Reliable Visual Question Answering: Abstain Rather Than Answer Incorrectly

Vedaad Shakib
Spencer Whitehead
Suzanne Petryk
Joseph Gonzalez, Ed.
Trevor Darrell, Ed.
Anna Rohrbach, Ed.
Marcus Rohrbach, Ed.

Electrical Engineering and Computer Sciences
University of California, Berkeley

Technical Report No. UCB/EECS-2022-137
http://www2.eecs.berkeley.edu/Pubs/TechRpts/2022/EECS-2022-137.html

May 18, 2022

Reliable Visual Question Answering: Abstain Rather Than Answer Incorrectly

by Vedaad Shakib

Research Project

Submitted to the Department of Electrical Engineering and Computer Sciences, University of California at Berkeley, in partial satisfaction of the requirements for the degree of **Master of Science, Plan II**.

Approval for the Report and Comprehensive Examination:

Committee:

Joseph Gonzalez

Professor Joseph Gonzalez
Research Advisor

5/16/2022

(Date)

* * * * * * *

Trevor Darrell

Professor Trevor Darrell
Second Reader

5/18/2022

(Date)

Reliable Visual Question Answering: Abstain Rather Than Answer Incorrectly

Spencer Whitehead[1*] **Suzanne Petryk**[1,2*] **Vedaad Shakib**[2] **Joseph Gonzalez**[2]

Trevor Darrell[2] **Anna Rohrbach**[2] **Marcus Rohrbach**[1]

[1]Facebook AI Research (FAIR) [2]UC Berkeley

Abstract

Machine learning has advanced dramatically, narrowing the accuracy gap to humans in multimodal tasks like visual question answering (VQA). However, while humans can say "*I don't know*" when they are uncertain (i.e., *abstain* from answering a question), such ability has been largely neglected in multimodal research, despite the importance of this problem to the usage of VQA in real settings. In this work, we promote a problem formulation for *reliable VQA*, where we prefer abstention over providing an incorrect answer. We first enable abstention capabilities for several VQA models, and analyze both their *coverage*, the portion of questions answered, and *risk*, the error on that portion. For that we explore several abstention approaches. We find that although the best performing models achieve over 71% accuracy on the VQA v2 dataset, introducing the option to abstain by directly using a model's softmax scores limits them to answering less than 8% of the questions to achieve a low risk of error (i.e., 1%). This motivates us to utilize a multimodal selection function to directly estimate the correctness of the predicted answers, which we show can triple the coverage from, for example, 5.0% to 16.7% at 1% risk. We also explore probabilistic calibration of VQA models, which improves coverage but less so than a multimodel selection function. While it is important to analyze both coverage and risk, these metrics have a trade-off which makes comparing VQA models challenging. To address this, we also propose an *Effective Reliability* metric for VQA that places a larger cost on incorrect answers compared to abstentions. This new problem formulation, metric, and analysis for VQA provide the groundwork for building effective and reliable VQA models that have the self-awareness to abstain if and only if they don't know the answer.

1 Introduction

Visual Question Answering (VQA) is an important task, and technology for it can offer a great benefit to users in trying to understand visual information. For instance, one core application of VQA is to provide a multimodal assistant, such as one that can answer questions to help with daily tasks for a user with visual impairments [6, 30]. To provide such utility, users must be able to trust the output of these tools as they may be basing decisions or actions on the output [7, 28, 56, 58]. While improving the accuracy of approaches may be an important factor for trusting models, models are imperfect and will inevitably produce some incorrect answers. In many scenarios, there is a price associated with a model giving an inaccurate answer as it may mislead the user and cause them to make a mistake that could be anywhere from mildly inconvenient to very serious. This is especially true for the example of helping visually impaired users, since they likely do not have a method of verifying the outputs themselves.

[*]Equal contribution

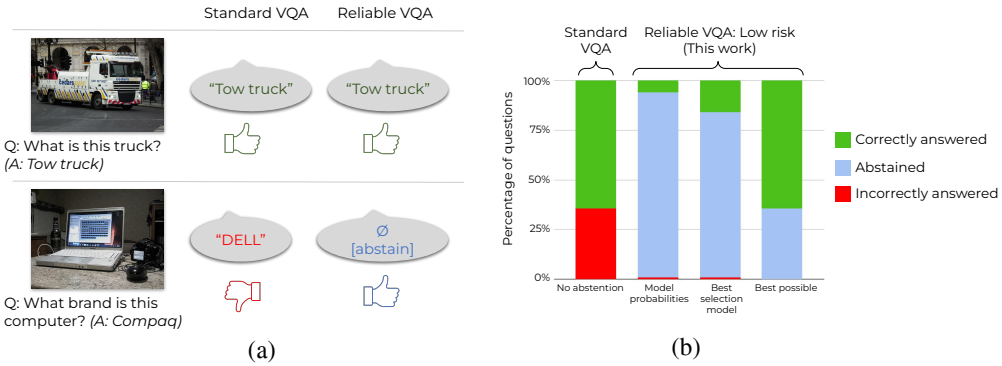

Figure 1: In the standard VQA problem, a model is required to provide an answer for all questions, even if it is likely to produce an error that can mislead a user, e.g. (a). A reliable VQA model, on the other hand, operates at *low risk* by having the option to abstain from answering if uncertain. As shown in (b), at 1% risk of error, a SoTA VQA model [72] can answer only about 5% of questions when using vanilla model probability estimates for determining when to abstain. Using a learned, multimodal selection function to estimate confidences triples the amount of questions the model can answer, yet there remains much room for improvement (Best possible, i.e., perfect abstention).

One way to avoid providing incorrect information and misleading users is to *abstain* from making a prediction, as in the framework of selective prediction [14, 21, 24, 25]. Consider Fig. 1(a): when a model is correct, we naturally would like it to give us an answer. However, when it is unable to do so (e.g., cannot "read" the brand name) or is very uncertain, we may prefer if the model communicated "*I don't know.*" In many applications, abstaining may be much more preferable compared to yielding an incorrect output [31, 44]. We say that VQA models are reliable, if they make highly accurate predictions when they choose to answer. Ideally, reliable models should also abstain as little as possible to be effective. Although reliability is often critical for the usage of VQA in real settings, this aspect has not received direct attention in the VQA literature aside from efforts to recognize unanswerable or false premise questions [12, 30, 40, 67]. Moreover, past efforts on selective prediction have not focused on the multimodal setting, where both an image and a question can be valid or in-distribution when considered independently, yet challenging in tandem.

In this work, we formalize and explore the notion of reliability in VQA. We propose to frame the task as a selective prediction problem [14, 21] in which models must either predict an answer or abstain from answering. This requires two techniques that have not been widely explored for VQA models: (1) gauging uncertainty of predictions and (2) learning when to abstain. To operationalize this framework, we measure performance with *coverage* (how many questions are answered) and *risk* (the error on these questions) [21, 42]. While low risk and high coverage are the goal, in practice there often is a trade-off between the two. To provide a scalar measure for this trade-off, we introduce a new *Effective Reliability* metric, which accounts for abstention while also introducing a cost for giving an incorrect answer. This also provides an alternative evaluation for domains where it may be more intuitive to specify the penalty for an individual error instead of a bound on risk.

Under this framework, we first show that existing VQA approaches leave much room for improvement. In particular, we demonstrate that, for a number of models, the common approach of using the maximum probability to determine abstention [34, 42] (by thresholding the softmax scores) limits the model to answering a small fraction of questions with a low risk of error (e.g., answering less than 8% of questions at 1% risk of error), despite having high standard VQA accuracy. This inability to answer a larger number of questions at low risk indicates low utility of the existing VQA models.

To address this, we explore two other approaches: calibration and training a multimodal selection function. We find that calibration leads to a better risk-coverage trade-off compared to using the original model probabilities. We further adapt calibration to the VQA setting, by adapting the calibration loss to use VQA soft targets, instead of a binary score. We improve beyond this by training a multimodal selection function that can better learn to predict if the model's answer is correct, based on intermediate representations as well as the answer from the VQA model. By training this selection function, we are able to consistently improve the coverage of different VQA models across varying risks of error, particularly for levels of low risk. However, we show that there is still room

to improve the effectiveness of these models (see Fig. 1(b)). Finally, we evaluate VQA models with our new effective reliability metric, and see that it correlates with risk/coverage in a meaningful way – the user-defined cost of an error impacts the risk at which the model operates.

In summary, our contributions are the following: (1) we are the first to analyze and operationalize reliability for multimodal VQA models; (2) we expose the issue of low coverage in VQA models when asked to operate at low risk levels; (3) we effectively adapt calibration techniques to the novel VQA setting; (4) we explore several methods for incorporating abstention, showing that a simple yet effective multimodal selection function outperforms other methods; (5) we propose a novel *Effective Reliability* metric for this problem, establishing a new benchmark for effective and reliable VQA models.

2 Related Work

VQA methods. Visual Question Answering (VQA) is a popular task with a plethora of methods proposed in recent years [2, 6, 11, 22, 23, 38, 39, 51, 53, 55, 72, 82, 83, 84]. To the best of our knowledge, there are no VQA models with a built-in abstention mechanism (i.e., they predict an answer for every image and question pair). We discuss a few exceptions with a non-standard problem statement in the following. Our work analyzes VQA models' reliability. Specifically, we introduce an ability to abstain into several prominent VQA models [39, 51, 55, 72].

Detecting intrinsic difficulty. Some prior work on VQA involves the categorization and detection of questions that are intrinsically difficult to answer, regardless of model ability. For example, the VizWiz VQA dataset contains labels for questions which are unanswerable [30] and reasons for annotation entropy, such as low image quality or question ambiguity [8]. [16] defines a similar categorization of unanswerable questions in VQA. Other work focuses on detecting whether the question incorrectly describes the visual semantics [40, 52, 57, 67]. Identifying intrinsically difficult examples has important implications in active learning, where such examples can stifle the ability of different methods to select useful examples to train on [43]. In this work, we focus on predicting uncertainty specific to a model as opposed to the intrinsic difficulty from data itself. However, in Sec. 5.6, we find that a subset of questions on which a model abstains from answering are ambiguous or unanswerable.

Calibration. In a classification setting, calibration typically refers to probabilistic calibration, where the predicted confidence for a given class should be representative of the probability of the prediction being correct. One popular parametric method is Platt scaling [64], in which a logistic regression model is trained on classifier outputs on the validation set to return calibrated probabilities, often by optimizing the negative log-likelihood [29, 62]. [48] improves upon Platt scaling by replacing the logistic regression model with a parametrized mapping based on the beta distribution. [29] extends Platt scaling to temperature scaling, a single-parameter variant of Platt scaling which is easier to implement on modern neural networks. [29] also introduces vector scaling and matrix scaling, two multi-class extensions to Platt scaling. [78] presents work on calibrating a model by training with Mixup, a data augmentation technique where new samples are generated by combining images and their associated labels. [59] introduces a new loss function called focal loss which replaces cross-entropy loss and leads to better calibration. [49] uses deep ensembles to produce accurate uncertainty estimates. In our work, we focus on a variant of vector scaling for improving selective prediction performance.

