art level. Particularly, the improvements are most significant in the baseline UpDn model (*e.g.*, 12.61% absolute performance gains on HM).

[6]Paraphrasing is an supplementary DA tricks proposed by SimpleAug [31]. Specifically, for each original sample (I_i, Q_i, a_i), if question Q_j is similar to Q_i (predicted by pre-trained BERT [20]), they construct a new augmented training sample (I_i, Q_j, a_i).

Base	Models	VQA-CP v2 test				VQA v2 val				HM
		All	Y/N	Num	Other	**All**	Y/N	Num	Other	
UpDn [5]	Baseline	39.74	42.27	11.93	46.05	63.48	81.18	42.14	55.66	48.88
	Baseline*	39.85	42.66	12.18	45.98	**63.30**	**81.06**	**42.46**	**55.32**	48.91
	KDDAug	**60.24**$_{+20.39}$	**86.13**	**55.08**	**48.08**	62.86$_{-0.44}$	80.55	41.05	55.18	**61.52**$_{+12.61}$
LMH [19]	Baseline	52.05	—	—	—	—	—	—	—	—
	Baseline*	53.87	73.31	44.23	46.33	61.28	76.58	**55.11**	40.69	57.24
	KDDAug	**59.54**$_{+5.67}$	**86.09**	**54.84**	**46.92**	62.09$_{+0.81}$	**79.26**	40.11	**54.85**	**60.79**$_{+3.55}$
RUBi [11]	Baseline	44.23	—	—	—	—	—	—	—	—
	Baseline*	46.84	70.05	44.29	11.85	52.83	54.74	**41.56**	**54.38**	49.66
	KDDAug	**59.25**$_{+12.41}$	**84.16**	**54.12**	**47.61**	**60.25**$_{+7.42}$	**74.97**	40.29	54.35	**59.75**$_{+10.09}$
CSS$^+$ [18]	Baseline	59.54	83.37	52.57	48.97	59.96	73.69	40.18	54.77	59.75
	Baseline*	59.19	83.54	51.29	**48.59**	58.91	71.02	39.76	**54.77**	59.05
	KDDAug	**61.14**$_{+1.95}$	**88.31**	**56.10**	48.28	**62.17**$_{+3.26}$	**79.50**	**40.57**	54.71	**61.65**$_{+2.60}$

Table 1. Accuracies (%) on VQA-CP v2 and VQA v2 of different VQA architectures. * indicates the results from our reimplementation using official codes.

Furthermore, when KDDAug is applied to another DA-based model CSS$^+$, KDDAug can still improve the performance on both OOD and ID benchmarks, and achieve the best performance (*e.g.*, 61.65% on HM).

4.3 Comparisons with State-of-the-Arts

Settings. We incorporated the KDDAug into model UpDn [5], LMH [19] and CSS$^+$ [18], and compared them with the SOTA VQA models both on VQA-CP v2 and VQA v2. According to the model framework design, we group them into: 1) *Non-DA Methods*: UpDn [5], AReg [45], MuRel [10], GRL [25], CF-VQA [40], GGE-DQ [26], D-VQA [52], IntroD [41], CSS+CL [36], and LMH [19]. 2) *DA Methods*: CVL [1], Unshuffling [47], CSS [16], CSS$^+$ [18], RandImg [48], SSL [56], MUTANT [22], SimpleAug [31], and ECD [32]. All results are reported in Table 2.
Results. Compared with all existing DA methods, KDDAug achieves the best OOD and trade-off performance on two datasets. For UpDn backbone, KDDAug improves the OOD performance of UpDn with a 20% absolute performance gain (60.24% vs. 39.74%) and improves accuracies on all different question categories. For LMH backbone, KDDAug boosts the performance on both ID and OOD benchmarks. Compared with other non-DA methods, KDDAug still outperforms most of them. It is worth noting that our KDDAug can also be incorporated into these advanced non-DA models to further boost their performance.

