Discovering the Unknown Knowns:
Turning Implicit Knowledge in the Dataset into Explicit Training Examples for Visual Question Answering

Jihyung Kil, Cheng Zhang, Dong Xuan, Wei-Lun Chao

The Ohio State University, Columbus, OH, USA

`{kil.5, zhang.7804, xuan.3, chao.209}@osu.edu`

Abstract

Visual question answering (VQA) is challenging not only because the model has to handle multi-modal information, but also because it is just so hard to collect sufficient training examples — there are too many questions one can ask about an image. As a result, a VQA model trained solely on human-annotated examples could easily over-fit specific question styles or image contents that are being asked, leaving the model largely ignorant about the sheer diversity of questions. Existing methods address this issue primarily by introducing an auxiliary task such as visual grounding, cycle consistency, or debiasing. In this paper, we take a drastically different approach. We found that many of the "unknowns" to the learned VQA model are indeed "known" in the dataset implicitly. For instance, questions asking about the same object in different images are likely paraphrases; the number of detected or annotated objects in an image already provides the answer to the "how many" question, even if the question has not been annotated for that image. Building upon these insights, we present a simple data augmentation pipeline SIMPLEAUG to turn this "known" knowledge into training examples for VQA. We show that these augmented examples can notably improve the learned VQA models' performance, not only on the VQA-CP dataset with language prior shifts but also on the VQA v2 dataset without such shifts. Our method further opens up the door to leverage weakly-labeled or unlabeled images in a principled way to enhance VQA models. Our code and data are publicly available at https://github.com/heendung/simpleAUG.

1 Introduction

"A picture is worth a thousand words," which tells how expressive an image can be, but also how challenging it is to teach a machine to understand an image like we humans do. Visual question answering (VQA) (Antol et al., 2015; Goyal et al., 2017; Zhu
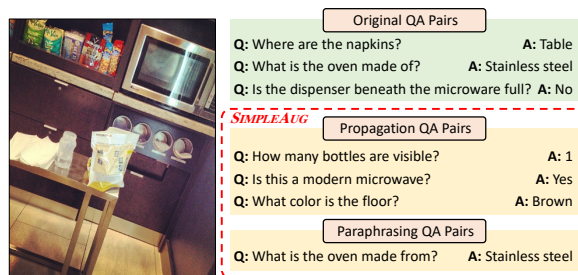

Figure 1: **Illustration of our approach SIMPLEAUG**. We show a *training* image and its corresponding question-answer pairs in VQA v2 (Goyal et al., 2017), and our generated pairs. A VQA model (Anderson et al., 2018) trained on the original dataset *just cannot answer these new questions on the training image correctly*, and we use them to improve model training.

et al., 2016) is a principled way to measure such an ability of a machine, in which given an image, a machine has to answer the image-related questions in natural language by natural language. While after years of effort, the state-of-the-art machine's performance is still behind what we expect (Hu et al., 2018; Yu et al., 2018; Anderson et al., 2018; Lu et al., 2019; Hudson and Manning, 2019).

Several key bottlenecks have been identified. In particular, a machine (*i.e.*, VQA model) learned in the conventional supervised manner using human-annotated image-question-answer (IQA) triplets is shown to overlook the image or language contents (Agrawal et al., 2016; Goyal et al., 2017; Chao et al., 2018a), over-fit the language bias (Agrawal et al., 2018), or struggle in capturing the diversity of human language (Shah et al., 2019; Chao et al., 2018b). Many recent works thus propose to augment the original VQA task with auxiliary tasks or losses such as visual grounding (Selvaraju et al., 2019; Wu and Mooney, 2019), de-biasing (Cadene et al., 2019; Clark et al., 2019; Ramakrishnan et al., 2018), or (cycle-)consistency (Shah et al., 2019; Gokhale et al., 2020a) to address these issues.

Intrigued by these findings and solutions, we investigate the bottlenecks further and argue that they may result from a more fundamental issue — there

6346

are simply not enough training examples (*i.e.*, IQA triplets). Concretely, most of the existing VQA datasets annotate each image with around ten questions, which are much fewer than what we humans can ask about an image. Take the popular VQA v2 dataset (Goyal et al., 2017), for instance, the trained VQA model can answer most of the training examples (on average, six questions per image) correctly. However, if we ask some more questions about the training images — *e.g.*, by borrowing relevant questions from other training images — the same VQA model fails drastically, *even if the model has indeed seen these images and questions during training* (see Figure 1). Namely, the VQA model just has not learned enough through the human-annotated examples, leaving the model unaware of the huge amount of visual information in an image and how it can be asked via natural language.

At first glance, this seems to paint a grim picture for VQA. However, in this paper we propose to take advantage of this weakness to strengthen the VQA model: we turn implicit information already in the dataset, such as unique questions and the rich contents in the training images, into explicit IQA triplets which can be directly used by VQA models via conventional supervised learning.

We propose a simple data augmentation method SIMPLEAUG, which relies on (i) the original image-question-answer triplets in the dataset, (ii) mid-level semantic annotations available on the training images (*e.g.*, object bounding boxes), and (iii) pretrained object detectors (Ren et al., 2016)[1]. Concretely, we build upon the aforementioned observations — *questions annotated for one image can be valuable add-ons to other relevant images* — and design a series of mechanisms to "propagate" questions from one image to the others. More specifically, we search images that contain objects mentioned in the question and identify the answers using information provided by (ii) and (iii), such as numbers of objects, their attributes, and existences.

SIMPLEAUG requires no question generation step via templates or language models (Kafle et al., 2017), bypassing the problems of limited diversity or artifacts. Besides, SIMPLEAUG is completely detached from the training phase of a VQA model and is therefore model-agnostic, making it fairly simple to use to improve VQA models.

We validate SIMPLEAUG on two datasets, VQA v2 (Goyal et al., 2017) and VQA-CP (Agrawal et al., 2018). The latter is designed to evaluate VQA models' generalizability under language bias shifts. With SIMPLEAUG, we can not only achieve comparable gains to other existing methods on VQA-CP, but also boost the accuracy on VQA v2, demonstrating the applicability of our method. We note that many of the prior works designed for VQA-CP indeed degrade the accuracy on VQA v2, which does not have language bias shifts between training and test data. SIMPLEAUG further justifies that mid-level vision tasks like object detection can effectively benefit high-level vision tasks like VQA.

In summary, our contributions are three-folded:

- We propose SIMPLEAUG, a simple and model-agnostic data augmentation method that turns information already in the datasets into explicit IQA triplets for training VQA models.

- We show that SIMPLEAUG can notably improve VQA models' accuracy on both VQA v2 (Goyal et al., 2017) and VQA-CP (Clark et al., 2019).

- We provide comprehensive analyses on SIMPLEAUG, including its applicability to weakly-labeled and unlabeled images.

2 Related Work

VQA datasets. More than a dozen datasets have been released (Lin et al., 2014; Antol et al., 2015; Zhu et al., 2016; Goyal et al., 2017; Krishna et al., 2017; Hudson and Manning, 2019; Gurari et al., 2019). Most of them use natural images from large-scale image databases, *e.g.*, MSCOCO (Lin et al., 2014). For each image, human annotators are asked to generate questions (Q) and provide the corresponding answers (A). Doing so, however, is hard to cover all the knowledge in the visual contents.

Leveraging side information for VQA. A variety of side information beyond the IQA triplets has been used to improve VQA models. For example, human attentions are used to enhance the explainability and visual grounding of VQA models (Patro and Namboodiri, 2018; Selvaraju et al., 2019; Das et al., 2017; Wu and Mooney, 2019). Image captions contain substantial visual information and can be used as an auxiliary task (*i.e.*, visual captioning) to strengthen VQA models' visual and language understanding (Wu et al., 2019; Kim and Bansal, 2019; Wang et al., 2021; Banerjee et al., 2020; Karpathy and Fei-Fei, 2015). Several papers

---

[1] We note that (ii) is commonly provided in existing VQA datasets like VQA v2 (Goyal et al., 2017), and (iii) has been widely used in the feature extraction stage of a VQA model (Anderson et al., 2018).

leveraged scene graphs and visual relationships as auxiliary knowledge for VQA (Johnson et al., 2017; Zhang et al., 2019; Hudson and Manning, 2019; Shi et al., 2019). A few works utilized mid-level vision tasks (*e.g.*, object detection and segmentation) to benefit VQA (Gan et al., 2017; Kafle et al., 2017). Most of these works use side information by defining auxiliary learning tasks or losses to the original VQA task. In contrast, we directly turn the information into IQA triplets for training.

