Super-CLEVR: A Virtual Benchmark to Diagnose Domain Robustness in Visual Reasoning

Zhuowan Li[1] Xingrui Wang[2] Elias Stengel-Eskin [1]
Adam Kortylewski[3, 4] Wufei Ma[1] Benjamin Van Durme[1] Alan Yuille[1]
[1] Johns Hopkins University [2] University of Southern California
[3] Max Planck Institute for Informatics [4] University of Freiburg

Abstract

Visual Question Answering (VQA) models often perform poorly on out-of-distribution data and struggle on domain generalization. Due to the multi-modal nature of this task, multiple factors of variation are intertwined, making generalization difficult to analyze. This motivates us to introduce a virtual benchmark, Super-CLEVR, where different factors in VQA domain shifts can be isolated in order that their effects can be studied independently. Four factors are considered: visual complexity, question redundancy, concept distribution and concept compositionality. With controllably generated data, Super-CLEVR enables us to test VQA methods in situations where the test data differs from the training data along each of these axes. We study four existing methods, including two neural symbolic methods NSCL [45] and NSVQA [59], and two non-symbolic methods FiLM [50] and mDETR [29]; and our proposed method, probabilistic NSVQA (P-NSVQA), which extends NSVQA with uncertainty reasoning. P-NSVQA outperforms other methods on three of the four domain shift factors. Our results suggest that disentangling reasoning and perception, combined with probabilistic uncertainty, form a strong VQA model that is more robust to domain shifts. The dataset and code are released at https://github.com/Lizw14/Super-CLEVR.

1. Introduction

Visual question answering (VQA) is a challenging task that assesses the reasoning ability of models to answer questions based on both visual and linguistic inputs. Current VQA methods are typically developed on standard benchmarks like VQAv2 [16] or GQA [25], with the implicit assumption that testing data comes from the same underlying distribution as training data. However, as has been widely studied in computer vision [15, 36, 51], algorithms trained on one domain often fail to generalize to other domains. Moreover, having learned the distributional prior of training data, models often struggle on out-of-distribution tests. This has been studied in VQA from the perspective of domain transfer [8,57,62], dataset bias [2,11,48], counter-factual diagnosis [9,47], and out-of-distribution benchmarking [30].

The multi-modal nature of VQA gives rise to multiple intertwined factors of variation, making domain shift an especially difficult problem to study. For example, [8] suggests that VQA domain shifts are a combination of differences in images, questions or answers; and [39] reveals a gap between synthetic and real VQA datasets by differences in the over-specification of questions and the underlying distribution of concepts. However, despite a wealth of research on domain generalization in VQA [3,26,57,62], there is no systematic analysis of the contributing factors in domain shifts.

To this end, we introduce a virtual benchmark, Super-CLEVR, which enables us to test VQA algorithms in situations where the test data differs from the training data. We decompose the domain shift into a set of isolated contributing factors, so that their effects can be diagnosed independently. We study four factors: visual complexity, question redundancy, concept distribution, and concept compositionality. These are illustrated in Fig. 1 and described in Sec. 3.1. With controllable data generation using our Super-CLEVR virtual benchmark, we are able to isolate the different factors in VQA domain shifts so that their effects can be studied independently. Compared with the original CLEVR dataset [27], Super-CLEVR contains more complicated visual components and has better controllability over the domain shift factors. As shown in Fig. 1, the Super-CLEVR dataset contains images rendered from 3D graphical vehicle models in the UDA-Part dataset [40], paired with questions and answers automatically generated from templates. The objects and questions are sampled based on the specified underlying probability distribution, which can be controlled to produce distribution shifts in different factors.

With Super-CLEVR, we diagnose the domain robustness of current VQA models. Four representative models are studied: for the classic two-stream feature fusing

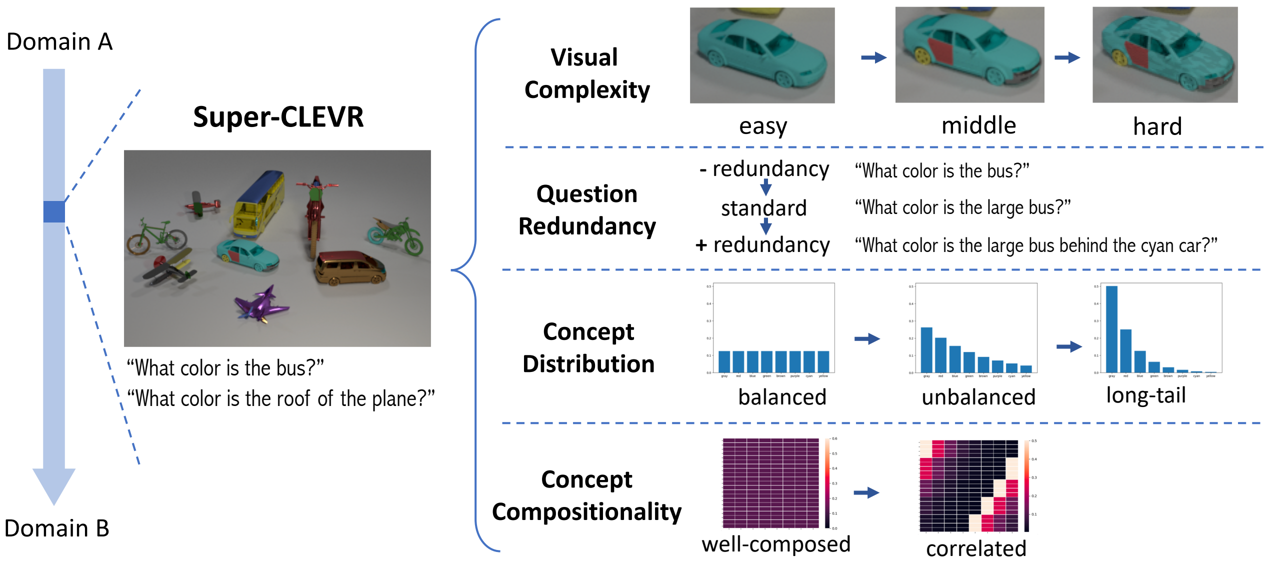

Figure 1. We decompose VQA domain shifts into four contributing factors: visual complexity, question redundancy, concept distribution and concept compositionality. The domain shifts along each factor can be independently studied with the proposed Super-CLEVR dataset.

architecture, we choose FiLM [50]; for a large-scale pretrained model we take mDETR [29]; we use NSCL [45] and NSVQA [59] as representative neuro-symbolic methods. We observe that all these models suffer from domain shifts to varying degrees of sensitivity. We analyze each factor separately to examine the influence of different model designs. Specifically, we find that the step-by-step design of neural modular methods enhances their robustness to changes in question redundancy compared with non-modular ones; however, the non-modular models are more robust to visual complexity. Furthermore, thanks to its decomposed reasoning and perception, NSVQA is more robust to concept distribution shifts.

