Logical Implications for Visual Question Answering Consistency

Sergio Tascon-Morales Pablo Márquez-Neila Raphael Sznitman
University of Bern
{sergio.tasconmorales, pablo.marquez, raphael.sznitman}@unibe.ch

Abstract

Despite considerable recent progress in Visual Question Answering (VQA) models, inconsistent or contradictory answers continue to cast doubt on their true reasoning capabilities. However, most proposed methods use indirect strategies or strong assumptions on pairs of questions and answers to enforce model consistency. Instead, we propose a novel strategy intended to improve model performance by directly reducing logical inconsistencies. To do this, we introduce a new consistency loss term that can be used by a wide range of the VQA models and which relies on knowing the logical relation between pairs of questions and answers. While such information is typically not available in VQA datasets, we propose to infer these logical relations using a dedicated language model and use these in our proposed consistency loss function. We conduct extensive experiments on the VQA Introspect and DME datasets and show that our method brings improvements to state-of-the-art VQA models while being robust across different architectures and settings.

1. Introduction

Visual Questioning Answering (VQA) models have drawn recent interest in the computer vision community as they allow text queries to question image content. This has given way to a number of novel applications in the space of model reasoning [8, 29, 54, 56], medical diagnosis [21,37,51,60] and counterfactual learning [1,2,11]. With the ability to combine language and image information in a common model, it is unsurprising to see a growing use of VQA methods.

Despite this recent progress, however, a number of important challenges remain when making VQAs more proficient. For one, it remains extremely challenging to build VQA datasets that are void of bias. Yet this is critical to ensure subsequent models are not learning spurious correlations or shortcuts [49]. This is particularly daunting in applications where domain knowledge plays an important role (*e.g.*, medicine [15, 27, 33]). Alternatively, ensuring that responses of a VQA are coherent, or *consistent*, is paramount as well. That is, VQA models that answer differently about similar content in a given image imply inconsistencies in how the model interprets the inputs. A number of recent methods have attempted to address this using logic-based approaches [19], rephrashing [44], question generation [18, 40, 41] and regularizing using consistency constraints [47]. In this work, we follow this line of research and look to yield more reliable VQA models.

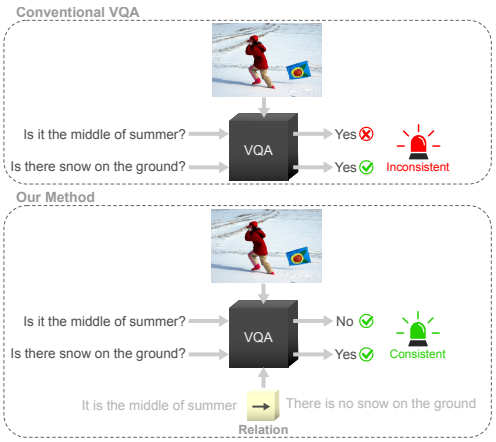

Figure 1. Top: Conventional VQA models tend to produce inconsistent answers as a consequence of not considering the relations between question and answer pairs. Bottom: Our method learns the logical relation between question and answer pairs to improve consistency.

We wish to ensure that VQA models are consistent in answering questions about images. This implies that if multiple questions are asked about the same image, the model's answers should not contradict themselves. For instance, if one question about the image in Fig. 1 asks "Is there snow on the ground?", then the answer inferred should be consistent with that of the question "Is it the middle of summer?" As noted in [43], such question pairs involve reasoning and perception, and consequentially lead the authors to define inconsistency when the reasoning and perception questions are answered correctly and incorrectly, respectively. Along this line, [47] uses a similar definition of inconsistency to

regularize a VQA model meant to answer medical diagnosis questions that are hierarchical in nature. What is critical in both cases, however, is that the consistency of the VQA model depends explicitly on its answers, as well as the question and true answer. This hinges on the assumption that perception questions are sufficient to answer reasoning questions. Yet, for any question pair, this may not be the case. As such, the current definition of consistency (or inconsistency) has been highly limited and does not truly reflect how VQAs should behave.

To address the need to have self-consistent VQA models, we propose a novel training strategy that relies on logical relations. To do so, we re-frame question-answer (QA) pairs as propositions and consider the relational construct between pairs of propositions. This construct allows us to properly categorize pairs of propositions in terms of their logical relations. From this, we introduce a novel loss function that explicitly leverages the logical relations between pairs of questions and answers in order to enforce that VQA models be self-consistent. However, datasets typically do not contain relational information about QA pairs, and collecting this would be extremely laborious and difficult. To overcome this, we propose to train a dedicated language model that infers logical relations between propositions. Our experiments show that we can effectively infer logical relations from propositions and use them in our loss function to train VQA models that improve state-of-the-art methods via consistency. We demonstrate this over two different VQA datasets, against different consistency methods, and with different VQA model architectures. Our code and data are available at `https://github.com/sergiotasconmorales/imp_vqa`.

2. Related work

Since its initial presentation in Antol *et al*. [4], VQA has thoroughly advanced. Initial developments focused on multimodal fusion modules, which combine visual and text embeddings [8, 36]. From basic concatenation and summation [4] to more complex fusion mechanisms that benefit from projecting the embeddings to different spaces, numerous approaches have been proposed [6, 16, 32]. The addition of attention mechanisms [8, 31, 36] and subsequently transformer architectures [50] has also contributed to the creation of transformer-based vision-language models, such as LXMERT, which have shown state-of-the-art performances [46].

More recently, methods have proposed to improve other aspects of VQA, including avoiding shortcut learning and biases [12, 25], improving 3D spatial reasoning [5], Out-Of-Distribution (OOD) generalization [9, 49], improving transformer-based vision-language models [57,61], external knowledge integration [14, 17] and model evaluation with visual and/or textual perturbations [22,52]. With the awareness of bias in VQA training data, some works have also addressed building better datasets (*e.g*., v2.0 [20], VQA-CP [3], CLEVR [30] and GCP [28]).

Furthermore, these developments have now given rise to VQA methods in specific domains. For instance, the VizWiz challenge [10,23,24] aims at creating VQA models that can help visually impaired persons with routine daily tasks, while there is a growing number of medical VQA works with direct medicine applications [21,37,51,60].