Data augmentation for VQA. Several existing works investigate data augmentation. One stream of works creates new triplets by manipulating images or questions (Chen et al., 2020; Agarwal et al., 2020; Tang et al., 2020; Gokhale et al., 2020a). See § 4.1.4 for some more details. The other creates more questions by using a learned language model to paraphrase sentences (Ray et al., 2019; Shah et al., 2019; Whitehead et al., 2020; Kant et al., 2021; Banerjee et al., 2020) or by learning a visual question generation model (Kafle et al., 2017; Li et al., 2018; Krishna et al., 2019).

The closest to ours is the pioneer work of data augmentation by Kafle et al. (2017), which also creates new questions by using mid-level semantic information annotated by humans (*e.g.*, object bounding boxes). Their question generation relies on either pre-defined templates or a learned language model, which may suffer limited diversity or labeling noise. In contrast, we directly reuse questions already in the dataset and show that they are sufficient to augment high-quality questions for other images. Besides, we further explore machine generated annotations (*e.g.*, via an object detector (Ren et al., 2016)), opening the door to augment triplets using extra unlabeled images. Furthermore, we benchmark our method on the popular VQA v2 (Goyal et al., 2017) and challenging VQA-CP (Agrawal et al., 2018) datasets, which are released after the publication of Kafle et al. (2017). *Overall, we view our paper as an attempt to revisit simple data augmentation like Kafle et al. (2017) for VQA, and show that it is indeed quite effective.*

Robustness of VQA models. Goyal et al. (2017); Chao et al. (2018a); Agrawal et al. (2018) pointed out the existence of superficial correlations (*e.g.*, language bias) in the datasets and showed that a VQA model can simply exploit them to answer questions. Existing works to address this can be categorized into three groups. The first group attempts to reduce the language bias by designing new VQA models or learning strategies (Agrawal et al., 2018; Ramakrishnan et al., 2018; Cadene et al., 2019; Clark et al., 2019; Grand and Belinkov, 2019; Clark et al., 2019; Jing et al., 2020; Niu et al., 2021; Shrestha et al., 2020; Gat et al., 2020). For example, RUBi and Ensemble (Cadene et al., 2019; Clark et al., 2019) explicitly modeled the question-answer correlations to encourage VQA models to explore other patterns in the data that are more likely to generalize. The second group leverages side information to facilitate visual grounding (Wu and Mooney, 2019; Selvaraju et al., 2019; Teney and Hengel, 2019). For example, Wu and Mooney (2019) used extra visual or textual annotations to determine important regions where a VQA model should focus on. The third group implicitly or explicitly augments the VQA datasets, *e.g.*, via self-supervised learning, counterfactual sampling, adversarial training, or image and question manipulation (Abbasnejad et al., 2020; Zhu et al., 2020; Teney et al., 2020; Chen et al., 2020; Gokhale et al., 2020a; Liang et al., 2020; Gokhale et al., 2020b; Li et al., 2020; Ribeiro et al., 2019; Selvaraju et al., 2020). SIMPLEAUG belongs to the third group but is simpler in terms of methodology. Besides, SIM-PLEAUG is completely detached from VQA model training and thus model-agnostic. Moreover, SIM-PLEAUG can improve on both VQA-CP (Agrawal et al., 2018) and VQA v2 (Goyal et al., 2017).

3 SIMPLEAUG for Data Augmentation

3.1 Implicit information

SIMPLEAUG leverages three sources of information that implicitly suggest extra IQA triplets beyond those provided in a VQA dataset. The **first** one is the original IQA triplets in the dataset. We find that for two similar images that locally share common objects or globally share common layouts, their corresponding annotated questions can either be treated as paraphrases or extra questions for each other. The **second** one is the object instance labels like object bounding boxes that are annotated on images, *e.g.*, MSCOCO images (Lin et al., 2014). These labels provide accurate answers to "how many" or some of the "what" questions, and many VQA datasets are built upon MSCOCO images. The **third** one is an object detector pre-trained on an external densely-annotated dataset like Visual Genome (VG) (Krishna et al., 2017). This detector can provide information not commonly annotated on images, such as attributes or fine-grained class
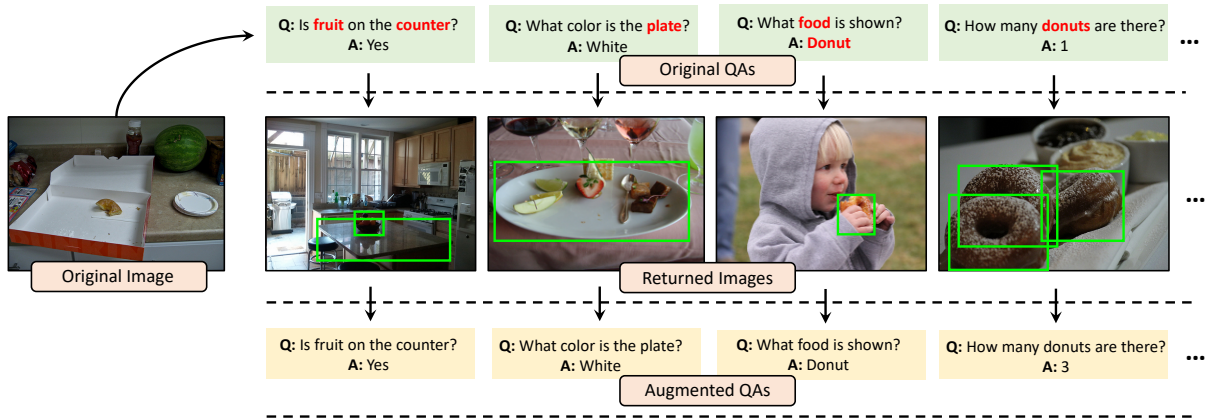
Figure 2: **The SIMPLEAUG pipeline.** We show four original question-answer pairs of the image on the left in VQA v2, and how they are propagated to other images. The green boxes are annotated in MSCOCO or detected by Faster R-CNN; each of them is associated with an object name and/or attribute. We only show boxes matched by nouns or used to derive answers.

names. We note that since the seminal work by Anderson et al. (2018), many following-up VQA models use the Faster R-CNN detector (Ren et al., 2016) pre-trained on VG for feature extraction.

3.2 The SIMPLEAUG pipeline

SIMPLEAUG processes each annotated IQA triplet (i, q, a) in term, and propagates q to other relevant images. To begin with, SIMPLEAUG extract meaningful words from the question, similar to (Chen et al., 2020; Wu and Mooney, 2019). We leverage a spaCy part-of-speech (POS) tagger (Honnibal and Montani, 2017) to extract "nouns" and tokenize their singular and plural forms. We remove words such as "picture" or "photo", which appear in many questions but are not informative for VQA[2].

Given the meaningful "nouns" of a question, we then retrieve relevant images and derive the answers, using MSCOCO annotations or Faster R-CNN detection. Concretely, we split questions into four categories and develop specific *question propagation rules*. Figure 2 illustrates the pipeline.

Yes/No questions. We apply the Faster R-CNN detector trained on VG to each image i' beside image i. The detector returns a set of bounding boxes and their labels. We ignore images whose object labels have no overlap with the nouns of question q, and assign answer "yes" or "no" to the remaining images as follows.

- "Yes": if the labels of image i' cover all nouns

of question q, we create (i', q, yes).

- "No": if the labels of image i' only cover some of the nouns of q, we create (i', q, no). For instance, if the question is "Is there a cat on the pillow?" but the image only contains "pillow" but no "cat", then the answer is "no".

