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], and Marcus Rohrbach[1]

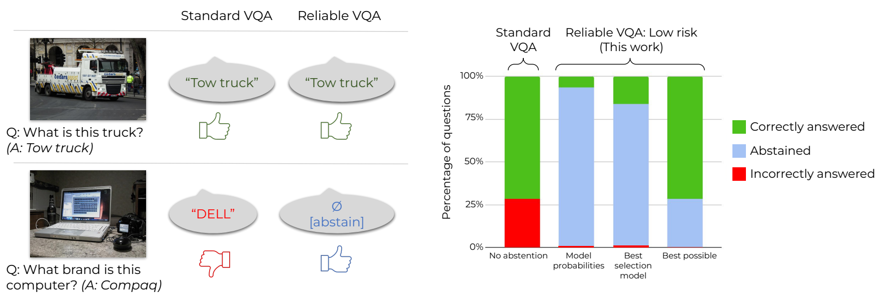
[1] Meta AI
[2] UC Berkeley

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 increase the coverage by, for example, 2.4× from 6.8% to 16.3% at 1% risk. 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]

1 Introduction

Visual Question Answering (VQA) is an important task and 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 [3, 24]. 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 [4, 22, 44, 46]. While improving the accuracy of approaches may be an important factor for trusting models,

⋆ Equal contribution
[1] Code and Models: `https://github.com/facebookresearch/reliable_vqa`



(a) (b)

Fig. 1: In the standard VQA problem, a model must answer all questions, even if it is likely to produce errors that could 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. In (b), at 1% risk of error, a SoTA model [55] can answer only ∼7% of questions when using vanilla model probabilities to choose when to abstain. Using a learned, multimodal selection function to estimate confidences can more than double the amount of questions answered, yet there remains much room for improvement (best possible, i.e., perfect abstention).

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 users with visual impairments, since they likely do not have a method of verifying the outputs themselves.

One way to avoid providing incorrect information and misleading users is to *abstain* from making a prediction, as in the framework of selective prediction [10, 15, 18, 19]. 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, in many application we may prefer if the model communicated "*I don't know.*", i.e., abstain [25, 37]. 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 difficult, unanswerable, or false premise questions [8, 24, 33, 52, 58]. 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 [10, 15] 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) [15, 35]. 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 that captures this trade-off and allows for clearer model comparisons, 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 [27, 35] (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 often leads to a better risk-coverage trade-off compared to using the original model probabilities. We improve beyond this by training a multimodal selection function that can better learn to predict if a the model's answer is correct, based on intermediate representations as well as the answer from the VQA model. This selection function consistently improves the coverage of different VQA models across varying risks of error, particularly for low levels of 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: (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 explore several methods for incorporating abstention, showing that a simple yet effective multimodal selection function outperforms other methods; (4) 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, 3, 7, 16, 17, 30, 31, 40, 42, 43, 55, 64–66]. 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 by introducing the ability to abstain into several prominent VQA models [31, 40, 43, 55].

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 [24] and reasons for annotation entropy, such as low image quality or question ambiguity [5]. [12] define a similar categorization of unanswerable questions in VQA. [58] compute precision/recall based on VQA model confidences and show that these can be reflective of the ambiguities of the ground truth answers. Other work focuses on detecting whether the question incorrectly describes the visual semantics [33, 41, 45, 52]. 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 [36]. In this work, we focus on predicting uncertainty specific to a model as opposed to the intrinsic difficulty from data itself. However, in Sec. 5.5, we find that a subset of questions on which a model abstains from answering are ambiguous or unanswerable.