Consistency in VQA Consistency in VQA can be defined as the ability of a model to produce answers that are not contradictory. This is, given a pair of questions about an image, the answers predicted by a VQA model should not be contrary (*e.g*. answering "Yes" to "Is it the middle of summer?" and "Winter" to "What season is it?"). Due to its significance in reasoning, consistency in VQA has become a focus of study in recent years [19, 29, 41, 43, 44]. Some of the first approaches for consistency enhancement focused on creating re-phrasings of questions, either by dataset design or at training time [44]. Along this line, entailed questions were proposed [19, 41], such that a question generation module was integrated into a VQA model [18,40], used as a benchmarking method to evaluate consistency [59] or as a rule-based data-augmentation technique [41]. Other approaches tried to shape the embedding space by imposing constraints in the learned representations [48] and by imposing similarities between the attention maps of pairs of questions [43]. Another work [47] assumed entailment relations between pairs of questions to regularize training. A more recent approach attempts to improve consistency by using graph neural networks to simulate a dialog in the learning process [29].

While these approaches show benefits in some cases, they typically only consider that a subset of logical relationships exists between pairs of question-answers or assume that a single relation holds for all QA pairs. Though true in the case of re-phrasings, other question generation approaches cannot guarantee that the produced questions preserve unique relations or that grammatical structure remains valid. Consequently, these methods often rely on metrics that either over or under-estimate consistency by relying on these assumptions. In the present work, we propose a strategy to alleviate these limitations by considering all logical relations between pairs of questions and answers.

Entailment prediction Natural Language Inference (NLI), or Recognizing Textual Entailment (RTE), is the task of predicting how two input sentences (namely *premise* and *hypothesis*) are related, according to three pre-established categories: entailment, contradiction and neutrality [35]. For example, if the premise is "A soccer game with multiple males playing" and the hypothesis is

"Some men are playing a sport," then the predicted relation should be an entailment, because the hypothesis logically follows from the premise. Several benchmarking datasets (*e.g.*, SNLI [58], MultiNLI [55], SuperGLUE [53], WIKIFACTCHECK [42] and ANLI [38]) have contributed to the adaption of general-purpose transformer-based models like BERT [13], RoBERTa [34] and DeBERTa [26] for this task. In this work, we will leverage these recent developments to build a model capable of inferring relations between propositions.

3. Method

Given an image $\mathbf{x} \in \mathcal{I}$, a question $\mathbf{q} \in \mathcal{Q}$ about the image and a set $\mathcal{A} = \{a_1, \ldots, a_K\}$ of possible answers to choose from, a VQA model is expected to infer the answer $\hat{a} \in \mathcal{A}$ that matches the true answer a^*. This can be formulated as,

$$\hat{a} = \underset{a \in \mathcal{A}}{\arg\max}\, p(a|\mathbf{x}, \mathbf{q}; \theta), \tag{1}$$

where θ represents the parameters of the VQA model.

In this context, we observe that two QA pairs $(\mathbf{q}_i, a_i)$ and $(\mathbf{q}_j, a_j)$ for the same image $\mathbf{x}$ can have different kinds of logical relations. In the simplest case, the two pairs may be unrelated, as with the pairs ("Is it nighttime?", "Yes") and ("Is there a bench in the image?", "No"). Knowing that one of the pairs is true gives no information about the truth value of the other.

On the other hand, two pairs may be related by a logical implication, as in the pairs ("Is the horse brown?", "No") and ("What is the color of the horse?", "White"). Knowing that the second pair is true implies that the first pair must be true as well. Conversely, if the first pair is false (*the horse is brown*), it implies that the second pair must also be false. In this case, the first pair is a necessary condition for the second one or, equivalently, the second pair is a sufficient condition for the first one.

Finally, it can be that two QA pairs are related by a double logical implication, as with the pairs ("Is this a vegetarian pizza?", "Yes") and ("Does the pizza have meat on it?", "No"). The veracity of the former implies the veracity of the latter, but the veracity of the latter also implies the veracity of the former. In this case, each pair is simultaneously a necessary and sufficient condition for the other pair, and both pairs are then equivalent.

Note that the logical implication existing between two QA pairs is an intrinsic property of the QA pairs, and does not depend on the correctness of the predictions coming from a VQA model. If a VQA model considers a sufficient condition true and a necessary condition false, it is incurring an *inconsistency* regardless of the correctness of its predictions.

Since logical implications are the basis of reasoning, we propose to explicitly use them when training a VQA model to reduce its inconsistent predictions. Unfortunately, doing so requires overcoming two important challenges: (1) a strategy is needed to train VQA models with logical relations that leverage consistency in a purposeful manner. Until now, no such approach has been proposed; (2) VQA datasets do not typically contain logical relations between pairs of QA. Acquiring these manually would, however, be both time-consuming and difficult.

We address these challenges in this work by formalizing the idea of consistency and treating QA pairs as logical propositions from which relations can comprehensively be defined. Using this formalism, we first propose a strategy to solve (1) and train a VQA model more effectively using logical relations and the consistency they provide (Sec. 3.2). We then show in Sec. 3.3 how we infer relations between pairs of propositions, whereby allowing standard VQA datasets to be augmented with logical relations.

3.1. Consistency formulation

We begin by observing that QA pairs $(\mathbf{q}, a)$ can be considered and treated as logical propositions. For instance, the QA ("Is it winter?", "Yes") can be converted to "It is winter," which is a logical proposition that can be evaluated as *true* or *false* (*i.e.*, its *truth value*). Doing so allows us to use a broad definition of consistency, namely one that establishes that two propositions are inconsistent if both cannot be true at the same time [7]. In the context of this work, we assume the truth value of a proposition $(\mathbf{q}, a)$ is determined by an agent (either a human annotator or the VQA model) after observing the information contained in an image $\mathbf{x}$.

