Improving Selective Visual Question Answering by Learning from Your Peers

Corentin Dancette[1,3†*] Spencer Whitehead[1*] Rishabh Maheshwary[1] Ramakrishna Vedantam[1]
Stefan Scherer[2] Xinlei Chen[1] Matthieu Cord[3,4] Marcus Rohrbach[1]

[1]FAIR, Meta AI [2]Reality Labs Research, Meta [3]Sorbonne Université [4]Valeo.ai

Abstract

Despite advances in Visual Question Answering (VQA), the ability of models to assess their own correctness remains under-explored. Recent work has shown that VQA models, out-of-the-box, can have difficulties abstaining from answering when they are wrong. The option to abstain, also called Selective Prediction, is highly relevant when deploying systems to users who must trust the system's output (e.g., VQA assistants for users with visual impairments). For such scenarios, abstention can be especially important as users may provide out-of-distribution (OOD) or adversarial inputs that make incorrect answers more likely. In this work, we explore Selective VQA in both in-distribution (ID) and OOD scenarios, where models are presented with mixtures of ID and OOD data. The goal is to maximize the number of questions answered while minimizing the risk of error on those questions. We propose a simple yet effective Learning from Your Peers (LYP) *approach for training multimodal selection functions for making abstention decisions. Our approach uses predictions from models trained on distinct subsets of the training data as targets for optimizing a Selective VQA model. It does not require additional manual labels or held-out data and provides a signal for identifying examples that are easy/difficult to generalize to. In our extensive evaluations, we show this benefits a number of models across different architectures and scales. Overall, for ID, we reach 32.92% in the selective prediction metric coverage at 1% risk of error ($\mathcal{C}$@1%) which doubles the previous best coverage of 15.79% on this task. For mixed ID/OOD, using models' softmax confidences for abstention decisions performs very poorly, answering <5% of questions at 1% risk of error even when faced with only 10% OOD examples, but a learned selection function with LYP can increase that to 25.38% $\mathcal{C}$@1%.*

*Equal contribution.
†Work primarily done during internship at FAIR.
Code: https://github.com/facebookresearch/selective-vqa_ood

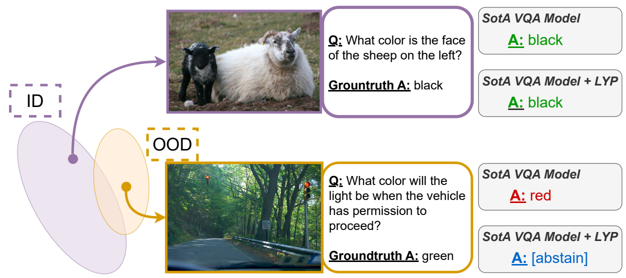

Figure 1. VQA Models are able to answer straightforward ID questions, as in the top example where a SotA model [62] with and without our Learning from Your Peers (LYP) approach answers correctly. However, difficult OOD examples can arise, like the bottom example. With LYP, the model is able to abstain from answering to avoid outputting the incorrect answer, whereas the existing model is overconfident and outputs the answer anyways.

1. Introduction

Recent successes of deep learning models for multimodal tasks have created the potential for many exciting real-world applications that require a large degree of reliability, such as providing assistance to users with visual impairments [23, 51]. However, with these novel, high-stakes applications come responsibilities towards the users, as well as the need to revise problem setups and the general approach to evaluating model performance. One particularly important consideration when developing models for real-world applications is *reliability*, i.e., the ability of the model to avoid making errors when facing uncertainty.

One way to approach reliability is to frame the problem as a selective prediction task [9, 14, 63]. In selective prediction, models are able to either output an answer or abstain from answering (i.e., effectively saying "*I don't know*") based on the model's confidence/uncertainty in order to avoid making incorrect predictions. A prevalent cause of such incorrect predictions in real-world settings is distribution shifts [13, 20, 42], where the test environment may differ from the training environment and models could encounter a wide variety of input examples at test time that may not satisfy the independent and identically distributed assumption often made by practitioners when developing models. This is especially true in open-ended tasks like Visual Question Answering (VQA) where models may receive adversarial, out-of-distribution (OOD) inputs that are difficult to answer correctly. For example, in Fig. 1, a model is asked a question that requires background knowledge that it simply does not possess. While the ability to answer open-ended questions has been a point of focus in VQA, having a model perfectly answer all questions, ID and OOD, is likely unattainable [19, 29]. Therefore, framing this problem as a selective prediction task provides an avenue to handle such OOD examples more gracefully as the model can abstain from answering on many of these inputs, while still attempting to answer as many questions as possible. Doing this requires models to recognize OOD examples for abstention decisions (OOD detection) and generalize to OOD examples (OOD generalization) in order to make predictions on examples that the model will get right.

However, previous evaluations for selective prediction in VQA [63] have been done on ID data, where the questions and images all come from the VQA v2 dataset [21]. In NLP, there are efforts on selective prediction with OOD examples [29, 59], although they tend to not address some practical considerations, such as assuming access to OOD data or threshold generalization. More broadly, selective prediction and OOD generalization have largely been studied as independent problems in the literature [58].

In this work, we explore selective prediction for VQA with distribution shifts, where we present models with mixtures of both ID and OOD examples, and measure the ability of different approaches to optimize answering as many questions as possible while maintaining a low risk of error (or high accuracy) on those questions. We perform extensive experiments on VQA v2 [21] as our ID data and AdVQA [50] as our adversarial, OOD data.

We evaluate a number of state-of-the-art approaches to this problem and find that existing models' softmax probabilities are generally poor confidence estimates for abstention decisions on OOD data, leading models to answer $<5\%$ of questions to achieve 1% risk of error in some settings. Further, we show that training a selection function [63] improves performance ID and OOD, but integrating features from OOD detection methods as well as augmenting with known-OOD data (i.e., OOD data different from the unknown target distribution) do not improve beyond simply training this selection function on ID data. However, we observe that existing methods for training multimodal selection functions can require a held-out dataset in order to be most effective.

Therefore, we propose a Learning from Your Peers (LYP) approach that alleviates the need for held-out data while also allowing both the VQA model and selection function to learn from the additional data that would have been withheld. LYP works by breaking the training data into N subsets and training different VQA models on distinct combinations of $N-1$ subsets, leaving one subset out at a time. Our approach then uses these trained models to predict answers on their respective N^{th} left-out subsets. We recombine this data into an updated training set that has predictions from the different models. We utilize these predictions and the associated accuracies as labels to train a multimodal selection function, which learns to predict the accuracies. By using predictions on the training data from models that have not seen these examples, our approach provides a signal for which examples in the training data can be generalized to for a given model class, and which are too hard and should be abstained on.