We develop two verification strategies at the end of this subsection to filter out outlier cases.

Color questions. To prevent ambiguous cases, we only consider questions with a single noun (besides the word "color"). We again apply the Faster R-CNN detector, which returns for each image a set of object labels that may also contain attributes like colors. We keep images whose labels cover the noun of question q. For each such image i', we create a triplet $(i', q, \hat{a})$, where $\hat{a}$ is the color attribute provided by the detector. As there are likely some other object-color pairs in i', we investigate replacing the noun in q by each detected object name and create some more IQA triplets about colors.

Number questions. We again focus on questions with a single noun (besides the word "number"). We use MSCOCO annotations, which give each image a set of object bounding boxes and labels. We find all the images whose labels cover the noun of question q. For each such image i', we derive the answer by counting annotated instances of that noun and create a triplet $(i', q, \hat{a})$, where $\hat{a}$ is the count. Some of the nouns (*e.g.*, "animal") are super-categories of sub-category objects (*e.g.*, "dog" or "cat"). Thus, if the noun of q is a super-category (*e.g.*, q is "How many animals are there?"), we follow the category hierarchy provided by MSCOCO and count all its sub-category instances.

Other questions. We focus on "what" questions

---

[2]For example, in a question, "What is the person doing in the picture?", the word "picture" refers to the image itself, not an object within it. We found 8% of the questions like this in VQA v2 (Goyal et al., 2017). Some questions really refer to "pictures" or "photos" within an image (*e.g.*, "How many pictures on the wall?"), but there are <1% such questions.

with a single noun and use MSCOCO annotations. We find all the images whose labels cover the noun of question q. (We also take the super-category cases into account.) For each such image i', we check whether its MSCOCO labels contain the answer a of question q, *i.e.*, according to the original (i, q, a) triplet. For instance, if q is "What animal is this?" and a is "sheep", then we check if image i''s labels contain "sheep". If yes, we create a triplet $(i', q, \hat{a} = a)$. This process essentially discovers "what" questions that can indeed be asked about i'.

Verification. The above rules simplify a question by only looking at its nouns, so they may lead to triplets whose answers are incorrect. To mitigate this issue, we develop two verification strategies.

The first strategy performs self-verification on the original (i, q, a) triplet, checking if our rules can reproduce it. That is, it applies the aforementioned rules to image i to derive the new triplet $(i, q, \hat{a})$. If $\hat{a}$ does not match a, *i.e.*, using the rules creates a different answer, we skip this question q.

The second strategy verifies our rules using IQA triplets annotated on retrieved image i'. For example, if image i' has an annotate triplet (i', q', a') whose q' has the same category and nouns as q, then we compare it to the created triplet $(i', q, \hat{a})$. If $\hat{a}$ does not match a', then we disregard $(i', q, \hat{a})$.

3.3 Paraphrasing by similar questions

Besides the four question propagation rules that look at the image contents, we also investigate a simple paraphrasing rule by searching similar questions in the dataset. Concretely, we apply the averaged word feature from BERT (Devlin et al., 2018) to encode each question as it better captures the object-level semantics for searching questions mentioning the same objects. Two questions are similar if their cosine similarity is above a certain threshold (0.98 in the experiments). If two IQA triplets (i, q, a) and (i', q', a') have similar questions, we create two extra triplets (i, q', a) and (i', q, a') by switching their questions as paraphrasing.

We choose a high threshold 0.98 to avoid false positives. On average, each question finds 11.4 similar questions, and we only pick the top-3 questions. We found that with this design, an extra verification step, like checking if the image i' contains nouns of the paraphrasing question q, does not further improve the overall VQA accuracy. Thus, we do not include an extra verification step for paraphrasing.

4 Experiments

4.1 Experimental setup

4.1.1 VQA datasets and evaluation metrics

We validate SIMPLEAUG on two popular datasets. See Appendix for the results on other datasets.

VQA v2 (Goyal et al., 2017) collects images from MSCOCO (Lin et al., 2014) and uses the same training/validation/testing splits. On average, six questions are annotated for each image. In total, VQA v2 has 444K/214K/448K training/validation/test IQA triplets.

VQA-CP v2 (Agrawal et al., 2018) is a challenging *adversarial* split of VQA v2 designed to evaluate the model's capability of handling language bias/prior shifts between training and testing. For instance, "white" is the most frequent answer for questions that start with "what color..." in the training set whereas "black" is the most common one in the test set. Such *prior changes* also reflect in individual questions, *e.g.*, the most common answer for "What color is the banana?" changes from "yellow" during training to "green" during testing. VQA-CP v2 has 438K/220K training/test IQA triplets.

Evaluation metrics. We follow the standard evaluation protocol (Antol et al., 2015; Goyal et al., 2017). For each test triplet, the predicted answer is compared with answers provided by ten human annotators in a leave-one-annotator-out fashion for robust evaluation. We report the averaged scores over all test triplets as well as over test triplets of Yes/No, number, or other answer types.

4.1.2 Implicit knowledge sources

MSCOCO annotations (Lin et al., 2014). MSCOCO is the most popular benchmark nowadays for object detection and instance segmentation, which contains 80 categories (*e.g.*, "cat") as well as the corresponding super categories (*e.g.*, "animal"). Object instances of all 80 categories are exhaustively annotated in all images, leading to approximately 1.2 million instance annotations.

Faster R-CNN detection (Anderson et al., 2018). We use the object detection results from a Faster R-CNN (Ren et al., 2016) pre-trained with Visual Genome (VG) (Krishna et al., 2017). This pre-trained detector can provide object attributes (*e.g.*, color and material) whereas MSCOCO annotations only contain object names (*e.g.*, "person" and "bicycle"). We use the detector provided by Anderson et al. (2018), which detects 36 objects per image.

4.1.3 Base VQA models

SIMPLEAUG is model-agnostic, and we evaluate it by using its generated data to augment the training set for training three base VQA models.

Bottom-Up Top-Down (UpDn) (Anderson et al., 2018). UpDn is a widely used VQA model. It first detects objects from an image and encodes them into visual feature vectors. Given a question, UpDn uses a question encoder to produce a set of word features. Both visual and language features are then fed into a multi-modal attention network to predict the answer.

Learned-Mixin+H (LMH) (Clark et al., 2019). LMH is a learning strategy to de-bias a VQA model, *e.g.*, UpDn. During training, LMH uses an auxiliary question-only model to encourage the VQA model to explore visual-question related information. During testing, only the VQA model is used. LMH is shown to largely improve the performance on VQA-CP v2 but can hurt that on VQA v2.

LXMERT (Tan and Bansal, 2019). We also study SIMPLEAUG with a stronger, transformer-based VQA model named LXMERT. LXMERT leverages multi-modal transformers to extract multi-modal features, and exploits a masking mechanism to better (pre-)train the model. While such a masking mechanism can be viewed as a way of data augmentation, SIMPLEAUG is fundamentally different from it in two aspects. First, SIMPLEAUG generates new triplets while masking manipulates existing triplets. Second, SIMPLEAUG is detached from model training and is therefore compatible with masking. As will be shown in the experimental results, SIMPLEAUG can provide solid gains to LXMERT on both VQA v2 and VQA-CP.

4.1.4 Compared data augmentation methods

We compare SIMPLEAUG with three existing data augmentation methods for VQA.

Template-based augmentation proposed by Kafle et al. (2017) generates new question-answer pairs using MSCOCO annotations (cf. § 2). We re-implement the method following the paper.

Counterfactual Samples Synthesizing (CSS) (Chen et al., 2020) generates counterfactual triplets by masking critical objects in images or words in questions and assigning different answers. These new training examples force the VQA model to focus on those critical objects and words, improving both visual explainability and question sensitivity.

Table 1: **Statistics on VQA-CP v2 training data.** Miss-answered: the number of SIMPLEAUG examples that a UpDn model trained on the original dataset cannot answer correctly.