Let $\mathcal{D} = \mathcal{I} \times \mathcal{Q} \times \mathcal{A}$ be a VQA dataset that contains triplets $(\mathbf{x}^{(n)}, \mathbf{q}_i^{(n)}, a_i^{(n)})$, where $\mathbf{x}^{(n)}$ is the n-th image and $(\mathbf{q}_i^{(n)}, a_i^{(n)})$ is the i-th question-answer pair about $\mathbf{x}^{(n)}$. In the following, we omit the index n for succinctness. For a given image $\mathbf{x}$, we can consider a pair of related question-answers as $(\mathbf{q}_i, a_i)$ and $(\mathbf{q}_j, a_j)$ as a pair of propositions. Following propositional logic notation, if both propositions are related in such a way that $(\mathbf{q}_i, a_i)$ is a sufficient condition for the necessary condition $(\mathbf{q}_j, a_j)$, we write that $(\mathbf{q}_i, a_i) \rightarrow (\mathbf{q}_j, a_j)$. For convenience, this arrow notation can be adapted to indicate different orderings between the necessary and sufficient conditions:

- $(\mathbf{q}_i, a_i) \leftarrow (\mathbf{q}_j, a_j)$ if the proposition $(\mathbf{q}_i, a_i)$ is a necessary condition for $(\mathbf{q}_j, a_j)$.
- $(\mathbf{q}_i, a_i) \leftrightarrow (\mathbf{q}_j, a_j)$ if the propositions $(\mathbf{q}_i, a_i)$ and $(\mathbf{q}_j, a_j)$ are equivalent, *i.e.*, both are simultaneously necessary and sufficient. Note that this is just notational convenience for the double implication $(\mathbf{q}_i, a_i) \rightarrow (\mathbf{q}_j, a_j) \wedge (\mathbf{q}_j, a_j) \rightarrow (\mathbf{q}_i, a_i)$, and in the following derivations the double arrow will be always considered as two independent arrows.

- Finally, we will write $(\mathbf{q}_i, a_i) - (\mathbf{q}_j, a_j)$ if the propositions $(\mathbf{q}_i, a_i)$ and $(\mathbf{q}_j, a_j)$ are not related.

If a VQA model is asked questions $\mathbf{q}_i$ and $\mathbf{q}_j$ about an image $\mathbf{x}$ and there exists a relation $(\mathbf{q}_i, a_i) \rightarrow (\mathbf{q}_j, a_j)$, the answers of the model will be inconsistent whenever it provides answers $\hat{a}_i = a_i$ and $\hat{a}_j \neq a_j$ (*i.e.*, the model evaluates the first proposition as true and the second proposition as false). More generally, for a pair of necessary and sufficient conditions, the agent will be inconsistent if it evaluates the necessary condition as false and the sufficient condition as true [7]. In what follows, we exploit these ideas to quantify model inconsistencies in our experiments and to develop a new loss function that encourages logically consistent VQA models.

3.2. Logical implication consistency loss

The core aim of our method is to encourage the VQA model to avoid inconsistent answers. When training, assume that the model receives an image $\mathbf{x}$ from $\mathcal{D}$ and two associated propositions $(\mathbf{q}_1, a_1)$ and $(\mathbf{q}_2, a_2)$ that are related by a logical implication $(\mathbf{q}_1, a_1) \rightarrow (\mathbf{q}_2, a_2)$. We define,

$$\pi_i = \pi\left((\mathbf{q}_i, a_i), \mathbf{x}\right) = p(a_i \mid \mathbf{x}, \mathbf{q}_i, \theta), \tag{2}$$

as the probability assigned by the VQA model that the proposition $(\mathbf{q}, a)$ is true for the image $\mathbf{x}$. The model has a high probability of incurring an inconsistency if it simultaneously gives a high probability π_1 to the sufficient condition and a low probability π_2 to the necessary condition.

We thus define our consistency loss as a function,

$$\begin{aligned}\mathcal{L}_{\text{cons}}(\mathbf{x}, (\mathbf{q}_1, a_1), (\mathbf{q}_2, a_2)) = -(1-\pi_2)\log(1-\pi_1) \\ -\pi_1 \log(\pi_2),\end{aligned} \tag{3}$$

that takes an image and a pair of sufficient and necessary propositions, and penalizes predictions with a high probability of inconsistency. As illustrated in Fig. 2, $\mathcal{L}_{\text{cons}}$ is designed to produce maximum penalties when $\pi_1 = 1$ and $\pi_2 < 1$ (*i.e.*, when the sufficient condition is absolutely certain but the necessary condition is not), and when $\pi_2 = 0$ and $\pi_1 > 0$ (*i.e.*, when the necessary condition can never be true but the sufficient condition can be true). At the same time, $\mathcal{L}_{\text{cons}}$ produces minimum penalties when either $\pi_1 = 0$ or $\pi_2 = 1$, as no inconsistency is possible when the sufficient condition is false or when the necessary condition is true. Interestingly, despite its resemblance, $\mathcal{L}_{\text{cons}}$ is not a cross-entropy, as it is not an expectation over a probability distribution.

Our final loss is then a linear combination of the consistency loss and the cross-entropy loss $\mathcal{L}_{\text{VQA}}$ typically used to train VQA models. Training with this loss then optimizes,

$$\min_{\theta} \mathbb{E}_{\mathcal{D}}[\mathcal{L}_{\text{VQA}}] + \lambda \mathbb{E}_{\substack{((\mathbf{x}_i, \mathbf{q}_i, a_i), (\mathbf{x}_j, \mathbf{q}_j, a_j)) \sim \mathcal{D}^2 \\ \mathbf{x}_i = \mathbf{x}_j, (\mathbf{q}_i, a_i) \rightarrow (\mathbf{q}_j, a_j)}} [\mathcal{L}_{\text{cons}}], \tag{4}$$

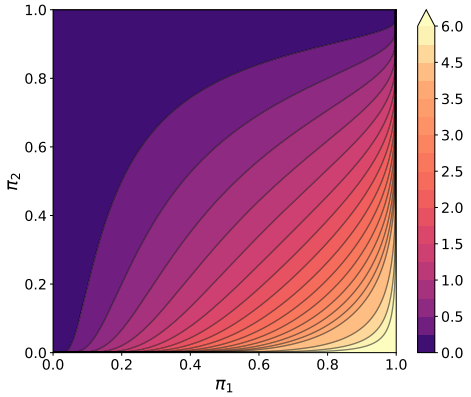

Figure 2. Consistency loss $\mathcal{L}_{\text{cons}}$ as a function of the estimated probabilities for the sufficient, π_1, and necessary, π_2, conditions. Note that the loss diverges to ∞ when $\pi_1 = 1, \pi_2 < 1$ and when $\pi_1 > 0, \pi_2 = 0$.

where the first expectation is taken over the elements of the training set $\mathcal{D}$ and the second expectation is taken over all pairs of necessary and sufficient propositions from $\mathcal{D}$ defined for the same image. In practice, we follow the sampling procedure described in [43, 47], where mini-batches contain pairs of related questions. The hyperparameter λ controls the relative strength between the VQA loss and the consistency term.