While existing models suffer from domain shifts with different characteristics, we make a technical improvement over NSVQA which enables it to significantly outperform existing models on three of the four factors. In particular, we inject probabilities into the deterministic symbolic executor of NSVQA, empowering it to take into account the uncertainty of scene understanding. We name our model *probabilistic NSVQA* (P-NSVQA), and show that its performance improvement in both the in-domain and out-of-domain settings. With superior results of P-NSVQA, we suggest that disentangling reasoning from vision and language understanding, together with probabilistic uncertainty, gives a strong model that is robust to domain shifts.

Our contributions are as follows. (1) We introduce the Super-CLEVR benchmark to diagnose VQA robustness along four different factors independently. This benchmark can also be used for part-based reasoning. (2) We enhance a neural-symbolic method by taking the uncertainty of visual understanding into account in reasoning. (3) We conduct detailed analysis of four existing methods, as well as our novel approach to study the influence of model designs on distinct robustness factors. We conclude that disentangled reasoning and perception plus explicit modeling of uncertainty leads to a more robust VQA model.

2. Related work

Visual question answering (VQA). Popular VQA methods fall into three categories. Two-stream methods extract features for image and questions using CNN and LSTM respectively, then enable interaction between the two modalities with different feature fusing methods [4, 13, 24, 32, 37, 50, 60]. Neural symbolic methods, on the other hand, use a parse-then-execute pipeline where the question is parsed into a functional program, which is then executed on the image using neural modules [45, 59]. Recently, transformers-based models have achieved impressive performance on various vision-and-language tasks by pretraining on large scale dataset then finetuning for downstream tasks [29, 38, 44, 55, 63]. We choose FiLM [50], NSCL [45] and mDETR [29] as category representatives.

VQA datasets. Datasets containing real images and human-written questions have been widely used to benchmark VQA models, *e.g*. VQA [5], VQAv2 [16], Visual 7w [65], VizWiz [18], Visual Genome [35], COCO QA [52], etc. However, subsequent work has revealed the strong prior and bias in those datasets which might be exploited by models to correctly predict the answers without reasoning [1, 16, 17, 30, 31, 47, 48]. Attempts to address this problem include better balancing datasets [2] and creating counterfactual examples [9, 11]. To assess a model’s true reasoning ability, the CLEVR dataset [27] proposes to generate complex multi-step questions on synthetic images, which is then extended to various vision-and-language tasks [6,23,34,41,53,58,61]. The GQA dataset [25] extends CLEVR-style questions to real images. Our benchmark is distinct from existing ones because we introduce more complex visual scenes into CLEVR and provide controllability to study domain robustness on isolated factors.

Domain shift in VQA. Domain shift is a long-standing challenge in computer vision, explore in prior works in domain adaptation [14, 15, 22, 43] and domain generalization [36, 51]. Recent works have focused on domain shifts in VQA. [8, 57] improves model adaptation between datasets by feature learning. [62] analyze domain shifts between nine popular VQA datasets and proposes an unsupervised method to bridge the gaps. [39] generalize symbolic reasoning from synthetic to real dataset. [3] introduce a question-answer generation module that simulates the domain shifts. [26] propose a training scheme X-GGM to improve out-of-distribution generalization. [6] assess models generalization on the CLOSURE of linguistic components. In contrast to prior works we study each of the different domain shift factors independently with our virtual benchmark.

3. Super-CLEVR

3.1. Motivation: domain shift factors

Visual complexity. A major difference between different VQA datasets is visual complexity. For example, in the CLEVR dataset, objects are simple, atomic shapes while in real-world data, objects are more complex and have hierarchical parts. While hard to quantify, visual complexity is related to various factors, such as object variety, object size, background, texture, lighting, occlusion, view point, etc. In our work, we control visual complexity by introducing more challenging objects that can have distinct attributes associated with their parts, and by optionally pasting various textures onto objects. Examples of generated images with different complexity levels are shown in Fig. 1.

Question redundancy. Question redundancy refers to the amount of over-specified or redundant information in the question, which can be in the form of either *attributes* or *relationships*. For example, in Fig. 1, "what color is the large bus behind the cyan car", `large` (*attributes*) and `behind the cyan car` (*relationship*) are redundant because there is only one bus in the image. As observed in linguistics and cognitive science [12, 33, 49, 54], human speakers may include over-specified information when identifying a target object, which has also been studied in referring expression generation [46]. For VQA, as analyzed in [39], a significant difference between synthetic and real datasets is that real questions contain some redundant information, which sometimes is a distraction leading to model prediction errors. Therefore, in this work, we generate questions with different redundancy levels and study the effect of question redundancy on model behaviors.

Concept distribution. The distributions of *concept*, *i.e. objects* (*e.g.* car) and *attributes* (*e.g.* large), are distinct across different VQA datasets. For example, while colors are well-balanced in CLEVR dataset, in the GQA dataset, the color distribution is long-tailed where "white" appears >50 times more frequently than "gold". Long-tailed distributions have been a challenge in many computer vision tasks [20, 28, 42, 56, 64]. In VQA, the long-tailed concept distribution not only hinders the learning of infrequent concepts due to few training samples, but also introduces strong biases and priors in the dataset that may mislead the models. For example, "tennis" is the correct answer to most questions with "what sport is ..." [2]. With strong priors in data, it is hard to assess the true reasoning capacity of current models. While previous works address this problem by carefully re-balancing datasets [2], in our work, we controllably vary the concept distribution in our dataset and study model robustness to concept distribution shifts.