Selective prediction. This refers to when models have the option to abstain from providing a prediction. It is also known by other terms like sample rejection [14] or selective classification [21]. Early works describe a formalized framework for sample rejection rules [13, 14]. [17] introduces a framework for evaluating a rejection function when assigning cost coefficients to misclassified, abstained, and correctly classified samples. Other works integrate a reject option in multi-stage networks or ensembles [15, 45, 65, 80]. [81] shows that a simple error regularization during training can improve selective prediction for NLP tasks. Furthermore, [24, 25] present a line of work where selective models optimize for specific coverage levels. An alternative direction is to measure the agreement between the classifier and a modified nearest-neighbor classifier on the testing example [37]. We explore learned selection functions, but in the multimodal VQA setting, where the complex interaction between modalities must be modeled and more than one output may be considered correct to varying degrees, as opposed to unimodal image or language classification or extractive tasks. Additionally,

for VQA, there is a clear motivation for abstention when applying these models in practice, yet this is also an unaddressed and largely understudied problem. Some efforts explore selective prediction performance on out-of-distribution data [27, 42]. In particular, [42] focus on selective prediction for text-based question answering. However, they show that their method does not generalize to questions from the same domain which are intrinsically unanswerable, whereas this represents an important portion of difficult VQA samples. Lastly, somewhat relevant in the multimodal space, [33] address gender bias in image captioning (i.e., the cost of misclassifying gender should be high). The proposed model can "abstain" from outputting gendered words, predicting "person" when it is uncertain. In our framework and with our proposed metric, the cost of error can be defined by a user, and potentially even be made class-specific to weigh certain errors higher.

Conformal prediction. This aims to predict a set of outputs, with a guarantee that the set contains the correct output with a specified probability [79, 69]. So far this has been primarily studied in classification settings, such as ImageNet [18], where classes are distinct [3]. In contrast, in VQA, two distinct answers might be both be true (*"yellow","brown"*) for *"What color are the bananas?"*, but others sets might be contradictory (*"yes","no"*). Further research is required on how to best convey answer sets to users in VQA and how semantic similarity of answers should be modeled (e.g., {*"yellow","brown"*} vs {*"yes","no"*}), especially if the sets become large. More generally, the field of risk control, which does not require variable-size output sets, provides theoretical guarantees that a given error measure is below a tolerance level with some specified probability [4, 36]. [5] describes how to choose a prediction threshold to satisfy a guarantee on error bound. [36] relates these guarantees to test sample accuracy based on training sample density. We view these probabilistic guarantees on error bounds as complementary to our framework, with opportunities for future work to incorporate them both.

3 Visual Question Answering with Abstention

Visual question answering is currently formulated and evaluated in the literature [6, 26, 30, 35] as *always* predicting an answer from the answer space, $\mathcal{A}$, annotated in the dataset. So, a model f predicts an answer $a \in \mathcal{A}$ for each input $x = (v, q) \in \mathcal{X}$, with the visual input v and question q:

$$f : \mathcal{X} \mapsto \mathcal{A} \tag{1}$$

This problem formulation forces the model to answer even if it is likely wrong, thus providing unreliable answers. To address this issue, we propose to extend the VQA problem formulation so that a model is given the option to *abstain* from answering a question (i.e., effectively saying "*I don't know*"). In settings outside of VQA, this problem formulation has also been referred to as "*classification with a reject option*" [13, 17, 25, 31, 65], or "*selective prediction/classification*" [21, 24]. We first discuss the problem definition in Sec. 3.1, and then the metrics to evaluate this problem in Sec. 3.2.

3.1 Problem Definition

We extend the standard VQA formulation (Eq. 1) to the setting where a model can either provide an answer from $\mathcal{A}$ or choose to abstain, denoted by $\emptyset$:

$$h : \mathcal{X} \mapsto \mathcal{A} \cup \{\emptyset\}. \tag{2}$$

We refer to h as a *selective model*.

One way to formulate and achieve this is by decomposing h into two functions, f and g, which jointly comprise a selective model [21, 24, 25]. f denotes the VQA model that predicts answers (Eq. 1), and $g : \mathcal{X} \mapsto \{0, 1\}$ is the selection function that determines whether the model answers or abstains from answering:

$$h(x) = (f, g)(x) = \begin{cases} f(x) & \text{if } g(x) = 1, \\ \emptyset & \text{if } g(x) = 0. \end{cases} \tag{3}$$

Given an input x, the selective model yields an output from f when the selection function predicts that an answer should be given, or abstains if the selection function predicts that the model should not answer. One straightforward way to formulate the selection function g is based on a threshold γ,

where the function $g' : \mathcal{X} \mapsto [0, 1]$ predicts a confidence in the correctness[2] of the model $f(x)$ [42]:

$$g(x) = \begin{cases} 1 & \text{if } g'(x) \geq \gamma, \\ 0 & \text{if } g'(x) < \gamma. \end{cases} \tag{4}$$

In Sec. 4, we will discuss how to define this function $g'(x)$.

3.2 Evaluation Metrics

To evaluate a VQA model with an ability to abstain, we consider two types of evaluation and discuss how we adapt them for VQA: first, *coverage* and *risk* [21] and, second, a cost-based metric for balancing the two.

3.2.1 Risk and Coverage

Coverage is the portion of questions that the model opted to answer, while *risk* is the error on that portion of questions [21]. Ideally, a reliable model should exhibit high coverage at low levels of risk, meaning it answers many questions with high accuracy and abstains on others. More concretely, coverage for dataset $\mathcal{D}$ with inputs x_i and ground truth answers y_i is given by

$$\mathcal{C}(g) = \frac{1}{|\mathcal{D}|} \sum_{(x_i, y_i) \in \mathcal{D}} g(x_i) \tag{5}$$

and risk is defined as

$$\mathcal{R}(f, g) = \frac{\frac{1}{|\mathcal{D}|} \sum_{(x_i, y_i) \in \mathcal{D}} \ell(f(x_i), y_i) \cdot g(x_i)}{\mathcal{C}(g)}, \tag{6}$$

where ℓ is a cost function that measures the error between the predicted answer $f(x_i)$ and the corresponding ground truth answer y_i. Assuming g follows Eq. 4, if the threshold γ decreases, coverage will increase, but risk will increase as well. Hence, there is a risk-coverage trade-off that models can aim to optimize.

Applying this to VQA, the composite function (f, g) becomes our VQA model, where f produces an answer and g decides whether to abstain. However, the open-ended nature of the VQA task requires careful consideration for designing the risk-coverage metrics. A given question might have multiple possible answers which could all be considered correct to varying degrees. As a result, the error for a prediction on a given input is not necessarily binary.

When calculating risk, we must use a cost function that accurately represents this multi-class nature. We follow [6] to define VQA accuracy for a given model answer $f(x)$ as $Acc(f(x), y) = \min\left(\frac{\text{\# annotations that match } f(x)}{3}, 1\right)$ and average these accuracies over all 10 choose 9 subsets of human annotated answers for the input question, similar to other VQA evaluations [26, 30, 75]. Under this, an answer is considered fully correct if it matches at least four of the human annotations, and receives partial credit for selecting an answer with one, two, or three humans in agreement. Thus, our risk measurement becomes:

$$\mathcal{R}(f, g) = \frac{\frac{1}{|\mathcal{D}|} \sum_{(x_i, y_i) \in \mathcal{D}} (1 - Acc(f(x_i), y_i)) \cdot g(x_i)}{\mathcal{C}(g)}. \tag{7}$$

In practice, the level of risk in model predictions that a user is willing to tolerate depends highly on the scenario. Therefore, we evaluate model reliability by computing the coverage at a range of risk levels ($\mathcal{C}@\mathcal{R}$), such as coverage at 1% or 10% risk. We can also summarize this over the distribution of risk levels by plotting coverage versus corresponding risk, and computing the area under the risk-coverage curve (AUC) [42]. Moreover, for an evaluation that controls for how the threshold γ of the selection function is chosen, we compute the maximum coverage for a desired risk tolerance, allowing for a more direct comparison of the selection function design.

[2]While we define the output space of g' as $[0, 1]$ as is the case for the common softmax, one can similarly define an output space which covers, e.g., all real values $\mathbb{R}$.

3.2.2 Effective Reliability

Recall the trade-off between risk and coverage: a standard VQA model may have high risk at 100% coverage, yet a reliable model may have low risk yet abstain on a large portion of questions (see Fig. 1(b)). In practice, for a model to be effective, it should achieve both low risk and high coverage. To jointly measure these two desirable qualities, we define a metric which assigns a reward to questions which are answered correctly, a penalty to those answered entirely incorrectly, and zero reward to those abstained on. We refer to this as *Effective Reliability*, or Φ_c for a given penalty c, inspired by the "effectiveness function" introduced by [17]. Besides providing a scalar metric to summarize the risk-coverage trade-off, Effective Reliability also provides an alternative evaluation setting for domains where it may be easier or more intuitive to define a cost for an incorrect answer as opposed to a target level of risk.

Formally, we define Effective Reliability as $\Phi_c(x)$ (Eq. 8), where c is the cost for answering incorrectly, g is the selection function, and Acc is a measure of a model's correctness. In our particular case, Acc is the VQA accuracy [6].

$$\Phi_c(x) = \begin{cases} Acc(x) & \text{if } g(x) = 1 \text{ and } Acc(x) > 0, \\ -c & \text{if } g(x) = 1 \text{ and } Acc(x) = 0, \\ 0 & \text{if } g(x) = 0. \end{cases} \tag{8}$$

Note that this formulation assigns a reward to answers which are at least partially correct (i.e., $Acc(x) > 0$) – an important property of the VQA accuracy, where answers can have varying levels of correctness based on the number of human annotators in agreement. The choice of c depends on the deployment-specific cost of providing an incorrect answer. In Sec. 5.3, we report Φ_c with cost values of 1, 10, and 100. While [17] suggests setting $\Phi_c(x) < 0$ for $g(x) = 0$, we set $\Phi_c(x) = 0$ (i.e., a score of 0 when abstaining). This enables our formulation to have the clear upper bound for models which abstain perfectly:

Lemma 1. *The Effective Reliability score is equal to the VQA Accuracy ($\Phi_c(x) = Acc(x)$) if a model abstains ($g(x) = 0$)* iff *it is incorrect ($Acc(x) = 0$).*

We provide a simple proof for this in Appendix B. It is also confirmed in our experiments in Tab. 3.

We choose an abstention threshold which optimizes Φ_c on a validation set to compute a model's Effective Reliability with the form of the selection function g defined in Eq. 4. Additionally, the Effective Reliability score Φ_c can be evaluated for any model, even those which do not incorporate the option to abstain from providing a prediction (i.e., $g(x)$ is always 1).

4 Selection Functions

We investigate three promising directions to extend VQA models to abstain by exploring different options for $g'(x)$ introduced in Sec. 3.1. The first is to use the maximum softmax probability a model produces, the second is calibration of the softmax probabilities to better reflect the likelihood of correctness, and the last is training a specific selector. In general, a good function $g'(x)$ for abstention should yield high values when $f(x)$ is correct and low values when it is incorrect.

MaxProb. Without any additional training, a model can be extended to abstain by defining g' as the softmax probability of the model's predicted class (i.e., maximum probability) and is thus refered to as MaxProb [34, 42, 50]. Essentially, MaxProb trusts that if the model gives a high probability to one class, it is quite certain that the answer is correct and should be given: $g'_{\text{MaxProb}}(x) = \max(f'(x))$, where $f'(x)$ represents the answer probabilities.