Overall, our contributions are: We present an evaluation benchmark for Selective VQA on both ID and OOD data. We show that model and data scaling are important factors for selective prediction and evaluate multiple baselines from prior works. Finally, we propose LYP and demonstrate that it can benefit performance over standard selection function training in both ID and mixed ID/OOD settings.

2. Related Work

Visual Question Answering. VQA is a popular multimodal task that requires an understanding of both vision and language modalities to predict answers. There are many standard datasets [5, 21, 23, 26] and models for this task [4, 27, 28, 35, 40, 49, 55, 62]. In our work, we employ recent state-of-the-art models [49, 62] as our backbones to explore selective VQA.

OOD VQA. Multiple works have demonstrated that VQA models often rely on shortcuts and do not generalize well on OOD data. VQA-CP [2] shows that VQA models rely on superficial correlation and lack image grounding. GQA-OOD [31] introduces an OOD benchmark that increases question diversity by including questions from various groups. VQA-CE [11] takes a step further and considers biases on both questions and images. AdVQA [50] and A-VQA [34] are recently introduced VQA benchmarks that comprise adversarial questions using human and model-in-the-loop procedures to generate adversarial examples. Other datasets require different abilities, such as TextVQA [52] which contains questions requiring reading text in the image, or OK-VQA [41] which requires external knowledge. Methods to overcome difficulties related to OOD data include [7], which tackles unimodal biases, and [47], which improves image grounding using adversarial regularization. Recently, [3] performs cross-dataset evaluations where VQA models exhibit poor generalization.

Selective prediction & reliability. Recently, [63] explore Selective Prediction for VQA with ID data. They experiment with different selectors on top of the base VQA model for improving their reliability on the VQA task. [59] investigates selective prediction approaches across several NLP tasks in ID, OOD, and adversarial settings. Specifically, they trained a selector (MLP) on top of the base model on a held-out split and used the selector's confidence scores to either answer or abstain from answering and improved risk, and coverage metrics compared to MaxProb. [10] studies failure prediction in deep neural networks by training a confidence model on top to provide confidence measures for the model prediction.

OOD selective prediction. [18] proposes SelectiveNet that incorporates a selection head on the top of the base model, which is optimized with a selective loss to reject samples that the model is uncertain about. [29] trains a calibrator on top of an existing NLP model to generalize to unknown OOD data at test time. Specifically, it trains the calibrator on a mixture of some held-out ID data and 'known' OOD data. The final model is used for the evaluation of the unknown OOD data at test time.

OOD detection. Earlier works [25] rely on the maximum class probability (MaxProb) to detect OOD samples. [36] proposes ODIN that combines temperature scaling and image perturbation to achieve better separation in softmax scores for OOD and ID images. Another line of work uses distance [33] or energy [37, 39, 61] scores for OOD detection. [60] introduces VIM that detects OOD samples by fusing the logits and feature information obtained from the model. [6, 53, 57] computes nearest-neighbor distances in the feature dimension to detect OOD data.

Image OOD detection & reliability. [44] investigates the effect of dataset distribution shift on accuracy and calibration. [32] uses deep ensembles to quantify uncertainty estimates of classification models. [1, 16] extensively review of uncertainty estimation methods in deep learning literature.

3. Selective VQA with ID and OOD Data

In this section, we discuss the problem formulation of Selective VQA in Sec. 3.1, and how we evaluate in the ID (in-distribution) scenario (Sec. 3.2) and in the mixed ID+OOD (out-of-distribution) scenarios (Sec. 3.3).

3.1. Problem Formulation

The primary setting for VQA is to learn a function $f : \mathcal{Q}, \mathcal{V} \mapsto \mathcal{A}$ to predict an answer $a \in \mathcal{A}$ to a question $q \in \mathcal{Q}$ about a given image $v \in \mathcal{V}$ [5, 21, 23]. However, when exposing models to the real world they might encounter hard questions, OOD data points, or even adversarial questions by users and we cannot expect that models are able to answer all questions in these scenarios correctly. Therefore, we instead would like to identify inputs $x = (v, q) \in \mathcal{X}$ that models cannot correctly answer and abstain in those cases. This is the setting of Selective Prediction [14], which has also recently been studied for ID VQA [63] and OOD text-only question answering [29]. In this work, we advocate for this selective prediction setting for ID and OOD scenarios. We closely follow the formalism introduced in [63] for VQA, though it is very similar to setups outside of VQA (e.g., "classification with a rejection option" [8, 12, 18, 24, 45], or "selective prediction/classification" [14, 17]). Specifically, the output space is extended to allow for an abstention option (denoted by $\emptyset$): $h : \mathcal{X} \mapsto \mathcal{A} \cup \{\emptyset\}$. Such a *Selective Model* h can be realized by decomposing h into two functions, a VQA model f and selection function $g : \mathcal{X} \mapsto \{0, 1\}$ [14, 17, 18, 63]:

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

For a given image-question pair $x = (v, q)$, the Selective VQA model h only predicts an answer from the VQA model f if the selection function g decides that an answer should be given. Otherwise, the Selective VQA model h abstains. The selection function g can be formulated based on a function $g' : \mathcal{X} \mapsto \mathbb{R}$ that scores the correctness of the model's prediction $f(x)$ [18, 29, 63], and a threshold $\gamma \in \mathbb{R}$. Then, for a given γ, the model outputs the answer $f(x)$ if $g'(x) \geq \gamma$ and abstains otherwise. Ideally, g' should yield higher values if $f(x)$ is correct and lower if it is incorrect. However, as we show in the experiments this is a hard task.

3.2. Evaluation

Beyond accuracy, we evaluate using the metrics designed for models with abstention options following [63]:

Risk and coverage. For a dataset D, model f, and a selection function g, *coverage* is the proportion of answered questions:

$$\mathcal{C}(g) = \frac{1}{|D|} \sum_{x \in D} g(x),$$

while *risk* is the average error on the covered subset

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

where Acc is VQA accuracy [5] and y_i is the corresponding ground truth answer. We measure the maximum coverage at a specific risk tolerance, denoted ($\mathcal{C}@\mathcal{R}$), by determining the largest consecutive subset of questions that can be answered with at most $\mathcal{R}$ risk. Further, we also compute the Area Under the Curve (AUC) for the risk-coverage curve [29] for a summary of performance across different coverage levels.

