Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 1496–1508
Marseille, 20-25 June 2022

Fine-tuning vs From Scratch: Do Vision & Language Models Have Similar Capabilities on Out-of-Distribution Visual Question Answering?

Kristian Nørgaard Jensen[♢] **Barbara Plank**[♢♣]
[♢]Department of Computer Science, IT University of Copenhagen, Denmark
[♣]Center for Information and Language Processing (CIS), LMU Munich, Germany
krnj@itu.dk bplank@cis.uni-muenchen.de

Abstract

Fine-tuning general-purpose pre-trained models has become a de-facto standard, also for Vision and Language tasks such as Visual Question Answering (VQA). In this paper, we take a step back and ask whether a fine-tuned model has superior linguistic and reasoning capabilities than a prior state-of-the-art architecture trained from scratch on the training data alone. We perform a fine-grained evaluation on out-of-distribution data, including an analysis on robustness due to linguistic variation (rephrasings). Our empirical results confirm the benefit of pre-training on overall performance and rephrasing in particular. But our results also uncover surprising limitations, particularly for answering questions involving boolean operations. To complement the empirical evaluation, this paper also surveys relevant earlier work on 1) available VQA data sets, 2) models developed for VQA, 3) pre-trained Vision+Language models, and 4) earlier fine-grained evaluation of pre-trained Vision+Language models.

Keywords: Visual Question Answering, Multimodal Transformers, Vision+Language, Fine-grained Evaluation

1. Introduction

Large general-purpose pre-trained machine learning models have become ubiquitous. While new models emerge at a rapid pace, only limited research exists into the robustness of such models (Li et al., 2020b; Cao et al., 2020). In particular, it remains an open question whether pre-training is the best way for Language+Vision tasks. With a growing number of general-purpose pre-trained Vision+Language (V+L) models, which all suggest that they are better than the previous one, it is important to take a step back, and investigate and analyse models.

In this paper, we perform a fine-grained evaluation of two models on Visual Question Answering (VQA). We compare a fine-tuned general-purpose V+L model to an earlier task-specific model trained from scratch on the standard VQA v2 data set (Goyal et al., 2017). We examine their linguistic and reasoning capabilities, as well as their robustness to language variation on three out-of-distribution data sets in a zero-shot fashion. Moreover, we also review early Visual Question Answering models, VQA data sets, recent pre-trained V+L models and survey related work on their evaluation. Our analysis focuses on BAN (Kim et al., 2018) and LXMERT (Tan and Bansal, 2019), each state-of-the-art VQA model at their respective times. Our results show that their capabilities diverge substantially, and while fine-tuned LXMERT performs the best on most setups, it is not always the case. Strikingly, BAN does perform better on datasets involving boolean compositions (VQA-Compose and VQA-Supplement). LXMERT is particularly susceptible to questions involving OR operations or multiple boolean comparisons in contrast to BAN, suggesting that even large pre-trained models perform insufficiently and answer priors have a major impact on model accuracy.

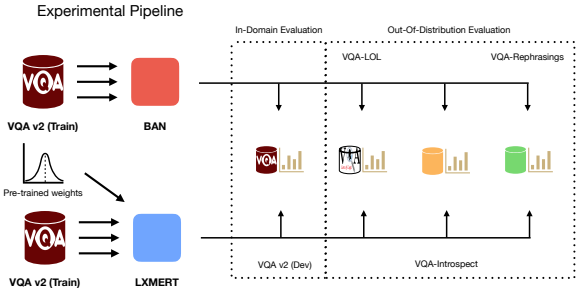


Figure 1: **Illustration of fine-grained evaluation**. We train BAN and LXMERT on the VQA v2 data set, and evaluate both on In-Domain data (VQA v2) and Out-Of-Distribution data (VQA-LOL, VQA-Introspect and VQA-Rephrasings). The LXMERT model is initialized using pre-trained weights.