Calibration. A model's confidence score for an output often does not correlate well with the actual correctness/accuracy [29]. Calibration [64] aims to ensure that the predicted probability for an output is representative of the likelihood of that output being correct. These techniques tune the absolute confidence values [64], while selective prediction has more to do with the relative confidence rankings [21]. However, this still offers a useful comparison as a poorly calibrated model may also imply poor confidence rankings [42]. Temperature scaling [29, 64] is a popular calibration method, but optimizing the single softmax temperature parameter makes it a monotonic function, which does not change the relative confidence rankings between examples and has no effect on the risk-coverage curve. That is, if a sample has a lower confidence score than another sample, it will still have a lower

confidence score after temperature scaling. Consequently, we do not consider it in this work. We instead consider a variant called vector scaling [29, 64], which applies the linear transformation $\mathbf{W}\mathbf{z} + \mathbf{b}$ to the logits $\mathbf{z}$, where $\mathbf{W}$ is a diagonal matrix, essentially having a temperature parameter and a bias term for each class.

Our implementation of vector scaling does not exactly follow the implementation detailed in [29]. The vector scaling detailed in [29] assumes a multi-class classification problem where the labels are one-hot vectors. They use a softmax activation after the last layer followed by a classic negative-log-likelihood (NLL) loss, which cannot effectively utilize the VQA metric's soft labels. We replace this with a sigmoid layer followed by a binary cross-entropy loss on all of the classes, which is more suited towards the VQA setting. We then apply MaxProb on top of these calibrated logits. In Sec. 5.4 we further evaluate how well the models are calibrated and compare the old and new loss functions.

Multimodal selection function: Selector. Vector scaling essentially trains an additional component on top of the VQA model to refine the model confidence. We move beyond this by training a component (Selector) to predict whether the answer is correct [20, 42, 64]. Different from prior work on confidence estimation in other tasks [20, 25, 42, 80], the multimodal nature of VQA presents unique challenges where the model must consider the interaction between the image, question, and answer. For instance, in Fig. 4 (right) (i.e., answering the question "*What is he doing?*"), a model needs to localize the subject "*he*" and understand the visual semantics of the scene. To model this, we extract the image v, question q, multimodal r, and answer representations $f'(x)$ from the VQA model and input these to the Selector, which gives it access to representations of both the answer itself as well as the evidence on which the answer is based. The Selector is a multi-layered perceptron that takes these representations as input and predicts the correctness of an answer with respect to the image-question pair. To train this component, the simplest method may be to assign binary labels to answers that are correct/incorrect and treat this as a classification problem. However, this setup does not account for answers that may be partially correct, or where one answer may be more correct than another. Since the relative confidence rankings are important for selective prediction [21, 42], losing this information in our training objective may hinder the performance. Therefore, we propose to treat this correctness prediction as a regression task where the target value is the VQA accuracy, allowing us to scale confidence scores with correctness.

Additional implementation details on the selection functions can be found in Appendix D.2.

5 Experiments

5.1 Data and Models

We experiment on the VQA v2 [26] dataset and require annotations for evaluation. As annotations for the test-dev and test-std sets of VQA v2 are not publicly available, we use questions from the official validation split for our evaluation as is common [1, 70]. As a reminder, under our selective prediction setup, the VQA model represents the function f, the selection function is g, and the composition of the two form a selective model h. We train the VQA models (f) on the training set of VQA v2. Meanwhile, we split the 214k examples in the VQA v2 validation set into three subsets: a split with 86k examples (40%) for validating VQA models (f) as well as training selection functions (g), another with 22k examples (10%) for validating the selection functions, and a held out test split of 106k examples (50%) that we use strictly for evaluating the full models (h). We employ the selection functions introduced in Sec. 4 and benchmark them in combination with a range of VQA models with varying architectures and performance:

Pythia [39]: A previous state-of-the-art model that won the 2018 VQA challenge and is an optimization of the widely used bottom-up top-down (BUTD) VQA model [2], which utilizes object detection features from a FasterRCNN [68] trained on Visual Genome [47]. Pythia achieves a VQA v2 test-std accuracy of 70.24% [39].

ViLBERT [55]: A two-stream vision-and-language transformer model [9, 76] that also uses object detection features. It is pre-trained with masked image modeling as well as masked language modeling objectives, following [74], and then fine-tuned on the VQA v2 training data. ViLBERT has a VQA v2 test-std accuracy of 70.92% [55].

VisualBERT [51]: This model is a single-stream transformer architecture, like BERT [19], that gets 71.00% VQA v2 test-std accuracy [51]. This also utilizes object detection features to represent input images and is pre-trained with a masked lanugage modeling objective, following [74].

CLIP-ViL [72]: This represents a state-of-the-art model that is trained from scratch on the VQA data whose visual encoder is from the CLIP model [66]. The VQA architecture, MoVie+MCAN [61], is an ensemble of a transformer encoder-decoder [83] and modulated convolutional [61] model, which won the 2020 VQA challenge. This enhanced version with features from CLIP attains 74.17% accuracy on VQA v2 test-std [72].

Details of data, model training, and hyperparameters are in Appendix C and Appendix D.

5.2 Benchmarking Risk and Coverage

As discussed in Sec. 3.2, we first measure the maximum coverage for a given risk ($\mathcal{C}@\mathcal{R}$) as well as AUC under the risk-coverage curves. We also include the best possible performance on these metrics for each model, which would be a selective model that abstains if and only if the answer is predicted incorrectly. Lastly, we present the overall accuracy on our test split for each model. Note that MaxProb and the Selector do not change the accuracy, but Calibration may.

Selector outperforms other methods. From Tab. 1, we see that adding the Selector consistently outperforms MaxProb in coverage for all risk tolerances as well as AUC. The strongest improvements occur at lower risk tolerances (e.g., 1% and 5%), becoming smaller as the tolerance increases (e.g., 10% and 20%). Notably, CLIP-ViL with Selector can improve $\mathcal{C}@1\%$ to $3\times$ that of CLIP-ViL with MaxProb. Fig. 2 illustrates how, for low risk levels, the addition of the selector maintains noticeably better risk as coverage increases compared to MaxProb. Further, it generally appears that the more accurate a model is overall, the more it may potentially improve in coverage at low risk tolerances when using Selector. For instance, when adding the Selector, we observe the largest improvements in $\mathcal{C}@1\%$ and $\mathcal{C}@5\%$ with CLIP-ViL (11.75% and 4.89%, respectively), which also has the highest accuracy. Meanwhile, Pythia has the lowest accuracy and exhibits the smallest improvements with the Selector at these tolerances (3.14% and 2.53%, respectively). Fig. 2 depicts this between 0-5% risk, where the gap between MaxProb and Selector appears to widen as we move to more accurate models (left to right). Lastly, we observe that Calibration can improve coverage beyond MaxProb as well, but largely less so than the Selector, especially at low risk tolerances (e.g., 1%, 5%), and not as consistently. Because Calibration modifies the output logits, it also slightly changes model accuracy.

Better accuracy $\not\Rightarrow$ better coverage at low risk. While accuracy appears to positively correlate with a better risk-coverage trade-off, the results in Tab. 1 also imply that higher accuracy does not guarantee better coverage at low risk tolerances. For example, CLIP-ViL has 2.54% higher accuracy than ViLBERT, but, with default MaxProb, ViLBERT has 2.52% higher $\mathcal{C}@1\%$ than CLIP-ViL. These findings imply that improving upon this risk-coverage trade-off requires not only building more accurate models but also learning better abstention policies.

Still room for improvement. Though the evidence presented in Tab. 1 and Fig. 2 show that coverage at different risk tolerances can be improved, these approaches still fall short of the best possible. For example, in Tab. 1, the difference in $\mathcal{C}@1\%$ between each model with Selector and their respective best possibles is still $>$50%. Although achieving the best possible may not be realistic, more work is needed to have reliable models with high accuracy and wide coverage that shrink this gap further.

Thresholds generalize to test time. Thus far, we have evaluated the maximum coverage at an exact risk level, given a selective model. In practice, however, the threshold γ used at test time must be chosen on a validation set. Here, we evaluate how close the actual test-time risk is to the target risk when using the validation threshold ($\Delta\mathcal{R}$). Experimenting with VisualBERT, comparing MaxProb and Selector, we see in Tab. 2 that the differences in risk for both selection functions tend to be at most 0.25%. Likewise, we observe corresponding differences in achieved coverage between the validation threshold and the maximum coverage ($\Delta\mathcal{C}$). These relatively small differences show that these thresholds generalize reasonably well, which aligns with previous findings on other tasks [25]. However, because the actual risk levels are now slightly different between models, we can no longer compare the corresponding coverage levels directly. This motivates the use of Effective Reliability, where we can compare models based on a predefined cost for wrong answers as opposed to an exact risk level.

Model	g	Acc. ↑	$\mathcal{C}@\mathcal{R}$ in % ↑				AUC ↓
			$\mathcal{R} = 1\%$	$\mathcal{R} = 5\%$	$\mathcal{R} = 10\%$	$\mathcal{R} = 20\%$	in %
Pythia [39]	MaxProb	66.20	6.13	24.71	41.59	71.48	13.83
	Calibration	66.46	6.79	25.28	42.50	**73.64**	13.46
	Selector	66.20	**9.27**	**27.24**	**43.28**	73.43	**13.26**
	Best Possible ($\mathcal{C}$)	66.20	62.72	68.45	73.55	82.75	6.67
ViLBERT [55]	MaxProb	69.20	7.51	29.01	47.99	79.89	11.78
	Calibration	69.16	10.07	30.15	48.75	79.96	11.62
	Selector	69.20	**11.82**	**32.44**	**50.20**	**79.97**	**11.31**
	Best Possible ($\mathcal{C}$)	69.20	65.66	71.67	76.89	86.50	5.49
VisualBERT [51]	MaxProb	70.18	6.85	30.78	50.46	81.78	11.21
	Calibration	70.02	9.78	32.09	51.14	81.92	11.21
	Selector	70.18	**11.47**	**34.14**	**52.53**	**82.04**	**10.75**
	Best Possible ($\mathcal{C}$)	70.18	66.70	72.76	77.98	87.73	5.13
CLIP-ViL [72]	MaxProb	71.74	4.99	34.45	55.34	85.01	10.33
	Calibration	71.57	12.72	37.46	56.08	84.97	9.95
	Selector	71.74	**16.74**	**39.34**	**58.02**	**85.27**	**9.53**
	Best Possible ($\mathcal{C}$)	71.74	68.48	74.53	79.71	89.67	4.59

Table 1: Risk-coverage metrics for different selection functions. For coverage at risk ($\mathcal{C}@\mathcal{R}$) and VQA accuracy (Acc.), higher is better. For AUC, lower is better.

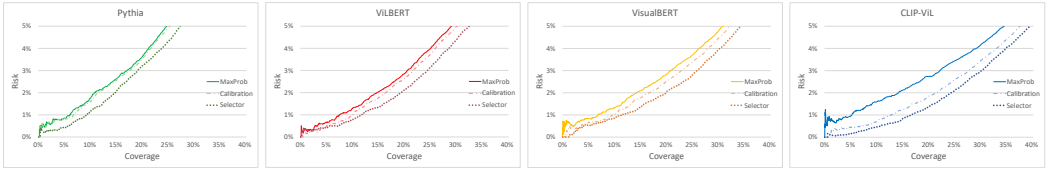

Figure 2: Risk-coverage plots for each model up to 5% risk.