Effective reliability Φ_c. This metric is introduced in [63] to better compare methods on the test set for a threshold selected on a validation set. This is especially important for OOD, as thresholds for a certain risk level don't generalize to the test scenario. Φ_c is a cost-based metric and jointly

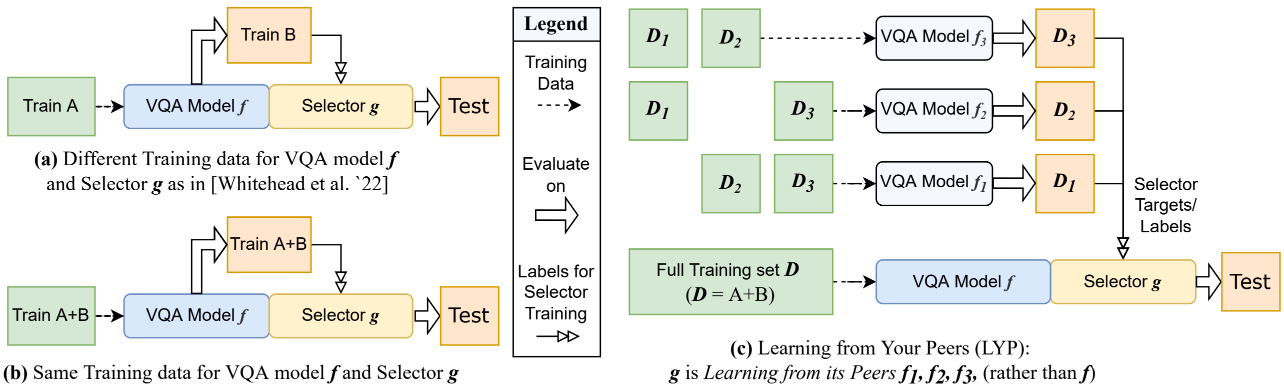

Figure 2. Comparison between Selector g training procedures. (a) shows the one in [63], (c) shows our LYP. See Sec. 4 for details.

measures the reliability and effectiveness of selective models in a single metric. It assigns a cost of c to every wrong answer that the model outputs (i.e., does not abstain on):

$$\Phi_c(x) = \begin{cases} Acc(x) & \text{if } g(x) = 1 \text{ and } Acc(x) > 0, \\ -c & \text{if } g(x) = 1 \text{ and } Acc(x) = 0, \\ 0 & \text{if } g(x) = 0. \end{cases} \quad (2)$$

The total score is $\Phi_c = \frac{1}{|D|} \sum_{x \in D} \Phi_c(x)$, a mean over all samples x. To compute this metric, we set the threshold γ on a validation set to maximize Φ_c. Then, we use this threshold for abstention decisions on the test set.

3.3. Evaluating with Mixed ID+OOD Data

As previously mentioned, we want to explore the setting where models encounter mixtures of ID and OOD data. More formally, we assume we are given $\mathcal{D}_{\text{train}}$ and $\mathcal{D}_{\text{test}}$ that are drawn from different distributions. In our setting, to simulate a setting closer to a real-world use case, the test data is sampled from a mixture of ID and OOD data. Similar to [29], we assume that our training data is drawn from P_{src} while our testing data is drawn from P_{tgt}, where $P_{\text{tgt}} = \alpha P_{\text{src}} + (1 - \alpha) P_{\text{unk}}$. Here, P_{unk} is an unknown distribution different from P_{src} from which we obtain our OOD examples. We obtain different mixtures of data by varying α and evaluate models across these using the metrics discussed in Sec. 3.2. Different from prior work in NLP [29], we assume we *do not* have access to known OOD data for training, meaning all models must be trained and thresholds must be chosen on ID data. However, we do compare to this setting in our experiments.

4. LYP: Learning from Your Peers

Prior work has established training a selection function (or Selector) g to predict the correctness of the outputs of a model f [18,29,63] as a method for selective prediction. As in [63], our Selector g learns to predict the VQA Accuracy of f. One option is to train f on one part of the training data (Train A) and g on a different, typically smaller, part (Train B), as shown in Fig. 2(a). Having separate training data for g can be crucial since if f has overfit to the training data, then training g on that same data will lead g to a solution that doesn't generalize well (e.g., always answering). We show some of these drawbacks in our experiments with observations similar to findings on stacked generalization [65]. However, withholding data from training f could reduce the overall performance of f, as it does not allow f to learn from this data. Likewise, g is unable to learn from the training data for f. This motivates training both f and g on the same data, e.g., as done in [18] (shown in Fig. 2(b)).

We propose a simple yet effective approach, called Learning from Your Peers (LYP), for training g that allows both f and g to utilize all the available training data. Inspired by work on collective outliers [30] and improving worst group performance [38], our approach aims to identify examples in the training data that are difficult to generalize to, for a given architecture and learning procedure. In particular, we want to provide more signal to g about which examples in the training data may not be generalizable and likely should be abstained on, despite the VQA model's potential ability to fit these examples during training.

Shown in Fig. 2(c), we first partition our full training set $\mathcal{D}$ into N disjoint subsets ($\mathcal{D}$ = Train A + Train B). For our VQA setting, we create our partitions by ensuring no images overlap between them. Next, we train N different models on combinations of the subsets in leave-one-out manner: we create a training set $\mathcal{D}^*_n = \mathcal{D} \setminus \mathcal{D}_n$ and train a VQA model f_n on $\mathcal{D}^*_n$. Once we have trained f_n, we use it to make predictions on $\mathcal{D}_n$, which it has not seen during its training. We use the ground truth annotations for $\mathcal{D}_n$ to obtain VQA accuracy for each prediction, which we treat as a label for the correctness of each prediction. After performing this operation for $n = 1, ..., N$, we can union the partitions to obtain an updated training set $\mathcal{D}^{\text{sel}}$ that additionally has correctness labels for each