Calibration. In classification settings, 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 [23, 27, 39, 48, 49]. One popular parametric method is Platt scaling [49], in which a logistic regression model is trained on classifier outputs on the validation set to return calibrated probabilities. In our work, we explore the effectiveness of vector scaling, a multiclass extension of Platt 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 as sample rejection [9, 10] or selective classification [15]. [13, 29, 59] propose various related evaluation metrics. [13] assigns cost coefficients to misclassified, abstained, and correctly classified samples. Concurrently with our work, [59] defines reliability as out-of-the-box performance for large-scale pretrained models across many unimodal vision or language tasks, including selective prediction. Other works integrate abstention in multi-stage networks or ensembles [6, 11, 38, 50, 61]. [32, 63] study selective prediction and transformer uncertainty within NLP tasks. [21, 35, 60] explore selective prediction performance on out-of-distribution data. [35] focuses 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. [18, 19] optimize selective models for specific coverage levels in image classification. 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. In the multimodal space, [26] addresses gender bias in image captioning, where the model can "abstain" by predicting gender-neutral words when it is uncertain. With our proposed metric, the cost of error (e.g., misclassifying gender) can be user-defined and potentially be made class-specific.

3 Visual Question Answering with Abstention

Visual question answering is currently formulated and evaluated in the literature [3, 20, 24, 28] as *always* predicting an answer from the answer space, $\mathcal{A}$, annotated in the dataset. So, a model $f : \mathcal{X} \mapsto \mathcal{A}$ predicts an answer $a \in \mathcal{A}$ for each input $x = (v, q) \in \mathcal{X}$, with image v and question q. This problem formulation forces the model to answer even if it is likely wrong, thus providing unreliable answers. To address this, 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*"). Outside VQA, this formulation has also been referred to as "*classification with a reject option*" [9, 13, 19, 25, 50] or "*selective prediction/classification*" [15, 18]. 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 to the setting where a model can either provide an answer from $\mathcal{A}$ or choose to abstain (denoted by $\emptyset$): $h : \mathcal{X} \mapsto \mathcal{A} \cup \{\emptyset\}$. 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 [15, 18, 19]. f denotes the VQA model that predicts answers and $g : \mathcal{X} \mapsto \{0, 1\}$ is the selection function that determines whether the model answers or abstains from answering:

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

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 $f(x)$ [35]:

$$g(x) = \begin{cases} 1 & \text{if } g'(x) \geq \gamma, \\ 0 & \text{if } g'(x) < \gamma. \end{cases} \tag{2}$$

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

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* [15] and, second, a cost-based metric for balancing the two.

[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 $\mathbb{R}$.

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 [15]. 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. Concretely, coverage for dataset $\mathcal{D}$ with inputs x_i and ground truth answers y_i is given by:

$$\mathcal{C}(g) = \frac{1}{|\mathcal{D}|} \sum_{(x_i,y_i)\in\mathcal{D}} g(x_i), \tag{3}$$

and risk is defined as:

$$\mathcal{R}(f,g) = \frac{\frac{1}{|\mathcal{D}|}\sum_{(x_i,y_i)\in\mathcal{D}} \ell(f(x_i),y_i)\cdot g(x_i)}{\mathcal{C}(g)}, \tag{4}$$

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. 2, 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 selective 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 [3] 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 [20, 24, 57]. Under this, an answer is considered fully correct if it matches at least four of the human annotations, and receives partial credit for predicting an answer with one, two, or three humans in agreement. Thus, our risk measurement becomes:

$$\mathcal{R}(f,g) = \frac{\frac{1}{|\mathcal{D}|}\sum_{(x_i,y_i)\in\mathcal{D}} (1 - Acc(f(x_i),y_i))\cdot g(x_i)}{\mathcal{C}(g)}. \tag{5}$$

In practice, the level of risk in model predictions that a user is willing to tolerate depends highly on the scenario. Therefore, we evaluate by computing 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 this risk-coverage curve (AUC) [35]. Moreover, for an evaluation that controls for how the threshold γ for g is chosen, we compute the maximum coverage for each risk level, allowing for a more direct comparison of the selection function design.

Effective Reliability. Recall the trade-off between risk and coverage: a standard VQA model may have high risk at 100% coverage, but 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 reliable and effective, it should ideally achieve both low risk and high coverage. To jointly measure these two desirable qualities, we define a metric which assigns a reward to questions that 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 [13].