5.3 Effective Reliability

We evaluate Effective Reliability (Φ_c) defined in Sec. 3.1, which assigns a cost to incorrect predictions, a reward to correct predictions, and zero to questions on which a model abstained from answering. This provides a single measure to jointly consider reliability (i.e., low risk) and effectiveness (i.e., high coverage). In Tab. 3, we choose cost values c set to 1, 10, or 100, to observe how models compare when the consequences for providing an incorrect prediction become high. Additionally, we can now directly compare to the original VQA formulation, where models do not have an option to abstain, denoted by a null selection function g. We also include Φ_c for the best possible g, where a model abstains exactly on those inputs which would result in incorrect predictions. As discussed in Sec. 3.1, this is equivalent to the model accuracy. Results are reported on the test set, with an abstention threshold selected to optimize Φ_c on the validation set. We include the corresponding risk and coverage for the selected threshold.

Selector still outperforms other methods. The Selector produces the highest effective reliability scores across all models and cost levels. As the penalty for wrong answers increases, the gap between the performance of Selector and the next best model generally increases as well. For example, the improvement of Selector over MaxProb for ViLBERT is 0.24 at a cost of 1, yet it is 3.74 at a cost of 100. Further, the gap between Selector and MaxProb at a cost of 100 increases as the VQA model itself has higher accuracy (or best possible performance). We observe a similar effect in Fig. 2, where more accurate models have larger gaps in risk between Selector and MaxProb at a given coverage.

Cost implicitly controls risk and coverage. When the penalty for a wrong answer is high, one might expect a selective model to operate in the low-risk regime. This is indeed reflected in Tab. 3, where the range of risk levels for selective models at $c = 100$ ($\Phi_c \approx$ 0.6-1.3) is much lower than the range of risk at $c = 1$ ($\Phi_c \approx$ 17-22). This directly translates to a similar trend in coverage, where selective models answer about 4–13% of questions at $c = 100$, and about 74–82% of questions at $c = 1$. This shows that effective reliability behaves intuitively around the influence of a user-selected cost on model risk and coverage.

Model	$\mathcal{R}=$	$\Delta\mathcal{R}$ 1%	5%	10%	20%	$\Delta\mathcal{C}$ 1%	5%	10%	20%
MaxProb		+0.12	−0.14	+0.17	−0.09	+0.92	−0.55	+0.81	−0.20
Selector		+0.14	+0.25	+0.17	−0.23	+2.00	+1.09	+0.59	−0.49

Table 2: Generalization of abstention thresholds γ from validation to test, with VisualBERT [51]. $\Delta\mathcal{R}$ and $\Delta\mathcal{C}$ are the differences in risk and coverage, respectively, when using γ selected for the target risk $\mathcal{R}$ on validation vs. γ with maximum $\mathcal{C}@\mathcal{R}$.

Model	g	c=1 $\Phi_c\uparrow$	$\mathcal{R}\downarrow$	$\mathcal{C}\uparrow$	c=10 $\Phi_c\uparrow$	$\mathcal{R}\downarrow$	$\mathcal{C}\uparrow$	c=100 $\Phi_c\uparrow$	$\mathcal{R}\downarrow$	$\mathcal{C}\uparrow$
Pythia [39]	—	38.53	33.80	100	−210.4	33.80	100	−2699.61	33.80	100
	MaxProb	47.39	21.72	76.40	15.08	5.13	25.17	1.81	0.77	4.00
	Calibration	48.16	20.90	75.62	15.53	4.92	25.05	2.14	0.80	5.09
	Selector	**48.21**	20.29	74.04	**17.55**	5.02	27.32	**4.12**	0.72	7.18
	Best Possible (Φ_c)	66.20	8.49	72.34	66.20	8.49	72.34	66.20	8.49	72.34
ViLBERT [55]	—	44.57	30.80	100	−177.05	30.80	100	−2393.23	30.80	100
	MaxProb	52.41	20.01	79.92	18.00	6.26	34.50	1.67	1.33	10.18
	Calibration	52.51	19.53	78.93	18.29	6.10	34.24	2.92	1.12	10.47
	Selector	**52.65**	19.37	78.60	**21.02**	5.56	34.57	**5.41**	0.90	11.06
	Best Possible (Φ_c)	69.20	8.20	75.38	69.20	8.20	75.38	69.20	8.20	75.38
VisualBERT [51]	—	46.49	29.82	100	−166.77	29.82	100	−2299.33	29.82	100
	MaxProb	53.72	19.09	79.83	19.29	5.63	33.64	2.49	1.02	6.89
	Calibration	53.80	19.07	79.84	19.96	5.57	34.37	3.83	0.87	8.42
	Selector	**54.12**	18.72	79.34	**22.04**	5.13	34.61	**4.82**	1.00	11.34
	Best Possible (Φ_c)	70.18	8.02	76.30	70.18	8.02	76.30	70.18	8.02	70.18
CLIP-ViL [72]	—	49.38	28.26	100	−151.87	28.26	100	−2164.33	28.26	100
	MaxProb	55.72	18.80	82.41	21.45	5.86	38.30	1.82	1.26	7.26
	Calibration	55.66	17.28	79.00	23.20	3.92	32.59	5.78	0.86	11.62
	Selector	**56.43**	17.19	79.52	**25.98**	5.25	40.45	**8.76**	0.65	12.26
	Best Possible (Φ_c)	71.74	7.60	77.64	71.74	7.60	77.64	71.74	7.60	77.64

Table 3: Effective Reliability Φ_c for VQA models with and without abstention options. The best possible Φ_c is computed by only selecting correct predictions, and is equal to the model's VQA accuracy.

Human evaluation shows noise has little effect even with high cost values. For high costs (e.g., $c = 100$), models are strongly penalized for producing incorrect predictions. Given these strict penalties on errors, it becomes pertinent to ask to what degree noise in the annotations might be contributing to these penalties, though the potential impact of noise is certainly not unique to our evaluations and is a challenging problem in VQA [6, 41, 71]. To check if our results are significantly affected by annotation noise when setting $c = 100$, we manually examine each sample where the model predictions were counted as incorrect (and thus heavily penalized when computing Φ_c). We annotate cases where each model may have been unfairly penalized and recompute Φ_c when removing this penalty (see Tab. 9). We find that vast majority of incorrect predictions that contribute to these penalties are properly marked as incorrect. We also see that label noise does slightly change the Effective Reliability scores at high cost, but the rankings between models and selection functions are preserved. More details are in Appendix E.

All models without an abstention option perform poorly. When the cost of a wrong answer is equal to the reward of getting an answer entirely correct ($c = 1$), all models without a selection function g underperform their selective model counterparts. As c increases, this gap widens dramatically, with non-abstaining models reaching Φ_c values firmly in the negative range. Meanwhile, all selective models reach a positive Φ_c, even at high cost, illustrating the necessity of the abstention option for building models which are reliable and effective.

5.4 Effect of Model Calibration

We report the calibration performance of the vector scaling, with both the original NLL loss and the new loss adapated to the VQA setting, as mentioned in Sec. 4. Specifically, we measure the expected calibration error (ECE) [60, 29], which represents the expected difference between the model confidence and accuracy, $\mathbb{E}_{\hat{P}}[|\mathbb{P}(\hat{Y} = Y|\hat{P} = p) - p|]$. In practice, ECE is calculated by dividing the model confidences into M bins and calculating the difference between the average confidence and answer accuracy in each bin, weighted by the number of samples in each bin. The

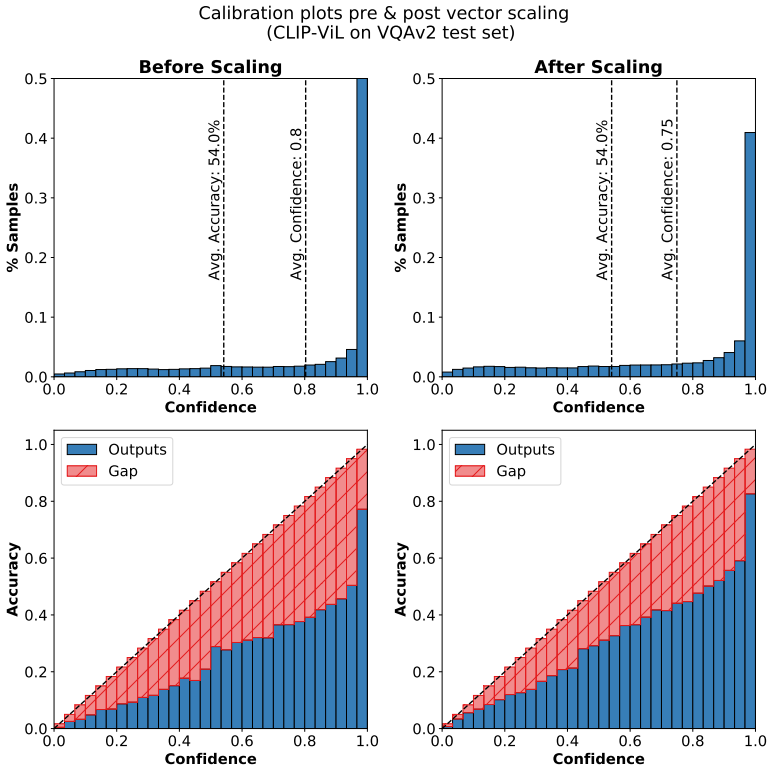

Figure 3: Top: % samples per confidence bin for CLIP-ViL; Bottom: average accuracy per confidence bin for CLIP-ViL

lower the ECE, the more that the model's confidence scores correspond to the actual accuracy of the predictions. Note that the ECE metric is designed for single label classification problems. To use the ECE metric for VQA, where there can be multiple possible answers for a question, we simply consider the most frequent human annotated answer as the ground truth for each question. Tab. 4 summarizes the ECE of the four models used before and after calibration, with the original NLL loss and the new loss tailored towards the VQA setting. As seen, the ECE using the original loss results in a better calibrated model, but the new loss still results in a well-calibrated model. This is unsurprising, as the ECE metric is also tailored towards a multi-class classification setting and not VQA, like the original loss function. However, we chose to use the original formulation of the ECE metric for consistency. We still prefer the new VQA loss because the adaptations result in better performance in our task of choosing whether to abstain or answer. In Fig. 3 is shown a visual of the calibration for CLIP-ViL with the VQA calibration loss.

In Tab. 5, we also compare the two calibration losses' effect on $\mathcal{C}@\mathcal{R}$ and AUC. Although the original NLL loss still beats MaxProb, our loss tailored towards the VQA soft labels performs better.

