

Benchmarking Out-of-Distribution Detection in Visual Question Answering

Xiangxi Shi

Oregon State University

shixia@oregonstate.edu

Stefan Lee

Oregon State University

leestef@oregonstate.edu

Abstract

When faced with an out-of-distribution (OOD) question or image, visual question answering (VQA) systems may provide unreliable answers. If relied on by real users or secondary systems, these failures may range from annoying to potentially endangering. Detecting OOD samples in single-modality settings is well-studied; however, limited attention has been paid to vision-and-language settings. In this work, we examine the question of OOD detection in the multimodal VQA task and benchmark a suite of approaches to identify OOD image-question pairs. In our experiments, we leverage popular VQA datasets to benchmark detection performance across a range of difficulties. We also produce composite datasets to examine impacts of individual modalities and of image-question agreement. Our results show that answer confidence alone is often a poor signal and that methods based on image-based question generation or examining model attention can lead to significantly better results. We find detecting ungrounded image-question pairs and small shifts in image distribution remain challenging.

1. Introduction

While the visual question answering (VQA) task is fairly open-ended and recent techniques have gained increasingly strong performance, their competency is typically restricted to the concepts and language seen during training – that is to say for in-distribution samples. On out-of-distribution examples where either the question or the image does not resemble the training set, VQA models may provide unreliable responses [53]. In real applications, these responses might range from annoying to potentially endangering – especially for potential use-cases involving the visually impaired. If these out-of-distribution (OOD) samples could be reliably identified, then the model could instead abstain from answering [5, 9, 18]. Currently, OOD detection in deep networks remains an active area of research even in single-modality settings like image or text classification, and few attempts have focused on multi-modality



Figure 1. In the wild, VQA models are likely to encounter out-of-distribution (OOD) image-question pairs that are not well-represented in their training sets due to (A) novel visuals, (C) novel language, or (B) novel combinations of in-distribution images and question. In this work, we study OOD detection in VQA.

tasks [27, 34]. In this work, we explore the problem of out-of-distribution detection in the multimodal visual question answering (VQA) setting by benchmarking several approaches to out-of-distribution (OOD) detection across different model architectures and popular datasets.

Given a training set of image-question pairs sampled from some joint distribution $P_{in}(I, Q)$, a new test-time sample may be out-of-distribution in a number of ways. It may contain a novel image i' that is not represented in the marginal $P_{in}(I)$ – i.e., a visual novelty. Likewise, it may contain a novel question q' that lacks support under $P_{in}(Q)$ – i.e., a linguistic novelty. Alternatively, both i' and q' may be in-distribution with respect to their corresponding marginals but are a novel combination without support under $P_{in}(I, Q)$ – i.e., a combination novelty. To examine these cases, we construct a benchmark for VQA OOD detection from six popular VQA datasets – taking VQAv2 [10] as the in-distribution set and evaluating OOD detection on samples from VizWiz [11], GQA [14], CLEVR [17], VQA Abstract Scenes [3] and QRPE [34]. In contrast to prior work [27], these OOD samples are image-question pairs and are drawn from data sources that range from very differ-

ent (CLEVR) to quite similar (GQA) to the in-distribution set. Taken together, these OOD settings allow us to examine OOD detection methods for VQA under a wide range of visual, linguistic, and combination novelties.

We examine unimodal, VQA model-based, and image-based question generation approaches to detecting novel image-question pairs – comparing common OOD baseline approaches. In total, we study twenty-eight different model/method configurations and find that question-generation-based OOD scores can result in strong performance in most settings. We analyze the impact of backbone task performance and pretraining for VQA model-based and question generation methods. All methods tested still have difficulty detecting subtle shifts in image distribution and novel combinations of known images and questions.

2. Related Work

Visual Question Answering (VQA). Proposed by Antol et al. [3], VQA tasks systems with providing answers to natural language questions about images. The topic has received significant interest with many follow-up datasets [3, 11, 14, 17] and increasingly powerful methods [2, 20, 50, 52, 55]. In this work, we are concerned with identifying out-of-distribution (OOD) question-image pairs that are sufficiently different from a source dataset. To characterize OOD detection performance, we sample representative VQA models. Epitomized by Anderson et al. [2], early VQA models were relatively small networks that performed question-guided image attention before fusing image and question representations. As transformer-based attention mechanisms [42] gained ground, many models began incorporating them as part of more complex attention schemes [50]. Recently, a contingent of methods pretrain large transformer-based methods on vision-and-language data from the web to learn broadly useful features [33, 52, 55] before being fine-tuned on the downstream VQA task. For our analysis, we examine [2], [50], and [52].

Out-of-Distribution Detection (OOD). OOD detection has been studied in computer vision [13, 22, 28, 30, 43, 51, 54] and natural language processing [21, 26, 31, 38] extensively, with much of the work focusing on classification tasks. These unimodal techniques often focus on predicted confidence [13, 28, 43], energy-based models [30], Bayesian methods [22], data density estimation, or reconstruction error [38, 51] to provide anomaly scores indicating if a given datapoint is OOD. Work in this space can be further divided by whether out-of-distribution samples are available during training or tuning (e.g. outlier exposure). In this work, we consider the OOD detection problem without outlier exposure in a complex multimodal setting.

OOD Detection in Visual Question Answering. Some prior work has examined specific types of out-of-

distribution samples. Mahendru et al. [34] examined detecting ungrounded samples where both the image and question were in-distribution but the question mentioned concepts not present in the image. Likewise, samples in the VizWiz dataset [11] may be unanswerable for multiple reasons [4] including severe image blur or image content not aligning with the question. We include a similar ungrounded setting in our benchmark. In both of these settings, a dataset of ungrounded (or unanswerable) image-question pairs is provided and leveraged to train a supervised model [4, 34]. In contrast, we consider the setting without outlier exposure.