Formally, we define Effective Reliability for an input x as $\Phi_c(x)$ (Eq. 6), where c is the cost for answering incorrectly, g is the selection function, and Acc is a measure of a model's correctness. In this case, Acc is the VQA accuracy [3].

$$\Phi_c(x) = \begin{cases} Acc(x) & \text{if } g(x) = 1 \text{ and } Acc(x) > 0, \\ -c & \text{if } g(x) = 1 \text{ and } Acc(x) = 0, \\ 0 & \text{if } g(x) = 0. \end{cases} \tag{6}$$

We define the total score $\Phi_c = \frac{1}{n} \sum_x \Phi_c(x)$, a mean over all n samples x. 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 the correctness of answers can vary 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 (Φ_1, Φ_{10}, Φ_{100}). While [13] suggest 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). We provide a simple proof for this in Appendix K. It is also confirmed in our experiments in Tab. 2.

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 ($Acc(x) = 0$).*

In our experiments, we choose a threshold γ 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. 2. 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).

Beyond its connection to VQA Accuracy (Lemma 1), Effective Reliability has several other advantages. We show that it meaningfully correlates with risk-coverage (Tab. 2), yet provides a single metric to compare models. This offers simpler comparisons that can be used to rank approaches (e.g., evaluating on a challenge server). It also provides an alternative evaluation for settings where it may be easier or more intuitive to define a cost for an incorrect answer as opposed to a target level of risk.

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. Additional implementation details for the selection functions can be found in Appendix I.2.

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 [27, 35, 39]. 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. Calibration techniques tune the absolute confidence values [49] to make the predicted probability for an output representative of the likelihood of that output being correct. Selective prediction has more to do with relative confidence rankings [15], but, nevertheless, a poorly calibrated model might also imply poor confidence rankings [35]. Temperature scaling [23, 49] is a popular calibration method, but it does not change the confidence rankings between examples and has no effect on the risk-coverage curve. Thus, we do not consider it in this work, but instead use vector scaling [23, 49] to calibrate the model logits. We then apply MaxProb on top of these calibrated logits. Appendix G has evaluations of how well the scores are calibrated.
Multimodal selection function: Selector. Vector scaling essentially trains an additional component on top of the VQA model to refine the model confidences. We move beyond this by training a component (Selector) to predict whether the answer is correct [14, 35, 49]. Different from prior work on confidence estimation in other tasks [14, 19, 35, 61], the multimodal nature of VQA presents unique challenges where the model must consider the interaction between the image, question, and answer. To model this, we extract the image v, question q, multimodal r, and answer $f'(x)$ representations 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 treat this as a binary classification problem (correct or incorrect). However, this does not account for answers that may be partially correct, or where one answer may be more correct than another, as is the case with VQA. Therefore, we propose to treat correctness prediction as a regression task where the target value is the VQA accuracy, allowing us to scale confidence scores with correctness.

5 Experiments

5.1 Data and Models

We experiment on the VQA v2 dataset [20] 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, 53, 62]. As a reminder, under our selective prediction setup, the VQA model is 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

Model f	Selection function g	VQA Acc. ↑	$\mathcal{C}@\mathcal{R}$ ↑ $\mathcal{R}=1\%$	$\mathcal{R}=5\%$	$\mathcal{R}=10\%$	$\mathcal{R}=20\%$	AUC ↓
Pythia [31]	MaxProb	66.17	6.00	24.71	40.99	71.45	13.88
	Calibration	66.45	6.50	25.07	41.95	**73.44**	13.52
	Selector	66.17	**8.79**	**26.92**	**43.24**	73.40	**13.30**
	Best Possible ($\mathcal{C}$)	66.17	62.67	68.41	73.52	82.71	6.68
ViLBERT [43]	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 [40]	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 [55]	MaxProb	71.75	6.78	34.69	55.72	85.13	10.23
	Calibration	71.71	13.12	37.06	56.06	85.23	9.91
	Selector	71.75	**16.34**	**39.48**	**58.16**	**85.37**	**9.52**
	Best Possible ($\mathcal{C}$)	71.75	68.49	74.55	79.72	89.69	4.58