# of samples	All	Y/N	Num	Other
Original	438K	183K	52K	202K
SIMPLEAUG	5,457K	2,062K	1,937K	1,458K
Miss-answered	3,081K	974K	1,489K	618K

MUTANT (Gokhale et al., 2020a) is a state-of-the-art data augmentation method by manipulating images and questions. For example, it applies a GAN-based inpainting network to change the object's color to create extra color questions; it manipulates object numbers using MSCOCO annotations; it masks or negates words to mutate questions.

Comparison. SIMPLEAUG is different from CSS and MUTANT in two aspects. First, CSS and MUTANT can only manipulate already annotated questions for an image, while we can create new questions for an image by borrowing them from other images. Second, CSS needs a pre-trained attention-based VQA model to identify critical objects/words while MUTANT's best version requires additional loss terms for training. In contrast, SIMPLEAUG is completely detached from model training.

4.2 Implementation details

Data augmentation by SIMPLEAUG. We speed up the implementation by grouping IQA triplets of the same unique question and only propagating the question once. We remove redundant triplets if the retrieved image already has the same question. To prevent creating too many triplets from paraphrasing (§ 3.3), for each question q we only search for its top-3 similar questions q' and only create (i', q, a') if a' is a rare answer to q — we define a' to be a rare answer if there are fewer than five (q, a') pairs in the dataset. *We emphasize that we only apply* SIMPLEAUG *to IQA triplets in the training set and search images in the training set.*

VQA models. For the base VQA models, we use the released code from corresponding papers. Please see Appendix for more details.

Training with SIMPLEAUG triplets. We explore three ways to train with the original ($\mathcal{O}$) triplets and augmented triplets ($\mathcal{A}$). The first is to train with both from the beginning ($\mathcal{A} + \mathcal{O}$); the second is to train with $\mathcal{O}$ first and then with both ($\mathcal{O} \rightarrow \mathcal{A} + \mathcal{O}$); the third is to train with $\mathcal{O}$ first, then with $\mathcal{A}$, and then with $\mathcal{O}$ again ($\mathcal{O} \rightarrow \mathcal{A} \rightarrow \mathcal{O}$). The rationale of training with multiple stages is to prevent the augmented data from dominating the

Table 2: **Performance on VQA v2 val set and VQA-CP v2 test set.** Our method SIMPLEAUG (cyan background) consistently improves all answer types for different base models on both VQA v2 and VQA-CP. Note that MUTANT (loss) (Gokhale et al., 2020a) (gray color) applies extra loss terms besides data augmentation.

Base	Method	VQA v2 val				VQA-CP test			
		All	Y/N	Num	Other	All	Y/N	Num	Other
UpDn	Baseline (Anderson et al., 2018)	63.48	81.18	42.14	55.66	39.74	42.27	11.93	46.05
	AdvReg (Ramakrishnan et al., 2018)	62.75	79.84	42.35	55.16	41.17	65.49	15.48	35.48
	RUBi (Cadene et al., 2019)	61.16	–	–	–	44.23	67.05	17.48	39.61
	CF-VQA (SUM) (Niu et al., 2021)	63.54	82.51	43.96	54.30	53.55	91.15	13.03	44.97
	SimpleReg (Shrestha et al., 2020)	62.60	–	–	–	48.90	69.80	11.30	47.80
	HINT (Selvaraju et al., 2019)	63.38	81.18	42.99	55.56	46.73	67.27	10.61	45.88
	SCR+VQA-X (Wu and Mooney, 2019)	62.20	78.80	41.60	54.50	49.45	72.36	10.93	48.02
	RandImg (Teney et al., 2020)	57.24	76.53	33.87	48.57	55.37	83.89	41.60	44.20
	Template-based (Kafle et al., 2017)	63.83	81.61	41.98	56.10	39.75	43.03	14.98	44.83
	CSS (Chen et al., 2020)	63.47	80.81	43.33	55.62	41.16	43.96	12.78	47.48
	MUTANT (plain) (Gokhale et al., 2020a)	–	–	–	–	50.16	61.45	35.87	50.14
	MUTANT (loss) (Gokhale et al., 2020a)	62.56	82.07	42.52	53.28	61.72	88.90	49.68	50.78
	SIMPLEAUG (paraphrasing)	63.66	81.44	42.56	55.72	52.57	86.56	13.52	45.47
	SIMPLEAUG (propagation)	64.37	81.91	44.13	56.40	52.27	65.15	45.32	47.42
	SIMPLEAUG (propagation + paraphrasing)	64.34	81.97	43.91	56.35	52.65	66.40	43.43	47.98
LMH	Baseline (Clark et al., 2019)	56.34	65.05	37.63	54.68	52.01	72.58	31.11	46.96
	RMFE (Gat et al., 2020)	–	–	–	–	54.44	74.03	49.16	45.82
	CSS (Chen et al., 2020)	59.91	73.25	39.77	55.11	58.95	84.37	49.42	48.21
	CSS+CL (Liang et al., 2020)	57.29	67.27	38.40	54.71	59.18	86.99	49.89	47.16
	MUTANT (loss) (Gokhale et al., 2020a)	–	–	–	–	55.38	90.99	39.74	40.99
	SIMPLEAUG (paraphrasing)	61.67	78.70	40.21	54.41	53.29	74.12	33.06	47.93
	SIMPLEAUG (propagation)	62.67	79.24	41.44	55.70	53.58	73.58	37.07	47.63
	SIMPLEAUG (propagation + paraphrasing)	62.63	79.31	41.71	55.48	53.70	74.79	34.32	47.97
LXMERT	Baseline (Tan and Bansal, 2019)	73.06	88.30	56.81	65.78	48.66	47.49	22.24	56.52
	Template-based (Kafle et al., 2017)	72.30	85.36	54.47	67.10	49.63	49.96	36.33	53.10
	MUTANT (plain) (Gokhale et al., 2020a)	–	–	–	–	59.69	73.19	32.85	59.29
	MUTANT (loss) (Gokhale et al., 2020a)	70.24	89.01	54.21	59.96	69.52	93.15	67.17	57.78
	SIMPLEAUG (paraphrasing)	74.37	88.78	57.95	67.76	59.09	73.17	28.72	60.04
	SIMPLEAUG (propagation)	74.96	89.00	60.00	68.25	61.82	68.39	53.35	60.69
	SIMPLEAUG (propagation + paraphrasing)	74.98	89.04	59.98	68.25	62.24	69.72	53.63	60.69

training process (see Table 1 for the statistics). We note that, there is a huge number of SIMPLEAUG examples that a VQA model trained with $\mathcal{O}$ only cannot answer. Thus, when training with multiple stages, we remove SIMPLEAUG examples that the model can already answer. We mainly report results using $\mathcal{O} \rightarrow \mathcal{A} \rightarrow \mathcal{O}$, but compare the three ways in § 4.4.

4.3 Main results on VQA v2 and VQA-CP v2

Table 2 summarizes the main results on VQA v2 val and VQA-CP v2 test. We experiment SIMPLEAUG with different base VQA models and compare it to state-of-the-art methods. SIMPLEAUG achieves consistent gains against the base models on all answer types (columns). When paired with LXMERT, SIMPLEAUG obtains the highest accuracy on both datasets, except MUTANT (loss) which applies extra losses besides data augmentation.

SIMPLEAUG improves all answer types. On **VQA-CP v2**, SIMPLEAUG boosts the overall accuracy of UpDn from 39.74% to 52.65%, outper-

forming all but three methods. One key strength of SIMPLEAUG is that it improves all the answer types, including a ~2% gain on "Other" where many methods suffer. Specifically, compared to CF-VQA (Niu et al., 2021) and RandImg (Teney et al., 2020) which have higher overall accuracy than SIMPLEAUG, SIMPLEAUG outperforms them on the challenging "Num" and "Other". On **VQA v2**, SIMPLEAUG achieves the highest accuracy using UpDn, improving +0.86% on "All", +0.79% on "Yes/No", +1.77% on "Num", and +0.69% on "Other". Other methods specifically designed for VQA-CP v2 usually degrade on VQA v2.