4.4 Ablation Studies

We validate the effectiveness of each component of KDDAug by answering the following questions: **Q1**: Does KDDAug assign more robust answers than existing methods? **Q2**: Does KDDAug mainly rely on the rule-based initial answers? **Q3**: Does KDDAug only benefit from much more training samples? **Q4**: Does the multi-teacher design help to improve pseudo ground-truth answers quality? **Q5**: Is the diversity of question types important for the data augmentation model?

Models	DA	VQA-CP v2 test				VQA v2 val				HM
		All	Y/N	Num	Other	**All**	Y/N	Num	Other	
UpDn [5]$_{\text{CVPR'18}}$		39.74	42.27	11.93	46.05	63.48	81.18	42.14	55.66	48.88
+AReg [45]$_{\text{NeurIPS'18}}$		41.17	65.49	15.48	35.48	62.75	79.84	42.35	55.16	49.72
+MuRel [10]$_{\text{CVPR'19}}$		39.54	42.85	13.17	45.04	—	—	—	—	—
+GRL [25]$_{\text{ACL'19}}$		42.33	59.74	14.78	40.76	51.92	—	—	—	46.64
+CF-VQA [40]$_{\text{CVPR'21}}$		53.55	91.15	13.03	44.97	63.54	82.51	43.96	54.30	58.12
+GGE-DQ [26]$_{\text{ICCV'21}}$		57.32	87.04	27.75	49.59	59.11	73.27	39.99	54.39	58.20
+D-VQA [52]$_{\text{NeurIPS'21}}$		61.91	88.93	52.32	50.39	64.96	82.18	44.05	57.54	63.40
+CVL [1]$_{\text{CVPR'20}}$	✓	42.12	45.72	12.45	48.34	—	—	—	—	—
+Unshuffling [47]$_{\text{ICCV'21}}$	✓	42.39	47.72	14.43	47.24	61.08	78.32	42.16	52.81	50.05
+CSS [16]$_{\text{CVPR'20}}$	✓	41.16	43.96	12.78	47.48	—	—	—	—	—
+CSS$^+$ [18]$_{\text{arXiv'21}}$	✓	40.84	43.09	12.74	47.37	—	—	—	—	—
+RandImg [48]$_{\text{NeurIPS'20}}$	✓	55.37	83.89	41.60	44.20	57.24	76.53	33.87	48.57	56.29
+SSL [56]$_{\text{IJCAI'20}}$	✓	57.59	86.53	29.87	50.03	63.73	—	—	—	60.50
+MUTANT† [22]$_{\text{EMNLP'20}}$	✓	50.16	61.45	35.87	50.14	—	—	—	—	—
+SimpleAug [31]$_{\text{EMNLP'21}}$	✓	52.65	66.40	43.43	47.98	**64.34**	**81.97**	**43.91**	**56.35**	57.91
+**KDDAug**	✓	**60.24**	**86.13**	**55.08**	**48.08**	62.86	80.55	41.05	55.18	**61.52**
LMH* [19]$_{\text{EMNLP'19}}$		53.87	73.31	44.23	46.33	61.28	76.58	55.11	40.69	57.24
+IntroD [41]$_{\text{NeurIPS'21}}$		51.31	71.39	27.13	47.41	62.05	77.65	40.25	55.97	56.17
+CSS+CL [36]$_{\text{EMNLP'20}}$		59.18	86.99	49.89	47.16	57.29	67.27	38.40	54.71	58.22
+CSS+IntroD [41]$_{\text{NeurIPS'21}}$		60.17	89.17	46.91	48.62	62.57	78.57	41.42	56.00	61.35
+CSS [16]$_{\text{CVPR'20}}$	✓	58.95	84.37	49.42	48.21	59.91	73.25	39.77	55.11	59.43
+CSS$^+$ [18]$_{\text{arXiv'21}}$	✓	59.54	83.37	52.57	48.97	59.96	73.69	40.18	54.77	59.75
+SimpleAug [31]$_{\text{EMNLP'21}}$	✓	53.70	74.79	34.32	47.97	**62.63**	79.31	**41.71**	**55.48**	57.82
+ECD [32]$_{\text{WACV'22}}$	✓	59.92	83.23	52.29	**49.71**	57.38	69.06	35.74	54.25	58.62
+**KDDAug**	✓	59.54	86.09	54.84	46.92	62.09	79.26	40.11	54.85	60.79
+**CSS$^+$+KDDAug**	✓	**61.14**	**88.31**	**56.10**	48.28	62.17	**79.50**	40.57	54.71	**61.65**

Table 2. Accuracies (%) on VQA-CP v2 and VQA v2 of SOTA models. "**DA**" denotes the data augmentation methods. * indicates the results from our reimplementation. "MUTANT†" denotes MUTANT [22] only trained with `XE` loss.