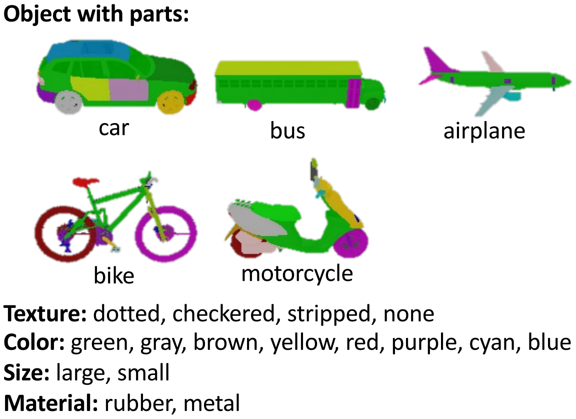
Object with parts:
car
bus
airplane
bike
motorcycle
Texture: dotted, checkered, stripped, none
Color: green, gray, brown, yellow, red, purple, cyan, blue
Size: large, small
Material: rubber, metal

Figure 2. Super-CLEVR contains 21 vehicle models belonging to 5 categories, with controllable attributes.

Concept compositionality. Concept compositionality refers to how different concepts (shapes, attributes) compose and co-occur with each other, *e.g.* roses are usually red while violets are usually blue [30]. Concept compositionality can be viewed as a conditional concept distribution in the context of other concepts. Shifts in concept compositionality impede the generalization of VQA models. For example, the model may fail to recognize a *green* banana because most bananas are *yellow* in the training data [7]. Previous works evaluate the out-of-distribution performance by collecting counterfactual testing examples [47]. In our work, we control the compositionality of *shapes* and *colors* with an intuitive motivation: if, for example, in the training data, bicycles are red and cars are blue, will the models be able to recognize blue bicycles and red cars in testing?

3.2. Dataset generation

Super-CLEVR follows a similar data generation pipeline as CLEVR, but with more complex visual components and better control of domain gap factors. We describe the generation procedure below.

Objects with parts. To improve the visual complexity of CLEVR scenes, we replace the simple shapes (*e.g*., *cube*, *sphere*) in CLEVR dataset with vehicles from UDA-Part dataset [40]. There are 21 vehicle models, belonging to 5 categories: *car*, *motorbike*, *aeroplane*, *bus*, and *bicycle*. Each 3D model comes with part annotations, *e.g*., *left front wheel* or *left right door* for *car*. Examples for the vehicle models are shown in Fig. 2. We remove or merge small parts from the original annotations to avoid severe difficulty in visual understanding. The full object and parts list is in the supplementary material.

Attributes. Besides the attributes in the original CLEVR dataset, *i.e*. *color*, *material*, *size*, we optionally add *texture* as an additional attribute to increase visual complexity. Note that in order to enable part-based questions, the attributes (color or material) of object parts can be different from that of the object. For example, a blue car can have a red wheel or a green door. In this case, the attribute of the holistic object refers to the attribute of its main body (*e.g*. the blue car has blue frame).

Scene rendering. Following CLEVR, each scene contains 3 to 10 objects. The objects are placed onto the ground plane with random position and orientation. When placing the objects, we ensure that the objects do not overlap with each other and we avoid severe occlusion by thresholding the number of visible pixels for each object. Random jitters are added to lamp and camera positions. When rendering, we also save the ground-truth bounding boxes and segmentation masks for each of the objects and their parts, which are required when training some of the models.

Question generation. Super-CLEVR follows similar question generation pipeline in CLEVR, which instantiates question templates using the underlying reasoning program that can be operated on the scene graph. For example, the program `select_shape(truck)` → `query_color(·)` can be instantiated as question "what is the color of the truck". Therefore, redundancy level of questions can be controlled by removing or adding redundant reasoning steps in the underlying reasoning program.

3.3. Controlling the dataset

To study domain generalization, we generate several variants of the dataset for each of the domain shift factors. The variants of the datasets serve as different data domains to test the model robustness. Here we describe the method for controllably generating the dataset variants.

Visual complexity. We generate three variants of the dataset with different levels of visual complexity: *easy*, *mid* (middle) and *hard*. The only difference between the 3 versions is visual complexity: for the easy version, objects with different sizes, colors and materials are placed into the scene; for the middle version, we choose 3 parts on each object that are visible and randomly change their attributes; for the hard version, we further add random textures to the objects and parts. An example of the 3 dataset versions can be found in Fig. 1. Note that the scene layout and the questions are shared, so that the influence of visual complexity can be isolated and studied independently.

Question redundancy. Three variants of the dataset with different redundancy levels are generated: *rd-*, *rd* (default), *rd+*. By default (*rd*), as in original CLEVR dataset, the questions contain some redundant *attributes* resulting from random sampling, while all redundant *relationships* are removed. In *rd-*, we also remove all redundant attributes from the questions, leading to no redundancy in the questions. In *rd+*, we add all possible attributes and relationships into the question, so that questions contain a high level of redundancy. For all the variants, the questions are ensured to be valid.

Concept distribution. We generate three dataset variants with different concept distributions: *bal* (balanced), *slt* (slightly unbalanced) and *long* (long-tail distributed). More specifically, we change the distribution of *shapes*, *colors* and *materials* while the distribution of *size* is kept fixed in order to keep visual complexity consistent, since objects with smaller sizes are visually harder to recognize. By default (*bal*), the shapes and attributes are randomly sampled, leading to a balanced distribution. For *slt* and *long*, the concept distribution $\mathbf{d}$ is generated by $d_i = a^{-i}$, where i is the index of the concept. a is a hyper-parameter controlling the length of the tail. A larger a leads to more imbalanced distribution and $a = 1$ leads to flat distribution (cf. Fig. 1). For *slt*, $a = 1.3$; for *long*, $a = 2.0$. In addition, to better analyze the performance on the frequent and rare concepts, we generate three variants for testing purpose only: *head* (frequent concepts in the long-tail distribution), *tail* (infrequent/rare concepts), and *oppo* (opposite to the long-tail distribution). We test each model on those three variants to analyze the performance on concepts with different degrees of frequency.