SIMPLEAUG is model-agnostic. SIMPLEAUG can directly be applied to other VQA models. Besides UpDn, in Table 2 we show that SIMPLEAUG can lead to consistent gains for two additional VQA models. LMH is a de-biasing method for UpDn, which however hurts the accuracy on VQA v2. With SIMPLEAUG, LMH can largely improve on VQA v2. LXMERT is a strong transformer-based VQA model, and SIMPLEAUG can also improve

Table 3: SIMPLEAUG (propagation) w/ or w/o verification (cf. § 3.2) on VQA-CP v2, using the UpDn model.

Method	Verification	All	Y/N	Num	Other
UpDn	–	39.74	42.27	11.93	46.05
SIMPLEAUG	✗	51.96	64.02	44.44	47.70
	✓	52.27	65.15	45.32	47.42

upon it, achieving the highest accuracy on VQA v2 (all answer types) and on VQA-CP v2 ("Other").

Comparison to data augmentation baselines. SIMPLEAUG notably outperforms the **template-based method** (Kafle et al., 2017), the closest method to ours. We attribute this to the question diversity via question propagation and paraphrasing. Compared to **CSS** (Chen et al., 2020; Liang et al., 2020), SIMPLEAUG performs better on all answer types on both datasets, using UpDn. While CSS outperforms SIMPLEAUG on VQA-CP v2 using the de-biasing LMH, its improvement on VQA v2 is smaller than SIMPLEAUG. Since LXMERT is a general VQA method like UpDn, we expect that SIMPLEAUG will outperform CSS. Finally, compared to **MUTANT** (Gokhale et al., 2020a), SIMPLEAUG achieves better results on VQA-CP v2 against the version without extra loss terms (*i.e.*, MUTANT(plain)). It is worth noting that while CSS and MUTANT both generate extra data, they cannot improve but degrade on VQA v2 (when using UpDn or LXMERT). In contrast, SIMPLEAUG improves on all cases, suggesting it as a more general data augmentation method for VQA.

4.4 Ablation studies of SIMPLEAUG

Question propagation vs. paraphrasing. SIMPLEAUG leverages the original IQA triplets by propagating questions to other images (§ 3.2) or by paraphrasing question using similar questions (§ 3.3). Propagation can ask more questions about an image. For example, the propagated questions in Figure 1 and Figure 3 ask about image contents different from the original questions. In contrast, paraphrasing only paraphrases the original questions of that image. As shown in Table 2, question propagation generally leads to better performance, especially on "Num" and "Other" answers, suggesting the importance of creating additional questions to cover image contents more exhaustively.

On verification for question propagation. Table 3 compares SIMPLEAUG (propagation) with and without the verification strategies (cf. § 3.2). Verification improves accuracy at nearly all cases.

Table 4: **A comparison of training strategies on VQA-CP v2 with the UpDn model.** $\mathcal{O}$: original triplets. $\mathcal{A}$: augmented triplets by SIMPLEAUG.

Method	Strategy	All	Y/N	Num	Other
UpDn	$\mathcal{O}$	39.74	42.27	11.93	46.05
	$\mathcal{O} \rightarrow \mathcal{O} \rightarrow \mathcal{O}$	39.47	43.11	11.75	45.16
SIMPLEAUG	$\mathcal{A} + \mathcal{O}$	47.50	59.76	38.18	43.63
	$\mathcal{O} \rightarrow \mathcal{A} + \mathcal{O}$	49.73	59.67	36.58	48.12
	$\mathcal{O} \rightarrow \mathcal{A} \rightarrow \mathcal{O}$	52.65	66.40	43.43	47.98

Table 5: **Effects of different augmention types (cf. § 3.2).** We report results on VQA-CP v2, using the UpDn model.

Method	Aug Type	All	Y/N	Num	Other
UpDn	–	39.74	42.27	11.93	46.05
SIMPLEAUG	Y/N	47.20	**68.63**	12.12	45.68
	Num	44.62	42.87	**43.80**	45.77
	Color	40.97	43.11	12.39	47.68
	Other	41.22	43.17	12.19	**48.16**
	All	**52.65**	66.40	43.43	47.98

Multiple-stage training. In Table 4, we compare the three training strategies with original triplets ($\mathcal{O}$) and augmented triplets ($\mathcal{A}$). We also train on $\mathcal{O}$ for multiple stages (*i.e.*, more epochs) for a fair comparison. $\mathcal{O} \rightarrow \mathcal{A} \rightarrow \mathcal{O}$ in general outperforms others, and we attribute this to the clear separation of clean and noisy data — the last training stage may correct noisy information learned in early stages (Zhang et al., 2021).

Training with SIMPLEAUG triplets alone. We further investigate training the UpDn model with augmented triplets alone ($\mathcal{A}$). On VQA-v2, we get 39.62% overall accuracy, worse than the baseline trained with original data (63.48%). This is likely due to the noise in the augmented data. On VQA-CP, we get 51.60%, much better than the baseline (39.74%) but worse than training with both augmented and original triplets (52.65%). We surmise that SIMPLEAUG triplets help mitigate the language bias shifts in VQA-CP. Please see Appendix for a human study on assessing the quality of the triplets augmented by SIMPLEAUG.

Effects of augmentation types. We experiment with propagating each question type alone on VQA-CP v2, using UpDn as the base model. In Table 5, we show the separate results of SIMPLEAUG with different question types. The augmented questions notably improve the corresponding answer type.

4.5 SIMPLEAUG in additional scenarios

We explore SIMPLEAUG in the scenarios where there are (i) limited questions per image, and (ii)

Table 6: **Learning with limited IQA triplets on VQA-CP v2.** We keep a certain fraction of QA pairs per image.

Method	Fraction	All	Y/N	Num	Other
UpDn	1.00	39.74	42.27	11.93	46.05
SIMPLEAUG	1.00	52.65	66.40	43.43	47.98
	0.50	47.67	57.65	37.77	45.17
	0.25	46.03	52.01	37.96	45.12
	0.10	42.91	45.09	30.24	45.25

extra weakly-labeled or unlabeled images. For (ii), both have no IQA triplets but the weakly-labeled ones have human-annotated object instances.

Learning with limited triplets. We randomly keep a fraction of annotated QA pairs for each training image on VQA-CP v2. Table 6 shows that even under this annotation-scarce setting (*e.g.*, only 10% of QA pairs are kept), SIMPLEAUG can already be effective, outperforming the baseline UpDn model trained with all data. This demonstrates the robustness of SIMPLEAUG on dealing with the challenging setting with limited triplets.

Learning with weakly-labeled or unlabeled images. We simulate the scenarios by keeping the QA pairs for a fraction of images (*i.e.*, labeled data) and removing the QA pairs entirely for the other images. Conventionally, a VQA model cannot benefit from the images without QA pairs, but SIMPLEAUG could leverage them by propagating questions to them. Specifically, for images without QA pairs, we consider two cases. We either keep their MSCOCO object instance annotations (*i.e.*, weakly-labeled data) or completely rely on object detectors (*i.e.*, unlabeled data). Table 7 shows the results, in which we only apply SIMPLEAUG to the weakly-labeled and unlabeled images. As shown, SIMPLEAUG yields consistent improvements, opening up the possibility of leveraging additional images to improve VQA.

4.6 Qualitative results

We show a training image and its augmented QA pairs by SIMPLEAUG in Figure 3. A VQA model trained on the original IQA triplets cannot answer many of the newly generated questions, even if the image is in the training set, showing the necessity to include them for training a stronger model. More qualitative results can be found in Appendix.

5 Conclusion

We proposed SIMPLEAUG, a data augmentation method for VQA that can turn information already

Table 7: **Learning with weakly-labeled and unlabeled images for VQA v2.** Fraction: the portion of images with annotated QA pairs. GT: MSCOCO ground truth annotations. OD: Faster R-CNN object detection. ✗: supervised training with only labeled VQA training examples.