Model	ECE (MaxProb)	ECE (NLL loss)	ECE (VQA loss)
Pythia	0.135	**0.0524**	0.0777
ViLBERT	0.146	**0.0880**	0.104
VisualBERT	0.146	**0.0807**	0.105
CLIP-ViL	0.198	**0.105**	0.146

Table 4: ECE of models, before and after vector scaling (lower is better)

Model	g	$\mathcal{C}@\mathcal{R}$ in % ↑ $\mathcal{R}=1\%$	$\mathcal{R}=5\%$	$\mathcal{R}=10\%$	$\mathcal{R}=20\%$	AUC ↓
Pythia [39]	MaxProb	6.13	24.71	41.59	71.48	13.83
	Calibration (NLL Loss)	**7.12**	24.71	41.74	72.94	13.67
	Calibration (VQA Loss)	6.79	**25.29**	**42.50**	**73.64**	**13.46**
ViLBERT [55]	MaxProb	7.51	29.01	47.99	79.89	11.78
	Calibration (NLL Loss)	**10.12**	**30.77**	48.60	79.88	**11.60**
	Calibration (VQA Loss)	10.07	30.15	**48.75**	**79.96**	11.62
VisualBERT [51]	MaxProb	6.85	30.78	50.46	81.78	11.21
	Calibration (NLL Loss)	9.60	**33.11**	**51.25**	81.80	**10.97**
	Calibration (VQA Loss)	**9.78**	32.09	51.14	**81.92**	11.21
CLIP-ViL [72]	MaxProb	4.99	34.45	55.34	**85.01**	10.33
	Calibration (NLL Loss)	7.54	37.18	55.92	84.79	10.05
	Calibration (VQA Loss)	**12.72**	**37.46**	**56.08**	84.97	**9.95**

Table 5: Risk-coverage metrics for calibrator with NLL loss and the new VQA loss. For coverage at risk ($\mathcal{C}@\mathcal{R}$), higher is better. For AUC, lower is better.

Features	Unimodal	Loss	$\mathcal{C}@\mathcal{R}$ in % ↑ $\mathcal{R}=1\%$	$\mathcal{R}=5\%$	$\mathcal{R}=10\%$	$\mathcal{R}=20\%$	AUC ↓
$\tilde{v}$	✓	Regression	0.04	0.04	0.04	16.09	23.23
q	✓	Regression	0.02	11.03	35.88	79.70	13.39
$f'(x)$		Regression	5.24	36.10	56.30	84.79	10.08
v		Regression	11.60	36.43	53.74	83.51	10.32
r		Regression	**13.42**	34.69	53.90	82.95	10.43
$f'(x)$+$\tilde{v}$		Regression	3.67	36.40	56.33	84.79	10.07
$f'(x)$+q		Regression	10.67	37.41	56.95	84.76	9.86
$f'(x)$+r		Regression	12.02	37.44	<u>57.68</u>	<u>84.93</u>	9.81
$f'(x)$+v		Regression	13.24	**38.51**	57.44	84.92	<u>9.76</u>
$f'(x)$+q+v+r		Classification	6.64	35.80	57.29	84.18	10.06
$f'(x)$+q+v+r		Regression	<u>13.32</u>	<u>38.02</u>	**58.16**	**85.03**	**9.73**

Table 6: Ablations of Selector with CLIP-ViL [72] on the selection function validation set using the same setup as Tab. 1. The overall best performance is in bold and second best is underlined. $f'(x)$, q, $\tilde{v}$, and r are the answer, question, image, and multimodal representations, respectively. Note, v is a question conditioned image representation that is not unimodal (see Appendix A for details).

5.5 Selection Function Ablations

Tab. 6 provides ablations for the selection function design. In the following, we distill the main observations, additional discussion can be found in Appendix A.

Selector requires multimodal input. The results in Tab. 6 show the importance of using multimodal information for coverage at low risk levels. When using each representation in isolation, we see that multimodal representations (r, v, and $f'(x)$) yield much stronger $\mathcal{C}@1\%$ and $\mathcal{C}@5\%$ than unimodal representations (image $\tilde{v}$ or question q). For highly reliable models ($\mathcal{C}@1\%$), unimodal selection functions fail (coverage $<0.05\%$), suggesting that building reliable and effective VQA models is a truly multimodal problem. Using the combination of all features is the best or close to the best according to Tab. 6, so we use this setup in all experiments.

Regressing to VQA accuracy is important. We find that formulating the objective as a regression of the answer accuracy, rather than classifying whether the answer is correct, offers very significant improvements (Tab. 6), especially at low risk. This is likely because predicting the fine-grained accuracy allows the model to account for partially correct answers and learn to rank answers that are more correct higher, as opposed the classification setup where the distinction between partially correct answers is lost.

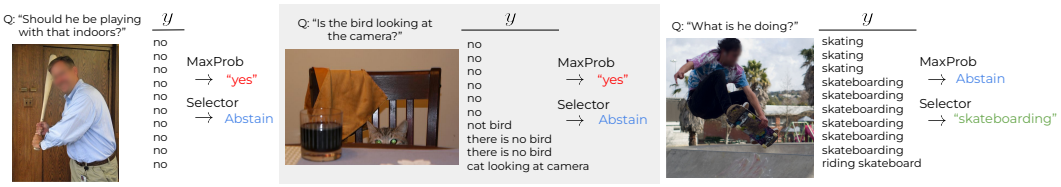

Figure 4: Qualitative test set examples with CLIP-ViL selective model predictions, when optimized for Φ_c at $c = 100$ on validation.

5.6 Qualitative Analysis

In Fig. 4, we visualize MaxProb and Selector decisions with CLIP-ViL for several examples on the test set (see appendix for more). The abstention threshold is chosen to maximize Φ_c for $c = 100$ on validation. Fig. 4 (left) shows an example of a question that requires commonsense reasoning to answer that the VQA model may not be certain of (and gets wrong), so Selector abstains. Similarly, in Fig. 4 (middle), we see a false premise question [67] where Selector abstains again as the question does not make sense for the image, while MaxProb yields an incorrect answer. Meanwhile, Fig. 4 (right) presents an example where the model is correct yet MaxProb chose to abstain and Selector chose to answer. Here, there are a number of synonymous answers, but, in the classification setup of the VQA model, they are simply distinct classes so the confidence may reflect this ambiguity. Whereas, Selector may be able to adjust the confidence to yield an answer as it may have learned the similarity/co-occurrence of these during training and to predict the correctness accordingly. These examples contribute to the higher coverage at low risk observed quantitatively in our experiments.

6 Conclusion

The standard VQA problem formulation does not include an option for models to abstain from answering if they are uncertain. However, for many applications, it is important that the model only provides an answer if there is a low risk of error. In this work, we promote a problem formulation for VQA which includes an option to abstain and discuss how to evaluate this, including a metric that rewards correct predictions but expects models to abstain if they are incorrect. We benchmark several VQA models in combination with approaches for answer abstention. If we want a reliable model with 1% risk of error, we find that a state-of-the-art VQA model [72] only answers less than 5% of the questions when using its softmax probabilities as estimates of model confidence. Using calibration can improve this, especially if the calibration is further adapted to the VQA setting. We find that the best results are consistently achieved by training a multimodal selection function to estimate performance directly. In contrast to standard VQA, where a unimodal model can get you a long way, unimodal selection functions have little utility (<0.05% coverage at 1% risk), it is thus critical to include multimodal information. Furthermore, the best selection function is obtained by *regressing* to the VQA Accuracy in its optimization. This increases the coverage from 4.99% to 16.74%. While this is an marked improvement, one has to consider that this state-of-the-art model achieves 71.74% standard VQA accuracy on the same set of data. With our *Effective Reliability* metric, the performance of this model drops from 71.74 (for perfect abstention) to 8.76 (our best abstention baseline) when penalizing wrong answers with a cost of 100. We believe this new framework and metric for VQA will encourage the community to build VQA models which are both reliable and effective. This will be an opportunity for many new exciting directions for future work, including incorporating ideas from out-of-distribution detection or joint optimization of the VQA model and selection function to improve self-awareness of models.

Acknowledgements

We would like to thank Anastasios Angelopoulos and Kurt Shuster for helpful discussions. Authors, as part of their affiliation with UC Berkeley, were supported in part by the NSF CISE Expeditions Award CCF-1730628; DoD, including DARPA's LwLL, and/or SemaFor programs; the Berkeley Artificial Intelligence Research (BAIR) industrial alliance program as well as gifts from Amazon

Web Services, Ant Group, Ericsson, Facebook, Futurewei, Google, Intel, Microsoft, Scotiabank, and VMware.

References

[1] Agrawal, A., Batra, D., Parikh, D., Kembhavi, A.: Don't just assume; look and answer: Overcoming priors for visual question answering. In: CVPR (2018)

[2] Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L.: Bottom-up and top-down attention for image captioning and visual question answering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6077–6086 (2018)

[3] Angelopoulos, A., Bates, S., Malik, J., Jordan, M.I.: Uncertainty sets for image classifiers using conformal prediction. In: ICLR (2021)

[4] Angelopoulos, A.N., Bates, S.: A gentle introduction to conformal prediction and distribution-free uncertainty quantification. arXiv preprint arXiv:2107.07511 (2021)

[5] Angelopoulos, A.N., Bates, S., Candès, E.J., Jordan, M.I., Lei, L.: Learn then test: Calibrating predictive algorithms to achieve risk control. arXiv preprint arXiv:2110.01052 (2021)

[6] Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Lawrence Zitnick, C., Parikh, D.: Vqa: Visual question answering. In: Proceedings of the IEEE international conference on computer vision. pp. 2425–2433 (2015)

[7] Asan, O., Bayrak, A.E., Choudhury, A., et al.: Artificial intelligence and human trust in healthcare: focus on clinicians. Journal of medical Internet research **22**(6), e15154 (2020)

[8] Bhattacharya, N., Li, Q., Gurari, D.: Why does a visual question have different answers? In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4271–4280 (2019)

[9] Cao, J., Gan, Z., Cheng, Y., Yu, L., Chen, Y.C., Liu, J.: Behind the scene: Revealing the secrets of pre-trained vision-and-language models. In: European Conference on Computer Vision. pp. 565–580. Springer (2020)

[10] Chen, X., Fang, H., Lin, T.Y., Vedantam, R., Gupta, S., Dollár, P., Zitnick, C.L.: Microsoft COCO captions: Data collection and evaluation server. arXiv preprint arXiv:1504.00325 (2015)

[11] Chen, Y.C., Li, L., Yu, L., El Kholy, A., Ahmed, F., Gan, Z., Cheng, Y., Liu, J.: UNITER: Universal image-text representation learning. In: ECCV. ECCV (2020)

[12] Chiu, T.Y., Zhao, Y., Gurari, D.: Assessing image quality issues for real-world problems. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3646–3656 (2020)

[13] Chow, C.: On optimum recognition error and reject tradeoff. IEEE Transactions on information theory **16**(1), 41–46 (1970)

[14] Chow, C.K.: An optimum character recognition system using decision functions. IRE Transactions on Electronic Computers **EC-6**(4), 247–254 (1957)

[15] Corbière, C., Thome, N., Bar-Hen, A., Cord, M., Pérez, P.: Addressing failure prediction by learning model confidence. Advances in Neural Information Processing Systems **32** (2019)

[16] Davis, E.: Unanswerable questions about images and texts. Frontiers in Artificial Intelligence **3**, 51 (2020)

[17] De Stefano, C., Sansone, C., Vento, M.: To reject or not to reject: that is the question-an answer in case of neural classifiers. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) **30**(1), 84–94 (2000). https://doi.org/10.1109/5326.827457