Concept compositionality. We generate 3 versions of the dataset, *co-0*, *co-1* and *co-2*, with different compositions of the 21 *shapes* (from 5 categories) and the 8 *colors*. The compositionality of the dataset is controlled with the co-distribution matrix $M \in \mathbb{R}^{21\times 8}$, where each entry M_{ij} is the probability an object of the i-th shape has the j-th color. Entries in each row of M sum up to 1. In the version *co-0*, M is a flat matrix so that the shapes and colors are randomly composed. In *co-1*, each shape in one category has a different color distribution, *e.g*. truck and sedan, while shapes from different categories may share the same color distribution *e.g*. sedan and airliner. Oppositely, in *co-2*, we make the shapes in same category have the same color distribution, whiles shapes from different categories have different distributions. The motivation is that since shapes from the same category are visually similar, the difference in *co-1* and *co-2* will help analyze the difference in model predictions on visually similar objects and dissimilar objects when composed with different color distributions.

Dataset Statistics. Every dataset variant contains 30k images, including 20k for training, 5k for validation and 5k for testing. Each image is paired with 10 object-based and 10 part-based questions. By *default*, the dataset refers to the version with *mid* visual complexity level (*i.e.* objects are untextured and has up to 3 parts with distinct attributes), *rd* redundancy level, balanced (*bal*) concept distribution and random (*co-0*) compositionality. More dataset statistics are in supplementary materials.

4. Evaluated methods

4.1. Simple baselines

Random, Majority. These simple baselines pick a random or the most frequent answer for each question type in the training set as the predicted answer.

LSTM. This question-only baseline encodes the question word embeddings with LSTM [21] and predicts answer with an MLP on top of the final hidden states of the LSTM.

CNN+LSTM. The image is represented with features extracted by CNN and question is encoded by LSTM. An MLP predicts answer scores based on the concatenation of image and question features.

4.2. Existing models

FiLM. We choose *Feature-wise Linear Modulation* [50] as a representative of classic two-stream feature merging methods. The question features extracted with GRU [10] and image features extracted with CNN are fused with the proposed FiLM module.

mDETR. mDETR [29] is a transformer-based detector trained to detects objects in an image conditioned on a text query. The model is pretrained with 1.3M image and text pairs and can be finetuned for various downstream tasks like referring expression understanding or VQA.

NSCL. The *Neuro-Symbolic Concept Learner* [45] is a representative neural symbolic method. NSCL executes neural modules on the scene representation based on the reasoning program, during which the modules learns embeddings of each concept with the answer supervision.

NSVQA. *Neural-Symbolic VQA* [59] is a neural symbolic method composed of three components: A scene parser (Mask-RCNN [19]) that segments an input image and recovers a structural scene representation, a question parser that converts a question from natural language into a program and a program executor that runs the program on the structural scene representation to obtain the answer. Notably, compared to NSCL, the individual components of NSVQA can be learned separately, hence, for example the scene parser can be learned from data that does not necessarily have Visual-Question annotations.

4.3. Probabilistic NSVQA (P-NSVQA)

Since the program executor in NSVQA is a collection of deterministic, generic functional modules, it can be augmented with a probabilistic reasoning process that takes into account the confidence of the predictions of the scene parser. This allows the model to execute the program that has the largest joint likelihood, instead of only taking the maximal likelihood execution at each step of the program. The experiment results demonstrate a significant performance improvement of this probabilistic approach over the deterministic NSVQA model proposed in [59].

In particular, we interpret the confidence of the Mask-RCNN output as a likelihood function for all detected object classes p_{object} and their attributes p_{att}. Moreover, we define a likelihood $p_{spatial}$ for the spatial relations between objects (behind, in front, left, right) that is proportional to the distance between the centers of two bounding boxes. Given a reasoning program containing multiple reasoning steps, we execute each step based on the scene parsing likelihood and produce an step-wise output with confidence. Finally, we use a factorized model, multiplying the output for all the steps to get the final answer prediction. We refer readers to the Appendix for more details.

4.4. Implementation details

Training mDETR requires ground-truth grounding of question tokens to image regions, which is available in Super-CLEVR. NSCL requires bounding box of objects, which can predicted using a trained Faster RCNN, and the reasoning program, which can be parsed using a trained parser. Similarly, ground-truth programs are used for training NSVQA and P-NSVQA. Note that we empirically find that the question-to-program parsing is a relatively easy task ($> 99\%$ accuracy using a simple LSTM), so we focus more on models' reasoning ability in our analysis.

Unless specified, the models are trained with default setting as in the official implementation. FiLM is trained for 100k iterations with batch size 256. mDETR is trained for 30 epochs with batch size 64 using 2 GPUs for both the grounding stage and the answer classification stage. NSCL is trained for 80 epochs with batch size 32. For NSVQA and P-NSVQA, we first train the object parser (Mask RCNN [19]) for 30k iterations with batch size 16, then train the attribute extraction model (using the Res50 backbone) for 100 epochs with batch size 64. For P-NSVQA, when counting the objects or determining whether objects exist in the scene, we use a threshold (0.7) to obtain the final selected objects. Early stopping is used based on validation accuracy. All the models are trained with 200k questions. We repeat experiments on *default* split for 3 times with different random seeds and get *std* of ± 0.10 (P-NSVQA) and ± 0.40 (NSVQA), showing the statistical significance of our results, then we only run other experiments once.

5. Results and analysis

In this section, we first show evaluation results for in-domain setting, then provide results and analysis on out-of-domain evaluation. Finally, we describe additional studies and future works.