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

set into three subsets: a split with 86k examples (40%) for validating VQA models 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 benchmark the selection functions introduced in Sec. 4 in combination with VQA models with varying architectures and performance (test-std VQA v2 accuracy in parentheses): **Pythia** [31] (70.24%), an optimization of the widely used bottom-up top-down VQA model [2]; **ViLBERT** [43] (70.92%), a two-stream transformer, and **VisualBERT** [40] (71.00%), a single-stream transformer, both of which use multimodal pretraining [56]; **CLIP-ViL** [55] (74.17%), which is the MoVie+MCAN [47] model with a visual encoder from CLIP [51].

In Tab. 1, Tab. 2, and Fig. 2, we report mean results over 10 random seeds for Pythia and CLIP-ViL (standard deviations in Appendix J), while we report single runs for ViLBERT and VisualBERT using existing pretrained and fine-tuned models. All other results are single runs from the same randomly chosen seed. Details of data and model setups are in Appendix H and Appendix I.

5.2 Benchmarking Risk and Coverage

As discussed in Sec. 3.2, we measure the maximum coverage for a given risk ($\mathcal{C}@\mathcal{R}$) as well as AUC for the risk-coverage curves and overall accuracy for each model. We include the best possible performance on these metrics for each model, which would be a selective model that abstains only when the prediction is incorrect. Results are reported on the test test.

Selector outperforms other methods. From Tab. 1, we see that adding the Selector consistently outperforms MaxProb in coverage for all risk tolerances as

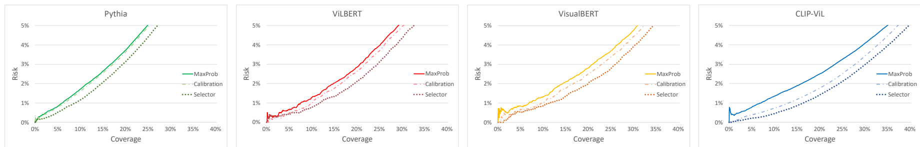


Fig. 2: Risk-coverage plots for each model up to 5% risk.

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 2.4× 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. 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 (9.56% and 4.79%, respectively), which also has the highest accuracy. Meanwhile, Pythia has the lowest accuracy and exhibits the smallest improvements with the Selector at these tolerances (2.79% and 2.21%, 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. For example, CLIP-ViL has 2.55% higher accuracy than ViLBERT, but, with default MaxProb, ViLBERT has 0.73% higher $\mathcal{C}$@1% than CLIP-ViL. Appendix B also shows that augmenting the VQA model training data with the selection function training data and using MaxProb still has worse coverage at low risk than when using this data for Selector training, despite having higher accuracy. These results imply that improving upon the 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. In practice, however, a threshold γ must be chosen, e.g., on a validation set, and used at test-time. We evaluate how close

Model f	Selection function g	c=1			c=10			c=100		
		$\Phi_1 \uparrow$	$\mathcal{R} \downarrow$	$\mathcal{C} \uparrow$	$\Phi_{10} \uparrow$	$\mathcal{R} \downarrow$	$\mathcal{C} \uparrow$	$\Phi_{100} \uparrow$	$\mathcal{R} \downarrow$	$\mathcal{C} \uparrow$
Pythia [31]	—	38.49	33.83	100	-210.62	33.83	100	−2701.68	33.83	100
	MaxProb	47.28	21.62	76.03	15.15	5.24	25.62	2.27	0.85	4.89
	Calibration	48.06	21.21	76.18	15.23	5.85	28.06	2.19	0.94	5.88
	Selector	**48.16**	20.67	74.84	**17.12**	5.99	30.16	**3.84**	0.94	8.23
	Best Possible (Φ_c)	66.17	8.51	72.32	66.17	8.51	72.32	66.17	8.51	72.32
ViLBERT [43]	—	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 [40]	—	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 [55]	—	49.41	28.25	100	-151.70	28.25	100	-2162.80	28.25	100
	MaxProb	55.82	19.22	83.45	22.03	5.59	37.67	2.85	0.96	6.97
	Calibration	56.03	18.30	81.61	23.24	4.95	36.82	5.30	0.73	9.97
	Selector	**56.45**	17.44	80.09	**26.06**	5.03	39.59	**8.01**	0.55	11.38
	Best Possible (Φ_c)	71.75	7.60	77.66	71.75	7.60	77.66	71.75	7.60	77.66