2. Related work

2.1. Visual Question Answering models

Before the advent of pre-training, early neural models consisted of a Convolutional Neural Network (CNN) to encode the image and a Recurrent Neural Network (RNN), to encode the question, and predicting an answer. These models were later equipped with attention mechanisms to ground the text in the image (Xu et al., 2015; Yang et al., 2016). For example, Malinowski et al. (2015) presented Neural-Image-QA as an end-to-end formulation to the VQA task. They use a CNN to encode the image, and subsequently use its features to encode the question using a RNN. The same RNN is used to predict an answer given the input. They evaluated their model on the DAQUAR data set (Malinowski

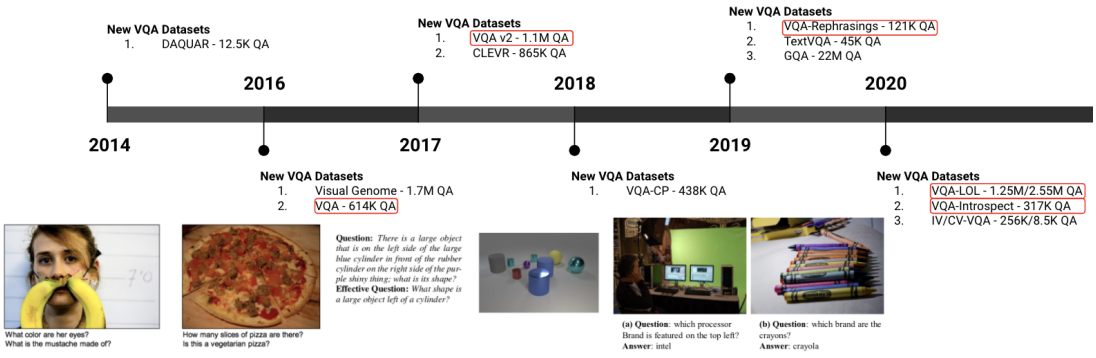


Figure 2: **Timeline of major VQA data set releases** with approximate number of question answers (QA). Highlighted (red) are the ones investigated in this work (Agrawal et al., 2016; Johnson et al., 2017; Singh et al., 2019).

and Fritz, 2014) performing more than twice as well compared to previous methods.

Yang et al. (2016) developed the Stacked Attention Network (SAN) which uses attention to help with reasoning in images. SAN uses generic CNN image features generated by a VGG model (Simonyan and Zisserman, 2015). Anderson et al. (2018) redefine their spatial features in terms of bounding boxes and by using a Faster-RCNN model as a bottom-up attention model, by attending to objects present in a given image. Given these redefined spatial features, they use a top-down attention mechanism to weight the image features, along with the question features. This bottom-up top-down (up-down) model was later improved by Jiang et al. (2018), and participated in the 2018 VQA Challenge. They implemented subtle improvements to the architecture, such as data augmentation and fine-tuning image features. The small changes gave an improvement of $\sim 5\%$ on the VQA v2 data set.

While some focus on how to develop sophisticated ways of producing attention mechanisms for questions and image features, Yu et al. (2018) thought of a way to pool together feature vectors from different modalities. They developed the multi-modal Factorized High-Order Pooling (MFH) mechanism which takes two feature vectors, and iteratively pools the features using elementwise multiplication and sum pooling. In the end they use attention modules to provide better predictions, but use the MFH to combine the image attention and the question attention features.

Continuing to use Faster R-CNN features as visual input, Kim et al. (2018) developed the *Bilinear Attention Networks* (BAN). BAN is a generalization of a bilinear model for two multi-channel inputs. It first generates G bilinear attention maps. These attention maps are then used with the two inputs to generate a rich set of features which are used for classifying answers. We use BAN as an early model in our empirical comparison, and compare it to a more recent V+L model.

2.2. Visual Question Answering data sets

One of the first VQA data sets, as illustrated in Figure 2, was the DAQUAR data set (Malinowski and Fritz, 2014). The DAQUAR data set is built on top of the NYU-Depth V2 data set (Silberman et al., 2012) which consists of 1,449 images. DAQUAR contains human-generated and synthetic question-answer pairs.

Two years later, Krishna et al. (2017) developed the seminal Visual Genome data set. The data set consists of a large amount of human annotations for each individual image, with $108,077$ real world images, an average of 21 annotated objects per image, and a total of 1.7 million question-answer pairs.

The same year, the VQA data set was published, and a year later the larger VQA v2 data sets (Agrawal et al., 2016; Goyal et al., 2017). Purposely built for VQA, they became common benchmarks, besides Visual Genome. VQA v2 consists of of $204,721$ images from the MSCOCO data set (Lin et al., 2015; Chen et al., 2015), $\sim$ 1.1 million questions with 10 answers each provided by humans. In this study we use the VQA v2 data set as the initial training data.