5.1. In-domain results

In-domain evaluation refers to the setting where the training and testing data come from the same domain (the *default* dataset variant in this case). We compare the in-domain results on Super-CLEVR and CLEVR. The results are shown in Fig. 3.

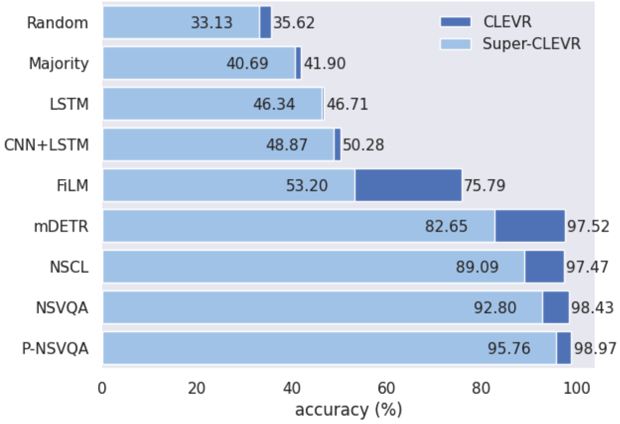

Figure 3. Comparison of models' accuracy on Super-CLEVR and the original CLEVR dataset.

For all the models, the performance is lower on Super-CLEVR than CLEVR, suggesting that Super-CLEVR is a more challenging benchmark. The scenes with vehicles are much harder visually for the models to understand compared with the simpler shapes from CLEVR. Note that the performance gap on two datasets for simple baselines (Random, Majority, LSTM, CNN+LSTM) is smaller than for the other better models. This is because Super-CLEVR contains more object types, and therefore the performance of simply guessing is lower than on CLEVR.

When comparing the performance of different models, we find that the neural modular methods, *i.e*. NSCL, NSVQA, P-NSVQA, perform much better than non-modular ones. This is not surprising given their nearly perfect performance on the original CLEVR dataset, which shows its strong ability to model synthetic images. The large-scale pretrained grounding model mDETR, which is a leading model on both real and synthetic images, also achieves good performance (82.7%) on Super-CLEVR. The two-stream method FiLM does not achieve very strong performance (53.2%), but is still much better than the other simple baselines.

Our proposed P-NSVQA outperforms all the other models. In particular, on Super-CLEVR, it outperforms its deterministic counterpart, NSVQA, by 2.96%. This shows the advantage of taking into account the probabilities when the scenes are challenging thus the model's uncertainty of predictions can be utilized.

5.2. Out-of-domain results

In this section, we train and test the five models (FiLM, mDETR, NSCL, NSVQA and P-NSVQA) on different dataset variants, and diagnose their domain robustness on each of the four domain shift factors. Please refer to Sec. 3.3 for a description of different variants. The validation accuracy is used for analysis here and the results are shown in Tab. 1.

All the methods suffer from domain shifts. The results show that the best performance mostly occurs in situations where the model is tested on the same dataset variant as it is trained on, *i.e*. the bold or underlined numbers fall mostly on the diagonals in Tab. 1.

We compare the domain robustness of the five models by measuring the relative performance decrease when the testing data differs from the training data, *i.e*. smaller performance drop on different testing domains means better robustness. Based on this intuition, for easier understanding of Tab. 1, we propose a measurement metric for domain robustness named ***Relative Degrade (RD)*** to better analyze the results. We define *Relative Degrade* as the percentage of accuracy decrease when the model is tested under a domain shift, *i.e*. the accuracy drop divided by the in-domain accuracy. Specifically, if a model gets accuracy a under in-domain testing (*i.e*. testing with the same dataset variant as training) and accuracy b under out-of-domain testing (*i.e*. testing with a different dataset variant from training), then $RD = (a - b)/a$. Since we train each model on three data variants, the RD's for the three models are averaged to measure its domain robustness.[1]

Tab. 2 shows the *Relative Degrade* of the five models on the four factors. We see that P-NSVQA outperforms other models by a significant margin on three of the four factors, indicating that it has better overall domain robustness. In the following, we take a closer look at the results on each of the factors separately, to diagnose the influence of different model designs.

Question redundancy. Neural modular methods are much more robust to question redundancy shifts than non-modular ones. The relative degrades for modular methods are less than 2%, while one-modular ones degrade for around 20%. Due to the step-by-step design of the reasoning in modular methods, each reasoning step is independent of the others so that the models are less likely to learn the spurious correlation between question and answers. There-

[1] For concept distributions, we compute relative degrade with a slight change: we compute the accuracy drop from *head* to *tail* and the drop from *long* to *oppo*, take their average, and divide by the accuracy on *bal*.