VQA Model f		Selection func. g			Acc ↑	$\mathcal{C}@\mathcal{R}$ in % ↑			AUC ↓	Φ_1	Φ_{10}	Φ_{100}
Name	Train Set	Name	Train Set	Targets		$\mathcal{C}$@1%	$\mathcal{C}$@5%	$\mathcal{C}$@10%				
CLIP-ViL	A	MaxProb	-	- [63]	69.98	4.97	33.79	53.62	10.92	54.67	21.40	1.32
		Selector	B	Self [63]	69.98	15.79	37.79	55.65	10.21	55.44	25.82	8.74
	A+B	MaxProb	-	-	70.72	5.54	34.84	55.04	10.49	55.93	22.81	2.59
		Selector	A+B	Self	70.72	6.45	34.26	56.07	10.48	56.07	22.99	2.39
		Selector	A+B	LYP	70.72	**18.40**	**38.65**	**57.40**	**9.76**	**56.53**	**26.45**	**9.74**
OFA-Base	A	MaxProb	-	-	74.87	3.45	45.60	66.61	7.99	62.52	30.57	6.81
		Selector	B	Self	74.87	23.78	49.16	69.00	7.32	63.03	34.39	12.53
	A+B	MaxProb	-	-	75.18	14.88	46.15	67.51	7.79	63.04	30.13	7.29
		Selector	A+B	Self	75.18	26.64	50.80	69.56	7.10	63.66	34.92	12.92
		Selector	A+B	LYP	75.18	**27.71**	**51.64**	**70.20**	**6.98**	**63.88**	**36.29**	**16.30**
OFA-Large	A	MaxProb	-	-	77.53	20.57	53.99	75.18	6.42	66.68	36.12	8.21
		Selector	B	Self	77.53	30.86	58.05	76.65	5.81	67.34	41.43	17.58
	A+B	MaxProb	-	-	77.79	16.31	53.83	75.27	6.43	66.96	36.06	6.29
		Selector	A+B	Self	77.79	31.47	58.80	77.14	5.69	67.82	41.43	16.08
		Selector	A+B	LYP	77.79	**32.92**	**59.43**	**77.52**	**5.60**	**68.02**	**42.83**	**18.78**

Table 1. Risk-coverage metrics and effective reliability on ID data (i.e., VQA v2 test split from [63]).

example $(x_i^{(n)}, y_i^{(n)}, f_n(x_i^{(n)}), \xi_i^{(n)})$ for $(x_i^{(n)}, y_i^{(n)}) \in \mathcal{D}_n$, where $\xi_i^{(n)} = \mathrm{Acc}(f_n(x_i^{(n)}), y_i^{(n)})$.

We train our VQA model f on all of $\mathcal{D}$ and then, with the obtained correctness labels, we train our Selector g on top of f using the full $\mathcal{D}^{\text{sel}}$ dataset. For training g, we follow [63] and optimize it using a regression objective with the correctness labels as the target. Note, our setup is similar to that of [63] in that we use a regression objective, but, importantly, the source of our targets is not the model f itself but, rather, the subset models $\{f_n\}_{n=1}^N$ (i.e., the *peers* of f). The idea behind this is that if a model trained on the remainder of the training data $\mathcal{D}_n^*$ cannot generalize to an example in $\mathcal{D}_n$, then that may be a challenging example that g should choose to abstain on as the model f is unlikely to generalize reliably to such an example at test time, even if it has fit it during training. Essentially, these correctness labels may provide a signal for which examples are difficult and might require abstention *more generally* rather than with respect to a specific model as in prior work [63]. Moreover, we show in our experiments that this allows f and g to learn from the entire training data, which can boost overall accuracy as well as abstention performance. Our method requires training N models, which can be done in parallel, but, unlike ensembling, we have a single model for inference.

5. Experiments

5.1. Setup

Data. We require both ID and OOD data that has annotations available for evaluation. Therefore, we utilize the splits of the VQA v2 dataset [21] made available by [63] as our ID data. The entire VQA v2 train set (call it split **A**) is used for training VQA models (f). Meanwhile, the VQA v2 validation set is split into 3 parts: 86k examples (40%) for training selection functions g (call it split **B**); 22k examples (10%) for validating models; 106k examples (50%) as a test split for evaluating full selective models $h = (f, g)$. LYP does not require different sets for training f and g, so we train them both with the combination of A and B (**A+B**). For OOD data, we use AdVQA [50], which is an adversarial dataset constructed by asking human annotators to create questions that are difficult to answer for existing VQA models trained on VQA v2. The images in AdVQA and VQA v2 overlap with each other, so we only use images from AdVQA that appear in the test split. While AdVQA is not OOD in terms of the images, one can still consider this as adversarial, OOD since the questions are designed to fall outside the training distribution of VQA v2. This is similar to other OOD VQA datasets like VQA-CP [2], VQA-CE [11], or other VQA generalization benchmarks [3, 64]. However, for our setting, we create mixtures of VQA v2 and AdVQA to serve as our evaluation data, where each mixture contains different amounts of ID/OOD data.

VQA models. We use two different VQA architectures: **CLIP-ViL** [49], which is an ensemble of MCAN [67] and MoVie [43] with a CLIP [46] image encoder, and the recent **OFA** model [62], which is a transformer encoder-decoder model that performs multiple tasks and achieves state-of-the-art accuracy on VQA v2. For OFA, we explore 2 different sizes of the model: Base and Large. CLIP-ViL is a strong VQA model that treats VQA as a classification task over a large set of answers [56], while OFA is a large-scale pre-trained model that treats VQA as a generative task[1].

[1] While OFA is a generative model, it uses a trie-based decoding method for VQA that restricts the generated sequences to an answer vocabulary, as opposed to open-ended generation [62].

VQA Model f		Selection function g			Acc ↑	$\mathcal{C}@\mathcal{R}$ in % ↑			AUC ↓	Φ_1	Φ_{10}	Φ_{100}
Name	Train Set	Name	Train Set	Targets		$\mathcal{C}@1\%$	$\mathcal{C}@5\%$	$\mathcal{C}@10\%$				
CLIP-ViL	A	MaxProb	-	- [63]	66.35	0.00	24.16	43.53	13.55	49.12	14.39	-4.64
		Selector	B	Self [63]	66.35	12.69	31.12	46.96	12.47	50.36	20.15	5.22
	A+B	MaxProb	-	-	67.12	2.60	26.13	45.25	12.97	50.49	16.59	-0.93
		Selector	A+B	Self	67.12	2.97	26.70	46.19	12.80	50.89	18.19	-0.65
		Selector	A+B	LYP	67.12	**15.22**	**32.58**	**49.18**	**11.90**	**51.43**	**22.09**	**7.12**
OFA-Base	A	MaxProb	-	-	71.59	0.01	36.07	56.49	10.10	57.49	23.15	-0.34
		Selector	B	Self	71.59	18.32	41.48	59.74	9.19	57.97	27.22	9.09
	A+B	MaxProb	-	-	72.00	1.74	37.02	57.57	9.78	58.11	22.09	0.53
		Selector	A+B	Self	72.00	19.72	42.70	60.84	8.90	58.90	28.05	2.88
		Selector	A+B	LYP	72.00	**21.58**	**44.09**	**61.69**	**8.74**	**59.11**	**28.79**	**10.88**
OFA-Large	A	MaxProb	-	-	74.56	4.76	44.53	66.06	8.21	61.90	28.20	0.21
		Selector	B	Self	74.56	23.53	50.17	68.76	7.33	62.96	34.43	9.88
	A+B	MaxProb	-	-	74.79	1.30	43.70	65.95	8.26	62.24	27.09	-2.46
		Selector	A+B	Self	74.79	22.68	50.27	69.27	7.32	63.03	33.50	4.92
		Selector	A+B	LYP	74.79	**25.38**	**51.07**	**69.74**	**7.17**	**63.41**	**34.85**	**10.34**