KDDAug vs. SimpleAug [31] (Q1). To answer **Q1**, we compared the answers assigned by KDDAug and SimpleAug. Due to different composition strategies for "`Yes/No`" questions, we firstly removed the "`Yes/No`" questions from $D_{\text{aug}}^{\text{basic}}$. We use a pretrained CLIP [43] (denoted as $\text{CLIP}_{\text{rank}}$ and more details are left in the appendix.) to rank the quality of all SimpleAug assigned answers. For more comprehensive comparisons, we divided the augmented samples into three subsets according to the ranks: 1) All augmented samples (100%), 2) Top-50% samples ($\uparrow$ 50%), and 3) Bottom-50% samples ($\downarrow$ 50%). We compared KDDAug and SimpleAug on these three subsets, and results are in Table 3.

Results for Q1. From Table 3, we can observe: 1) The performance of SimpleAug on three subsets varies significantly, *e.g.*, bottom-50% samples lead to a huge drop in HM (-0.87%). 2) In contrast, KDDAug on different subsets achieves similar decent performance. 3) With the same augmented image-question pairs, KDDAug consistently outperforms the corresponding SimpleAug by a significant margin (over 8% on HM), which proves the robustness of our answers.

Influence of Rule-based Initial Answers (Q2). To evaluate the quality of directly automatically assigned answers, we again used $\text{CLIP}_{\text{rank}}$ to divide the augmented samples (except "`Yes/No`" samples) into two parts. The augmented

Models	VQA-CP v2 test				VQA v2 val				HM
	All	Yes/No	Num	Other	**All**	Yes/No	Num	Other	
UpDn [5]	39.74	42.27	11.93	46.05	63.48	81.18	42.14	55.66	48.88
+SimpleAug [31]	52.65	66.40	43.43	47.98	64.34	81.97	43.91	56.35	57.91
+SimpleAug*	48.56	58.83	35.41	46.78	60.67	75.65	40.45	54.65	54.57
+SimpleAug$^-$ (100%)	45.38	45.59	37.63	47.40	62.62	80.49	40.68	54.85	52.62
+SimpleAug$^-$ ($\uparrow$ 50%)	46.35	47.72	39.36	47.55	62.51	80.64	40.08	54.67	53.23
+SimpleAug$^-$ ($\downarrow$ 50%)	45.06	44.84	38.44	46.99	62.49	80.38	40.24	54.78	52.36
+**KDDAug**	60.24	86.13	55.08	48.08	62.86	80.55	41.05	55.18	61.52
+**KDDAug**$^-$ (100%)	59.96	84.95	54.98	48.23	62.72	80.07	40.90	55.30	61.31
+**KDDAug**$^-$ ($\uparrow$ 50%)	59.94	84.78	54.70	48.36	62.67	79.86	41.06	55.32	61.27
+**KDDAug**$^-$ ($\downarrow$ 50%)	59.97	85.11	55.13	48.13	62.60	80.13	40.78	55.04	61.26

Table 3. Accuracies (%) on different augmented subsets. * indicates our reimplementation. For fair comparisons, SimpleAug* didn't leverage human annotation and didn't remove examples that can be answered. $^-$ denotes without "`Yes/No`" questions.

	Baseline	KDDAug									
	(UpDn)	0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
VQA-CP v2	39.74	53.99	54.37	54.98	56.17	57.51	58.80	59.27	59.68	60.02	60.19	60.24
VQA v2	63.48	63.26	63.24	63.20	63.20	63.19	63.10	63.08	63.04	63.01	62.83	62.86
HM	48.88	58.26	58.47	58.80	59.48	60.22	60.87	61.12	61.31	61.48	61.48	61.52

Table 4. Accuracies (%) on VQA-CP v2 test set and VQA v2 valset of different δ.

samples with top-δ% ranks use rule-based initial answers (*cf.* Eq. (4)) and left samples use question-type *bias* (*cf.* Eq. (2)). All results are shown in Table 4.