Most related to our work, Lee et al. [27] examine a more general OOD detection setting – creating synthetic OOD examples by replacing images or questions in image-question pairs. Specifically, Lee et al. used images from low-resolution image classification datasets like MNIST and non-question sentences dataset like IMDB. The synthetic distribution shifts may be too severe to adequately assess the shifts in real-world application. In their setting, [27] demonstrate that answer confidence alone may not be a strong signal and proposed the average maximum attention probability (MAP) method which was shown to be effective for simple settings with outlier exposure. In this work, we consider OOD in VQA by bridging across different VQA datasets – ensuring the OOD sets contain grounded image-question pairs. Further, we do not use any outlier exposure to set parameters as in [27].

Selective Prediction in VQA. In selective prediction setting, a model may abstain from samples for which its answer is likely to be wrong [8, 16, 18, 46, 49]. A selective prediction model consists of a learned function for the target task and a prediction selector to determine if a sample should be abstained. Whitehead et al. [45] explore the problem in the VQA domain with in-distribution data. Dancette et al. [5] proposed to split training data into subsets and evaluate them with the VQA backbone trained with the complement to achieve a better generalization for the prediction selector. They then evaluate their method on OOD data to measure generalization. Compared to selective prediction, we consider OOD detection in general but note that our approach could be paired with a VQA model to perform the selective prediction task.

Robustness to Shortcut Learning in VQA. When associations between questions and their answers are very consistent in training, prior work has shown that VQA models may ignore the image content entirely [1, 19, 41]. For instance, if the vast majority of images contain green grass (rather than yellowed) during training, then VQA models may learn to “shortcut” the actual visual reasoning and reply to “What color is the grass?” with “green” regardless of the image content. Benchmarks like VQA-CP [1] and GQA-OOD [19] that intentionally amplify these answer-bias effects have been developed to study shortcut learn-

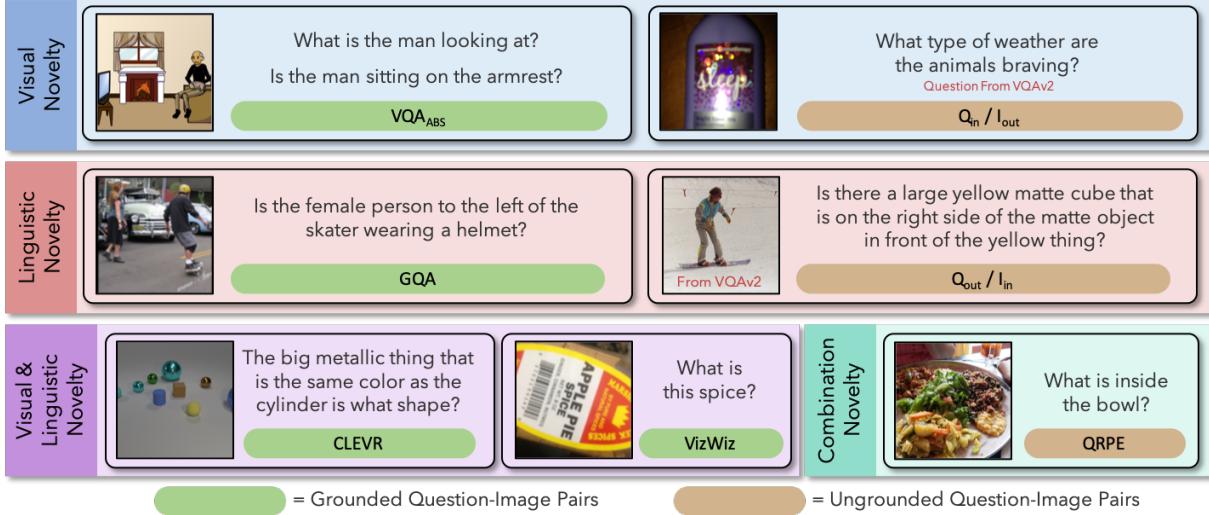


Figure 2. Novelty types in our benchmark, example out-of-distribution instances, and associated datasets. We note whether a dataset provided “grounded” pairs – i.e., whether sample question appropriately refer to entities in their associated image.

ing in VQA. These benchmarks represent distribution shifts where concepts seen during training are presented at different rates during testing – e.g. yellowed grass being rare-but-present in training and frequent in test. Unlike our setting, this means that there is not a notion of any individual question-image pair being out-of-distribution for these benchmarks. Further, approaches developed for these tasks have focused on avoiding shortcut learning due to answer bias rather than detecting the distributional shift.

3. Out-of-Distribution Detection in VQA

Given a set of question-image examples D_{in} drawn from a distribution $P_{in}(Q, I)$, we are interested in mechanisms to determine whether a new sample (q, i) is also drawn from P_{in} or from some unknown distribution P_{out} . A typical paradigm for OOD detection tasks is to define some scoring function $f(q, i)$ based on D_{in} such that in-distribution and out-of-distribution samples receive different scores. In this benchmark, we evaluate a set of scoring functions across a range of different definitions of D_{out} with varying similarity to the in-distribution dataset. In this section, we define the experimental setting and demonstrate the effect of out-of-distribution data on answer quality. We describe our scoring functions in the following section.