Table 2. Results on the mixed ID/OOD scenario composed of 90% VQA v2 and 10% AdVQA examples.

Selection functions. We explore **MaxProb** [17, 22, 25, 29, 63] as a baseline as it is a natural comparison to the VQA model out-of-the-box since the confidence scores are simply the output probabilities of the model. We also employ the **Selector** developed by [63] as it attains the strongest performance for selective VQA. Selector is a two-layer MLP that takes in a combination of image, question, multimodal, and answer representations from the VQA model in order to predict a confidence score. We apply LYP to train Selector and compare to training with the original approach in [63] that utilizes held-out data. For each approach, we set a threshold on the output confidence scores to make abstention decisions (Sec. 3.1). Unless specified, we use by default $N = 10$ disjoint subsets to partition our A+B data.

All results for the strongest VQA model, OFA-Large, are averaged over 5 runs, while all other results are single runs. More experimental details are in the appendix.

5.2. In-Distribution Experiments

We first experiment with only in-distribution data to compare with prior work. Discussed in Sec. 3.1, we evaluate using maximum coverage at different risk levels ($\mathcal{C}@\mathcal{R}$), AUC for the risk-coverage curve, and effective reliability at different costs (Φ_c). We also present accuracy to give an idea of the question-answering performance of each model.

ID performance consistently improves with LYP. Tab. 1 shows that across all model architectures the top scores are achieved using LYP. For instance, we see improvements in $\mathcal{C}@1\%$ over both MaxProb (A+B) and Selector (B) with OFA-Large of 16.61% and 2.06%, respectively. Likewise, Φ_{100} increases with LYP by 12.49 and 1.20 over MaxProb (A+B) and Selector (B), respectively, for OFA-Large. The improvements are sustained at higher risk levels and lower costs (e.g., +0.63% $\mathcal{C}@10\%$ for Selector with LYP for CLIP-ViL compared to Selector trained on held-out data). These observations hold across each model we experiment with on ID data. Lastly, we see that all Selector models outperform all MaxProb models on every metric, just as in [63].

LYP helps VQA models and Selector learn from the same data. We see that training Selector and CLIP-ViL on the same data (A+B) performs poorly, achieving $\mathcal{C}@\mathcal{R}$ and Φ_c similar to its MaxProb counterpart. Conversely, the OFA models and Selector are able to be trained on the same data and reap the benefits of training on more data. We conjecture that this is due to the overfitting issue discussed in Sec. 4 as CLIP-ViL has a training accuracy of 87.40% whereas, e.g., OFA-Base has a training accuracy of 82.92% while also having higher accuracy on the test split. However, when using LYP, CLIP-ViL and Selector can be trained on the same data and improve beyond the model of [63] by, e.g., 2.61% $\mathcal{C}@1\%$. Further, although training on the same data can be done for the OFA models and Selector, it does not perform quite as well as when LYP is used. For example, with OFA-Base, training both the VQA model and Selector on A+B has $\mathcal{C}@1\%$ of 26.64% compared to 23.78% when the VQA model is trained on A and Selector is trained on B. Meanwhile, using LYP with OFA-Base attains 27.71% $\mathcal{C}@1\%$. Overall, these results suggest LYP helps better utilize the training data with Selector, improving ID performance.

5.3. OOD Evaluation

For our OOD evaluations, we build mixed datasets comprised of 10%, 33%, 50%, and 66% OOD examples. All mixtures contain 5K examples from AdVQA as OOD examples, and the rest are randomly sampled from the ID

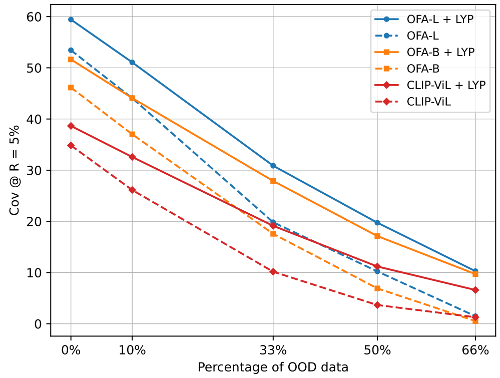

Figure 3. $\mathcal{C}$@5% for various mixtures of VQA v2 + AdVQA. OFA-L stands for Large, OFA-B for Base. Baseline is MaxProb.

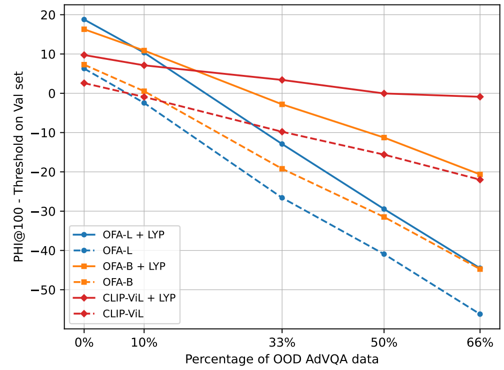

Figure 4. Φ_{100} for various mixtures of VQA v2 + AdVQA. OFA-L stands for Large, OFA-B for Base. Baseline is MaxProb.

	Acc.	$\mathcal{C}$@1%	$\mathcal{C}$@5%	$\mathcal{C}$@10%	AUC
Selector	71.25	19.05	41.83	59.55	9.29
Selector + KNN	71.25	19.92	41.78	59.75	9.27
Selector + SSD	71.25	18.99	41.90	59.27	9.27

Table 3. OOD Detection baselines. Scores are reported on the mixed ID/OOD data composed of 90% VQA v2 / 10% AdVQA.