Results for Q2. From the results, we can observe that KDDAug achieves consistent gains against the baseline (UpDn) on all proportions. The performance is best when all samples use rule-based initial samples. Even if all the samples' answers are assigned without any initial answers, KDDAug still gains significant improvement gains (58.26% vs. 48.88% on HM), and is better than SimpleAug.

Influence of Number of Augmented Samples (Q3). We set different α values in the CLIP-based filtering to control the number of augmented samples.

Results for Q3. From the results in Table 5, we have several observations: 1) When more samples are used, the model performs better, which reflects the robustness of generated pseudo answers. 2) Even if a small number of samples are used (*e.g.*, $\alpha\% = 10\%$), KDDAug still achieves decent performance, and is better than SimpleAug (60.85% vs. 57.91%). 3) By adjusting the value of α, we can easily achieve a trade-off between training efficiency and model performance.

Effects of Different Teachers (Q4). To show the effectiveness of multi-teacher KD strategy, we compared three teachers with different ID teacher weight w_{id} and OOD teacher weight w_{ood}: 1) *Averaged Teacher* (Simple Avg.): $w_{id} = w_{ood} = 0.5$. 2) *Only ID Teacher* (ID-distill.): $w_{id} = 1, w_{ood} = 0$. 3) *Only OOD Teacher* (OOD-distill.): $w_{id} = 0, w_{ood} = 1$. All results are shown in Table 6.

Results for Q4. From Table 6, we have several observations: 1) Learning from all teachers can improve the performance over baseline. 2) Learning from fixed-weight teachers (*i.e.*, single-teacher KD) can't achieve good performance on both ID and OOD settings simultaneously, *e.g.*, "OOD-distill." obtains OOD

Models (α%)	VQA-CP v2 test					VQA v2 val					HM
	All	Y/N	Num	Other	#Samples	**All**	Y/N	Num	Other	#Samples	
UpDn [5]	39.74	42.27	11.93	46.05	438K	63.48	81.18	42.14	55.66	444K	48.88
+KDDAug (100%)	60.24	86.13	55.08	48.08	+4,088K	62.86	80.55	41.05	55.18	+2,279K	61.52
+KDDAug (90%)	60.19	86.09	55.13	48.00	+3,679K	62.83	80.53	41.14	55.12	+2,051K	61.48
+KDDAug (70%)	60.12	85.96	55.09	47.96	+2,861K	62.82	80.52	40.99	55.14	+1,595K	61.44
+KDDAug (50%)	60.13	86.18	54.81	47.94	+2,044K	62.71	80.41	40.98	55.00	+1,139K	61.39
+KDDAug (30%)	59.92	86.06	54.74	47.64	+1,226K	62.51	80.35	40.56	54.76	+684K	61.19
+KDDAug (10%)	59.41	85.81	54.85	46.82	+409K	62.37	80.47	40.98	54.28	+228K	60.85
+SimgpleAug [31]	52.65	66.40	43.43	47.98	+3,081K	64.34	81.97	43.91	56.35	—	57.91
+SimgpleAug*	48.56	58.83	35.41	46.78	+4,702K	60.67	75.65	40.45	54.65	+2,358K	54.57

Table 5. Accuracies (%) of different α in CLIP-based filtering. SimpleAug* is the same as Table 3. "#Samples" denotes the number of total training samples.