Table 2: 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. All in %.

the actual test-time risk is to the target risk when using the validation threshold with VisualBERT, with results in Appendix F. We find relatively small differences in risk, showing that the thresholds generalize reasonably well. This aligns with prior findings on other tasks [19]. However, since the actual risks are now slightly different between models, we can no longer compare the corresponding coverages directly. This motivates Effective Reliability, which compares models based on a predefined cost for wrong answers as opposed to an exact risk level.

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. 2, we choose cost values c of 1, 10, and 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% for Φ_1, yet it is 3.74% for Φ_{100}. Further, the gap between Selector and MaxProb for Φ_{100} generally 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. 2, where the range of risk levels for selective models at Φ_{100} ($\mathcal{R} \approx 0.5$–1.3%) is much lower than the range of risk at Φ_1 ($\mathcal{R} \approx 17$–22%). This directly translates to a similar trend in coverage, where selective models answer about 5–11% of questions at Φ_{100}, and about 76–83% of questions at Φ_1. This shows that Effective Reliability behaves intuitively around the influence of a user-selected cost on model risk and coverage.

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 [3, 34, 54]. To see if our results for Φ_{100} are significantly affected by annotation noise, in Appendix C, we manually examine each sample where the model predictions were marked incorrect (and thus heavily penalized when computing Φ_{100}). We annotate cases where models may have been unfairly penalized and recompute Φ_{100} when removing this penalty. 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.

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 Selection Function Ablations

Tab. 3 provides ablations for the selection function design. In the following, we distill the main observations. Additional discussion is in Appendix A.

Selector requires multimodal input. Tab. 3 shows 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%, $\mathcal{C}$@5%, Φ_{10}, and Φ_{100} than unimodal representations

Features	Unimodal	Loss	$\mathcal{C}@\mathcal{R}$ ↑ $\mathcal{R}=1\%$	$\mathcal{R}=5\%$	$\mathcal{R}=10\%$	$\mathcal{R}=20\%$	AUC ↓	Φ_c ↑ c=1	c=10	c=100
$\tilde{v}$	✓	Regression	0.00	0.00	0.00	16.09	23.23	48.83	0.00	0.00
q	✓	Regression	0.02	11.03	35.88	79.70	13.39	52.99	10.36	1.33
$f'(x)$		Regression	5.24	36.10	56.30	84.79	10.08	56.03	23.14	5.88
v		Regression	11.60	36.43	53.74	83.51	10.32	54.84	23.91	6.10
r		Regression	**13.42**	34.69	53.90	82.95	10.43	54.35	22.34	**7.77**
$f'(x)$+$\tilde{v}$		Regression	3.67	36.40	56.33	84.79	10.07	55.97	23.63	4.60
$f'(x)$+q		Regression	10.67	37.41	56.95	84.76	9.86	56.01	24.35	5.32
$f'(x)$+r		Regression	12.02	37.44	<u>57.68</u>	<u>84.93</u>	9.81	56.07	24.28	5.51
$f'(x)$+v		Regression	13.24	**38.51**	57.44	84.92	<u>9.76</u>	**56.20**	**25.11**	7.03
$f'(x)$+q+v+r		Classification	6.64	35.80	57.29	84.18	10.06	55.61	23.23	4.36
$f'(x)$+q+v+r		Regression	<u>13.32</u>	<u>38.02</u>	**58.16**	**85.03**	**9.73**	<u>56.09</u>	<u>24.85</u>	<u>7.32</u>

Table 3: Ablations of Selector with CLIP-ViL [55] on our selection function validation set. 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). All in %.