VQA v2 test split. We report the results on the 10% OOD mixture in Tab. 2. More details, results (e.g., on other mixtures), and qualitative examples are in the appendix.

MaxProb can be overconfident on OOD data. Across all models, we see that MaxProb has <5% $\mathcal{C}$@1% and Φ_c scores <1. This suggests that MaxProb can be overconfident on OOD examples, on which the model is more likely to be incorrect. While improving the VQA accuracy of the model improves MaxProb performance, training a Selector still remains the most effective approach and consistently.

LYP maintains improvements over other methods in the 90%/10% setting. Similar to the pure ID setting, LYP continues to outperform other methods on the 90%/10% mixed setting as shown in Tab. 2. Although, we see decreases in all metrics across each of the different methods, demonstrating the challenge of this task even with just 10% OOD data.

The more OOD data, the more challenging. We display $\mathcal{C}$@5% and Φ_{100} for the various mixtures of ID/OOD data in Figs. 3 and 4. For both plots, we show each of the three models with a LYP-trained Selector versus the performance of MaxProb, each trained on the full A+B data. Across all OOD levels, LYP largely outperforms the baseline for all three models, for both $\mathcal{C}$@5% and Φ_{100} metrics. However, we observe that performances degrade quickly with a high OOD level. At the highest level (i.e., 33.3%/66.7% ID/OOD), all Maxprob models have <2% $\mathcal{C}$@5%, while LYP has around 10% coverage. For Φ_{100}, most models are below zero. We see that scaling alone is not sufficient to ensure high performances: while OFA-Large (MaxProb) has good performances on ID data, and is above OFA-Base + LYP, this is no longer true with OOD data. Our LYP Selector is effective at mitigating this loss in performance on OOD data. However, for OFA-Large, we note that the LYP-trained Selector is not always better than the Selector trained with held-out data for higher OOD levels. We discuss this and a mitigation strategy in the appendix. Combining the observations in Figs. 3 and 4, we see the potential performance that models could achieve, based on $\mathcal{C}@\mathcal{R}$ which is irrespective of the threshold chosen, versus the realized performance when choosing a threshold as one would do in practice, shown by Φ_{100}. This shows that more work is needed to help generalize to such OOD data.

OOD detection features do not necessarily help. Inspired by [15], we train Selector with out-of-distribution detection scores computed with KNN [54] or SSD [48] as added features. We find that these features do not bring significant improvements to our evaluation metrics (Tab. 3). More details about those experiments can be found in the appendix.

Augmenting Selector training with known OOD data also does not improve. As discussed in Sec. 3.3, we also try training Selector on the B set, along with some known OOD datasets similar to [29]. This may help learn to discard hard examples which are very far from its training distribution. For this experiment, we use the training sets of OK-VQA [41], which has the same image distribution but a different question distribution, and of VizWiz [23], which has both image and question distribution shifts compared to VQA v2. We see in Tab. 4 that this method is not very successful at improving selective prediciton in this OOD evaluation setting. Contrary to the findings of [29] for text-only question answering, on the Selective VQA task, adding this known OOD data during training decreases the performance of our Selector on unknown OOD data at test time. Overall, it appears that more traditional approaches for handling OOD examples may have difficulty generalizing to this multimodal setting.

Model	Selector	$\mathcal{C}$@1%	$\mathcal{C}$@5%	$\mathcal{C}$@10%	AUC
90% VQA v2, 10% AdVQA					
A	B	19.00	41.64	58.97	9.34
A	B + OOD	18.48	41.08	59.40	9.36
50% VQA v2, 50% AdVQA					
A	B	2.68	15.98	26.72	18.97
A	B + OOD	2.56	14.93	26.82	19.08

Table 4. Results with exposure to known OOD data for OFA-Base.

N	$\mathcal{C}$@1%	$\mathcal{C}$@5%	$\mathcal{C}$@10%	AUC
10	**27.71**	**51.64**	**70.20**	**6.98**
2	27.64	51.24	70.12	7.01

Table 5. Varying the number of splits N for LYP. Results are reported on the ID test split for OFA-Base, trained on A+B, with a selector trained on A+B.

5.4. Further Analysis

In addition to the following, we have more qualitative results, analysis, and evaluation on other tasks (visual entailment [66]) and datasets (VizWiz [23]) in the appendix.

Different numbers of splits/peers for LYP. We ablate the number of splits/peer models N of the training data $\mathcal{D}$ with OFA-Base. Ideally, the peer models in LYP should have predictions and failure modes similar to the full model, which suggests that more peer models may be better (i.e., each peer model is trained on more data). We see in Tab. 5 that for OFA-Base the number of peers has a small impact on the final results. However, we also find that the difference in accuracy of OFA-Base fine-tuned on 50% vs 100% of the training data is small (74.03% vs 75.18%). Therefore, the difference in signal from the labels of 2 vs 10 peers may be similar. For VQA models with less pre-training, this difference may be greater and more models may be needed. This suggests that the training requirements for LYP can be reduced while maintaining strong performance for large pre-trained models.

Effect of training data size. We show in Tab. 6 that the amount of data used for the Selector training is an important factor for its performance. Note that the Train B set, used by [63] to train their selector, has 86K examples, which is ~15% of the full Train A+B. The additional data, labeled with LYP, helps Selector generalize better to test examples.

Impact of scaling on selective prediction. Tab. 7 shows results for three OFA sizes: Medium, Base, and Large. We see that larger models, in addition to having a much higher accuracy on the testing set, have much better ID selective prediction performance when paired with a trained Selector.

% of A+B	$\mathcal{C}$@1%	$\mathcal{C}$@5%	$\mathcal{C}$@10%	AUC
100	**27.71**	**51.64**	70.20	**6.98**
75	27.48	51.11	**70.26**	7.01
50	26.84	51.04	70.04	7.06
25	26.03	50.15	69.65	7.16
10	23.30	47.97	68.03	7.44
5	22.62	46.10	66.10	7.71

Table 6. Varying the amount of training data for the Selector with LYP. The model is OFA-Base and results are on the ID test split.

Model	Method	Acc.	$\mathcal{C}$@1%	$\mathcal{C}$@5%	$\mathcal{C}$@10%	AUC
Med.	MaxProb	71.30	5.08	37.56	56.85	9.95
Base	MaxProb	74.70	3.45	45.60	66.61	7.99
Large	MaxProb	77.79	20.57	53.99	75.18	6.42
Med.	LYP	71.30	19.69	41.28	59.60	9.17
Base	LYP	74.70	27.71	51.64	70.20	6.98
Large	LYP	77.79	32.92	59.43	77.52	5.60

Table 7. Scaling results for OFA Medium (93M params), Base (180M params), and Large (470M params) on the ID test split.