As a response to the use of one data set (VQA v2) for evaluating VQA models, many follow up data sets have been released to test models on varying problems not contained in the original VQA v2 data set. These new challenges include changing the answer distribution to help models better generalize rather than rely on a specific distribution of answers (Agrawal et al., 2018), rephrasing the original questions to evaluate consistency (Shah et al., 2019), and composing new questions based on the original questions and boolean operators (Gokhale et al., 2020). Similarly, to help the models better answer reasoning questions contained in the VQA v2 data set, the VQA-Introspect data set proposes a number of perception questions for the same image (Selvaraju et al., 2020). Lastly, to investigate the robustness of the model to changes in the images rather than linguistic changes, Agarwal et al. (2020) developed the IV-VQA and CV-VQA data sets for In-Variant (IV) and CoVariant (CV) changes. The above mentioned data sets are all based on real world images, but to test the visual reasoning capabilities of the models in a controlled manner, the CLEVR data set (Johnson et al., 2017) was developed. It consists of images of generated objects, to study how well models reason

about specific objects in images.

Analogously to CLEVR, the GQA data set (Hudson and Manning, 2019) sets out to test the reasoning capabilities and linguistic compositionality in real world images using a scene graph to compose 22 million questions. Along with the data the paper also presents a range of new metrics to help better understand the performance of reasoning models.

A specialized field of VQA is regarding the text in images (Scene text). This can be text in signs, banners or otherwise present in images. To help with the research in this area, Singh et al. (2019) developed the TextVQA data set, which consists of question-answer pairs all relating to text within images.

2.3. Vision+Language models

General-purpose pre-trained V+L models emerged around 2019 and are all built on top of the transformer architecture (Vaswani et al., 2017). We highlight 22 V+L models in Table 1, with details such as their pre-training data, pre-training and evaluation tasks, and the type of visual features they use. The table is ordered by approximate release date, meaning that the VideoBERT (Sun et al., 2019) is the earliest work, and LaTr (Biten et al., 2021) being the latest.

At a glance it is clear that most of the Vision Language models today use the Faster R-CNN model (Ren et al., 2015) as visual features. However, newer models have started to use different visual embedding methods, such as linear embedding similar to what Dosovitskiy et al. (2021) uses for their Vision Transformer (ViT)(Kim et al., 2021; Lin et al., 2021a). Others fine-tune a full CNN to extract custom features from the images (Huang et al., 2020; Xu et al., 2021; Wang et al., 2021). Furthermore, it is also worth noting that the authors behind OSCAR (Li et al., 2020c) have tried to improve the performance of an object detector for vision-language models, they call this VinVL (Zhang et al., 2021) and show that it does indeed improve the performance of OSCAR when used instead of the default Faster-R-CNN.

Furthermore, the most prevailing architecture type is that of single-stream models. These models process both modalities using the same transformer neural network, whereas the two-stream approach first process the inputs individually, and subsequently have a cross-modal processing step. However, SemVLP (Li et al., 2021) implements both architecture type and switches between them depending on the downstream task.

Across the 22 models in Table 1 they use 13 different pre-training tasks. These include tasks that use either of the modalities, or span the two modalities. The first group is the language only tasks Masked Language Modelling (MLM), and Sequence to Sequence (Seq2Seq). The second group is image only tasks Masked Region Regression/Classification (MRR/MRC), which are similar to MLM in that they mask a region/patch, and then have to either classify the object in that are or have to reconstruct the feature vector. The last group are the multi-modal tasks, such as Image-Text Matching (ITM), and Word Region Alignment (WRA), which are used to help the models align the two modalities. There are also the ones that fall outside these categories, such as Masked Token Modelling (MTM) which is similar to MLM and MRC, however, it is masking input tokens from both modalities at the same time, rather individually. It is also worth noting that InterBERT (Lin et al., 2021b) and M6 (Lin et al., 2021a) have extended the goal of MLM and MRC, such that they mask multiple tokens (i.e. spans) that the models have to retrieve rather than individual tokens.

Another important part of the models is the data they are trained on. Most of the models use a mix of in-domain and out-of-domain pre-training data, such as Conceptual Captions (CC) (Sharma et al., 2018), SBU Captions (Ordonez et al., 2011), MSCOCO (Chen et al., 2015), and Visual Genome (Krishna et al., 2017). Do note, that domain here refers to the downstream tasks, such as VQA, and Image-Text Retrieval. However, a few works also use text only data such as C4 (Roberts et al., 2019). People do also collect their own image-text data, such as the LAIT data from Qi et al. (2020) and the M6-corpus from Lin et al. (2021a).

Extensions of previous architectures Not all are developing entire new architectures, some are also extending and improving some of the previously developed models.