3.1. Benchmark Datasets

We consider VQAv2 [10] as the in-distribution dataset and five other visual question answering datasets in our experiments – GQA [14], VizWiz [11], CLEVR [17], VQA Abstract Scene [3] and QRPE [34] to compose our OOD validation sets. An overview of the associated novelty types that we explore is shown in Fig. 2 above.

In-Distribution Dataset. We take the widely-used VQAv2 [10] dataset as our in-distribution set. VQAv2 consists of over 1 million question-answer pairs based on $\sim 205,000$ images from COCO [29]. By construction, these are consumer photographs containing at least one instance of 80 common objects including vehicles, animals, foods, household objects, road features, and sporting equipment. The images tend to be of real scenes, well-framed, and not blurry. Questions were generated via crowd-sourcing. Workers were provided an image and prompted to ask a question about the image that a human could easily answer but a smart robot might not be able to answer. We denote the training set as D_{in}^{train} and test-standard set as D_{in}^{test} .

Out-of-Distribution Sets. We draw OOD samples from the validation sets of the remaining datasets to represent different novelty types. Samples are shown in Fig. 2. We describe each dataset and its relation to VQAv2 below:

- **GQA** [14] consists of synthetic questions paired with natural images. The images are a subset of those from VQA and thus there is no visual novelty relative to VQA. The questions however are generated using a probabilistic grammar based on annotations of objects, attributes, and relationships depicted in the images. Compared with VQA, GQA questions are more detailed and focus on relations and attributes, i.e., “to the left of” and “small”, which are not frequently used in VQA questions.
- **VQA Abstract Scenes (VQA_{ABS})** [3] consists of cartoon images and human-annotated questions. The images are composed by humans from a dictionary of clipart and are significantly different from VQA images. However, VQA_{ABS} questions are quite similar to VQA, having been

collected in the same fashion under similar instructions.

- **VizWiz** [11] consists of questions asked by visually-impaired users via a cellphone application. These images are often blurry, low-resolution, or do not frame their content well. The questions focus heavily on identifying objects or writing on objects in the scene. As such VizWiz represents a departure in both image and question content compared to VQA.
- **CLEVR** [17] is a fully synthetic dataset consisting of rendered images and automatically generated questions. The images are generated programmatically as a collection of geometric primitives with limited types, sizes and textures in a gray scene. Like GQA, CLEVR questions are generated via a grammar based on image contents. CLEVR questions tend to be longer and focus on compositions of reasoning skills. As such, CLEVR differs significantly from VQA in image and question content.
- **QRPE** [34] is a VQA-based dataset constructed such that questions refer to entities not present in the images but are generally plausible given the depicted scene. As both questions and images come from VQAv2, QRPE samples are examples of challenging combination novelties.
- **I_{In}/Q_{Out}**, and **I_{Out}/Q_{In}**. We construct pairs where either the question or image from in-distribution VQA samples is replaced by a random question or image from one of the out-of-distribution datasets. These are useful to examine if the in-distribution component acts as a distractor or if novelty in one modality is easier to detect.

Evaluation and Metrics. For each D_{in}^{test}, D_{out} pair, we sample 50,000 examples from each to form a combined evaluation set. We will provide these instance indexes upon acceptance to aid replicability. Each question-answer pair is evaluated by the score function and higher values are interpreted as evidence of being in-distribution. Following existing work in unimodal OOD detection [22, 28, 30, 54], we report Area Under the Curve of Receiver Operating Characteristic (AUCROC), to compare performance of different methods. The scale of AUC ROC is $[0, 1]$ and larger AUC implies better performance.

3.2. Effect of OOD Data on VQA Task Performance

To further motivate our study of OOD detection, we report a cross-domain evaluation of an X-VLM [52] model trained on the VQA dataset. Without any finetuning, we evaluate the VQA-trained model on the GQA and CLEVR datasets and present results in Table. 1 (row 1). While the model performs well in the in-distribution VQA evaluation, performance is significantly degraded for GQA and CLEVR. These drops cannot just be explained by a change in dataset difficulty either, because the X-VLM

Methods	VQAv2	GQA	CLEVR
1 X-VLM (VQAv2 train)	78.4	55.8	32.9
2 In-Distribution SOTA	84.0 [44]	73.6 [35]	99.8 [48]

Table 1. Overall accuracy of X-VLM model tested on OOD datasets (grey). We find significant performance degradation.

cross-domain performance significantly underperforms in-distribution trained models (row 2). These results corroborate similar drops in performance identified in [53] while studying cross-domain transfer for VQA. This suggests out-of-distribution image-question pairs reduce the reliability of VQA models; however, if models were equipped with an OOD detection mechanism, it could possibly abstain from answering. In the following section, we explore a range of OOD models and techniques for VQA.

4. Out-of-Distribution Detection Methods

To perform OOD detection, we consider different scoring functions $f(q, i)$ in four broad categories – density-based, reconstruction-based, prediction-based, and feature-based scoring functions – reflecting common directions in OOD methods [47]. We take as convention that high $f(q, i)$ scores indicate that a sample is likely to be in-distribution. We overview these methods below and provide full details of each model in the supplementary materials.

4.1. Density-based Scoring Functions

Directly fitting a distribution to D_{in} is a straight-forward approach to OOD detection. Given the difficulties of fitting image density models, we do not present a model of $P_{in}(I)$ and instead target just the question distribution.