6. Conclusions

This is the first work to explore Selective Visual Question Answering in the realistic, and challenging, mixed ID+OOD scenario, where a model is exposed to samples from both the training distribution and also out-of-distribution (OOD) examples. We find that out-of-the-box, state-of-the-art VQA models [49,62] largely fail on this task at a low risk of error (e.g., 1%). When training a multimodal Selector [63] models significantly improve their abstention decisions, matching observations in the in-distribution (ID) scenario. However, a limitation of the multimodal Selector training is that it requires splitting the training data between the VQA model training and the Selector training to avoid over-fitting on the training data. In this work, we address this with our approach *Learning from Your Peers* (LYP), which allows us to train both the VQA model and the Selector on the full training data. We find that in the ID scenario as well as the mixed scenario of 90%/10% ID/OOD data, LYP consistently performs best across all VQA models and metrics, improving over baselines and prior work. Our best result doubles the $\mathcal{C}$@1% over prior work [63]. Overall, all models still have difficulties recognizing when they cannot answer OOD examples correctly and thus decrease in performance when the percentage of OOD data increases. Interestingly, we observe that the better a VQA model is ID, the more it loses if it has to also generalize the threshold for abstention from ID to OOD (as measured by Effective Reliability Φ_c). Thus, major challenges remain, both for improving the generalizing abilities of VQA models to OOD examples (i.e., answering OOD questions correctly) as well as identifying examples that the model cannot answer, whether they are in- or out-of-distribution.

References

[1] Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, Xiaochun Cao, Abbas Khosravi, U Rajendra Acharya, et al. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion*, 76:243–297, 2021. 3

[2] Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. Don't just assume; look and answer: Overcoming priors for visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4971–4980, 2018. 2, 5

[3] Aishwarya Agrawal, Ivana Kajić, Emanuele Bugliarello, Elnaz Davoodi, Anita Gergely, Phil Blunsom, and Aida Nematzadeh. Rethinking evaluation practices in visual question answering: A case study on out-of-distribution generalization. *arXiv preprint arXiv:2205.12191*, 2022. 2, 5

[4] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 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*, pages 6077–6086, 2018. 2

[5] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015. 2, 3

[6] Liron Bergman, Niv Cohen, and Yedid Hoshen. Deep nearest neighbor anomaly detection. *arXiv preprint arXiv:2002.10445*, 2020. 3

[7] Remi Cadene, Corentin Dancette, Matthieu Cord, Devi Parikh, et al. Rubi: Reducing unimodal biases for visual question answering. *Advances in neural information processing systems*, 32, 2019. 2

[8] C Chow. On optimum recognition error and reject tradeoff. *IEEE Transactions on information theory*, 16(1):41–46, 1970. 3

[9] Chi-Keung Chow. An optimum character recognition system using decision functions. *IRE Transactions on Electronic Computers*, EC-6(4):247–254, 1957. 1

[10] Charles Corbière, Nicolas Thome, Avner Bar-Hen, Matthieu Cord, and Patrick Pérez. Addressing failure prediction by learning model confidence. *Advances in Neural Information Processing Systems*, 32, 2019. 3

[11] Corentin Dancette, Remi Cadene, Damien Teney, and Matthieu Cord. Beyond question-based biases: Assessing multimodal shortcut learning in visual question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1574–1583, 2021. 2, 5

[12] C. De Stefano, C. Sansone, and M. Vento. 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. 3

[13] Alex J DeGrave, Joseph D Janizek, and Su-In Lee. Ai for radiographic covid-19 detection selects shortcuts over signal. *Nature Machine Intelligence*, 3(7):610–619, 2021. 1

[14] Ran El-Yaniv and Yair Wiener. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11:1605–1641, 2010. 1, 3

[15] Adam Fisch, Tommi Jaakkola, and Regina Barzilay. Calibrated selective classification. *arXiv preprint arXiv:2208.12084*, 2022. 7

[16] Jakob Gawlikowski, Cedrique Rovile Njieutcheu Tassi, Mohsin Ali, Jongseok Lee, Matthias Humt, Jianxiang Feng, Anna Kruspe, Rudolph Triebel, Peter Jung, Ribana Roscher, et al. A survey of uncertainty in deep neural networks. *arXiv preprint arXiv:2107.03342*, 2021. 3

[17] Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. *Advances in neural information processing systems*, 30, 2017. 3, 6

[18] Yonatan Geifman and Ran El-Yaniv. Selectivenet: A deep neural network with an integrated reject option. In *International Conference on Machine Learning*, pages 2151–2159. PMLR, 2019. 3, 4

[19] Atticus Geiger, Ignacio Cases, Lauri Karttunen, and Christopher Potts. Posing fair generalization tasks for natural language inference. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4485–4495, Hong Kong, China, Nov. 2019. Association for Computational Linguistics. 2

[20] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020. 1

[21] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6904–6913, 2017. 2, 3, 5

[22] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, pages 1321–1330. PMLR, 2017. 6

[23] Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3608–3617, 2018. 1, 2, 3, 7, 8

[24] Blaise Hanczar and Edward R Dougherty. Classification with reject option in gene expression data. *Bioinformatics*, 24(17):1889–1895, 2008. 3

[25] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *Proceedings of International Conference on Learning Representations*, 2017. 3, 6

[26] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019. 2

[27] Huaizu Jiang, Ishan Misra, Marcus Rohrbach, Erik Learned-Miller, and Xinlei Chen. In defense of grid features for visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10267–10276, 2020. 2

[28] Yu Jiang, Vivek Natarajan, Xinlei Chen, Marcus Rohrbach, Dhruv Batra, and Devi Parikh. Pythia v0. 1: the winning entry to the vqa challenge 2018. *arXiv preprint arXiv:1807.09956*, 2018. 2

[29] Amita Kamath, Robin Jia, and Percy Liang. Selective question answering under domain shift. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5684–5696, Online, July 2020. Association for Computational Linguistics. 2, 3, 4, 6, 7

[30] Siddharth Karamcheti, Ranjay Krishna, Li Fei-Fei, and Christopher Manning. 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)*, pages 7265–7281, Online, Aug. 2021. Association for Computational Linguistics. 4