Fraction	SIMPLEAUG	All	Y/N	Num	Other
1.00	✗	63.48	81.18	42.14	55.66
	✓	64.34	81.97	43.91	56.35
0.50	✗	60.93	78.45	40.74	52.96
	GT	61.47	78.92	41.43	53.50
	OD	61.47	78.93	41.42	53.50
0.25	✗	56.70	74.02	37.81	48.53
	GT	57.54	74.49	39.08	49.54
	OD	57.56	74.63	38.67	49.57
0.10	✗	51.06	69.18	33.46	41.93
	GT	52.18	69.76	35.95	43.10
	OD	52.27	69.98	35.07	43.34

Figure 3: **Qualitative results.** We show the training image and its QA pairs from VQA-CP, and the generated QA pairs by SIMPLEAUG. ✓/✗ indicates if the baseline VQA model (trained without SIMPLEAUG) answers correctly/incorrectly. In augmented QA pairs, the first three are from question propagation and the last one is by paraphrasing.

in the datasets into explicit IQA triplets for training. SIMPLEAUG is simple but by no means trivial. First, it justifies that mid-level vision tasks like object detection can effectively benefit VQA. Second, we probably will never be comprehensive enough in annotating data, and SIMPLEAUG can effectively turn what we have at hand (*i.e.*, "knowns") to examples a VQA model wouldn't have known (*i.e.*, "unknowns"). SIMPLEAUG can notably improve the accuracy of VQA models on both VQA v2 (Goyal et al., 2017) and VQA-CP v2 (Clark et al., 2019).

References

Ehsan Abbasnejad, Damien Teney, Amin Parvaneh, Javen Shi, and Anton van den Hengel. 2020. Counterfactual vision and language learning. In *Proceedings of CVPR*.

Vedika Agarwal, Rakshith Shetty, and Mario Fritz. 2020. Towards causal vqa: Revealing and reducing spurious correlations by invariant and covariant semantic editing. In *Proceedings of CVPR*.

Aishwarya Agrawal, Dhruv Batra, and Devi Parikh. 2016. Analyzing the behavior of visual question answering models. In *Proceedings of EMNLP*.

Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. 2018. Don't just assume; look and answer: Overcoming priors for visual question answering. In *Proceedings of CVPR*.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of CVPR*.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual question answering. In *Proceedings of ICCV*.

Pratyay Banerjee, Tejas Gokhale, Yezhou Yang, and Chitta Baral. 2020. Self-supervised vqa: Answering visual questions using images and captions. *arXiv preprint arXiv:2012.02356*.

Remi Cadene, Corentin Dancette, Hedi Ben-younes, Matthieu Cord, and Devi Parikh. 2019. RUBi: Reducing unimodal biases in visual question answering. In *Proceedings of NeurIPS*.

Wei-Lun Chao, Hexiang Hu, and Fei Sha. 2018a. Being negative but constructively: Lessons learnt from creating better visual question answering datasets. In *Proceedings of NAACL*.

Wei-Lun Chao, Hexiang Hu, and Fei Sha. 2018b. Cross-dataset adaptation for visual question answering. In *Proceedings of CVPR*.

Long Chen, Xin Yan, Jun Xiao, Hanwang Zhang, Shiliang Pu, and Yueting Zhuang. 2020. Counterfactual samples synthesizing for robust visual question answering. In *Proceedings of CVPR*.

Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. 2019. Don't take the easy way out: Ensemble based methods for avoiding known dataset biases. In *Proceedings of EMNLP*.

Abhishek Das, Harsh Agrawal, Larry Zitnick, Devi Parikh, and Dhruv Batra. 2017. Human attention in visual question answering: Do humans and deep networks look at the same regions? *Computer Vision and Image Understanding*, 163:90–100.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL*.

Chuang Gan, Yandong Li, Haoxiang Li, Chen Sun, and Boqing Gong. 2017. VQS: Linking segmentations to questions and answers for supervised attention in vqa and question-focused semantic segmentation. In *Proceedings of ICCV*.

Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2013. Ppdb: The paraphrase database. In *Proceedings of NAACL*.

Itai Gat, Idan Schwartz, Alexander Schwing, and Tamir Hazan. 2020. Removing bias in multi-modal classifiers: Regularization by maximizing functional entropies. In *Proceedings of NeurIPS*.

Tejas Gokhale, Pratyay Banerjee, Chitta Baral, and Yezhou Yang. 2020a. MUTANT: A training paradigm for out-of-distribution generalization in visual question answering. In *Proceedings of EMNLP*.

Tejas Gokhale, Pratyay Banerjee, Chitta Baral, and Yezhou Yang. 2020b. Vqa-lol: Visual question answering under the lens of logic. In *Proceedings of ECCV*.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of CVPR*.

Gabriel Grand and Yonatan Belinkov. 2019. Adversarial regularization for visual question answering: Strengths, shortcomings, and side effects. In *Proceedings of ACL Workshop*.

Danna Gurari, Qing Li, Chi Lin, Yinan Zhao, Anhong Guo, Abigale Stangl, and Jeffrey P Bigham. 2019. Vizwiz-priv: A dataset for recognizing the presence and purpose of private visual information in images taken by blind people. In *Proceedings of CVPR*.

Matthew Honnibal and Ines Montani. 2017. spacy 2: Natural language understanding with bloom embeddings. *convolutional neural networks and incremental parsing*, 7(1).

Hexiang Hu, Wei-Lun Chao, and Fei Sha. 2018. Learning answer embeddings for visual question answering. In *Proceedings of CVPR*.

Drew A Hudson and Christopher D Manning. 2019. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the CVPR*.

Chenchen Jing, Yuwei Wu, Xiaoxun Zhang, Yunde Jia, and Qi Wu. 2020. Overcoming language priors in vqa via decomposed linguistic representations. In *Proceedings of AAAI*.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Judy Hoffman, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. 2017. Inferring and executing programs for visual reasoning. In *Proceedings of CVPR*.

Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David Shamma, Michael Bernstein, and Li Fei-Fei. 2015. Image retrieval using scene graphs. In *Proceedings of CVPR*.

Kushal Kafle, Mohammed Yousefhussien, and Christopher Kanan. 2017. Data augmentation for visual question answering. In *Proceedings of International Conference on Natural Language Generation*.

Yash Kant, Abhinav Moudgil, Dhruv Batra, Devi Parikh, and Harsh Agrawal. 2021. Contrast and classify: Alternate training for robust vqa. In *Proceedings of ICCV*.

Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of CVPR*.

Hyounghun Kim and Mohit Bansal. 2019. Improving visual question answering by referring to generated paragraph captions. *Proceedings of ACL*.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of ICLR*.

Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. 2019. Information maximizing visual question generation. In *Proceedings of CVPR*.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73.

Linjie Li, Zhe Gan, and Jingjing Liu. 2020. A closer look at the robustness of vision-and-language pre-trained models. *arXiv preprint arXiv:2012.08673*.

Yikang Li, Nan Duan, Bolei Zhou, Xiao Chu, Wanli Ouyang, Xiaogang Wang, and Ming Zhou. 2018. Visual question generation as dual task of visual question answering. In *Proceedings of CVPR*.

Zujie Liang, Weitao Jiang, Haifeng Hu, and Jiaying Zhu. 2020. Learning to contrast the counterfactual samples for robust visual question answering. In *Proceedings of EMNLP*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Proceedings of ECCV*.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Proceedings of NIPS*.

George A Miller. 1995. WordNet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. 2021. Counterfactual vqa: A cause-effect look at language bias. In *Proceedings of CVPR*.

Badri Patro and Vinay P Namboodiri. 2018. Differential attention for visual question answering. In *Proceedings of CVPR*.

Sainandan Ramakrishnan, Aishwarya Agrawal, and Stefan Lee. 2018. Overcoming language priors in visual question answering with adversarial regularization. In *Proceedings of NeurIPS*.

Arijit Ray, Karan Sikka, Ajay Divakaran, Stefan Lee, and Giedrius Burachas. 2019. Sunny and dark outside?! improving answer consistency in vqa through entailed question generation. In *Proceedings of EMNLP*.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2016. Faster r-cnn: towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137–1149.