[18] Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. pp. 248–255. IEEE (2009)

[19] Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)

[20] Dong, L., Quirk, C., Lapata, M.: Confidence modeling for neural semantic parsing. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 743–753. Association for Computational Linguistics, Melbourne, Australia (Jul 2018). https://doi.org/10.18653/v1/P18-1069, https://aclanthology.org/P18-1069

[21] El-Yaniv, R., Wiener, Y.: On the foundations of noise-free selective classification. Journal of Machine Learning Research **11**, 1605–1641 (2010)

[22] Fukui, A., Park, D.H., Yang, D., Rohrbach, A., Darrell, T., Rohrbach, M.: Multimodal compact bilinear pooling for visual question answering and visual grounding. In: EMNLP (2016)

[23] Gao, P., Jiang, Z., You, H., Lu, P., Hoi, S.C., Wang, X., Li, H.: Dynamic fusion with intra-and inter-modality attention flow for visual question answering. In: CVPR (2019)

[24] Geifman, Y., El-Yaniv, R.: Selective classification for deep neural networks. Advances in neural information processing systems **30** (2017)

[25] Geifman, Y., El-Yaniv, R.: Selectivenet: A deep neural network with an integrated reject option. In: International Conference on Machine Learning. pp. 2151–2159. PMLR (2019)

[26] Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., Parikh, D.: Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6904–6913 (2017)

[27] Guillory, D., Shankar, V., Ebrahimi, S., Darrell, T., Schmidt, L.: Predicting with confidence on unseen distributions. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1134–1144 (2021)

[28] Gulshan, V., Peng, L., Coram, M., Stumpe, M.C., Wu, D., Narayanaswamy, A., Venugopalan, S., Widner, K., Madams, T., Cuadros, J., Kim, R., Raman, R., Nelson, P.C., Mega, J.L., Webster, D.R.: Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA **316**(22), 2402–2410 (12 2016). https://doi.org/10.1001/jama.2016.17216, https://doi.org/10.1001/jama.2016.17216

[29] Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: International Conference on Machine Learning. pp. 1321–1330. PMLR (2017)

[30] Gurari, D., Li, Q., Stangl, A.J., Guo, A., Lin, C., Grauman, K., Luo, J., Bigham, J.P.: Vizwiz grand challenge: Answering visual questions from blind people. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3608–3617 (2018)

[31] Hanczar, B., Dougherty, E.R.: Classification with reject option in gene expression data. Bioinformatics **24**(17), 1889–1895 (2008)

[32] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)

[33] Hendricks, L.A., Burns, K., Saenko, K., Darrell, T., Rohrbach, A.: Women also snowboard: Overcoming bias in captioning models. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 771–787 (2018)

[34] Hendrycks, D., Gimpel, K.: A baseline for detecting misclassified and out-of-distribution examples in neural networks. In: Proceedings of International Conference on Learning Representations (2017)

[35] Hudson, D.A., Manning, C.D.: 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. pp. 6700–6709 (2019)

[36] Ji, X., Pascanu, R., Hjelm, D., Lakshminarayanan, B., Vedaldi, A.: Test sample accuracy scales with training sample density in neural networks. arXiv preprint arXiv:2106.08365 (2021)

[37] Jiang, H., Kim, B., Guan, M., Gupta, M.: To trust or not to trust a classifier. In: Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 31. Curran Associates, Inc. (2018), https://proceedings.neurips.cc/paper/2018/file/7180cffd6a8e829dacfc2a31b3f72ece-Paper.pdf

[38] Jiang, H., Misra, I., Rohrbach, M., Learned-Miller, E., Chen, X.: In defense of grid features for visual question answering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10267–10276 (2020)

[39] Jiang, Y., Natarajan, V., Chen, X., Rohrbach, M., Batra, D., Parikh, D.: Pythia v0. 1: the winning entry to the vqa challenge 2018. arXiv preprint arXiv:1807.09956 (2018)

[40] Kafle, K., Kanan, C.: An analysis of visual question answering algorithms. In: ICCV (2017)

[41] Kafle, K., Kanan, C.: Visual question answering: Datasets, algorithms, and future challenges. Computer Vision and Image Understanding **163**, 3–20 (2017)

[42] Kamath, A., Jia, R., Liang, P.: Selective question answering under domain shift. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 5684–5696. Association for Computational Linguistics, Online (Jul 2020). https://doi.org/10.18653/v1/2020.acl-main.503, https://aclanthology.org/2020.acl-main.503

[43] Karamcheti, S., Krishna, R., Fei-Fei, L., Manning, C.: Mind your outliers! investigating the negative impact of outliers on active learning for visual question answering. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). pp. 7265–7281. Association for Computational Linguistics, Online (Aug 2021). https://doi.org/10.18653/v1/2021.acl-long.564, https://aclanthology.org/2021.acl-long.564

[44] Khan, J., Wei, J.S., Ringner, M., Saal, L.H., Ladanyi, M., Westermann, F., Berthold, F., Schwab, M., Antonescu, C.R., Peterson, C., et al.: Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nature medicine **7**(6), 673–679 (2001)

[45] Khani, F., Rinard, M., Liang, P.: Unanimous prediction for 100% precision with application to learning semantic mappings. arXiv preprint arXiv:1606.06368 (2016)

[46] Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: Proceedings of the International Conference on Learning Representations (2015)

[47] Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., et al.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. International journal of computer vision **123**(1), 32–73 (2017)

[48] Kull, M., Filho, T.M.S., Flach, P.: Beyond sigmoids: How to obtain well-calibrated probabilities from binary classifiers with beta calibration. Electronic Journal of Statistics **11**(2), 5052 – 5080 (2017). https://doi.org/10.1214/17-EJS1338SI, https://doi.org/10.1214/17-EJS1338SI

[49] Lakshminarayanan, B., Pritzel, A., Blundell, C.: Simple and scalable predictive uncertainty estimation using deep ensembles. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 30. Curran Associates, Inc. (2017), https://proceedings.neurips.cc/paper/2017/file/9ef2ed4b7fd2c810847ffa5fa85bce38-Paper.pdf

[50] Lakshminarayanan, B., Pritzel, A., Blundell, C.: Simple and scalable predictive uncertainty estimation using deep ensembles. In: Advances in neural information processing systems. vol. 30 (2017)

[51] Li, L.H., Yatskar, M., Yin, D., Hsieh, C.J., Chang, K.W.: Visualbert: A simple and performant baseline for vision and language. In: Arxiv (2019)

[52] Li, M., Weber, C., Wermter, S.: Neural networks for detecting irrelevant questions during visual question answering. In: International Conference on Artificial Neural Networks. pp. 786–797. Springer (2020)

[53] Li, X., Yin, X., Li, C., Zhang, P., Hu, X., Zhang, L., Wang, L., Hu, H., Dong, L., Wei, F., et al.: Oscar: Object-semantics aligned pre-training for vision-language tasks. In: European Conference on Computer Vision. pp. 121–137. Springer (2020)

[54] Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)

[55] Lu, J., Batra, D., Parikh, D., Lee, S.: Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. Advances in neural information processing systems **32** (2019)

[56] Lütkenhöner, B., Basel, T.: Predictive modeling for diagnostic tests with high specificity, but low sensitivity: a study of the glycerol test in patients with suspected meniere's disease. PLoS One **8**(11), e79315 (2013)

[57] Mahendru, A., Prabhu, V., Mohapatra, A., Batra, D., Lee, S.: The promise of premise: Harnessing question premises in visual question answering. In: EMNLP (2017)

[58] Mcknight, D.H., Carter, M., Thatcher, J.B., Clay, P.F.: Trust in a specific technology: An investigation of its components and measures. ACM Transactions Management Information Systems **2**(2) (jul 2011). https://doi.org/10.1145/1985347.1985353, https://doi.org/10.1145/1985347.1985353

[59] Mukhoti, J., Kulharia, V., Sanyal, A., Golodetz, S., Torr, P., Dokania, P.: Calibrating deep neural networks using focal loss. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (eds.) Advances in Neural Information Processing Systems. vol. 33, pp. 15288–15299. Curran Associates, Inc. (2020), https://proceedings.neurips.cc/paper/2020/file/aeb7b30ef1d024a76f21a1d40e30c302-Paper.pdf

[60] Naeini, M.P., Cooper, G., Hauskrecht, M.: Obtaining well calibrated probabilities using bayesian binning. In: Twenty-Ninth AAAI Conference on Artificial Intelligence (2015)

[61] Nguyen, D.K., Goswami, V., Chen, X.: Movie: Revisiting modulated convolutions for visual counting and beyond. In: Proceedings of the International Conference on Learning Representations (2021)

[62] Niculescu-Mizil, A., Caruana, R.: Predicting good probabilities with supervised learning. In: Proceedings of the 22nd international conference on Machine learning. pp. 625–632 (2005)

[63] Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). pp. 1532–1543 (2014)

[64] Platt, J., et al.: Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. Advances in large margin classifiers **10**(3), 61–74 (1999)

[65] Pudil, P., Novovicova, J., Blaha, S., Kittler, J.: Multistage pattern recognition with reject option. In: Proceedings., 11th IAPR International Conference on Pattern Recognition. Vol.II. Conference B: Pattern Recognition Methodology and Systems. pp. 92–95 (1992). https://doi.org/10.1109/ICPR.1992.201729

[66] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning. pp. 8748–8763. PMLR (2021)

[67] Ray, A., Christie, G., Bansal, M., Batra, D., Parikh, D.: Question relevance in vqa: Identifying non-visual and false-premise questions. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. pp. 919–924 (2016)

[68] Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems **28** (2015)

[69] Shafer, G., Vovk, V.: A tutorial on conformal prediction. Journal of Machine Learning Research **9**(3) (2008)

[70] Shah, M., Chen, X., Rohrbach, M., Parikh, D.: Cycle-consistency for robust visual question answering. In: CVPR (2019)

[71] Sharma, H., Jalal, A.S.: A survey of methods, datasets and evaluation metrics for visual question answering. Image and Vision Computing **116**, 104327 (2021)

[72] Shen, S., Li, L.H., Tan, H., Bansal, M., Rohrbach, A., Chang, K.W., Yao, Z., Keutzer, K.: How much can clip benefit vision-and-language tasks? arXiv preprint arXiv:2107.06383 (2021)

[73] Singh, A., Goswami, V., Natarajan, V., Jiang, Y., Chen, X., Shah, M., Rohrbach, M., Batra, D., Parikh, D.: Mmf: A multimodal framework for vision and language research. https://github.com/facebookresearch/mmf (2020)

[74] Singh, A., Goswami, V., Parikh, D.: Are we pretraining it right? digging deeper into visio-linguistic pretraining. arXiv preprint arXiv:2004.08744 (2020)

[75] Singh, A., Natarjan, V., Shah, M., Jiang, Y., Chen, X., Parikh, D., Rohrbach, M.: Towards vqa models that can read. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8317–8326 (2019)

[76] Tan, H., Bansal, M.: LXMERT: Learning cross-modality encoder representations from transformers. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. pp. 5100–5111 (2019). https://doi.org/10.18653/v1/D19-1514, https://www.aclweb.org/anthology/D19-1514

[77] Teney, D., Anderson, P., He, X., Van Den Hengel, A.: Tips and tricks for visual question answering: Learnings from the 2017 challenge. In: CVPR (2018)

[78] Thulasidasan, S., Chennupati, G., Bilmes, J.A., Bhattacharya, T., Michalak, S.: On mixup training: Improved calibration and predictive uncertainty for deep neural networks. In: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 32. Curran Associates, Inc. (2019), https://proceedings.neurips.cc/paper/2019/file/36ad8b5f42db492827016448975cc22d-Paper.pdf

[79] Vovk, V., Gammerman, A., Shafer, G.: Algorithmic learning in a random world. Springer Science & Business Media (2005)

[80] Wang, X., Luo, Y., Crankshaw, D., Tumanov, A., Yu, F., Gonzalez, J.E.: Idk cascades: Fast deep learning by learning not to overthink. arXiv preprint arXiv:1706.00885 (2017)

[81] Xin, J., Tang, R., Yu, Y., Lin, J.: The art of abstention: Selective prediction and error regularization for natural language processing. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). pp. 1040–1051 (2021)

[82] Yang, Z., He, X., Gao, J., Deng, L., Smola, A.: Stacked attention networks for image question answering. In: CVPR (2016)

[83] Yu, Z., Yu, J., Cui, Y., Tao, D., Tian, Q.: Deep modular co-attention networks for visual question answering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6281–6290 (2019)

[84] Zhang, P., Li, X., Hu, X., Yang, J., Zhang, L., Wang, L., Choi, Y., Gao, J.: Vinvl: Revisiting visual representations in vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5579–5588 (2021)

Appendix to Reliable Visual Question Answering: Abstain Rather Than Answering Incorrectly

Sec. A provides more discussion on Selector ablations.