(image $\tilde{v}$ or question q). For highly reliable models ($\mathcal{C}$@1%, Φ_{100}), unimodal selection functions fail (coverage $\leq$0.02%, $\Phi_{100} < 2\%$), suggesting that building reliable and effective VQA models is a truly multimodal problem. Combining all representations generally performs best, 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 significant improvements (Tab. 3), 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 to classification where the distinction between partially correct answers is lost.
Selector Architecture. Appendix A presents results using different Selector architectures, where a less complex architecture can degrade performance, but a more complex one does not necessarily improve it. Together with Tab. 3, we find that, rather than the network layout, the *input* to the Selector and optimization target are more critical to the performance when using the Selector.

5.5 Qualitative Analysis

Fig. 3 visualizes MaxProb and Selector decisions with CLIP-ViL for several examples on the test set (more in Appendix E). The abstention threshold is chosen to maximize Φ_{100} on validation. Fig. 3 (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. 3 (middle), we see a false premise question [52] where Selector abstains again as the question does not make sense for the image, while MaxProb yields an incorrect answer. Fig. 3 (right) presents an example with synonymous answers where the

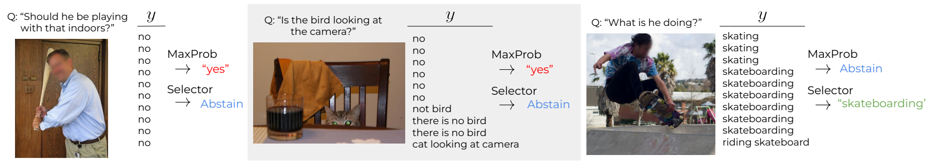


Fig. 3: Qualitative test set examples with CLIP-ViL selective model predictions.

model is correct yet MaxProb chooses to abstain and Selector chooses to answer. In a classification-based VQA model, synonyms can split the maximum softmax score used by MaxProb, whereas the Selector can potentially learn these answer similarities and adjust the confidence. These examples contribute to the higher coverage at low risk observed quantitatively in our experiments. We also find that MaxProb chooses to answer many simple questions, while Selector additionally chooses to answer more difficult, multimodal ones as well (see Appendix D).

6 Conclusion

The standard VQA 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 abstention. If we want a reliable model with 1% risk of error, we find that a state-of-the-art VQA model [55] only answers less than 7% of the questions when using its softmax probabilities as estimates of model confidence. Using calibration can improve this, but we find that the best results are consistently achieved by training a multimodal selection function to estimate correctness directly. This increases the coverage from 6.78% to 16.34%. While this is a marked improvement, one has to consider that this model achieves 71.75% standard VQA accuracy on the same set of data. With our *Effective Reliability* metric, the performance drops from 71.75% (for perfect abstention) to 8.01% (our best abstention baseline) with high penalties for wrong answers. We believe this new framework and metric for VQA will encourage the community to build VQA models which are both reliable and effective, as well as offer an opportunity for many exciting directions to improve the self-awareness of models.
Acknowledgements: We 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, PTG, 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. 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)
4. 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)
5. 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)
6. Black, E., Leino, K., Fredrikson, M.: Selective ensembles for consistent predictions. In: International Conference on Learning Representations (2022)
7. 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)
8. 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)
9. Chow, C.: On optimum recognition error and reject tradeoff. IEEE Transactions on information theory **16**(1), 41–46 (1970)
10. Chow, C.K.: An optimum character recognition system using decision functions. IRE Transactions on Electronic Computers **EC-6**(4), 247–254 (1957)
11. 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)
12. Davis, E.: Unanswerable questions about images and texts. Frontiers in Artificial Intelligence **3**, 51 (2020)
13. 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
14. 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`
15. El-Yaniv, R., Wiener, Y.: On the foundations of noise-free selective classification. Journal of Machine Learning Research **11**, 1605–1641 (2010)
16. 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)
17. 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)