Marco Tulio Ribeiro, Carlos Guestrin, and Sameer Singh. 2019. Are red roses red? evaluating consistency of question-answering models. In *Proceedings of ACL*.

Ramprasaath R Selvaraju, Stefan Lee, Yilin Shen, Hongxia Jin, Shalini Ghosh, Larry Heck, Dhruv Batra, and Devi Parikh. 2019. Taking a HINT: Leveraging explanations to make vision and language models more grounded. In *Proceedings of ICCV*.

Ramprasaath R Selvaraju, Purva Tendulkar, Devi Parikh, Eric Horvitz, Marco Tulio Ribeiro, Besmira Nushi, and Ece Kamar. 2020. SQuINTing at VQA Models: Introspecting vqa models with sub-questions. In *Proceedings of CVPR*.

Meet Shah, Xinlei Chen, Marcus Rohrbach, and Devi Parikh. 2019. Cycle-consistency for robust visual question answering. In *Proceedings of CVPR*.

Jiaxin Shi, Hanwang Zhang, and Juanzi Li. 2019. Explainable and explicit visual reasoning over scene graphs. In *Proceedings of CVPR*.

Robik Shrestha, Kushal Kafle, and Christopher Kanan. 2020. A negative case analysis of visual grounding methods for vqa. In *Proceedings of ACL*.

Hao Tan and Mohit Bansal. 2019. LXMERT: Learning cross-modality encoder representations from transformers. In *Proceedings of EMNLP*.

Ruixue Tang, Chao Ma, Wei Emma Zhang, Qi Wu, and Xiaokang Yang. 2020. Semantic equivalent adversarial data augmentation for visual question answering. In *Proceedings of ECCV*.

Damien Teney and Anton van den Hengel. 2019. Actively seeking and learning from live data. In *Proceedings of CVPR*.

Damien Teney, Kushal Kafle, Robik Shrestha, Ehsan Abbasnejad, Christopher Kanan, and Anton van den Hengel. 2020. On the value of out-of-distribution testing: An example of goodhart's law. In *Proceedings of NeurIPS*.

Zixu Wang, Yishu Miao, and Lucia Specia. 2021. Latent variable models for visual question answering. *arXiv preprint arXiv:2101.06399*.

Spencer Whitehead, Hui Wu, Yi Ren Fung, Heng Ji, Rogerio Feris, and Kate Saenko. 2020. Learning from lexical perturbations for consistent visual question answering. *arXiv preprint arXiv:2011.13406*.

Jialin Wu, Zeyuan Hu, and Raymond J Mooney. 2019. Generating question relevant captions to aid visual question answering. In *Proceedings of ACL*.

Jialin Wu and Raymond J Mooney. 2019. Self-critical reasoning for robust visual question answering. In *Proceedings of NeurIPS*.

Zhou Yu, Jun Yu, Chenchao Xiang, Jianping Fan, and Dacheng Tao. 2018. Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering. *IEEE Transactions on Neural Networks and Learning Systems*, (99):1–13.

Cheng Zhang, Wei-Lun Chao, and Dong Xuan. 2019. An empirical study on leveraging scene graphs for visual question answering. In *Proceedings of BMVC*.

Cheng Zhang, Tai-Yu Pan, Yandong Li, Hexiang Hu, Dong Xuan, Soravit Changpinyo, Boqing Gong, and Wei-Lun Chao. 2021. MosaicOS: A simple and effective use of object-centric images for long-tailed object detection. In *Proceedings of ICCV*.

Xi Zhu, Zhendong Mao, Chunxiao Liu, Peng Zhang, Bin Wang, and Yongdong Zhang. 2020. Overcoming language priors with self-supervised learning for visual question answering. In *Proceedings of IJCAI*.

Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. 2016. Visual7W: Grounded question answering in images. In *Proceedings of CVPR*.

Discovering the Unknown Knowns:
Turning Implicit Knowledge in the Dataset into Explicit Training Examples for Visual Question Answering
(Appendix)

In this appendix, we provide details and results omitted in the main text.

- Appendix A: additional implementation details. (§ 4.2 of the main paper)

- Appendix B: results on GQA dataset. (§ 4.1 of the main paper)

- Appendix C: human study for assessing the quality of triplets generated by SIMPLEAUG. (§ 4.4 of the main paper)

- Appendix D: additional qualitative results generated by SIMPLEAUG. (§ 4.6 of the main paper)

A Additional Implementation Details

A.1 Baseline VQA models.

We validate SIMPLEAUG with three VQA models in our experiments: Bottom-Up Top-Down (UpDn)[3] (Anderson et al., 2018), Learned-Mixin+H (LMH)[4] (Clark et al., 2019), and LXMERT[5] (Tan and Bansal, 2019). All baseline models are implemented using officially released codebase. More details of code and data are publicly available at `https://github.com/heendung/simpleAUG`.

A.2 Optimization

UpDn and LMH. We maintain the default settings in UpDn and LMH except for using the mini-batch size of 512 on VQA v2 and $1,024$ on VQA-CP v2. Following the official implementations, our visual features are the output of Faster R-CNN (Ren et al., 2016) object detector trained on Visual Genome (Krishna et al., 2017), provided by Anderson et al. (2018). We optimize UpDn and LMH using stochastic gradient descent (SGD) with Adamax (Kingma and Ba, 2015) and learning rate

---

2×10^{-4}. Training a baseline UpDn or LMH model on a single NVIDIA RTX A6000 takes around 2 hours for convergence.

LXMERT. Similar to UpDn and LMH models, LXMERT leverages the object features from the Faster R-CNN detection provided by Anderson et al. (2018). We train a LXMERT model using the mini-batch size of 256. Following Tan and Bansal (2019), we use Adam (Kingma and Ba, 2015) as the optimizer with a linear decayed learning rate schedule. Training a baseline LXMERT model on a single NVIDIA RTX A6000 takes around 8 hours for convergence.

Multi-stage training. As discussed in § 4.2 and § 4.4 of the main paper, we train the VQA models with a three-stage paradigm ($\mathcal{O} \to \mathcal{A} \to \mathcal{O}$): first with original triplets $\mathcal{O}$, then with the SIMPLEAUG triplets $\mathcal{A}$, and then with $\mathcal{O}$ again. *In each of these three stages, we follow the same optimization procedures as we train the baseline VQA models in the first stage.* We report the best results on VQA v2 validation set (Goyal et al., 2017) and VQA-CP v2 test set (Agrawal et al., 2018).

A.3 Additional details of SIMPLEAUG

As mentioned in the main paper, for each annotated IQA triplet (i, q, a) in the dataset, SIMPLEAUG propagates q to other relevant images. To begin with, we find unique questions by filtering out any duplicate sentences. We then extract meaningful words from the unique questions in line with (Chen et al., 2020; Wu and Mooney, 2019). Concretely, we remove the question type from q and then apply a spaCy part-of-speech (POS) tagger (Honnibal and Montani, 2017) to extract "nouns". To handle the synonyms, we further consider the singular/plural forms and super-categories of nouns[6]. Moreover, we remove non-informative words (*e.g.*, "picture" or "photo") in the sentence. For example, in a question, "What is the man doing in the picture?", the word "picture" refers to an image itself but not

---

[3]UpDn model implementation: `https://github.com/yanxinzju/CSS-VQA`.

[4]LMH model implementation: `https://github.com/chrisc36/bottom-up-attention-vqa`.

[5]LXMERT model implementation: `https://github.com/airsplay/lxmert`.

---

[6]Paraphrase database (Ganitkevitch et al., 2013) or Word-Net (Miller, 1995) could be used to handle other synonyms.

any specific object. There are around 8% of triplets like this in VQA v2 (Goyal et al., 2017). While it is possible that both sentence and image may contain such non-informative contents (*e.g.*, "How many pictures on the wall?"), there are <1% such questions.