Language Model (LangM). We train a simple language model $P_{\text{Transf}}(q)$ that approximates the question distribution $P_{in}(Q)$ with a 4-layer Transformer. This model is trained via cross-entropy loss to mimic in-distribution questions. We take the score function of question q to be

$$f_{\text{LangM}}(q, i) = \sqrt[|q|]{P_{\text{Transf}}(q)} \quad (1)$$

where $|q|$ denotes the length of q . This a geometric mean of per-word conditional probabilities and is preferred over taking P_{Transf} directly to reduce the bias against longer sentences which tends to occur in language models.

Image-to-Question Captioning Model (I2Q). To jointly model both modalities, we also fit an image-to-question (essentially a captioning model) which approximates $P_{in}(Q|I)$. We develop a transformer-based encoder-decoder model [42] that produces a question given spatial grid features from a ResNet101 [12] model pretrained on ImageNet [6]. This model is trained via cross-entropy loss to mimic in-distribution questions given the corresponding

image. For a new (q, i) pair, we compute a score as the geometric mean probability of q given a certain image i :

$$f_{\text{f2Q}}(q, i) = \sqrt[|q|]{p(q|i)} \quad (2)$$

4.2. Reconstruction-based Scoring Functions

Reconstruction-based methods assume that autoencoder style models trained on in-distribution methods are likely to reconstruct out-of-distribution inputs poorly.

RIAD [51] masks portions of an input image and then measures the error in image inpainting. A U-Net-based [37] image inpainting network $U_{\text{NET}}(\cdot)$ is trained to reproduce masked out regions in images from D_{in} . At evaluation, a new image is masked and the negative mean squared error of the reconstruction is the basis for the OOD score, i.e.

$$f_{\text{RIAD}}(q, i) = -\|i - U_{\text{NET}}(m \odot i)\|_2^2. \quad (3)$$

where m is a random mask and \odot an element-wise product.

Language Variational Autoencoder (Lang-VAE). For questions, we train a transformer-based Gaussian VAE. Inspired by Jiang et al. [15], the model encodes each token of a question q to a distribution over latent states $p(z_i|q)$ and then decodes the feature sequence $\mathbf{Z}' \sim p(\mathbf{Z}|q)$ back to q . The model is trained using the standard ELBO-based VAE objective [23]. For a given (q, i) pair, we measure the geometric mean of probability of reconstructing q as the score:

$$f_{\text{LangVAE}}(q, i) = \sqrt[|q|]{p(q|\mathbf{Z}')} \quad (4)$$

4.3. Prediction-based Scoring Functions

Maximum Softmax Probability (MSP). A common OOD score in classification tasks is simply to consider the predicted confidence of the classifier. For VQA, this means taking the probability of the predicted answer of a VQA model $M_{\text{VQA}}(\cdot)$ as the scoring function,

$$f_{\text{MSP}}(q, i) = \max_a M_{\text{VQA}}(a|q, i) \quad (5)$$

While simple, the maximum softmax probability (MSP) has proven surprisingly effective in image-based OOD detection when the base model is highly performant on the in-distribution task – even outperforming more complex methods [43]. In our experiments with MSP, we consider three VQA models described in the following section.

4.4. Feature-based Scoring Functions

Feature-base scores use intermediate feature representations of a trained model to score a sample; for instance, computing distances to in-distribution samples in some encoding of a pretrained network.

Negative Mahalanobis Distance (Maha). Given some feature encoder $z = e(\cdot)$ we can estimate a mean μ and covariance matrix Σ using data from D_{in}^{train} and then compute a negative Mahalanobis distance as

$$f_{\text{Maha}}(z_{test}) = -\sqrt{(z_{test} - \mu)^T \Sigma^{-1} (z_{test} - \mu)}, \quad (6)$$

where z_{test} is the representation of a given image, question, or image-question pair from $e(\cdot)$. We use Maha-L, Maha-V, and Maha-X to refer to distances computed from language, vision, or a multimodal embeddings.

Average Maximum Attention Probability (MAP) [27] considers intermediate states of a VQA model and is only applicable to models using attention mechanisms. Without loss of generality, an attention-augmented model contains some number of attention operations that given a set of k input feature vectors produces a k -dimensional attention distribution A . These features may represent image regions or question tokens and may be conditioned on some context vector. In the case of modern transformer-based models, there are many such distributions due to the many self-attention layers with multiple attention heads. Proposed in [27], the MAP score for a network with n cross-modal attention distributions (image \leftrightarrow query) $A_1(q, i), \dots, A_n(q, i)$ is computed as the average maximum attention probability across the attention distributions,

$$f_{\text{MAP}}(q, i) = \frac{1}{n} \sum_{j=1}^n \max A_j(q, i). \quad (7)$$

In our experiments, we refer to the original MAP score as MAP-X and extend it to Language-only/Vision-only variants MAP-L/MAP-V, where MAP is calculated from only question / image self-attention modules. We also denote the average of these three as MAP-A.

5. Benchmarking VQA OOD Detection

We apply the scoring functions from Section 4 to the datasets from Section 3 – taking VQAv2 as our in-distribution dataset and all others as out-of-distribution sets. Using testing splits, we report area under the ROC curve (AUCROC) for in- vs. out-of-distribution detection as our metric by thresholding the scoring function outputs.

Base Models. Some of our score functions require pre-trained VQA models or feature encoders (MSP, MAP, Maha). We examine a range of three VQA models of differing task performance and complexity:

- **BUTD** [2] (2017) is a representative pre-transformer era VQA model that deploys only a single round of cross-modal attention between LSTM-encoded question features and region features from a pretrained object detector. We use a VinVL-based object detection model

[55] pretrained on COCO [29], OpenImages [24], Objects365 [39], and Visual Genome [25] rather than the original BUTD features. The implementation we use achieves 66.98% VQA-accuracy on VQAv2 test-dev.