	FiLM			**mDETR**			**NSCL**			**NSVQA**			**Prob NSVQA**		
Visual Complexity															
	easy	mid	hard	easy	mid	hard	easy	mid	hard	easy	mid	hard	easy	mid	hard
easy	**59.96**	53.95	50.66	**93.36**	84.30	82.97	**95.13**	92.31	90.81	**95.19**	94.19	94.09	**96.76**	95.98	96.37
mid	**57.41**	53.28	50.18	**83.34**	82.36	81.27	84.5	**89.10**	86.33	81.99	92.80	**93.78**	86.25	**95.76**	95.11
hard	**55.95**	53.11	50.47	79.71	79.94	**80.71**	76.85	78.66	**85.08**	73.11	79.71	**92.65**	79.81	86.47	**95.36**
Question Redundancy															
	rd-	rd	rd+	rd-	rd	rd+	rd-	rd	rd+	rd-	rd	rd+	rd-	rd	rd+
rd-	51.42	52.54	**53.51**	**83.94**	80.37	66.28	88.64	88.82	**90.33**	**92.95**	92.94	92.67	95.66	**95.72**	95.43
rd	50.39	53.28	**54.78**	**82.77**	82.36	70.36	88.45	89.10	**91.45**	91.19	**92.78**	92.14	94.87	**95.72**	95.43
rd+	46.14	52.30	**71.47**	78.48	84.05	**90.42**	87.94	88.34	**91.16**	91.38	91.96	**92.80**	94.88	95.47	**95.72**
Concept Distribution															
	bal	slt	long	bal	slt	long	bal	slt	long	bal	slt	long	bal	slt	long
bal	50.47	53.04	**54.35**	**80.71**	75.79	74.54	**85.08**	83.79	75.10	**92.65**	90.82	83.74	**95.36**	94.89	89.88
long	49.43	54.75	**62.96**	79.06	80.29	**90.66**	85.33	89.42	**91.10**	92.73	**93.38**	92.53	96.31	**96.32**	95.25
head	48.60	58.06	**61.60**	80.75	79.60	**87.46**	84.58	88.39	**90.19**	93.87	**94.82**	92.48	96.42	**96.80**	95.92
tail	**51.80**	48.70	50.08	**81.50**	70.88	60.94	**86.10**	80.27	60.55	**90.26**	89.20	75.32	**94.08**	93.20	82.68
oppo	**49.06**	48.93	46.68	**79.13**	68.37	56.98	**85.07**	77.86	55.14	**91.22**	88.65	71.32	**95.76**	94.09	79.74
Concept Compositionality															
	co-0	co-1	co-2	co-0	co-1	co-2	co-0	co-1	co-2	co-0	co-1	co-2	co-0	co-1	co-2
co-0	53.28	**57.00**	56.1	**83.36**	77.03	82.43	**89.1**	82.52	83.77	**92.80**	90.11	91.59	**95.76**	94.02	95.12
co-1	52.41	**60.57**	56.67	79.46	82.45	**83.93**	78.89	**87.18**	84.2	78.74	89.99	**90.67**	87.12	94.53	**94.78**
co-2	52.96	57.37	**60.53**	80.03	77.41	**87.24**	78.40	81.55	**88.84**	77.85	89.28	**92.23**	87.19	93.49	**95.61**

Table 1. Accuracy of models trained and tested on different domains. Column headings indicate *training* settings, while rows indicate the dataset variant for *testing*. The best performance in each row (*i.e*. the best training setting) is marked in **bold** and best performance in each column (*i.e*. the best testing setting) is underlined. Description for different splits is in Sec. 3.3 and analysis is in Sec. 5.2.

	Visual	**Redund.**	**Dist.**	**Comp.**
FiLM	**4.03**	21.33	28.46	9.04
mDETR	9.81	19.05	36.34	9.45
NSCL	15.57	0.92	37.44	15.40
NSVQA	17.48	1.72	20.92	11.44
Prob NSVQA	12.88	**0.84**	**13.72**	**7.00**

Table 2. *Relative Degrade* under domain shifts, *i.e*. the percentage of accuracy decrease when the model is tested with domain that differs with training. Lower *RD* means better robustness.

fore the modular methods are less vulnerable to change in question/program length.

Visual complexity. Different from our findings on question redundancy, for domain shifts in visual complexity, non-modular methods are more robust compared to modular ones. As shown in Tab. 2, while FiLM and mDETR gets less than 10% degrade, NSCL and (P-)NSVQA degrade for more than 12%. The reason might be that the simple reasoning modules in modular methods can not process the visual signals as well as the dense non-modular models.

Comparing P-NSVQA with NSVQA, we find that injecting probability into deterministic symbolic reasoning greatly improves the robustness on visual complexity (4.04% decrease in *RD*). This suggests that some errors in visual understanding can be corrected and recovered by taking into account the uncertainty of visual parsing and combining the results of each reasoning step with probability.

Concept distribution. While all the four existing models suffer a lot (larger than 20% *RD*) on domain shifts in concept distribution, we see that the symbolic method NSVQA is better than the other three (by more than 7.5%). With the disentangled reasoning and visual understanding components in NSVQA, the distribution priors in the images and the programs/answers cannot intertwine with each other, which prevent the model heavily relying on the priors. With uncertainty, we can further boost the robustness of NSVQA with a large margin (from 21% to 14% *RD*).

Moreover, the head-tail results suggests that the overall accuracy, which is commonly used to measure VQA performance, should be taken with cautious. When the testing split is imbalanced, the seemingly high accuracy is misleading because the head concepts dominates the testing while the tail ones are not well-reflected. For example, for NSCL, although it gets high accuracy (91%) on the long-tailed data, its performance is only 60.6% on the tail concepts. In real-world datasets, the data are usually not well-balanced, which suggests the value of synthetic testing.

Concept compositionality. Comparing the existing methods, we find that the non-modular methods seems to be more robust than modular methods NSCL or NSVQA. However, with uncertainty, P-NSVQA improves the result of NSVQA, which even outperforms the non-modular methods. This suggest the large potential of better robustness of modular methods by improving current models.

In summary, while non-modular methods are more robust to visual complexity shifts, the modular symbolic methods (improved with uncertainty) are more robust on the other three factors. By disentangling reasoning with visual understanding, separately executing every each reasoning step then merging the results of the steps using probabilities based on uncertainty, our P-NSVQA outperforms all the existing models in question redundancy, concept distribution and compositionality. Therefore, we suggest that symbolic reasoning with uncertainty leads to strong VQA models that are robust to domain shifts.

5.3. More analysis and future work

Synthetic-to-real transfer. We provide an additional proof-of-concept study to show that the findings drawn from Super-CLEVR dataset can transfer to real datasets. In the following experiments, we show our finding that neuro-symbolic methods (NSCL, NSVQA, P-NSVQA) are more robust than mDETR on question redundancy also holds true on the real GQA dataset [25]. More precisely, we progressively removed the redundant operations from the reasoning program in GQA testdev split, and then regenerated questions using a program-to-question generator. Using the change of models' testing accuracy as the redundant operations are removed, we can evaluate the models' robustness towards question redundancy. The results are show in Tab. 3.[2] We observe that the performance drop of mDETR is much larger than neuro-symbolic methods as the redundant information is progressively removed, which indicates that symoblic methods have better question redundancy than mDETR on GQA dataset. This is consistent with our findings on Super-CLEVR.