3.3. Inferring logical implications

By and large, VQA datasets do not include annotations with logical relations between question-answers pairs, which makes training a VQA with $\mathcal{L}_{\text{cons}}$ infeasible. To overcome this, we propose to train a language model to predict logical implications directly and use these predictions instead. We achieve this in two phases illustrated in Fig. 3 and refer to our approach as the Logical-Implication model (LI-MOD).

First, we pre-train BERT [13] on the task of Natural Language Inference using the SNLI dataset [58], which consists of pairs of sentences with annotations of entailment, contradiction or neutrality. In this task, given two sentences, a language model must predict one of the mentioned categories. While these categories do not exactly match the logical implication relevant to our objective, they can be derived from the entailment category. To this end, given two propositions $(\mathbf{q}_i, a_i)$ and $(\mathbf{q}_j, a_j)$, we evaluate them using the finetuned NLI model in the order $(\mathbf{q}_i, a_i), (\mathbf{q}_j, a_j)$, and then repeat the evaluation by inverting the order, to evaluate possible equivalences or inverted relations. If the relation is predicted as neutral in both passes, the pair is considered to be unrelated.

Then, we finetune the NLI model on a sub-set of annotated pairs from the VQA dataset Introspect [43]. In prac-

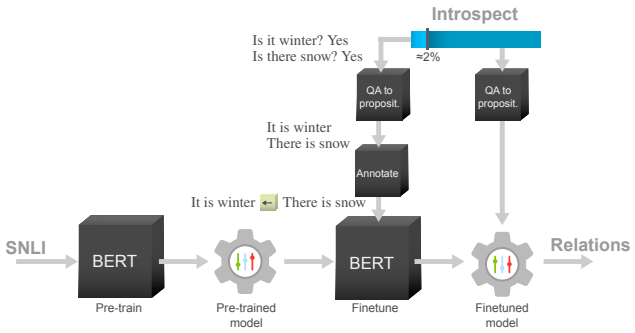

Figure 3. LI-MOD: Approach to predict logical relations between pairs of propositions. A BERT-based NLP model is first pre-trained on the SNLI dataset [58] to solve a Natural Language Inference task and subsequently fine-tuned with annotated pairs from a subset of Introspect dataset [43]. The resulting model is used to predict the relations of the remaining part of the dataset.

tice, we use a subset of binary QA pairs that were manually annotated with logical implications. Even though the relation need not be limited to binary questions (*i.e.*, yes/no questions), we chose to do so because the relation annotation is simpler than for open-ended questions. Since BERT expects sentences and not QA pairs, these were first converted into propositions using Parts Of Speech (POS) tagging [39] and simple rules that apply to binary questions (*e.g.*, to convert "Is it winter?," "Yes" we invert the first two words of the question and remove the question mark). After finetuning the model, the relations were predicted for the remaining part of the dataset. Further implementation details on this are given in Sec. 4.3.

4. Experiments

We evaluate our proposed consistency loss function on different datasets and using a variety of VQA models.

4.1. Datasets

Introspect [43]: Contains perception questions (or sub-questions) created by annotators for a subset of reasoning questions (or main questions) of the VQA v1.0 and v2.0 datasets [4,20]. It contains 27,441 reasoning questions with 79,905 sub-questions in its training set and 15,448 reasoning questions with 52,573 sub-questions for validation. For images that have the same sub-question repeated multiple times, we remove duplicates in the sub-questions for every image in both the train and validation sets.

DME Dataset [47]: Consists of retinal fundus images for the task of Diabetic Macular Edema (DME) staging. It contains 9,779 QA pairs for training, 2,380 QA pairs for validation and 1,311 QA pairs for testing. There are three types of questions in the dataset: main, sub, and independent questions. Main questions ask about diagnosis information (*i.e.* the stage of the disease) and sub-questions ask about the presence and location of biomarkers. Sub-questions are further subdivided into grade questions, questions about the whole image, questions about a region of the eye called macula, and questions about random regions in the image. To deal with questions about image regions, we follow the procedure described in [47], whereby only the relevant region is shown to the model.

4.2. Baseline methods and base models

We consider 3 different consistency enhancement baseline methods. To ensure fair comparisons, all methods use the same VQA base models and only differ in the consistency method used. These consist in:

- *None:* Indicating that no consistency preserving method is used with the VQA model. This corresponds to the case where $\lambda = 0$.
- *SQuINT [43]:* Optimizes consistency by maximizing the similarity between the attention maps of pairs of questions. As such, it requires a VQA model that uses guided attention.
- *CP-VQA [47]:* Assumes entailment relations and uses a regularizer to improve consistency.

VQA architectures: We show experiments using three VQA models depending on the dataset used. For experiments on Introspect, we make use of the BAN model [31], as its structure with guided attention allows the use of SQuINT. In addition, we evaluate the vision-language architecture LXMERT [46] on this dataset to evaluate improvement in state-of-the-art, transformer-based VQA models. For experiments on the DME dataset, we use the base model described in [47], which we denote by MVQA.

4.3. Implementation details

LI-Model We first pre-train BERT on SNLI for 5 epochs until it reaches a maximum accuracy of 84.32% on that dataset. For this pre-training stage, we initialize BERT with the *bert-base-uncased* weights and use a batch size of 16. We use a weight decay rate of 0.01 and the AdamW optimizer with a learning rate of $2 \cdot 10^{-5}$. The same setup was kept to finetune the model on a subset of 2'000 pairs of propositions from Introspect which were manually annotated (distribution of labels being: $\leftarrow$ 60%, $\leftrightarrow$ 17%, $-$12%, $\rightarrow$ 11%), and an additional 500 pairs were annotated for validation. Notice that LI-MOD is only necessary for the Introspect dataset since, for the DME dataset, the implications annotations are available.