18. Geifman, Y., El-Yaniv, R.: Selective classification for deep neural networks. Advances in neural information processing systems **30** (2017)
19. 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)
20. 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)
21. 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)
22. 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`
23. 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)
24. 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)
25. Hanczar, B., Dougherty, E.R.: Classification with reject option in gene expression data. Bioinformatics **24**(17), 1889–1895 (2008)
26. 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)
27. 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)
28. 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)
29. 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`
30. 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)
31. 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)
32. Kadavath, S., Conerly, T., Askell, A., Henighan, T., Drain, D., Perez, E., Schiefer, N., Dodds, Z.H., DasSarma, N., Tran-Johnson, E., et al.: Language models (mostly) know what they know. arXiv preprint arXiv:2207.05221 (2022)

33. Kafle, K., Kanan, C.: An analysis of visual question answering algorithms. In: ICCV (2017)
34. Kafle, K., Kanan, C.: Visual question answering: Datasets, algorithms, and future challenges. Computer Vision and Image Understanding **163**, 3–20 (2017)
35. 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
36. 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
37. 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)
38. Khani, F., Rinard, M., Liang, P.: Unanimous prediction for 100% precision with application to learning semantic mappings. arXiv preprint arXiv:1606.06368 (2016)
39. 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)
40. 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)
41. 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)
42. 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)
43. 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)
44. 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)
45. Mahendru, A., Prabhu, V., Mohapatra, A., Batra, D., Lee, S.: The promise of premise: Harnessing question premises in visual question answering. In: EMNLP (2017)
46. 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
47. 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)

48. 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)
49. 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)
50. 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
51. 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)
52. 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)
53. Shah, M., Chen, X., Rohrbach, M., Parikh, D.: Cycle-consistency for robust visual question answering. In: CVPR (2019)
54. Sharma, H., Jalal, A.S.: A survey of methods, datasets and evaluation metrics for visual question answering. Image and Vision Computing **116**, 104327 (2021)
55. 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)
56. Singh, A., Goswami, V., Parikh, D.: Are we pretraining it right? digging deeper into visio-linguistic pretraining. arXiv preprint arXiv:2004.08744 (2020)
57. 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)
58. Teney, D., Liu, L., van Den Hengel, A.: Graph-structured representations for visual question answering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1–9 (2017)
59. Tran, D., Liu, J., Dusenberry, M.W., Phan, D., Collier, M., Ren, J., Han, K., Wang, Z., Mariet, Z., Hu, H., Band, N., Rudner, T.G.J., Singhal, K., Nado, Z., van Amersfoort, J., Kirsch, A., Jenatton, R., Thain, N., Yuan, H., Buchanan, K., Murphy, K., Sculley, D., Gal, Y., Ghahramani, Z., Snoek, J., Lakshminarayanan, B.: Plex: Towards reliability using pretrained large model extensions (2022). https://doi.org/10.48550/ARXIV.2207.07411, `https://arxiv.org/abs/2207.07411`
60. Varshney, N., Mishra, S., Baral, C.: Investigating selective prediction approaches across several tasks in IID, OOD, and adversarial settings. In: Findings of the Association for Computational Linguistics: ACL 2022. pp. 1995–2002 (2022). https://doi.org/10.18653/v1/2022.findings-acl.158, `https://aclanthology.org/2022.findings-acl.158`
61. 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)
62. Whitehead, S., Wu, H., Ji, H., Feris, R., Saenko, K.: Separating skills and concepts for novel visual question answering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5632–5641 (2021)

63. 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)
64. Yang, Z., He, X., Gao, J., Deng, L., Smola, A.: Stacked attention networks for image question answering. In: CVPR (2016)
65. 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)
66. 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)