Reasoning with part-object hierarchies. In addition to evaluating domain generalization, Super-CLEVR can be extended for broader purposes, *e.g.* part-based reasoning. We can ask questions like "what is the color of the front wheel of the bike?", "what is the color of the vehicle that has a yellow wheel", etc. Those questions require the model to correctly understanding the part-object hierarchy, which is an ability that current VQA models lack.

	0%	**14%**	**32%**	**70%**	**91%**	**100%**
mDETR	0	-4.82	-8.46	-13.16	-13.88	-14.56
NSCL	0	-0.14	-0.34	-1.09	-1.71	-2.59
NSVQA	0	-3.47	-4.80	-7.01	-7.02	-7.02
P-NSVQA	0	-1.93	-3.15	-5.73	-5.91	-5.78

Table 3. Accuracy drop on the GQA dataset when redundant information is progressively removed.

Limitations. The main limitations of our work lie in the synthetic nature of our dataset. Future efforts can be made in collecting better controlled and balanced real datasets for model diagnosis. We emphasize that the purpose of the dataset is for model diagnosis and that models should also be tested on real data.

6. Conclusion

We diagnose domain shifts in visual reasoning using a proposed virtual benchmark, Super-CLEVR, where distinct factors can be independently studied with controlled data generation. We evaluate four existing methods and show that all of them struggle with domain shifts, highlighting the importance of out-of-domain testing. Among the evaluated methods, neural modular methods are more robust towards question redundancy. In particular, NSVQA with disentangled perception and reasoning shows better robustness towards distribution and compositionality shifts. We further propose P-NSVQA, which improves NSVQA with uncertainty in the reasoning modules. We show that P-NSVQA outperforms all the existing methods in both in-domain testing and out-of-domain testing. With detailed analysis, our study suggests that disentangling reasoning and perception, combined with probabilistic uncertainty, form a strong VQA model that is more robust to domain shifts. We hope our analysis may facilitate better understanding of strengths and weaknesses of VQA models and, more broadly, future work might explore using the Super-CLEVR benchmark for other tasks like part-based reasoning.

Acknowledgements

We would like to thank Nils Holzenberger, Kate Sanders, Chenxi Liu, Zihao Xiao, Qing Liu, Reno Kriz, and David Etter for their helpful comments and suggestions, as well as the anonymous reviewers. Zhuowan Li is supported by ONR N00014-21-1-2812 and grant IAA80052272 from the Institute for Assured Autonomy at JHU. Elias Stengel-Eskin is supported by an NSF GRFP. A. Kortylewski acknowledges support via his Emmy Noether Research Group funded by the German Science Foundation (DFG) under Grant No.468670075.

[2] For implementation of NSCL on a real-word dataset, we use the model in [39] (the version without calibration). The model accuracies on the original not-perturbed GQA testdev split are as following: mDETR (61.67%), NSCL (56.13%), NSVQA (39.58%), P-NSVQA (39.66%).

References

[1] Aishwarya Agrawal, Dhruv Batra, and Devi Parikh. Analyzing the behavior of visual question answering models. *arXiv preprint arXiv:1606.07356*, 2016. 2

[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. 1, 2, 3

[3] Arjun Akula, Soravit Changpinyo, Boqing Gong, Piyush Sharma, Song-Chun Zhu, and Radu Soricut. Crossvqa: scalably generating benchmarks for systematically testing vqa generalization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2148–2166, 2021. 1, 3

[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

[6] Dzmitry Bahdanau, Harm de Vries, Timothy J O'Donnell, Shikhar Murty, Philippe Beaudoin, Yoshua Bengio, and Aaron Courville. Closure: Assessing systematic generalization of clevr models. *arXiv preprint arXiv:1912.05783*, 2019. 2, 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. 3

[8] Wei-Lun Chao, Hexiang Hu, and Fei Sha. Cross-dataset adaptation for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5716–5725, 2018. 1, 3

[9] Long Chen, Xin Yan, Jun Xiao, Hanwang Zhang, Shiliang Pu, and Yueting Zhuang. Counterfactual samples synthesizing for robust visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10800–10809, 2020. 1, 2

[10] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014. 5

[11] Corentin Dancette, Rémi Cadène, Damien Teney, and Matthieu Cord. Beyond question-based biases: Assessing multimodal shortcut learning in visual question answering. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1554–1563, 2021. 1, 2

[12] William Ford and David Olson. The elaboration of the noun phrase in children's description of objects. *Journal of Experimental Child Psychology*, 19(3):371–382, 1975. 3

[13] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. *arXiv preprint arXiv:1606.01847*, 2016. 2

[14] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR, 2015. 3

[15] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030, 2016. 1, 3

[16] 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 *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1, 2

[17] Vipul Gupta, Zhuowan Li, Adam Kortylewski, Chenyu Zhang, Yingwei Li, and Alan Yuille. Swapmix: Diagnosing and regularizing the over-reliance on visual context in visual question answering. *arXiv preprint arXiv:2204.02285*, 2022. 2

[18] 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. 2

[19] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 5

[20] Ruifei He, Jihan Yang, and Xiaojuan Qi. Re-distributing biased pseudo labels for semi-supervised semantic segmentation: A baseline investigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6930–6940, 2021. 3

[21] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 5

[22] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *International conference on machine learning*, pages 1989–1998. PMLR, 2018. 3

[23] Yining Hong, Li Yi, Josh Tenenbaum, Antonio Torralba, and Chuang Gan. Ptr: A benchmark for part-based conceptual, relational, and physical reasoning. *Advances in Neural Information Processing Systems*, 34, 2021. 2

[24] Drew A Hudson and Christopher D Manning. Compositional attention networks for machine reasoning. 2018. 2

[25] 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. 1, 2, 8

[26] Jingjing Jiang, Ziyi Liu, Yifan Liu, Zhixiong Nan, and Nanning Zheng. X-ggm: Graph generative modeling for out-of-distribution generalization in visual question answering. In

Proceedings of the 29th ACM International Conference on Multimedia, pages 199–208, 2021. 1, 3