B Results on GQA Dataset

We further conduct a preliminary study of SIMPLEAUG on the popular GQA dataset (Hudson and Manning, 2019), which focuses on compositional VQA tasks and consists of 22M questions about various day-to-day images. Each image in GQA is associated with a scene graph (Johnson et al., 2015) which consists of the objects, attributes, and relationships.

We focus on binary questions (35% of all questions) and propagate a question q to an image i according to the image's scene graph. Particularly, we leverage the semantic type of the question (*e.g.*, "attribute", "relation") and the scene graph to generate the answer. For example, suppose q asks if an object contains a certain "attribute", we check the scene graph's node of that object to determine the answer. SIMPLEAUG can improve the accuracy of UpDn from 56.06% to 56.52%, justifying its generalizability and applicability.

C Human Evaluation on SIMPLEAUG Triplets

SIMPLEAUG requires no sentence/image generation steps, and thus all examples are natural annotations from humans, largely alleviating the artificial noise that the previous methods may have. To further evaluate the quality of the augmented triplets, we randomly select 500 images and pick 5 augmented QA pairs per image from each type (4 by propagation Y/N, Num, Other, Color and 1 by paraphrasing). For those 2,500 triplets, we ask 5 different crowd workers to evaluate "relatedness (1/0)" of the augmented questions and "correctness (1/0)" of the answer given the question and image. That is, if the question makes sense for the corresponding image, rate 1, otherwise 0; if the answer is correct, rate 1, otherwise 0 (see Figure 4). Table 8 shows the human study results. The average relatedness / correctness are 86.75% / 67.60% for propagation and 80.80% / 64.40% for paraphrasing.

We note that these generated data are based on human-annotated questions in the dataset. Therefore, there are no artifacts in the questions. More-

Table 8: **Human evaluation.** The Relatedness and Correctness are shown on different types / question types.

SIMPLEAUG	Type	Relatedness (%)	Correctness (%)
Propagation	Y/N	82.60	52.20
	Color	87.20	77.20
	Num	89.20	60.80
	Other	88.00	80.20
	Overall	86.75	67.60
Paraphrasing	Overall	80.80	64.40

over, these generated data are to augment the original data. Thus, even if they contain noise, they can consistently improve the model's performance.

D Additional Qualitative Results by SIMPLEAUG

Figure 5 provides more qualitative results. We note that the baseline model has still suffered in learning many IQA relationships. Even for simple cases (*e.g.*, What color is the elephant?), the model is not able to answer them correctly. Thus, augmenting IQA triplets with the implicit information in the dataset can notably improve the model's performance on both VQA v2 and VQA-CP v2.
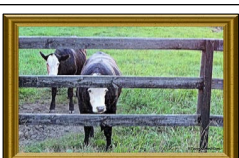
Image	Type	Question	Answer	Relatedness (1 / 0)	Correctness (1 / 0)
	Y/N	• Is the ramp completely covered in snow?	No		
	Color	• What color is his shirt?	Blue		
	Num	• How many people are standing on there surfboard?	1		
	Other	• What sport is depicted?	Skateboarding		
	Paraphrasing	• Is this man wearing any safety gear?	yes		
	Y/N	• Are there waves shown in this picture?	Yes		
	Color	• What color is the sky?	Blue		
	Num	• How many people are gathered around?	1		
	Other	• What sport is represented in this scene?	Surfing		
	Paraphrasing	• What is the man walking on?	Surfboard		
	Y/N	• Is this photo pulling into a station?	Yes		
	Color	• What color is his shirt?	Black		
	Num	• How many trains can you see?	2		
	Other	• What vehicle is shown?	Train		
	Paraphrasing	• Is the boy walking?	Yes		
	Y/N	• Is the water clean and safe?	Yes		
	Color	• What is the color of the grass?	Green		
	Num	• How many people might live here?	3		
	Other	• What are the yellow vehicle?	Car		
	Paraphrasing	• Is that the ocean?	No		
	Y/N	• Are these sheep marked?	Yes		
	Color	• What is the color of the grass?	Green		
	Num	• How many cows are stacked?	2		
	Other	• Which animal is it?	Cow		
	Paraphrasing	• What color is the photo frame?	Yellow		
	Y/N	• Is the elephant crying?	Yes		
	Color	• What is the color of the grass?	Green		
	Num	• How many animal is there in the picture?	2		
	Other	• What type of animal is behind them?	Elephant		
	Paraphrasing	• Which elephant is bigger?	Right		
	Y/N	• Is it on display?	Yes		
	Color	• What color is the button?	Black		
	Num	• How many cats are in the picture?	1		
	Other	• What kind of animal is pictured?	Cat		
	Paraphrasing	• What color are the cat's eyes?	Blue		

Figure 4: **Examples in human study.** For each image, we pick 5 triplets created by SIMPLEAUG (4 by propagation Y/N, Num, Other, Color and 1 by paraphrasing) and ask crowd workers to evaluate the IQA triplets by the question's *relatedness (1 / 0)* to the image and the answer's *correctness (1 / 0)* to the image and question.
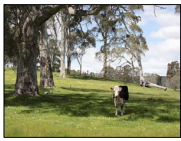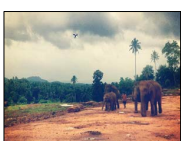
Image	Augmented QA Pairs		Original QA Pairs	
	Question	**Answer**	**Question**	**Answer**
	• *How many giraffe?* • *What color is his eyes?* • *What color is the giraffe?* • *What color is the sign?*	*3* ✗ *Black* ✓ *Brown* ✗ *Red* ✗	• *Are the giraffes resting their heads?* • *Are the humans on the ground?* • *What is the fence made of?*	*Yes* *Yes* *Wood*
	• *How many animals are in this?* • *How many cows are visible?* • *Are the cows hornless?* • *What color is the cow?* • *What color are the rocks?*	*1* ✗ *1* ✗ *Yes* ✗ *White* ✗ *Gray* ✗	• *Which animals are seen?* • *Is this a farm?* • *Is the cow under a tree?*	*Cow* *No* *Yes*
	• *How many of these animals are laying down?* • *How many cats are pictured?* • *What color is the suitcase?* • *How many suitcases are they?*	*1* ✗ *1* ✗ *Brown* ✗ *1* ✗	• *Is this a good place for the cat to sleep?* • *Is this an old suitcase?* • *What kind of cat?* • *What are the cats laying on?*	*Yes* *Yes* *Black* *Suitcase*
	• *How many people are clearly visible in this picture?* • *How many people are standing around?* • *How many people are actually in the photo?* • *What color is the sky?*	*2* ✗ *2* ✗ *2* ✗ *Blue* ✗	• *How many tires are there?* • *What sport is the equipment for?* • *About what time of day is it?*	*6* *Biking* *Daytime*
	• *What color is the elephant?* • *What color are his legs?*	*Gray* ✗ *Gray* ✗	• *Are the elephants mad?* • *How many animals are here?* • *What animals are shown?*	*Yes* *4* *Elephant*
	• *What color is the stove?* • *What color is the hair?* • *What color is the pants?* • *What color is the table?* • *How many people are actually in this photo?*	*White* ✓ *Black* ✗ *Black* ✗ *Gray* ✗ *2* ✗	• *How sanitary does the counter look?* • *How many clear glass bowls are on the counter?* • *What is the counter made of?*	*Clean* *5* *Steel*
	• *What color are the leaves on the tree?* • *What color is the field?* • *What is the color of the cloud?* • *How many elephants in the picture?* • *How many animals are seen?*	*Green* ✓ *Green* ✗ *White* ✗ *4* ✗ *5* ✗	• *Sunny or overcast?* • *Is the water fresh looking?* • *Is there a man in the picture?* • *Are any of the elephants on the dirt road?* • *What is the water on the ground?*	*Overcast* *No* *Yes* *Yes* *Mud*

Figure 5: **Additional qualitative results on VQA-CP.** We show the original image, the generated QA pairs by SIMPLEAUG, and the original QA pairs. ✓/✗ indicates if the baseline VQA model (trained without SIMPLEAUG) predicts correctly/incorrectly.