VQA models: For our base models, we use the official and publicly available implementations (BAN [45],

LXMERT [46] and MVQA [47]) with default configurations. We re-implemented SQuINT [43] and used the provided implementation of CP-VQA [47], reporting the best results, which were obtained with $\lambda = 0.1, \gamma = 0.5$ for BAN and $\lambda = 0.5, \gamma = 1$ for MVQA (parameters refer to original implementations). For SQuINT, we set the gain of the attention map similarity term to 0.5 for BAN and 1.0 for MVQA. For Introspect, we train 5 models with different seeds for each parameter set and for DME, we train 10 models with different seeds. To train LXMERT, BAN and MVQA, we use batch sizes of 32, 64 and 128, respectively. Regarding the VQA cross-entropy loss, we follow the original implementations and use soft scores for the answers in LXMERT and categorical answers for BAN and MVQA.

4.4. Quantifying consistency

Given a test set $\mathcal{T} = \{t_n\}_{n=1}^{|\mathcal{T}|}$, where $t_n = (\mathbf{x}, \mathbf{q}, a)$ is a test sample triplet, we wish to measure the level of consistency of a VQA model p. To this end, we define the set of implications $G(\mathcal{T}) \subset \mathcal{T}^2$ as the collection of all pairs of test samples $((\mathbf{x}_i, \mathbf{q}_i, a_i), (\mathbf{x}_j, \mathbf{q}_j, a_j))$ for which $(\mathbf{q}_i, a_i) \rightarrow (\mathbf{q}_j, a_j)$ and $\mathbf{x}_i = \mathbf{x}_j$, and the set of inconsistencies $I_p(\mathcal{T})$ produced by the VQA model as the subset of $G(\mathcal{T})$ that contains the pairs for which the model evaluated the sufficient condition as true and the necessary condition as false,

$$I_p(\mathcal{T}) = \{(t_i, t_j) \in G(\mathcal{T}) \mid e_p((\mathbf{q}_i, a_i), \mathbf{x}) \wedge \neg e_p((\mathbf{q}_j, a_j), \mathbf{x})\}. \quad (5)$$

The function e_p returns the truth value of the proposition $(\mathbf{q}, a)$ for image $\mathbf{x}$ evaluated by the VQA model p,

$$e_p((\mathbf{q}, a), \mathbf{x}) = [\hat{a} = a], \quad (6)$$

where $\hat{a}$ is the answer of maximum probability following Eq. (1). In other words, e_p returns whether the estimated answer for question $\mathbf{q}$ matches the answer of the proposition a. Finally, the consistency ratio c for model p on the test set $\mathcal{T}$ is the proportion of implications in $G(\mathcal{T})$ that did not lead to an inconsistency,

$$c_p(\mathcal{T}) = 1 - \frac{|I_p(\mathcal{T})|}{|G(\mathcal{T})|}. \quad (7)$$

5. Results

Performance comparison: For both datasets, we first compare the performance of our method against the baseline consistency methods in Tab. 1 and Tab. 2. In either case, we see that our method outperforms previous approaches by not only increasing overall prediction accuracy but also by increasing consistency. In Fig. 4 and Fig. 5, we show illustrative examples of our approach on the Introspect and DME datasets, respectively (see additional examples in the Supplementary materials).

In Tab. 1, we also show the performance of the state-of-the-art LXMERT VQA model when combined with our method. In this case, too, we see that our method provides increased performance via consistency improvements. Here we investigate the performance induced when flipping the answers of one of the members of each inconsistent pair at test time. Suppose implication labels are present, either by manual annotation or by LI-MOD. In that case, a trivial manner of correcting an inconsistent QA pair of binary answers is to flip or negate one of the answers. This is far simpler than our proposed method as it permits training the VQA model with the standard VQA loss. Having obtained the answers from the model when $\lambda = 0$, we identify the inconsistent pairs using the relations predicted by our LI-MOD and then flip the answers (1) either randomly, (2) of the first QA or (3) of the second QA. By including the flipping baselines, we confirm that the added complexity in training our method results in improved accuracy compared to merely correcting inconsistencies post-hoc. Increases in consistency at the expense of accuracy are explained by the fact that an inconsistent QA pair guarantees that one of the two answers is incorrect, but correcting the inconsistency does not necessarily fix the incorrect answer. This phenomenon is particularly noticeable in the flipping baselines, as they fix inconsistencies without considering their correctness.

In general, we observe that training LXMERT with our consistency loss provides performance gains. Indeed, while random flipping based on LI-MOD clearly deteriorates the performance of LXMERT, so does flipping the first or sec-

Model	Cons. Method	Acc.	Cons.
BAN	None	67.14±0.10	69.45±0.17
	SQuINT [43]	67.27±0.19	69.87±0.45
	CP-VQA [47]	67.18±0.24	69.52±0.45
	Ours ($\lambda = 0.01$)	**67.36±0.19**	70.38±0.39
LXMERT	None	75.10±0.10	76.24±0.63
	Random flip	69.67±1.24	75.99±3.91
	Flip first	73.81±0.47	71.94±2.82
	Flip second	65.82±1.03	87.56±2.51
	Ours	**75.17±0.08**	78.75±0.21

Table 1. Results of different consistency methods on the Introspect dataset using two different VQA models: (top) BAN and (bottom) LXMERT. In the case of LXMERT, we show the impact of randomly flipping the answer of either the first or the second question for pairs detected as inconsistent using the relations from LI-MOD. Similarly, *flip first* and *flip second* refer to flipping the answer to the first and second question in inconsistent pairs, respectively.

Model	Consis. Method	Accuracy					Consistency
		all	grade	whole	macula	region	
MVQA	None	81.15±0.49	78.17±2.07	83.44±1.87	87.25±1.20	80.38±2.02	89.95±3.20
	SQuINT [43]	80.58±0.78	77.48±0.40	82.82±0.74	85.34±0.87	80.02	89.39±2.12
	CP-VQA [47]	83.49±0.99	**80.69±1.30**	84.96±1.14	87.18±2.18	**83.16±1.09**	94.20±2.15
	Ours ($\lambda = 0.25$)	**83.59±0.69**	80.15±0.95	**86.22±1.67**	**88.18±1.07**	82.62±1.02	95.78±1.19

Table 2. Comparison of methods on the DME dataset with common MVQA backbone. Accuracy and consistency are reported for all questions, as well as for different medically relevant sub-question categories: grade, whole, macula and region.