[27] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2901–2910, 2017. 1, 2

[28] Lie Ju, Xin Wang, Lin Wang, Tongliang Liu, Xin Zhao, Tom Drummond, Dwarikanath Mahapatra, and Zongyuan Ge. Relational subsets knowledge distillation for long-tailed retinal diseases recognition. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 3–12. Springer, 2021. 3

[29] Aishwarya Kamath, Mannat Singh, Yann LeCun, Ishan Misra, Gabriel Synnaeve, and Nicolas Carion. Mdetr–modulated detection for end-to-end multi-modal understanding. *arXiv preprint arXiv:2104.12763*, 2021. 1, 2, 5

[30] 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. 1, 2, 3

[31] Corentin Kervadec, Theo Jaunet, Grigory Antipov, Moez Baccouche, Romain Vuillemot, and Christian Wolf. How transferable are reasoning patterns in vqa? *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4205–4214, 2021. 2

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

[33] Ruud Koolen, Martijn Goudbeek, and Emiel Krahmer. Effects of scene variation on referential overspecification. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 33, 2011. 3

[34] Satwik Kottur, José MF Moura, Devi Parikh, Dhruv Batra, and Marcus Rohrbach. Clevr-dialog: A diagnostic dataset for multi-round reasoning in visual dialog. *arXiv preprint arXiv:1903.03166*, 2019. 2

[35] 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. 2

[36] Da Li, Jianshu Zhang, Yongxin Yang, Cong Liu, Yi-Zhe Song, and Timothy M Hospedales. Episodic training for domain generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1446–1455, 2019. 1, 3

[37] Linjie Li, Zhe Gan, Yu Cheng, and Jingjing Liu. Relation-aware graph attention network for visual question answering. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10313–10322, 2019. 2

[38] Xiujun Li, Xi Yin, Chunyuan Li, Xiaowei Hu, Pengchuan Zhang, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. Oscar: Object-semantics aligned pre-training for vision-language tasks. *ECCV 2020*, 2020. 2

[39] Zhuowan Li, Elias Stengel-Eskin, Yixiao Zhang, Cihang Xie, Quan Hung Tran, Benjamin Van Durme, and Alan Yuille. Calibrating concepts and operations: Towards symbolic reasoning on real images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14910–14919, October 2021. 1, 3, 8

[40] Qing Liu, Adam Kortylewski, Zhishuai Zhang, Zizhang Li, Mengqi Guo, Qihao Liu, Xiaoding Yuan, Jiteng Mu, Weichao Qiu, and Alan Yuille. Learning part segmentation through unsupervised domain adaptation from synthetic vehicles. In *CVPR*, 2022. 1, 4

[41] Runtao Liu, Chenxi Liu, Yutong Bai, and Alan L Yuille. Clevr-ref+: Diagnosing visual reasoning with referring expressions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4185–4194, 2019. 2

[42] Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X Yu. Large-scale long-tailed recognition in an open world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2537–2546, 2019. 3

[43] Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Conditional adversarial domain adaptation. *Advances in neural information processing systems*, 31, 2018. 3

[44] 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

[45] Jiayuan Mao, Chuang Gan, Pushmeet Kohli, Joshua B. Tenenbaum, and Jiajun Wu. The Neuro-Symbolic Concept Learner: Interpreting Scenes, Words, and Sentences From Natural Supervision. In *International Conference on Learning Representations*, 2019. 1, 2, 5

[46] Margaret Mitchell, Kees Van Deemter, and Ehud Reiter. Generating expressions that refer to visible objects. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1174–1184, 2013. 3

[47] Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xiansheng Hua, and Ji-Rong Wen. Counterfactual vqa: A cause-effect look at language bias. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12695–12705, 2021. 1, 2, 3

[48] Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. Counterfactual vqa: A cause-effect look at language bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12700–12710, June 2021. 1, 2

[49] Thomas Pechmann. Incremental speech production and referential overspecification. 1989. 3

[50] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. 1, 2, 5

[51] Fengchun Qiao, Long Zhao, and Xi Peng. Learning to learn single domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12556–12565, 2020. 1, 3

[52] Mengye Ren, Ryan Kiros, and Richard Zemel. Exploring models and data for image question answering. *Advances in neural information processing systems*, 28, 2015. 2

[53] Leonard Salewski, A Koepke, Hendrik Lensch, and Zeynep Akata. Clevr-x: A visual reasoning dataset for natural language explanations. In *International Workshop on Extending Explainable AI Beyond Deep Models and Classifiers*, pages 69–88. Springer, 2022. 2

[54] Susan Sonnenschein. The development of referential communication skills: Some situations in which speakers give redundant messages. *Journal of Psycholinguistic Research*, 14(5):489–508, 1985. 3

[55] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019. 2

[56] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8769–8778, 2018. 3

[57] Yiming Xu, Lin Chen, Zhongwei Cheng, Lixin Duan, and Jiebo Luo. Open-ended visual question answering by multimodal domain adaptation. *arXiv preprint arXiv:1911.04058*, 2019. 1, 3

[58] Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B Tenenbaum. Clevrer: Collision events for video representation and reasoning. *arXiv preprint arXiv:1910.01442*, 2019. 2

[59] Kexin Yi, Jiajun Wu, Chuang Gan, Antonio Torralba, Pushmeet Kohli, and Joshua B Tenenbaum. Neural-Symbolic VQA: Disentangling Reasoning from Vision and Language Understanding. In *Advances in Neural Information Processing Systems (NIPS)*, 2018. 1, 2, 5

[60] 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 Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6281–6290, 2019. 2

[61] Chi Zhang, Feng Gao, Baoxiong Jia, Yixin Zhu, and Song-Chun Zhu. Raven: A dataset for relational and analogical visual reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5317–5327, 2019. 2

[62] 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, 3

[63] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Making visual representations matter in vision-language models. *CVPR 2021*, 2021. 2

[64] Xiao Zhang, Zhiyuan Fang, Yandong Wen, Zhifeng Li, and Yu Qiao. Range loss for deep face recognition with long-tailed training data. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5409–5418, 2017. 3

[65] Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. Visual7w: Grounded question answering in images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4995–5004, 2016. 2