VILLA (Gan et al., 2020) is an adversarial training paradigm that can be applied to all V+L models. They add adversarial noise to the embedding space to help the model be more robust. They use it as a regularization technique in both the pre-training and the fine-tuning stage of the V+L models.

12-in-1 (Lu et al., 2020) is a large multi-task model trained on 12 different data sets used in 4 different V+L tasks such as VQA. They extend the ViLBERT architecture with a number of output heads which handles the different tasks.

2.4. Evaluation of V+L models

Research on evaluating V+L models is limited, yet rising along with the development of new models and more challenging data sets. Next, we summarize recent works on evaluation of V+L models.

Li et al. (2020b) investigate 4 generic types of robustness: 1) linguistic variabilty, 2) logical reasoning, 3) visual content manipulation, 4) answer distribution shift. They also introduce a new approach called MANGO that introduces adversarial noise at the embedding level. For evaluating the UNITER model, VILLA model and their own MANGO adversarial training scheme, they test on different data sets representing the 4 types of robustness described before. They find that using their MANGO framework produces the most robust model across the 4 types of

Name	Architecture	Image Features	Data	Task	Evaluation	
VideoBERT (Sun et al., 2019)	Single	Full frame	YT	MTM + ITM	Action Clas. + Video Cap.	✗
ViLBERT (Lu et al., 2019)	Two	Faster R-CNN	CC	MTM + ITM	VQA + VCR + Ref. Expr. + (ZS) IR	✓
VisualBERT (Li et al., 2019)	Single	Faster R-CNN	MSCOCO	MLM + ITM	VQA + VCR + NLVR2 + IR/TR	✗
LXMERT (Tan and Bansal, 2019)	Two	Faster R-CNN	MSCOCO + VG	MRR + MRC + MLM + ITM + VQA	VQA + GQA + NLVR2	✓
ImageBERT (Qi et al., 2020)	Single	Faster R-CNN	LAIT + CC + SBU	MLM + MRC + MRR + ITM	IR/TR	✗
VL-BERT (Su et al., 2020)	Single	Faster R-CNN	CC + BookCorpus + Wiki	MLM + MRC	VCR + VQA + Ref. Expr.	✓
Unicoder-VL (Li et al., 2020a)	Single	Faster R-CNN	CC + SBU	MLM + MRC + ITM	VCR + IR/TR	✗
VLP (Zhou et al., 2020)	Single	Faster R-CNN	CC	MLM + Seq2Seq	VQA + Img. Cap.	✓
Pixel-BERT (Huang et al., 2020)	Single	ResNet	MSCOCO + VG	MLM + ITM	VQA + NLVR2 + IR/TR	✗
UNITER (Chen et al., 2020)	Single	Faster R-CNN	MSCOCO +VG + CC+ SBU	MLM + MRC + MRR + ITM + WRA	VQA + VCR + NLVR2 + SNLI-VE + (ZS) IR/TR + Ref. Expr.	✓
OSCAR (Li et al., 2020c)	Single	Faster R-CNN	MSCOCO + CC + SBU + Flicker30k + VQA + GQA + VG-QA	MLM + ITM	VQA + GQA + NLVR2 + IR/TR + Img. Cap. + Obj. Cap.	✓
ERNIE-ViL (Yu et al., 2020)	Two	Faster R-CNN	CC + SBU	SGP	VCR + VQA + Ref. Expr. + IR/TR	✗
DeVLBert (Zhang et al., 2020)	Single	Faster R-CNN	CC	MTM[1]	VQA + (ZS) IR	✓
LAMP (Guo et al., 2020)	Single	Faster R-CNN	MSCOCO + VG	MLM + ITM + MRC + MRR	VQA + GQA + NLVR2 + ZS IR	✗
SemVLP (Li et al., 2021)	Single+Two	Faster R-CNN	CC + SBU + MSCOCO + VG + VQA + GQA + VG-QA	MLM + ITM + MRC + VQA	VQA + IR/TR + NLVR2 + GQA	✗
InterBERT (Lin et al., 2021b)	Single	Faster R-CNN	CC + SBU + COCO	MLM^+ + MRC^+ + ITM	VCR + (ZS) IR	✗
E2E-VLP (Xu et al., 2021)	Single	ResNet	MSCOCO + VG	MLM + ITM + Img. Cap. + Obj. Det.	VQA + NLVR2 + Img Cap. + IR/TR	✗
ViLT (Kim et al., 2021)	Single	Linear	MSCOCO + VG + CC + SBU	MLM + ITM	VQA + NLVR2 + (ZS) IR/TR	✓
M6 (Lin et al., 2021a)	Single	Linear	M6-Corpus	MLM^+ + Seq2Seq(*) + Img. Cap.	VQA + ITM + Img. Cap. + Text Clas. + RC + Cloze	✗
SimVLM (Wang et al., 2021)	Single	ResNet	ALIGN + C4	PrefixLM	VQA + NLVR2 + SNLI-VE + Img. Cap.	✗
CMA-CLIP (Liu et al., 2021)	Two	ViT	WIT+ MRWPA + Food101 + Fashion-Gen	Img. Text Clas.	MRWPA + Food101 + Fashion-Gen	✗
LaTr (Biten et al., 2021)	Two	ViT	C4 + IDL	MLM	TextVQA + ST-VQA + OCR-VQA	✗