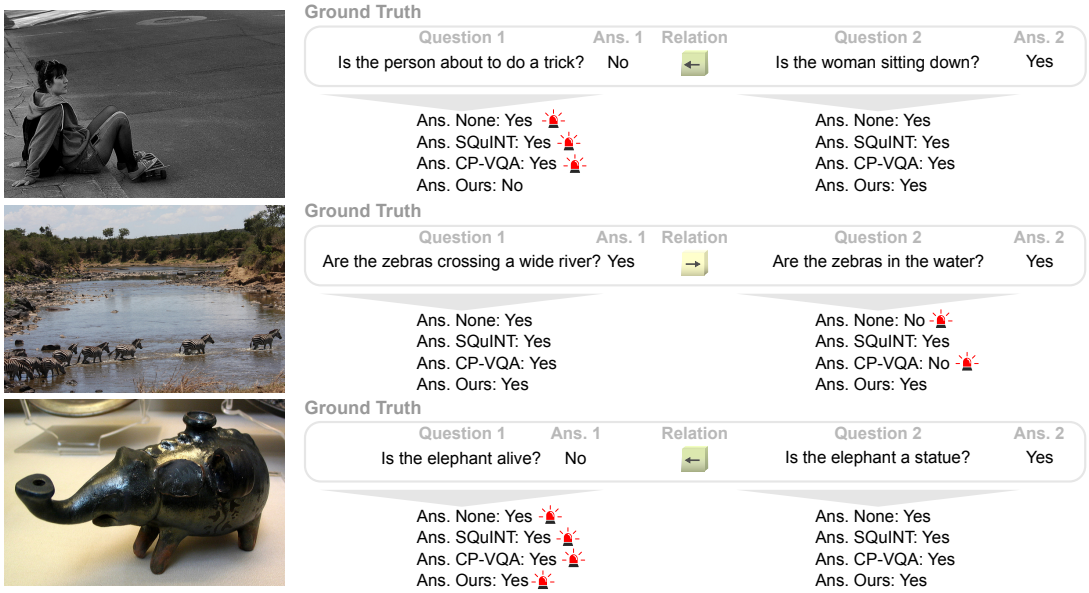

Figure 4. Qualitative examples from the Introspect dataset using BAN as backbone. Red siren symbols indicate inconsistent cases.

ond answers. This implies that our proposed method indeed leverages the predictions of LI-MOD to make LXMERT more consistent as it improves both model accuracy and consistency.

Sensitivity of λ: We now show the sensitivity of our method and its relation to λ. We evaluate the performance of our method for different values of λ to understand the behavior of the performance, both in terms of accuracy and consistency.

Fig. 6 shows the accuracy and consistency of LXMERT and MVQA for different values of λ. The difference in the ranges of the values is due to the relative magnitude of the loss function terms and depends on the used loss functions (*e.g.*, binary and non-binary cross-entropy) and the ground-truth answer format (*i.e.*, soft scores for LXMERT, as mentioned in Sec. 4.3).

In general, we observe very similar behavior for the accuracy, which increases and then slowly decreases as λ increases. We sustain that the maximum value the accuracy can reach is established by the number of related pairs that are still inconsistent after training with $\lambda = 0$. In other words, the limitations in size impose a limit on how much our method can improve the accuracy. For LXMERT on Introspect, for instance, our model corrected 4,553 (78.9%) of the 5'771 existing inconsistencies and introduced new inconsistencies by mistakenly altering 1,562 (3.5%) of the 44,111 consistent samples.

Regarding consistency, we observe a constant increase as λ increases. The simultaneous decrease in accuracy as λ increases suggests that the relative weight of the consistency loss dominates so that the model no longer focuses on optimizing the cross-entropy. Since it is possible to be consistent without answering correctly, the optimization process results in an increase in consistency at the expense of accuracy for higher values of λ. However, it is clear from these results that there is a set of λ values for which both metrics improve.

LI-MOD performance: We report that the finetuning of BERT on the subset of annotated relations from Introspect produced 78.67% accuracy in the NLI task. We analyze

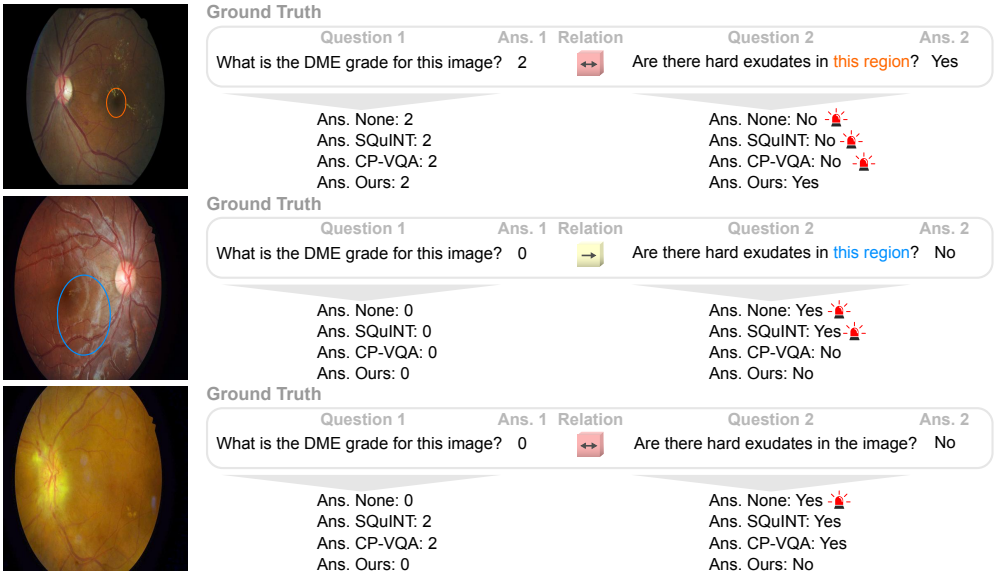

Figure 5. Examples from the DME dataset and comparison of methods. Red siren symbols indicate inconsistent cases. DME is a disease that is staged into grades (0, 1 or 2), which depend on the number of visual pathological features of the retina. *Top* and *middle:* Although all methods correctly predict the answer to the first question, some inconsistencies appear when a necessary condition is false. *Bottom*: Only the None baseline produces an inconsistency. Note that SQuINT and CP-VQA's answers do not produce inconsistent pairs because both questions were answered incorrectly, and those answers ("2" and "yes") respect all known relations.