Sec. B provides a proof of Lemma 1, providing a motivation for the definition of the Effective Reliability score Φ_c.

Sec. C provides additional details on the dataset splits.

Sec. D provides additional model details.

Sec. E provides a manual evaluation of the label noise.

Sec. F provides more qualitative results.

A Selector Design Ablations

Extending the discussion in Sec. 5.5, we are isolating the effects of different features/modalities on the risk-coverage trade-off when using Selector. In this direction, we experiment with different input representation variants from CLIP-ViL [72] in Tab. 6 by ablating the question q, multimodal r, and answer $f'(x)$ representations as well as different image representations. For image representations, we ablate the usage of the visual representation $\tilde{v}$ directly from the CLIP visual encoder [66], as well as the visual representation v that is the concatenation of the respective pooled outputs from MCAN's self-guided attention module [83] and MoVie's modulated convolutional bottleneck [61], which are visual representations that also contain multimodal information from the question. Question representations are taken from the output of MCAN's self-attention module. The multimodal representation is the concatentation of the multimodal representations that are used as inputs to the softmax output (i.e., classification) layer of CLIP-ViL. For the answer representation, we use the logits just before the softmax in the output layer. We measure the performance via maximum coverage (as in Tab. 1).

The results in Tab. 6 show the importance of using multimodal information for coverage at low risk levels. When comparing using each representation in isolation, we see that multimodal representations (r, v, and $f'(x)$) yield much stronger $\mathcal{C}$@1% and $\mathcal{C}$@5% than unimodal representations ($\tilde{v}$ and q). We also observe that the answer representation achieves the best performance for $\mathcal{C}$@10% and $\mathcal{C}$@20% when each input representation is used in isolation. Overall, we find that considering multimodal information (i.e., combinations of multimodal representations and unimodal representations from different modalities) to be most effective, with the top performers being the models that incorporate the answer representation alongside multimodal representations ($f'(x)$+r, $f'(x)$+v, and $f'(x)$+q+v+r).

B Proof of Lemma 1

Lemma 1 states that if a model abstains "perfectly", the introduced effective reliability score is equal to the VQA Accuracy. In this section we provide a proof of Lemma 1 in the main paper, which we repeat here for ease of understanding the proof:

Lemma 1. *The effective reliability score is equal to the VQA Accuracy ($\Phi_c(x) = Acc(x)$) if a model abstains ($g(x) = 0$)* iff *it is incorrect ($Acc(x) = 0$).*

Distilling this to the mathematical notation:

$$(g(x) = 0 \leftrightarrow Acc(x) = 0) \longrightarrow \Phi_c(x) = Acc(x) \tag{9}$$

Extending Eq. 8 to both cases, $Acc(x) = 0$ and $Acc(x) > 0$ (note, that Acc cannot be smaller than 0):

$$\Phi_c(x) = \begin{cases} Acc(x) & \text{if } g(x) = 1 \text{ and } Acc(x) > 0, \\ -c & \text{if } g(x) = 1 \text{ and } Acc(x) = 0, \\ 0 & \text{if } g(x) = 0 \text{ and } Acc(x) > 0, \\ 0 & \text{if } g(x) = 0 \text{ and } Acc(x) = 0. \end{cases} \tag{10}$$

To prove Lemma 1, we must show that the condition $(g(x) = 0 \leftrightarrow Acc(x) = 0)$ implies $\Phi_c(x) = Acc(x)$. The condition $(g(x) = 0 \leftrightarrow Acc(x) = 0)$ simplifies Eq. 10 as the second and third line contradict the condition:

$$\Phi_c(x) = \begin{cases} Acc(x) & \text{if } g(x) = 1 \text{ and } Acc(x) > 0, \\ 0 & \text{if } g(x) = 0 \text{ and } Acc(x) = 0. \end{cases} \tag{11}$$

As the $Acc(x) = 0$, the second line can be re-written as:

$$\Phi_c(x) = \begin{cases} Acc(x) & \text{if } g(x) = 1 \text{ and } Acc(x) > 0, \\ Acc(x) & \text{if } g(x) = 0 \text{ and } Acc(x) = 0. \end{cases} \tag{12}$$

Now, in both cases $\Phi_c(x) = Acc(x)$ □

C Additional Dataset Split Details

Source	Split Name	Usage	% src	#I	#Q	#A
VQA v2 train	Train	Train f	100%	82,783	443,757	4,437,570
VQA v2 val	Dev	Validate f / Train g	40%	16,202	86,138	861,380
	Val	Validate g	10%	4,050	21,878	218,780
	Test	Test h	50%	20,252	106,338	1,063,380

Table 7: Table of statistics for the dataset splits used for training as well as validating VQA models (f), training as well as validating selection functions (g), and testing full selective models ($h = (f, g)$). % src indicates the percentage of the source data (Source) that each split represents. #I, #Q, and #A indicate the number of images, questions, and answers, respectively.

We experiment on the VQA v2 dataset [26], which contains a large amount of human-annotated image-question-answer triplets. Tab. 7 lays out the data splits we use in our experiments. We create splits of the VQA v2 validation set since we require answer annotations to evaluate risk, coverage, and effective reliability. These splits are created such that no images (and therefore no question-answer annotations) are shared between them. Note that the data in the held out test set (Test in Tab. 7) is never seen during the training or validation of any component (f or g) and is only used for evaluations. All presented results are on our test set, except for the ablations, which are on our selection function validation set (Val in Tab. 7).

D Model Details

In this section, we present the details of the models we use in our experiments. We will release our implementations upon publication.

D.1 VQA Models

We use the open-source MMF framework [73] for all our experiments, which contains implementations of each VQA model.[3] For training VQA models, we follow the hyperparameters from MMF, which we list in Tab. 8. All models treat VQA as a classification task and are trained with VQA accuracy as soft target scores via a binary cross-entropy loss [77]. We briefly discuss input features and initializations used in our experiments, extending Sec. 5.1:

Pythia [39]: This model uses bottom-up top-down (BUTD) object detection features [2], but the features are extracted from a ResNext-152 based FasterRCNN [68]. Pythia's implementation further uses grid features from a ResNet-152 [32] as additional inputs to improve performance [39]. GloVe embeddings [63] are used to initialize the word representations. We train this model from scratch on the VQA v2 training data.

ViLBERT [55]: The same object detection features from Pythia are used, but without the addition of grid features. We use the pre-trained and fine-tuned model provided by MMF. The MMF version

[3] `https://mmf.sh/`

Hyperparameters	Pythia	ViLBERT†	VisualBERT†	CLIP-ViL
Batch Size	512	896	896	32
Hidden Size	5,000	1,024	768	1,024
# Layers	L-1, V-1	L-12, V-6	12	6 / 4
Optimizer	Adamax[46]	AdamW[54]	AdamW[54]	AdamW[54]
Adam ϵ	1e-8	1e-8	1e-8	1e-9
Adam β_1	0.9	0.9	0.9	0.9
Adam β_2	0.999	0.98	0.98	0.98
Learning rate	0.01	5e-5	5e-5	5e-5
Dropout	–	0.1	0.1	0.1
# Steps	22,000	88,000	88,000	236,000
# Warmup Steps	1,000	2,000	2,000	54,000
Max Grad. L2-Norm	0.25	–	–	5

Table 8: Hyperparameters of each model used in our experiments. Max Grad. L2-Norm is used for gradient clipping. L and V indicate language and vision layers, respectively. The 6 / 4 for CLIP-ViL indicates that the model has 6 MCAN layers and 4 MoVie layers. † indicates that the hyperparameters are reported directly from [74].

of this model is from [74] is pre-trained on the VQA v2 training data [26] using self-supervised objectives (masked language modeling and masked image modeling). The VQA model is initialized with the pre-trained encoder weights, and then fine-tuned on the VQA v2 training data.

VisualBERT [51]: Here, the setup is very similar to ViLBERT and we use the same visual features as ViLBERT. We again use the pre-trained and fine-tuned model provided by MMF. This MMF version of VisualBERT [74] is pre-trained on MSCOCO captions [10] using a masked language modeling objective. Just like ViLBERT, the VQA model is also initialized with the pre-trained encoder weights and fine-tuned on VQA v2.

CLIP-ViL [72]: The visual representations are grid features that are obtained from the visual encoder of the CLIP model [66]. We use the implementation provided by the authors of [72] to extract the visual features.[4] The VQA backbone is an ensemble of the MCAN [83] and MoVie [61] (please see [61] for more details). GloVe embeddings [63] are also used to initialize the word representations. Like Pythia, we train this model from scratch on VQA v2 training data.

D.2 Selection Functions

We detail the Calibration and Selector selection functions here. We do not cover MaxProb as no additional training is required. While training each selection function, we freeze the weights of the VQA model.

Calibration. The inputs to the calibration are the unnormalized answer logits (i.e., answer representation just before the softmax) of the VQA model, and the outputs are the calibrated logits. Since we use vector scaling [29, 64], we input the logits from the VQA model into a linear layer with a diagonal weight matrix and a bias term. During training, after the linear layer, we apply a sigmoid activation and use these as input to a binary cross entropy loss with the soft VQA labels [77]. We train the linear layer using the AdamW optimizer [54] with a learning rate of 0.01 and a weight decay of 1e-4. At test time, we use the output of this linear layer as our calibrated logits, apply a softmax, and use the same abstention procedure as MaxProb (Sec. 4).