Models	**OOD W.**	VQA-CP v2 test				VQA v2 val				**HM**
		All	Y/N	Num	Other	**All**	Y/N	Num	Other	
UpDn [5] (baseline)		39.74	42.27	11.93	46.05	63.48	81.18	42.14	55.66	48.88
ID-Teacher		36.93	36.56	12.82	43.73	63.01	80.76	42.30	55.01	46.57
OOD-Teacher		58.07	82.47	52.03	46.93	60.21	74.19	40.32	54.86	59.12
Simple Avg.	0.5	53.06	63.89	50.72	48.03	**63.45**	**81.33**	**42.48**	**55.41**	57.79
ID-distill.	0.0	43.10	42.33	29.34	47.28	62.90	81.10	41.05	54.84	51.15
OOD-distill.	1.0	58.40	82.27	53.16	47.33	61.50	77.08	41.62	54.92	59.91
KDDAug (Ours)	dynamic	**60.24**	**86.13**	**55.08**	**48.08**	62.86	80.55	41.05	55.18	**61.52**

Table 6. Effects of different teachers on pseudo answer assignment. "OOD W." denotes w_{ood}, and "dynamic" denotes w_{ood} is dynamically calculated by our strategy.

performance gains (+18.66%) while suffering from a significant drop on ID performance (-1.98%). 3) In contrast, our dynamic multi-teacher strategy increases OOD performance by 20.50% while ID performance drops slightly by 0.62%, and achieves the best trade-off performance, which proves its effectiveness.

Effects of Augmentation Diversity (Q5). To explore the effects of more diverse augmentation types, we compared KDDAug and KDDAug$^+$ with $D_{\text{aug}}^{\text{basic}}$ and $D_{\text{aug}}^{\text{extra}}$. For fair comparison, we didn't use initial answers and removed all paraphrasing samples from $D_{\text{aug}}^{\text{basic}}$ (denoted as $D_{\text{aug}^-}^{\text{basic}}$) since they are a subset of $\mathcal{D}_{extra}$. Meanwhile, we sampled same number of samples of the "`Other`" cateogry samples with $D_{\text{aug}^-}^{\text{basic}}$ from $D_{\text{aug}}^{\text{extra}}$[7]. We compared them on different size of samples (different α values in CLIP-based filtering). Results are shown in Table 7.

Results for Q5. From the results, we can observe that with more diverse augmented samples, KDDAug can consistently improve both ID and OOD performance for all different α, especially on "`Other`" category (*e.g.*, $> 0.78\%$ gains). In particular, even if we don't rely on any rule-based answers, KDDAug surpasses the baseline UpDn model on all categories on VQA-CP v2 when $\alpha = 50$ or 100.

4.5 Visualization Results

We show some augmented samples by KDDAug in Fig. 5. From Fig.5, we can observe that our KDDAug can compose potential reasonable VQ pairs and assign

[7] $D_{\text{aug}^-}^{\text{basic}}$ and $D_{\text{aug}}^{\text{extra}}$ have same "`Yes/No`" and "`Number`" category samples.

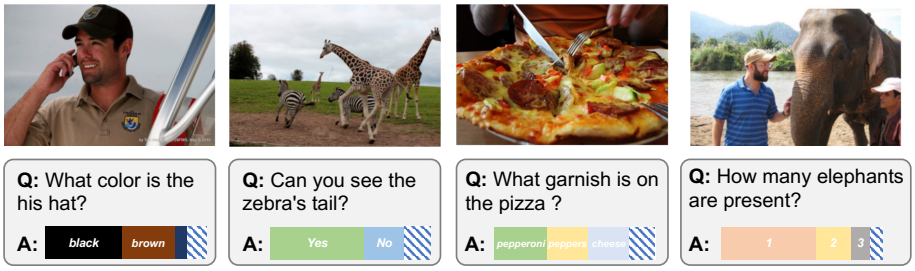

Fig. 5. Visualization results of some augmented samples by our KDDAug.