- **MCAN** [50] (2019) is an early transformer-based model. Images are encoded as in BUTD, but features are processed by multiple rounds of modality-specific self-attention and question-guided cross-modal attention. The implementation we use achieves 69.44% VQA-accuracy on VQAv2 test-dev.
- **X-VLM** [52] (2021) is a representative large transformer architecture wherein pretrained components are combined and further tuned on large-scale multimodal datasets. The visual encoder is initialized with a pre-trained Swin Transformer [32] and the question encoder and cross-modal layers by a pretrained BERT model [7]. These are then further pretrained with self-supervised objectives on a combined dataset including VQA [3], Visual Genome [25], SBU Captions [36] and Conceptual Captions [40] to learn multimodal representations. The model is then fine-tuned on the downstream VQA task. This end-to-end multimodal training is in contrast to the frozen pretrained elements in MCAN and BUTD. The implementation we use achieves 78.07% VQA-accuracy on VQAv2 test-dev. To study the influence of multimodal pretraining, we include a variant X-VLM* which has same initialization (Swin+BERT), but is then directly finetuned on the VQA task.

To study the role of pretrained vision and language encoders, we also consider the pretrained Swin [32] and BERT [7] models used to initialize the X-VLM model.

Features for Maha-*. When extracting features, we take the outputs from the final modality-specific layer for Maha-V and Maha-L variants and the final joint encoding for Maha-X variants. If more than one representation exists (e.g. for multiple visual or text tokens), we perform a mean-pool operation. The specific layers used for feature extraction are described in the supplemental material. Note – for the sake of space, we report results for Maha-* and MAP-* for BUTD and MCAN in the supplement only, but summarize the result in the following section.

AUCROC Upper Bounds for VizWiz and GQA. As both VizWiz and GQA contain real images paired with common sense questions, it is possible some samples should be considered in-distribution for VQAv2. As a result, the maximum achievable AUCROC might be less than 1. To establish an estimate for this upper bound, we train classifiers to distinguish training set instances from VQAv2 from those drawn from VizWiz or GQA. We adapt pre-trained X-VLM models for this binary classification task. We note that this outlier exposure setting assumes access to outlier samples – an oracle setting relative to our actual methods and bench-

mark. These models achieve 0.936 and 0.999 AUCROC for GQA and VizWiz respectively on our benchamrk, indicating that the sampled in- and out-of-distribution sets are distinguishable from each other.

5.1. Experimental Results

We report results for selected model-score combination across datasets in Table 2 due to space limitations, but the full set of results is available in the supplement. Where a model can admit multiple score functions, we denote the score function in parenthesis. We also note whether a method considers the image and question in column 2–3. In real-world usage, OOD VQA pairs may originate from a wide range of cases, we report average over the OOD settings we consider as an aggregate metric. Methods relying solely on single modalities must gain a score near 0.5 on mixed datasets which contains OOD data only in the unobserved modality. We include these for completeness but gray them out to avoid visual clutter.

We organize our discussion around specific questions and will refer to row numbers shown in the second column.

How do different scoring function categories compare? In general, we find that density (rows 1–2) and feature-based (8–19) tend to outperform reconstruction (3–4) and prediction-based (5–7) methods on average.

The average performance of reconstruction models is limited because we only consider unimodal models but multimodal novelties; however, we also see low performance in settings that rely heavily on the target modality. For instance, RIAD (3) performs poorly in the visually distinct CLEVR setting – likely because reconstructing the images with simple background and few objects is quite easy, confounding RIAD’s MSE-based scoring function.

For prediction-based methods using maximum softmax probability (5–7), we find that answer confidence is an unreliable predictor of I_{In}/Q_{Out} , and I_{Out}/Q_{In} which are partially in-distribution and ungrounded – corroborating prior findings for OOD detection of non-question sentences [27]. Interestingly, these methods outperform others on QRPE – suggesting the VQA models are able to provide reasonable confidence estimates when both the image and question are in-distribution. We note that improved VQA task performance does weakly correlate with improved OOD detection from MSP; however, the overall result suggest that model confidence alone is insufficient for strong OOD detection in VQA if novel images or questions are expected.

Both density and feature-based models contain variants that perform well, with the image-to-question (I2Q) model (2) serving as a strong baseline for settings requiring assessing both modalities. Despite its relative simplicity and limited training data (just VQA), I2Q shows stable performance across different OOD conditions and achieves the strongest average performance.