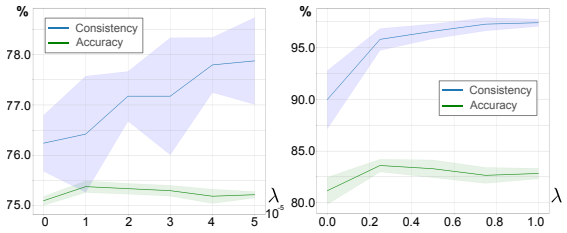

Figure 6. Behavior of the accuracy and consistency as a function of λ with 95% confidence intervals. *Left:* LXMERT trained on the Introspect dataset (5 models with random seeds for each value of λ). *Right:* MVQA trained on the DME dataset (10 models with random seeds for each λ).

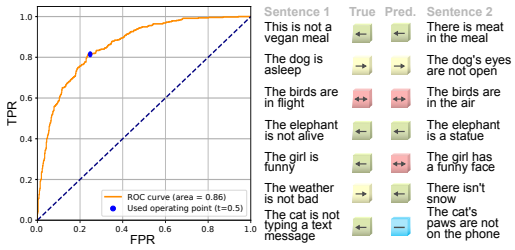

Figure 7. *Left:* Receiver Operating Characteristic (ROC) for the entailment class of our LI-MOD in validation. *Right:* Qualitative examples of LI-MOD's predictions.

the performance of this model for entailment and report an AUC value of 0.86, which indicates good generalization capability considering that only $\approx 2\%$ of the dataset was annotated with relations. In addition, the overlap in the QA pairs between the train and validation sets of the Introspect dataset is only 1.12% for binary questions. This shows that our LI-MOD is generalizing to variations in questions and to new combinations of QA pairs. Fig. 7 shows the ROC curve for entailment and examples of LI-MOD's predictions. Some of the observed sources of errors in LI-MOD include negations, unusual situation descriptions (*e.g.*, a cat typing a text message), and image-specific references (*e.g.*, "is *this* animal real?").

6. Conclusion and future work

In this paper, we propose a model-agnostic method to measure and improve consistency in VQA by integrating logical implications between pairs of questions in the training process. We also present a method to infer implications between QA pairs using a transformer-based language model. We conduct experiments to validate the generalizability and robustness of our consistency loss against several baselines and across different datasets. Our results show that our method reduces incoherence in responses and improves performance. Future work includes creating a larger dataset with human-annotated relations to use as a general-purpose relations database for VQA training.

Acknowledgements This work was partially funded by the Swiss National Science Foundation through grant 191983.

References

[1] Ehsan Abbasnejad, Damien Teney, Amin Parvaneh, Javen Shi, and Anton van den Hengel. Counterfactual vision and language learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10044–10054, 2020.

[2] Vedika Agarwal, Rakshith Shetty, and Mario Fritz. Towards causal vqa: Revealing and reducing spurious correlations by invariant and covariant semantic editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9690–9698, 2020.

[3] Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. Don't just assume; look and answer: Overcoming priors for visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4971–4980, 2018.

[4] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015.

[5] Pratyay Banerjee, Tejas Gokhale, Yezhou Yang, and Chitta Baral. Weakly supervised relative spatial reasoning for visual question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1908–1918, 2021.

[6] Hedi Ben-Younes, Rémi Cadene, Matthieu Cord, and Nicolas Thome. Mutan: Multimodal tucker fusion for visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2612–2620, 2017.

[7] R. Bradley and N. Swartz. *Possible Worlds: An Introduction to Logic and Its Philosophy*. B. Blackwell, 1979.

[8] Remi Cadene, Hedi Ben-Younes, Matthieu Cord, and Nicolas Thome. Murel: Multimodal relational reasoning for visual question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1989–1998, 2019.

[9] Qingxing Cao, Wentao Wan, Keze Wang, Xiaodan Liang, and Liang Lin. Linguistically routing capsule network for out-of-distribution visual question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1614–1623, 2021.

[10] Chongyan Chen, Samreen Anjum, and Danna Gurari. Grounding answers for visual questions asked by visually impaired people. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19098–19107, 2022.

[11] Long Chen, Xin Yan, Jun Xiao, Hanwang Zhang, Shiliang Pu, and Yueting Zhuang. Counterfactual samples synthesizing for robust visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10800–10809, 2020.

[12] Corentin Dancette, Remi Cadene, Damien Teney, and Matthieu Cord. Beyond question-based biases: Assessing multimodal shortcut learning in visual question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1574–1583, 2021.

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[14] Yang Ding, Jing Yu, Bang Liu, Yue Hu, Mingxin Cui, and Qi Wu. Mukea: Multimodal knowledge extraction and accumulation for knowledge-based visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5089–5098, 2022.

[15] Tuong Do, Binh X Nguyen, Erman Tjiputra, Minh Tran, Quang D Tran, and Anh Nguyen. Multiple meta-model quantifying for medical visual question answering. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 64–74. Springer, 2021.

[16] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. *arXiv preprint arXiv:1606.01847*, 2016.

[17] Feng Gao, Qing Ping, Govind Thattai, Aishwarya Reganti, Ying Nian Wu, and Prem Natarajan. Transform-retrieve-generate: Natural language-centric outside-knowledge visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5067–5077, 2022.

[18] Vatsal Goel, Mohit Chandak, Ashish Anand, and Prithwijit Guha. IQ-VQA: Intelligent visual question answering. In *International Conference on Pattern Recognition*, pages 357–370. Springer, 2021.

[19] Tejas Gokhale, Pratyay Banerjee, Chitta Baral, and Yezhou Yang. VQA-LOL: Visual question answering under the lens of logic. In *European conference on computer vision*, pages 379–396. Springer, 2020.