Models (α%)	**Extra**	VQA-CP v2 test				VQA v2 val				**HM**
		All	Y/N	Num	Other	**All**	Y/N	Num	Other	
UpDn [5]		39.74	42.27	11.93	46.05	63.48	81.18	42.14	55.66	48.88
+KDDAug‡ (100%)		53.03	**86.55**	**17.30**	45.26	61.59	80.43	42.33	52.36	56.99
+KDDAug$^{+\ddagger}$ (100%)	✓	**53.76**	86.41	17.10	**46.70**	**62.58**	**80.51**	**42.43**	**54.28**	**57.84**
+KDDAug‡ (50%)		52.91	85.59	17.01	45.64	62.01	80.42	41.98	53.31	57.10
+KDDAug$^{+\ddagger}$ (50%)	✓	**53.76**	**86.57**	**17.86**	**46.42**	**62.64**	**80.35**	**42.25**	**54.56**	**57.86**
+KDDAug‡ (10%)		51.26	82.28	17.13	44.37	62.03	80.47	**41.41**	53.47	56.13
+KDDAug$^{+\ddagger}$ (10%)	✓	**52.53**	**83.16**	**17.76**	**46.01**	**62.71**	**80.76**	41.34	**54.63**	**57.17**

Table 7. Accuracies (%) on VQA-CP v2 and VQA v2. ‡ denotes using $\mathcal{D}_{\text{aug}^-}^{\text{basic}}$. "**Extra**" denotes using $\mathcal{D}_{\text{aug}}^{\text{extra}}$.

satisfactory pseudo labels. Take the third question "`What garnish is on the pizza?`" as an example, there are multiple garnishes on the pizza, and KDDAug cleverly assigns multiple answers: "`pepperoni`", "`peppers`" and "`cheese`", which demonstrates the superiority of the "soft" pseudo labels generated by KDDAug.

5 Conclusions and Future Work

In this paper, we proposed a model-agnostic Knowledge Distillation based Data Augmentation (KDDAug) for VQA. KDDAug relaxes the requirements for pairing reasonable image-question pairs, and utilizes a multi-teacher KD to generate robust pseudo labels for augmented samples. KDDAug can consistently improve both ID and OOD performance of different VQA baselines. We validated the effectiveness of KDDAug through extensive experiments. Moving forward, we are going to 1) extend the KDDAug-like DA strategy to other visual-language tasks (*e.g.*, captioning [17,13,39] or grounding [15,14,37,53]); 2) design some specific training objectives (*e.g.*, contrastive loss) to further benefit from these augmented samples. 3) further improve the generalization ability of VQA models by incorporating other available large-scale datasets.

Acknowledgement. This work was supported by the National Key Research & Development Project of China (2021ZD0110700), the National Natural Science Foundation of China (U19B2043, 61976185), Zhejiang Natural Science Foundation (LR19F020002), Zhejiang Innovation Foundation(2019R52002), and the Fundamental Research Funds for the Central Universities (226-2022-00087).