#	Method (Score)	Q	I	VIZWIZ	GQA	CLEVR	VQA _{ABS}	I _{In} /Q _{Out}	I _{Out} /Q _{In}	QRPE	Average
Density-based	1 LangM		✓	0.768	0.869	0.983	0.606	0.913	0.500	0.439	0.725
	2 I2Q		✓	0.729	0.884	0.983	0.755	0.956	0.792	0.620	0.817
Reconst.-based	3 RIAD			✓	0.246	0.546	0.016	0.584	0.500	0.145	0.492
	4 LangVAE			✓	0.554	0.522	0.835	0.512	0.666	0.500	0.512
Prediction-based	5 BUTD (MSP)		✓	✓	0.775	0.512	0.700	0.608	0.580	0.529	0.698
	6 MCAN (MSP)		✓	✓	0.794	0.506	0.667	0.591	0.573	0.518	0.739
	7 X-VLM (MSP)		✓	✓	0.714	0.583	0.670	0.656	0.605	0.549	0.726
Feature-based	8 X-VLM (Maha-V)			✓	0.967	0.442	0.988	0.999	0.500	0.732	0.592
	9 X-VLM (Maha-L)			✓	0.593	0.686	0.940	0.530	0.875	0.500	0.432
	10 X-VLM (Maha-X)			✓	✓	0.852	0.534	0.784	0.705	0.685	0.640
	11 Swin (Maha-V)			✓	0.933	0.488	0.997	<u>0.983</u>	0.500	0.756	0.561
	12 BERT (Maha-L)			✓	0.645	0.836	0.942	0.496	0.872	0.500	0.390
	13 Swin (MAP-V)			✓	0.323	0.623	0.178	0.452	0.500	0.396	0.493
	14 BERT (MAP-L)			✓	0.449	0.782	0.977	0.519	0.848	0.500	0.550
	15 X-VLM (MAP-V)			✓	0.849	0.332	0.985	0.495	0.500	0.671	0.542
	16 X-VLM (MAP-L)			✓	0.960	<u>0.916</u>	<u>0.999</u>	0.570	0.999	0.500	0.605
	17 X-VLM (MAP-X)			✓	✓	0.930	0.578	0.857	0.528	0.922	0.816
	18 X-VLM (MAP-A)			✓	✓	0.953	0.880	1.000	0.562	<u>0.998</u>	0.630
	19 X-VLM* (MAP-A)			✓	✓	<u>0.962</u>	0.872	0.996	0.560	0.990	0.548

Table 2. AUCROC results of OOD detection on different OOD sets (higher is better). All models are trained on the VQAv2 dataset as in-distribution. The best performing methods are bolded and second best underlined. The Q and I columns denote if the question or image are considered by the model-score combination. The Average is the averaged score among all categories. Single-modality results are grayed for off-modality OOD settings – e.g., for a vision-only method evaluated on purely language novelty.

How effective are different modalities? Many of our model-score combinations have unimodal variants which we find can perform quite well on individual OOD datasets where the corresponding modality is more distinct. For instance, all language-only variants (1, 9, 12, 14, 16) perform well on the CLEVR dataset. Similarly, vision-only variants based on Swin transformer feature distances (8, 11) identify the cartoonish VQA_{ABS} images easily. However, these approaches can hardly identify QRPE samples representing novel combinations of in-distribution questions and images whereas the VQA-based models with MSP (6, 7) can achieve some non-trivial detection.

For VQA model-based methods, cross-modal score variants tend not to retain the performance of their unimodal counterparts. For instance, X-VLM (Maha-X) (10) underperforms either its vision (8) or language (9) variant (or both) on all datasets, but achieves relatively stronger performance on QRPE where assessing multimodal alignment is critical. We speculate this is a result of individual modality information becoming diluted during cross-modal fusion, such that samples which may be identifiable through one modality but not the other become more difficult to discern. On the other hand, compared with X-VLM (Maha-X) (10), the cross-modal X-VLM (MAP-X) (17) has stronger I_{In}/Q_{Out}, I_{Out}/Q_{In} performance. Our X-VLM (MAP-A)

(18) model that instead averages the MAP-V, MAP-L, and MAP-X scores seems to strike a balance between the two – maintaining more of the individual modality performance but sacrificing performance in the QRPE setting.

What effect does pretraining have? We introduce a version of our best performing model without multimodal pretraining. X-VLM (MAP-A) (18) and X-VLM* (MAP-A) (19) differ only in that the X-VLM* did not undergo self-supervised multimodal pretraining before it was fine tuned on the VQA dataset. Comparing these, we can see that the multimodal pretraining does not yield much improvements in OOD detection performance, much of the capability can be acquired from initialization and VQA finetuning.

As X-VLM is originally initialized with the Swin and BERT models, we can also examine the joint effect of multimodal pretraining and VQA task finetuning on these unimodal models. Comparing the vision (8) and language-based (9) Maha variants with the corresponding Swin (11) and BERT (12) models from which they were initialized, we see that feature distances do not significantly nor consistently change in their utility for unimodal OOD detection. This is in contrast to the behavior for MAP scores. Both X-VLM (MAP-V) and X-VLM (MAP-L) significantly outperform the Swin (MAP-V) and BERT (MAP-L) meth-

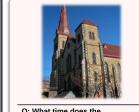
Top 1% OOD Score Samples								Bottom 1% OOD Score Samples							
	VQA	GQA	VQA _{ABS}	QRPE	VQA	GQA	VQA _{ABS}	QRPE		VQA	GQA	VQA _{ABS}	QRPE		
I2Q (GMP)	 Q: How many giraffes are there? S: 0.837	 Q: what color is the soap dispenser on the left? S: 0.701	 Q: what is on the bookshelf? S: 0.544	 Q: what color is the truck? S: 0.715	 Q: tell me sign in board? S: 0.000	 Q: Is the cd open? S: 0.027	 Q: Is the mom and baby have lunch outside? S: 0.002	 Q: What is at the dogs feet? S: 0.017							
X-VLM (MSP)	 Q: Is this outside? S: 1.000	 Q: Does the keyboard sit beside a mug? S: 0.992	 Q: What other object could the couple sit on? S: 1.000	 Q: What is the bench made out of? S: 0.959	 Q: What time does the clock say? S: 0.030	 Q: How tall is that flag? S: 0.130	 Q: Is she goth or just color coordinated? S: 0.124	 Q: What time does the clock read? S: 0.085							
X-VLM (Maha-X)	 Q: Are there clouds in the sky? S: -16.06	 Q: Are there any children in the photo? S: -17.10	 Q: Does the man feel like being bothered by the dog? S: -18.60	 Q: Will this chair damage the floor? S: -17.58	 Q: What is white? S: -50.05	 Q: What place is pictured? S: -47.80	 Q: What is the woman playing on? S: -48.36	 Q: What breed of cat is pictured? S: -46.82							
X-VLM (MAP-A)	 Q: How many people are in the picture? S: 0.189	 Q: Are there pizzas near the glass that is shown in this picture? S: 0.151	 Q: What is the white cat playing with? S: 0.189	 Q: What is in the bottles on the table? S: 0.169	 Q: Is this the ocean? S: 0.070	 Q: What's the snow in front of? S: 0.070	 Q: Is she drinking alone? S: 0.071	 Q: What is in the cup? S: 0.069							