Selector. The inputs to Selector are the answer, question, image, and multimodal representations. For each input, we have a specific 1-layer multi-layered perceptron network with a ReLU activation and hidden size of 512. We then concatenate the outputs of these layers and input them to a 2-layer multi-layered perceptron with ReLU activations and hidden size of 1,024, followed by a binary output layer to produce a confidence value. This architecture remains exactly the same for all models. However, if a model produces a set of representations for the image or question, then we max pool these features to collapse them to a single representation. For optimization, we employ the AdamW

[4] https://github.com/clip-vil/CLIP-ViL/tree/master/CLIP-ViL-Direct/vqa

optimizer [54] with a learning rate of 1e-4, a batch size of 256, and gradient clipping with a max gradient L2 norm of 0.25.

E Manual Evaluation of Label Noise

As discussed in Sec. 5.3, we provide further details on our manual annotation for label noise as well as Effective Reliability results for $c = 100$ when accounting for cases where the model may have been unfairly penalized. We specifically annotate image-question-answer triples, and discovered the following cases (Fig. 5 provides examples of each):

Incomplete Ground Truth: The ground truth is in some way incomplete and simply misses the predicted answer.

Semantic Match: The predicted answer is semantically correct but does not exactly match the ground truth.

Incomplete Prediction: The predicted answer is incomplete but has part of the correct answer.

Singular/Plural: The predicted answer is singular/plural while the ground truth is plural/singular (though only if providing the opposite singular/plural version is still correct).

We do these annotations for each considered VQA model and selection function trained to optimize Effective Reliability for $c = 100$ (i.e., the strongest penalty for wrong answers) and focus our efforts on questions with VQA accuracy of 0, meaning questions that contribute negatively to Φ_c. Once we have the annotations of unfairly penalized questions, we recompute the Effective Reliability score Φ'_c when counting those questions as either abstentions or as answered questions that achieved a VQA accuracy of 100%. Although the selection function decided to answer each of the unfairly penalized questions that we annotated, we compute Φ'_c under these two cases because it is unclear exactly how correct these non-matching answers should be considered. Counting them as abstentions serves as a lower bound for Φ'_c, whereas assigning a VQA accuracy of 100% is an upper bound.

We present the results before (Φ_c) and after (Φ'_c) controlling for noise in Tab. 9. We find that while this noise does contribute to some differences in performance, it does not affect the rankings between selection functions. For example, relative to each Φ_c with CLIP-ViL, Φ'_c yields an increase of 0.27 for MaxProb, 0.38 for Calibration, and 0.56 for Selector, yet the rankings remain the same. Qualitatively, we observe that there tends to be a very significant overlap in unfairly penalized examples between selection functions, which is likely part of why the rankings remain the same. Moreover, the amount of these label errors tends to be small, and the vast majority of questions contributing to the penalties in Φ_c across all models are properly marked as incorrect ($\sim$93%). Since the score for an incorrect sample (-100) is considerably lower than a sample marked as 100% correct (+1), there is also little difference in Φ'_c when considering these few unfairly penalized questions as abstentions versus as correct answers. These results imply that the comparisons between different selection functions at high cost (or low risk) for a given model are still meaningful despite the potential presence of noise.

F More Qualitative Analysis

In Fig. 6, we show several more examples of cases from our test split that illustrate Selector and MaxProb decisions, where we use CLIP-ViL with selection functions optimized for $c = 100$ on the validation set (same as Fig. 4). In particular, we show cases where the decisions of Selector and MaxProb differed — where Selector chooses to answer while MaxProb abstains, and vice-versa. We see some cases where the MaxProb decision to abstain may have been influenced by variability in possible answers that may cause model confidence values to be split, yet the annotations themselves have underlying semantic agreement (e.g., Fig. 6 top left, where "*sunny*" weather conditions are also described as "*nice*" or "*clear*"). On the other hand, we also see cases where the model was incorrect on questions which may have been unclear or surprising, and Selector chose to abstain whereas MaxProb chose to answer (e.g., the second example on row (c) asks the unusual question "*Is the bear wearing a helmet?*"). In these cases, we would expect a selective VQA model to abstain from answering to avoid providing an incorrect answer. Additionally, we show several failure cases of Selector, which chose to answer on an incorrect question while MaxProb chose to abstain.

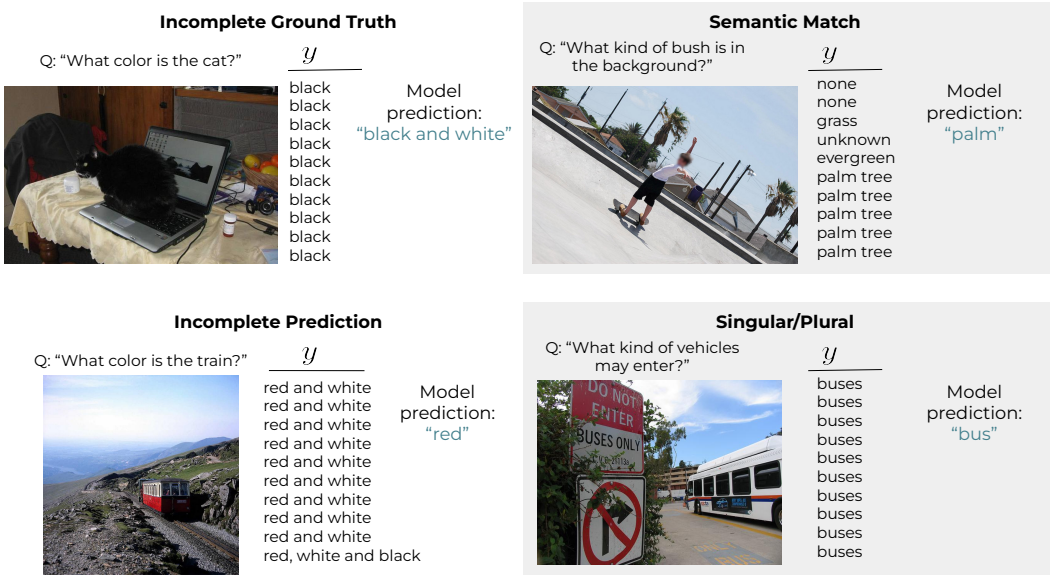

Figure 5: Example question, image, annotations, and model prediction from each category of label noise we discovered.

Model	g	% Correct GT	$\Phi_c \uparrow$	$\Phi'_c \uparrow$ *Abstain*	$\Phi'_c \uparrow$ *Correct*
Pythia [39]	MaxProb	91.30	1.81	2.00	2.00
	Calibration	93.55	2.14	2.32	2.33
	Selector	87.50	**4.12**	**4.49**	**4.50**
ViLBERT [55]	MaxProb	97.75	1.67	1.86	1.86
	Calibration	94.94	2.92	3.30	3.30
	Selector	88.14	**5.41**	**6.07**	**6.08**
VisualBERT [51]	MaxProb	100.00	2.49	2.49	2.49
	Calibration	97.92	3.83	3.93	3.93
	Selector	85.29	**4.82**	**5.77**	**5.78**
CLIP-ViL [72]	MaxProb	94.74	1.82	2.09	2.09
	Calibration	93.44	5.78	6.16	6.16
	Selector	87.23	**8.76**	**9.32**	**9.32**

Table 9: Effect of label noise on Effective Reliability at $c = 100$. % Correct GT indicates the percentage of answered samples with a VQA accuracy of 0, where the ground truth and resulting VQA accuracy was considered correct based on the question, image, annotations, and model prediction. Φ_c indicates the original score, whereas Φ'_c indicates the score when counting answered questions where label errors led to a VQA accuracy of 0 as abstentions (*Abstain*) or having a VQA accuracy of 100% (*Correct*) instead of being counted as incorrect. Although there is a small amount of label noise, it does not affect the ranking between selection functions with respect to effective reliability.

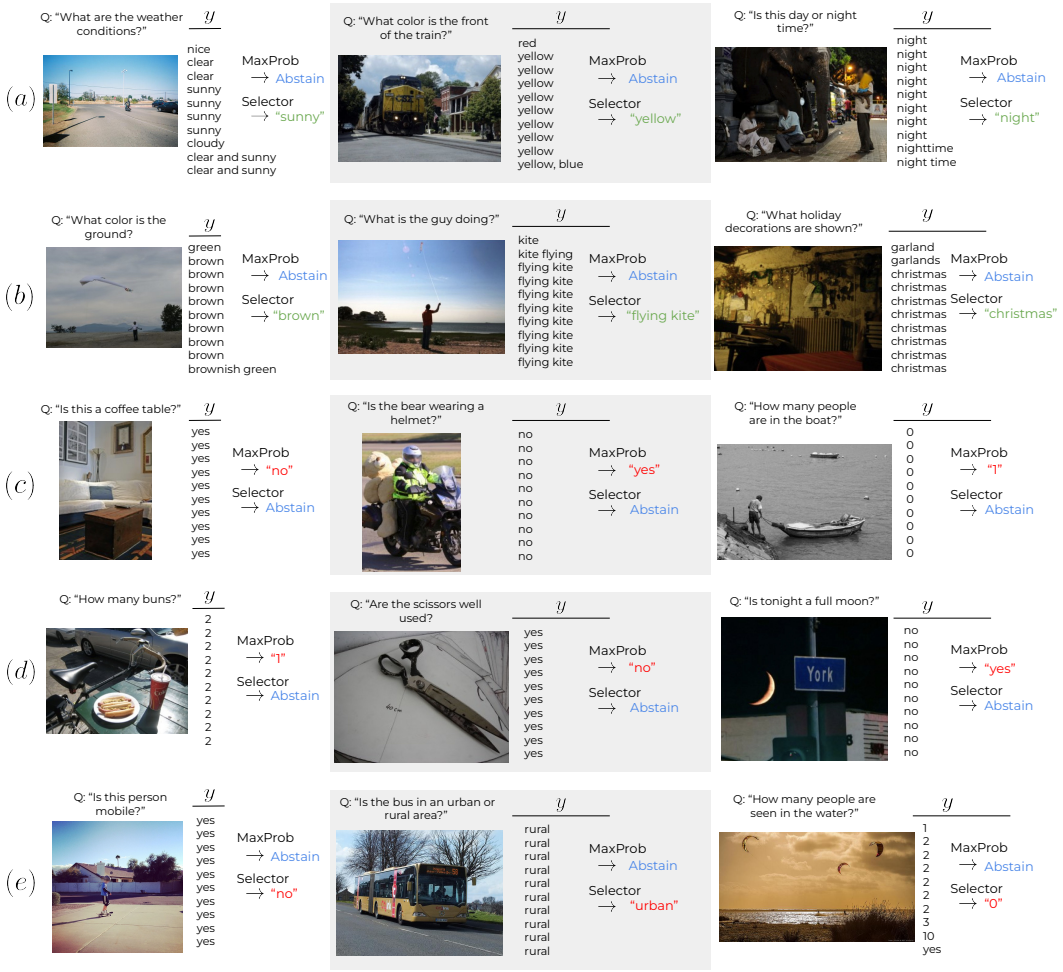

Figure 6: More qualitative test set examples with CLIP-ViL selective model predictions, when optimized for $c = 100$ on validation. Rows (a) and (b) show cases where the model was correct, yet MaxProb chose to abstain and Selector chose to answer. Rows (c) and (d) show examples of the opposite case, where the model was wrong, yet MaxProb chose to answer (contributing to the risk) and Selector chose to abstain. Row (e) shows failure cases of Selector, which chose to answer on an incorrect sample when MaxProb chose to abstain.