[31] Corentin Kervadec, Grigory Antipov, Moez Baccouche, and Christian Wolf. Roses are red, violets are blue... but should vqa expect them to? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2776–2785, 2021. 2

[32] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in neural information processing systems*, volume 30, 2017. 3

[33] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in neural information processing systems*, 31, 2018. 3

[34] Linjie Li, Jie Lei, Zhe Gan, and Jingjing Liu. Adversarial vqa: A new benchmark for evaluating the robustness of vqa models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2042–2051, 2021. 2

[35] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. In *Arxiv*, 2019. 2

[36] Shiyu Liang, Yixuan Li, and Rayadurgam Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. *arXiv preprint arXiv:1706.02690*, 2017. 3

[37] Ziqian Lin, Sreya Dutta Roy, and Yixuan Li. Mood: Multi-level out-of-distribution detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15313–15323, 2021. 3

[38] Evan Z Liu, Behzad Haghgoo, Annie S Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. Just train twice: Improving group robustness without training group information. In *International Conference on Machine Learning*, pages 6781–6792. PMLR, 2021. 4

[39] Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. *Advances in Neural Information Processing Systems*, 33:21464–21475, 2020. 3

[40] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019. 2

[41] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 3195–3204, 2019. 2, 7

[42] Gustav Mårtensson, Daniel Ferreira, Tobias Granberg, Lena Cavallin, Ketil Oppedal, Alessandro Padovani, Irena Rektorova, Laura Bonanni, Matteo Pardini, Milica G Kramberger, et al. The reliability of a deep learning model in clinical out-of-distribution mri data: a multicohort study. *Medical Image Analysis*, 66:101714, 2020. 1

[43] Duy-Kien Nguyen, Vedanuj Goswami, and Xinlei Chen. Movie: Revisiting modulated convolutions for visual counting and beyond. In *Proceedings of the International Conference on Learning Representations*, 2021. 5

[44] Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, David Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in neural information processing systems*, 32, 2019. 3

[45] P. Pudil, J. Novovicova, S. Blaha, and J. Kittler. Multistage pattern recognition with reject option. In *Proceedings., 11th IAPR International Conference on Pattern Recognition. Vol.II. Conference B: Pattern Recognition Methodology and Systems*, pages 92–95, 1992. 3

[46] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 5

[47] Sainandan Ramakrishnan, Aishwarya Agrawal, and Stefan Lee. Overcoming language priors in visual question answering with adversarial regularization. *Advances in Neural Information Processing Systems*, 31, 2018. 2

[48] Vikash Sehwag, Mung Chiang, and Prateek Mittal. Ssd: A unified framework for self-supervised outlier detection. *arXiv preprint arXiv:2103.12051*, 2021. 7

[49] Sheng Shen, Liunian Harold Li, Hao Tan, Mohit Bansal, Anna Rohrbach, Kai-Wei Chang, Zhewei Yao, and Kurt Keutzer. How much can clip benefit vision-and-language tasks? *arXiv preprint arXiv:2107.06383*, 2021. 2, 5, 8

[50] Sasha Sheng, Amanpreet Singh, Vedanuj Goswami, Jose Magana, Tristan Thrush, Wojciech Galuba, Devi Parikh, and Douwe Kiela. Human-adversarial visual question answering. *Advances in Neural Information Processing Systems*, 34:20346–20359, 2021. 2, 5

[51] Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. Textcaps: a dataset for image captioning with reading comprehension. In *European conference on computer vision*, pages 742–758. Springer, 2020. 1

[52] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326, 2019. 2

[53] Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-distribution detection with deep nearest neighbors. *arXiv preprint arXiv:2204.06507*, 2022. 3

[54] Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-distribution detection with deep nearest neighbors. *arXiv preprint arXiv:2204.06507*, 2022. 7

[55] Hao Tan and Mohit Bansal. 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*, pages 5100–5111, 2019. 2

[56] Damien Teney, Peter Anderson, Xiaodong He, and Anton Van Den Hengel. Tips and tricks for visual question answering: Learnings from the 2017 challenge. In *CVPR*, 2018. 5

[57] Jing Tian, Michael H Azarian, and Michael Pecht. Anomaly detection using self-organizing maps-based k-nearest neighbor algorithm. In *PHM Society European Conference*, volume 2, 2014. 3

[58] Dustin Tran, Jeremiah Liu, Michael W. Dusenberry, Du Phan, Mark Collier, Jie Ren, Kehang Han, Zi Wang, Zelda Mariet, Huiyi Hu, Neil Band, Tim G. J. Rudner, Karan Singhal, Zachary Nado, Joost van Amersfoort, Andreas Kirsch, Rodolphe Jenatton, Nithum Thain, Honglin Yuan, Kelly Buchanan, Kevin Murphy, D. Sculley, Yarin Gal, Zoubin Ghahramani, Jasper Snoek, and Balaji Lakshminarayanan. Plex: Towards reliability using pretrained large model extensions, 2022. 2

[59] Neeraj Varshney, Swaroop Mishra, and Chitta Baral. Investigating selective prediction approaches across several tasks in IID, OOD, and adversarial settings. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1995–2002, 2022. 2, 3

[60] Haoqi Wang, Zhizhong Li, Litong Feng, and Wayne Zhang. Vim: Out-of-distribution with virtual-logit matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4921–4930, 2022. 3

[61] Haoran Wang, Weitang Liu, Alex Bocchieri, and Yixuan Li. Can multi-label classification networks know what they don't know? *Advances in Neural Information Processing Systems*, 34:29074–29087, 2021. 3

[62] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International Conference on Machine Learning*, pages 23318–23340. PMLR, 2022. 1, 2, 5, 8

[63] Spencer Whitehead, Suzanne Petryk, Vedaad Shakib, Joseph Gonzalez, Trevor Darrell, Anna Rohrbach, and Marcus Rohrbach. Reliable visual question answering: Abstain rather than answer incorrectly. *arXiv preprint arXiv:2204.13631*, 2022. 1, 2, 3, 4, 5, 6, 8

[64] Spencer Whitehead, Hui Wu, Heng Ji, Rogerio Feris, and Kate Saenko. Separating skills and concepts for novel visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5632–5641, 2021. 5

[65] David H Wolpert. Stacked generalization. *Neural networks*, 5(2):241–259, 1992. 4

[66] Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. Visual entailment: A novel task for fine-grained image understanding. *arXiv preprint arXiv:1901.06706*, 2019. 8

[67] Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. Deep modular co-attention networks for visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6281–6290, 2019. 5