Figure 3. Sample high- and low-scoring samples under four multimodal model-score combinations. For each, we a high and low scoring sample from in VQA, GQA, VQA_{ABS}, and QRPE. We use background color when scores align with in/out-of-distribution status.

ods. Taken together, these results suggest that while the usefulness of representations does not change dramatically from multimodal task training, the dynamics of attention do.

Finally, the strong performance of our I2Q model (2) trained only on VQA suggests that large-scale multimodal pretraining may not be necessary for strong OOD detection.

5.2. Qualitative Analysis

Figure 3 shows examples that are scored as highly likely (within top-1% of score) or highly unlikely (within bottom 1% of score) to be in-distribution samples. We consider four of our methods (rows) that rely on both images and questions. We select samples to represent no-novelty (VQA), linguistic-novelty (GQA), visual-novelty (VQA_{ABS}), and combination-novelty (QRPE).

For I2Q (row 1), we find OOD samples with high scores tend to be simple, well-grounded questions; whereas, the low scoring examples tend to be not questions “tell me sign in board”, grammatically incorrect “is the mom and baby have lunch outside?” or poorly grounded. The “cd” being difficult to see in the GQA sample and the QRPE sample being unrelated entirely.

For X-VLM (MSP) (row 2), many of the high scoring examples have simple questions with strong answer priors – e.g., “Does the keyboard sit beside the mug?” or “What other object could the couple sit on?”. The model may place probability only on a small set of reasonable answers (yes/no or chair/bench/sofa), resulting in a high MSP score. In contrast, the lower scoring samples are all questions which VQA models find difficult – measuring heights,

telling time, or judging subjective questions. These samples suggest that MSP based methods may suffer from biases regarding the model’s actual capabilities. This effect may be particularly amplified by the prevalence of binary questions.

For X-VLM (Maha-X) (row 3), all high-scoring samples are binary questions while the low-scoring are diverse. This may be because binary questions make up the plurality of questions in VQA (38.17% [3]) – skewing the mean and covariance in the Mahalanobis distance to favor representations of binary questions. In fact, we observe binary questions to occur in the top 500 scored VQA example at roughly twice the rate expected by chance.

For X-VLM (MAP-A) (row 4), the top-scoring samples tend to have many references to easily-visible objects in the scene; whereas, the bottom-scoring samples tend to lack grounding or refer to background elements. This might be explained by easily-groundable references resulting in highly activated attention maps.

6. Conclusion

We investigated out-of-distribution detection in visual question answering – presenting a more realistic VQA OOD setting composed of 6 popular VQA datasets and 2 composite datasets. We benchmark OOD detection methods based on density estimation, reconstruction error, model prediction, and intermediate network features. Further, we examine how different modalities and pretraining schemes affect OOD detection. We find that the image-to-question (I2Q) model achieves strong results despite not benefiting from large-scale multimodal pretraining.

References

[1] Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Anirudh 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](#)

[2] 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](#)

[3] 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. [1, 2, 3, 6, 8](#)

[4] Nilavra Bhattacharya, Qing Li, and Danna Gurari. Why does a visual question have different answers? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4271–4280, 2019. [2](#)

[5] Corentin Dancette, Spencer Whitehead, Rishabh Maheshwary, Ramakrishna Vedantam, Stefan Scherer, Xinlei Chen, Matthieu Cord, and Marcus Rohrbach. Improving selective visual question answering by learning from your peers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24049–24059, 2023. [1, 2](#)

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. [4](#)

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. [6](#)

[8] Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. *Advances in neural information processing systems*, 30, 2017. [2](#)

[9] 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. [1](#)

[10] 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 *CVPR*, 2017. [1, 3](#)

[11] 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 *CVPR*, 2018. [1, 2, 3, 4](#)

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. [4](#)

[13] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016. [2](#)

[14] Drew A Hudson and Christopher D Manning. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, 2019. [1, 2, 3](#)

[15] Junyan Jiang, Gus G Xia, Dave B Carlton, Chris N Anderson, and Ryan H Miyakawa. Transformer vae: A hierarchical model for structure-aware and interpretable music representation learning. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 516–520. IEEE, 2020. [5](#)

[16] Wenming Jiang, Ying Zhao, and Zehan Wang. Risk-controlled selective prediction for regression deep neural network models. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2020. [2](#)

[17] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *CVPR*, 2017. [1, 2, 3, 4](#)

[18] Amita Kamath, Robin Jia, and Percy Liang. Selective question answering under domain shift. *arXiv preprint arXiv:2006.09462*, 2020. [1, 2](#)

[19] 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](#)

[20] Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. Bilinear attention networks. *Advances in Neural Information Processing Systems*, 31, 2018. [2](#)