Table 1: Masked Language Modelling (MLM), Masked Token Modelling (MTM), Masked Segment Modeling (MSM), Masked Region Regression (MRR), Masked Region Classification (MRC), Word-Region Alignment (WRA), Scene Graph Prediction (SGP). Reading Comprehension (RC). Note ($^+$) means that the model has to predict multiple words in sequence, and multiple object at the same time. Note (*) for M6, they do both regular text only seq2seq, but also seq2seq where the initial sequence consists of both visual and masked textual tokens. Note (^{1}) DeVLBert develops 2 invervention based methods, that can replace MTM, and 2 that are independent of MTM. Note ( ) shows whether the authors have made code and pre-trained models available in their papers.

robustness that they investigate.

Cao et al. (2020) designed a number of probing tasks for large-scale pre-trained V+L models called VALUE. They investigate the following 5 questions 1) How intertwined the multi-modal embeddings are, 2) The modality importance (How important are the textual modality versus the visual), 3) if any attention heads are cross-modal. For this they design a new probing task called Visual Coreference Resolution, 4) investigating if the models captures visual relations between image regions. For this they use a task called visual relation detection/classification, which evaluates the combined power of multiple attention heads in the models, 5) how much linguistic knowledge is encoded on the models. They find that pre-trained models tend to favor text over images, and that there exists a number of attention heads for cross-modal relations.

Parcalabescu et al. (2020) tests the capabilities of three V+L models using 2 different tasks. They test the LXMERT, ViLBERT and ViLBERT 12-in-1 models (Tan and Bansal, 2019; Lu et al., 2019; Lu et al., 2020). The tasks are 1) Image-Sentence Alignment Probe, 2) Counting Entities Probe. The first task is thought to be easy for the models, since they all three have been pre-trained using this task. They find that LXMERT and to some extend ViLBERT 12-in-1 suffers from catastrophic forgetting when pre-trained and fine-tuned. Moreover, they find that neither of the three models are good at counting, and mostly relies on the statistics of the data sets on which they are trained.

Li et al. (2020d) experiments with using LXMERT, VisualBERT, UNITER and PixelBERT for medical images and reports. They investigate whether these models can outperform previous best TieNet (CNN+RNN) and language embedding only models (BERT & ClinicalBERT (Alsentzer et al., 2019)). They see a clear increase in performance when applying the pre-trained models to the OpenI data set (Demner-Fushman et al., 2016). However, they find that when loading LXMERT, VisualBERT and UNITER with the ClinicalBERT parameters they do not see an increase in performance, even though those parameters should be better suited for the task. On top of this they experiment with freezing the visual backbone for PixelBERT. They find that freezing the backbone severely decreases performance, and they find it essential that the parameters should be updated during fine-tuning.

Hendricks et al. (2021) investigate the importance of pre-training data, attention and loss functions for V+L models for image retrieval. For pre-training they find that language-only or image-only pre-training does not improve performance. They also find that the multi-modal attention mechanism is important for the performance of these models, showing that smaller models with multi-modal attention perform better than larger models without this mechanism. On the contrary, for pre-training tasks they find that the masked-region modelling task does not provide information to the models, as pre-training without this task achieves comparable results.

3. Fine-grained VQA Evaluation

For the experiments in this paper we have chosen to work with the BAN and LXMERT models. We chose these two models, because they are developed in close succession (2018 and 2019 respectively). Besides the similarity in time they were developed, we also chose

them based on their code base and its ease of use. We found that many of the previous VQA models were built with a rather strict data set framework, which made zero-shot evaluation hard.