References

1. Abbasnejad, E., Teney, D., Parvaneh, A., Shi, J., Hengel, A.v.d.: Counterfactual vision and language learning. In: CVPR (2020) 4, 10, 11
2. Agarwal, V., Shetty, R., Fritz, M.: Towards causal vqa: Reveling and reducing spurious correlations by invariant and covariant semantic editing. In: CVPR (2020) 2
3. Agrawal, A., Batra, D., Parikh, D.: Analyzing the behavior of visual question answering models. In: EMNLP (2016) 2
4. Agrawal, A., Batra, D., Parikh, D., Kembhavi, A.: Don't just assume; look and answer: Overcoming priors for visual question answering. In: CVPR (2018) 2, 8
5. Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L.: Bottom-up and top-down attention for image captioning and visual question answering. In: CVPR (2018) 3, 9, 10, 11, 12, 13, 14
6. 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) 9
7. Askarian, N., Abbasnejad, E., Zukerman, I., Buntine, W., Haffari, G.: Inductive biases for low data vqa: A data augmentation approach. In: WACV. pp. 231–240 (2022) 2, 4
8. Bitton, Y., Stanovsky, G., Schwartz, R., Elhadad, M.: Automatic generation of contrast sets from scene graphs: Probing the compositional consistency of GQA. In: NAACL. pp. 94–105 (2021) 2
9. Boukhers, Z., Hartmann, T., Jürjens, J.: Coin: Counterfactual image generation for vqa interpretation. In: arXiv (2022) 2
10. Cadene, R., Ben-Younes, H., Cord, M., Thome, N.: Murel: Multimodal relational reasoning for visual question answering. In: CVPR (2019) 10, 11
11. Cadene, R., Dancette, C., Ben-younes, H., Cord, M., Parikh, D.: Rubi: Reducing unimodal biases in visual question answering. In: NeurIPS (2019) 4, 9, 10
12. Chen, G., Choi, W., Yu, X., Han, T., Chandraker, M.: Learning efficient object detection models with knowledge distillation. NeurIPS (2017) 4
13. Chen, L., Jiang, Z., Xiao, J., Liu, W.: Human-like controllable image captioning with verb-specific semantic roles. In: CVPR. pp. 16846–16856 (2021) 14
14. Chen, L., Lu, C., Tang, S., Xiao, J., Zhang, D., Tan, C., Li, X.: Rethinking the bottom-up framework for query-based video localization. In: AAAI. pp. 10551–10558 (2020) 14
15. Chen, L., Ma, W., Xiao, J., Zhang, H., Chang, S.F.: Ref-nms: Breaking proposal bottlenecks in two-stage referring expression grounding. In: AAAI. pp. 1036–1044 (2021) 14
16. Chen, L., Yan, X., Xiao, J., Zhang, H., Pu, S., Zhuang, Y.: Counterfactual samples synthesizing for robust visual question answering. In: CVPR. pp. 10800–10809 (2020) 2, 4, 5, 6, 9, 10, 11
17. Chen, L., Zhang, H., Xiao, J., Nie, L., Shao, J., Liu, W., Chua, T.S.: Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In: CVPR. pp. 5659–5667 (2017) 14
18. Chen, L., Zheng, Y., Niu, Y., Zhang, H., Xiao, J.: Counterfactual samples synthesizing and training for robust visual question answering. arXiv (2021) 2, 4, 6, 9, 10, 11
19. Clark, C., Yatskar, M., Zettlemoyer, L.: Don't take the easy way out: Ensemble based methods for avoiding known dataset biases. In: EMNLP (2019) 4, 9, 10, 11

20. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: NAACL. pp. 4171–4186 (2019) 9
21. Geman, D., Geman, S., Hallonquist, N., Younes, L.: Visual turing test for computer vision systems. PNAS pp. 3618–3623 (2015) 1
22. Gokhale, T., Banerjee, P., Baral, C., Yang, Y.: Mutant: A training paradigm for out-of-distribution generalization in visual question answering. In: EMNLP (2020) 2, 4, 10, 11
23. Gokhale, T., Banerjee, P., Baral, C., Yang, Y.: Vqa-lol: Visual question answering under the lens of logic. In: ECCV. pp. 379–396 (2020) 2, 4
24. 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) 2, 3, 4, 8
25. Grand, G., Belinkov, Y.: Adversarial regularization for visual question answering: Strengths, shortcomings, and side effects. In: ACLW (2019) 4, 10, 11
26. Han, X., Wang, S., Su, C., Huang, Q., Tian, Q.: Greedy gradient ensemble for robust visual question answering. In: ICCV (2021) 4, 10, 11
27. Honnibal, M., Montani, I.: spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. To appear (2017) 6
28. Johnson, J., Hariharan, B., van der Maaten, L., Fei-Fei, L., Lawrence Zitnick, C., Girshick, R.: Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In: CVPR (2017) 2
29. Kafle, K., Yousefhussien, M., Kanan, C.: Data augmentation for visual question answering. In: INLG. pp. 198–202 (2017) 2, 4
30. Kant, Y., Moudgil, A., Batra, D., Parikh, D., Agrawal, H.: Contrast and classify: Training robust vqa models. In: ICCV. pp. 1604–1613 (2021) 2, 4
31. 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) 2, 4, 5, 6, 8, 9, 10, 11, 12, 13
32. Kolling, C., More, M., Gavenski, N., Pooch, E., Parraga, O., Barros, R.C.: Efficient counterfactual debiasing for visual question answering. In: WACV. pp. 3001–3010 (2022) 2, 4, 10, 11
33. 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) 3
34. Li, X., Chen, L., Ma, W., Yang, Y., Xiao, J.: Integrating object-aware and interaction-aware knowledge for weakly supervised scene graph generation. In: ACM MM (2022) 4
35. Liang, Z., Hu, H., Zhu, J.: Lpf: A language-prior feedback objective function for de-biased visual question answering. In: ACM SIGIR. pp. 1955–1959 (2021) 4
36. Liang, Z., Jiang, W., Hu, H., Zhu, J.: Learning to contrast the counterfactual samples for robust visual question answering. In: EMNLP (2020) 4, 10, 11
37. Lu, C., Chen, L., Tan, C., Li, X., Xiao, J.: Debug: A dense bottom-up grounding approach for natural language video localization. In: EMNLP. pp. 5144–5153 (2019) 14
38. Mahabadi, R.K., Belinkov, Y., Henderson, J.: End-to-end bias mitigation by modelling biases in corpora. In: ACL. pp. 8706–8716 (2020) 4
39. Mao, Y., Chen, L., Jiang, Z., Zhang, D., Zhang, Z., Shao, J., Xiao, J.: Rethinking the reference-based distinctive image captioning. In: ACM MM (2022) 14