[21] Joo-Kyung Kim and Young-Bum Kim. Joint learning of domain classification and out-of-domain detection with dynamic class weighting for satisfying false acceptance rates. *arXiv preprint arXiv:1807.00072*, 2018. [2](#)

[22] Keunseo Kim, JunCheol Shin, and Heeyoung Kim. Locally most powerful bayesian test for out-of-distribution detection using deep generative models. *Advances in Neural Information Processing Systems*, 34, 2021. [2, 4](#)

[23] Diederik P Kingma and Max Welling. An introduction to variational autoencoders. *arXiv preprint arXiv:1906.02691*, 2019. [5](#)

[24] Ivan Krasin, Tom Duerig, Neil Alldrin, Vittorio Ferrari, Sami Abu-El-Haija, Alina Kuznetsova, Hassan Rom, Jasper Uijlings, Stefan Popov, Shahab Kamali, Matteo Malloci, Jordi Pont-Tuset, Andreas Veit, Serge Belongie, Victor Gomes, Abhinav Gupta, Chen Sun, Gal Chechik, David Cai, Zheyun Feng, Dhyane Narayanan, and Kevin Murphy. Openimages: A public dataset for large-scale multi-label and multi-class image classification. *Dataset available from https://storage.googleapis.com/openimages/web/index.html*, 2017. [6](#)

[25] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense

image annotations. *International journal of computer vision*, 123(1):32–73, 2017. 6

[26] Ian Lane, Tatsuya Kawahara, Tomoko Matsui, and Satoshi Nakamura. Out-of-domain utterance detection using classification confidences of multiple topics. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(1):150–161, 2006. 2

[27] Doyup Lee, Yeongjae Cheon, and Wook-Shin Han. Regularizing attention networks for anomaly detection in visual question answering. In *AAAI*, 2021. 1, 2, 5, 6

[28] 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. 2, 4

[29] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 3, 6

[30] 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. 2, 4

[31] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. 2

[32] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 6

[33] 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

[34] Aroma Mahendru, Viraj Prabhu, Akrit Mohapatra, Dhruv Batra, and Stefan Lee. The promise of premise: Harnessing question premises in visual question answering. *arXiv preprint arXiv:1705.00601*, 2017. 1, 2, 3, 4

[35] Binh X Nguyen, Tuong Do, Huy Tran, Erman Tjiputra, Quang D Tran, and Anh Nguyen. Coarse-to-fine reasoning for visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4558–4566, 2022. 4

[36] Vicente Ordonez, Girish Kulkarni, and Tamara Berg. Im2text: Describing images using 1 million captioned photographs. *Advances in neural information processing systems*, 24, 2011. 6

[37] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 5

[38] Seonghan Ryu, Seokhwan Kim, Junhwi Choi, Hwanjo Yu, and Gary Geunbae Lee. Neural sentence embedding using only in-domain sentences for out-of-domain sentence detection in dialog systems. *Pattern Recognition Letters*, 88:26–32, 2017. 2

[39] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8430–8439, 2019. 6

[40] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, 2018. 6

[41] Damien Teney, Ehsan Abbasnejad, Kushal Kafle, Robik Shrestha, Christopher Kanan, and Anton van den Hengel. On the value of out-of-distribution testing: An example of good-harts law. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 407–417. Curran Associates, Inc., 2020. 2

[42] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2, 4

[43] Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Open-set recognition: A good closed-set classifier is all you need. In *ICLR*, 2021. 2, 5

[44] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhi-liang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhajit Som, et al. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. *arXiv preprint arXiv:2208.10442*, 2022. 4

[45] Spencer Whitehead, Suzanne Petryk, Veda Shakib, Joseph Gonzalez, Trevor Darrell, Anna Rohrbach, and Marcus Rohrbach. Reliable visual question answering: Abstain rather than answer incorrectly. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVI*, pages 148–166. Springer, 2022. 2

[46] Ji Xin, Raphael Tang, Yaoliang Yu, and Jimmy Lin. 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)*, pages 1040–1051, 2021. 2

[47] Jingkang Yang et al. Generalized out-of-distribution detection: A survey. *arXiv*, 2021. 4

[48] Kexin Yi, Jiajun Wu, Chuang Gan, Antonio Torralba, Pushmeet Kohli, and Josh Tenenbaum. Neural-symbolic vqa: Disentangling reasoning from vision and language understanding. *Advances in neural information processing systems*, 31, 2018. 4

[49] Hiyori Yoshikawa and Naoaki Okazaki. Selective-lama: Selective prediction for confidence-aware evaluation of language models. In *Findings of the Association for*

Computational Linguistics: EACL 2023, pages 1972–1983, 2023. 2

- [50] 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. 2, 6
- [51] Vitjan Zavrtanik, Matej Kristan, and Danijel Skočaj. Reconstruction by inpainting for visual anomaly detection. *Pattern Recognition*, 112:107706, 2021. 2, 5
- [52] Yan Zeng, Xinsong Zhang, and Hang Li. Multi-grained vision language pre-training: Aligning texts with visual concepts. *arXiv preprint arXiv:2111.08276*, 2021. 2, 4, 6
- [53] Mingda Zhang, Tristan Maidment, Ahmad Diab, Adriana Kovashka, and Rebecca Hwa. Domain-robust vqa with diverse datasets and methods but no target labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7046–7056, 2021. 1, 4
- [54] Mingtian Zhang, Andi Zhang, and Steven McDonagh. On the out-of-distribution generalization of probabilistic image modelling. *Advances in Neural Information Processing Systems*, 34, 2021. 2, 4
- [55] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5579–5588, 2021. 2, 6