[20] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6904–6913, 2017.

[21] Deepak Gupta, Swati Suman, and Asif Ekbal. Hierarchical deep multi-modal network for medical visual question answering. *Expert Systems with Applications*, 164:113993, 2021.

[22] Vipul Gupta, Zhuowan Li, Adam Kortylewski, Chenyu Zhang, Yingwei Li, and Alan Yuille. Swapmix: Diagnosing and regularizing the over-reliance on visual context in visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5078–5088, 2022.

[23] Danna Gurari, Qing Li, Chi Lin, Yinan Zhao, Anhong Guo, Abigale Stangl, and Jeffrey P Bigham. Vizwiz-priv: A dataset for recognizing the presence and purpose of private visual information in images taken by blind people. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 939–948, 2019.

[24] Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham.

Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3608–3617, 2018.

[25] Xinzhe Han, Shuhui Wang, Chi Su, Qingming Huang, and Qi Tian. Greedy gradient ensemble for robust visual question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1584–1593, 2021.

[26] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*, 2020.

[27] Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. Pathvqa: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*, 2020.

[28] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019.

[29] Chenchen Jing, Yunde Jia, Yuwei Wu, Xinyu Liu, and Qi Wu. Maintaining reasoning consistency in compositional visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5099–5108, 2022.

[30] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2901–2910, 2017.

[31] Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. Bilinear attention networks. *Advances in neural information processing systems*, 31, 2018.

[32] Jin-Hwa Kim, Kyoung-Woon On, Woosang Lim, Jeonghee Kim, Jung-Woo Ha, and Byoung-Tak Zhang. Hadamard product for low-rank bilinear pooling. *arXiv preprint arXiv:1610.04325*, 2016.

[33] Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1):1–10, 2018.

[34] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[35] Bill MacCartney and Christopher D Manning. Modeling semantic containment and exclusion in natural language inference. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 521–528, 2008.

[36] Hyeonseob Nam, Jung-Woo Ha, and Jeonghee Kim. Dual attention networks for multimodal reasoning and matching. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 299–307, 2017.

[37] Binh D Nguyen, Thanh-Toan Do, Binh X Nguyen, Tuong Do, Erman Tjiputra, and Quang D Tran. Overcoming data limitation in medical visual question answering. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 522–530. Springer, 2019.

[38] Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. Adversarial nli: A new benchmark for natural language understanding. *arXiv preprint arXiv:1910.14599*, 2019.

[39] Slav Petrov, Dipanjan Das, and Ryan McDonald. A universal part-of-speech tagset. *arXiv preprint arXiv:1104.2086*, 2011.

[40] Arijit Ray, Karan Sikka, Ajay Divakaran, Stefan Lee, and Giedrius Burachas. Sunny and dark outside?! improving answer consistency in vqa through entailed question generation. *arXiv preprint arXiv:1909.04696*, 2019.

[41] Marco Tulio Ribeiro, Carlos Guestrin, and Sameer Singh. Are red roses red? evaluating consistency of question-answering models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6174–6184, 2019.

[42] Aalok Sathe, Salar Ather, Tuan Manh Le, Nathan Perry, and Joonsuk Park. Automated fact-checking of claims from wikipedia. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6874–6882, 2020.

[43] Ramprasaath R Selvaraju, Purva Tendulkar, Devi Parikh, Eric Horvitz, Marco Tulio Ribeiro, Besmira Nushi, and Ece Kamar. Squinting at vqa models: Introspecting vqa models with sub-questions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10003–10011, 2020.

[44] Meet Shah, Xinlei Chen, Marcus Rohrbach, and Devi Parikh. Cycle-consistency for robust visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6649–6658, 2019.

[45] Weijie Su. Pythia. `https://github.com/jackroos/pythia`, 2019.

[46] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019.

[47] Sergio Tascon-Morales, Pablo Márquez-Neila, and Raphael Sznitman. Consistency-preserving visual question answering in medical imaging. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 386–395. Springer, 2022.

[48] Damien Teney, Ehsan Abbasnejad, and Anton van den Hengel. On incorporating semantic prior knowledge in deep learning through embedding-space constraints. *arXiv preprint arXiv:1909.13471*, 2019.

[49] Damien Teney, Ehsan Abbasnejad, and Anton van den Hengel. Unshuffling data for improved generalization. *arXiv preprint arXiv:2002.11894*, 2020.

[50] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[51] Minh H Vu, Tommy Löfstedt, Tufve Nyholm, and Raphael Sznitman. A question-centric model for visual question an-

swering in medical imaging. *IEEE transactions on medical imaging*, 39(9):2856–2868, 2020.

[52] Matthew Walmer, Karan Sikka, Indranil Sur, Abhinav Shrivastava, and Susmit Jha. Dual-key multimodal backdoors for visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15375–15385, 2022.

[53] Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32, 2019.

[54] Peng Wang, Qi Wu, Chunhua Shen, Anton van den Hengel, and Anthony Dick. Explicit knowledge-based reasoning for visual question answering. *arXiv preprint arXiv:1511.02570*, 2015.

[55] Adina Williams, Nikita Nangia, and Samuel R Bowman. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426*, 2017.

[56] Yirui Wu, Yuntao Ma, and Shaohua Wan. Multi-scale relation reasoning for multi-modal visual question answering. *Signal Processing: Image Communication*, 96:116319, 2021.

[57] Xu Yang, Chongyang Gao, Hanwang Zhang, and Jianfei Cai. Auto-parsing network for image captioning and visual question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2197–2207, 2021.

[58] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014.

[59] Yuanyuan Yuan, Shuai Wang, Mingyue Jiang, and Tsong Yueh Chen. Perception matters: Detecting perception failures of vqa models using metamorphic testing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16908–16917, 2021.

[60] Li-Ming Zhan, Bo Liu, Lu Fan, Jiaxin Chen, and Xiao-Ming Wu. Medical visual question answering via conditional reasoning. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2345–2354, 2020.

[61] Yiyi Zhou, Tianhe Ren, Chaoyang Zhu, Xiaoshuai Sun, Jianzhuang Liu, Xinghao Ding, Mingliang Xu, and Rongrong Ji. Trar: Routing the attention spans in transformer for visual question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2074–2084, 2021.