40. 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) 4, 9, 10, 11
41. Niu, Y., Zhang, H.: Introspective distillation for robust question answering. In: NeurIPS (2021) 4, 7, 8, 9, 10, 11
42. Pan, B., Cai, H., Huang, D.A., Lee, K.H., Gaidon, A., Adeli, E., Niebles, J.C.: Spatio-temporal graph for video captioning with knowledge distillation. In: CVPR. pp. 10870–10879 (2020) 4
43. 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) 6, 11
44. Radosavovic, I., Dollár, P., Girshick, R., Gkioxari, G., He, K.: Data distillation: Towards omni-supervised learning. In: CVPR. pp. 4119–4128 (2018) 7
45. Ramakrishnan, S., Agrawal, A., Lee, S.: Overcoming language priors in visual question answering with adversarial regularization. In: NeurIPS (2018) 4, 10, 11
46. Tang, R., Ma, C., Zhang, W.E., Wu, Q., Yang, X.: Semantic equivalent adversarial data augmentation for visual question answering. In: ECCV. pp. 437–453 (2020) 2
47. Teney, D., Abbasnejad, E., Hengel, A.v.d.: Unshuffling data for improved generalization. In: ICCV (2021) 10, 11
48. Teney, D., Kafle, K., Shrestha, R., Abbasnejad, E., Kanan, C., Hengel, A.v.d.: On the value of out-of-distribution testing: An example of goodhart's law. In: NeurIPS (2020) 4, 10, 11
49. Wang, L., Yoon, K.J.: Knowledge distillation and student-teacher learning for visual intelligence: A review and new outlooks. IEEE TPAMI (2021) 4
50. Wang, T., Yuan, L., Zhang, X., Feng, J.: Distilling object detectors with fine-grained feature imitation. In: CVPR. pp. 4933–4942 (2019) 4
51. Wang, Z., Miao, Y., Specia, L.: Cross-modal generative augmentation for visual question answering. In: BMVC (2021) 2
52. Wen, Z., Xu, G., Tan, M., Wu, Q., Wu, Q.: Debiased visual question answering from feature and sample perspectives. In: NeurIPS (2021) 4, 10, 11
53. Xiao, S., Chen, L., Zhang, S., Ji, W., Shao, J., Ye, L., Xiao, J.: Boundary proposal network for two-stage natural language video localization. In: AAAI. pp. 2986–2994 (2021) 14
54. Zhang, P., Goyal, Y., Summers-Stay, D., Batra, D., Parikh, D.: Yin and yang: Balancing and answering binary visual questions. In: CVPR (2016) 2, 4
55. Zhang, Z., Shi, Y., Yuan, C., Li, B., Wang, P., Hu, W., Zha, Z.J.: Object relational graph with teacher-recommended learning for video captioning. In: CVPR. pp. 13278–13288 (2020) 4
56. Zhu, X., Mao, Z., Liu, C., Zhang, P., Wang, B., Zhang, Y.: Overcoming language priors with self-supervised learning for visual question answering. In: IJCAI (2020) 4, 10, 11