To evaluate the performance gain of a pre-trained model, we first train and evaluate LXMERT (Tan and Bansal, 2019) and BAN (Kim et al., 2018) on the VQA v2 (Goyal et al., 2017) data set. The LXMERT model has been pre-trained as described in (Tan and Bansal, 2019). BAN has not received any pre-training prior to being trained on VQA v2. For both models we use the default parameters and training parameters.

We want to test the models to see if pre-training helps the model being more robust towards changing the data set. To test this, we zero-shot evaluate on out-of-distribution data from the VQA-LOL (both compose and supplement), VQA-Introspect, and VQA-Rephrasings (Gokhale et al., 2020; Selvaraju et al., 2020; Shah et al., 2019) data sets. We define zero-shot evaluation as testing on a data set different from the one used for training without fine tuning. All the results can be seen in Tables 5, 6, 7, 8 and 9 and Figure 3.

Set	Operation	Question
C	Q1	Are trees visible?
	Q2	Are the streetlamps on?
	$\neg Q1$	Are not trees visible?
	$\neg Q2$	Are the streetlamps not on?
S	Q	Is this a creamy soup?
	$\neg Q$	Is not this a creamy soup?
	$Q \wedge B$	Is this a creamy soup and is there a bowl?
	$Q \wedge C$	Is this a creamy soup and is that a tasty bowl of ramen served for someone to enjoy?

Table 2: **Example questions from VQA-LOL.** C: VQA-Compose; S: VQA-Supplement.

3.1. Boolean Compositions

To experiment with boolean compositions we test each of the models on the VQA-Compose and the VQA-Supplement data sets. Examples of the data can be seen in Table 2. We evaluate the performance both with overall accuracy on the entire data set, and by investigating the performance across specific boolean operations (AND ($\wedge$), OR ($\vee$), NOT ($\neg$)). We furthermore evaluate whether the models are better at composed questions using only a single boolean operation or multiple.

3.2. Visual Reasoning

To evaluate each of the model's ability to answer reasoning questions versus perception questions, we test the models on the VQA-Introspect data set. The authors of the data set define reasoning questions as questions that require a synthesis of perception and prior knowledge / reasoning capabilities, and they define perception questions as the ones which can be answered by detecting the presence or physical properties of objects. Examples of questions from the data set can be seen in Table 3. To use the data set with the two models in a zero-shot setup, we had to filter out answers not contained in the classification head of LXMERT and BAN. This caused us to remove $2,100$ questions out of $94,507$, for which no answer was contained in the set of possible answers. We do note that the data defines main questions (reasoning questions) and sub questions (perception questions) in a hierarchy, however, in this experiment we treat them as equal and keep a mapping between them for evaluation purposes. To gauge how well the models are handling perception versus reasoning questions, we investigate four quadrants as defined in (Selvaraju et al., 2020). The four quadrants are:

- Both Main and Sub question correct (Main + Sub)
- Main Q correct, Sub question wrong (Only Main)
- Main Q wrong, Sub question correct (Only Sub)
- Both Main and Sub question wrong (Neither)

Each of these four quadrants can help us gauge whether the model tends to prefer perception questions, reasoning questions, both or neither. An optimal model has an accuracy of the first quadrant (Main + Sub) of 100% while all others are 0%.

	Question	Answer
MQ	Does this cause cavities?	Yes
SQ	What food is on the plate?	Dessert
	Is the boy eating whipped cream?	Yes
	Is his mouth wide open?	Yes

Table 3: **Example questions from VQA-Introspect**. MQ: Main Question; SQ: Sub Question

3.3. Linguistic Consistency

We evaluate how consistent each of the models are when dealing with rephrasings of the same question. Thus, we are evaluating each of the models on the VQA-Rephrasings data set. Examples of questions from the data set can be seen in Table 4. This data set provides $\sim$ 3 rephrasings per original question in a subset of the VQA v2 data set, in total providing ~ 158.500 questions. Similar to the VQA-Introspect data set we had to filter out answers not used in the original VQA v2 data set. This resulted in us having to remove $3,492$ questions out of $162,020$ total, because they had no answer available in the fixed list of answers. To evaluate the performance of the models, we use the consensus score presented in (Shah et al., 2019).

$$CS(k) = \sum_{Q' \subset Q, |Q'|=k} \frac{\mathcal{S}(Q')}{C_k} \quad (1)$$

$$\mathcal{S}(Q') = \begin{cases} 1, & \text{if } \forall q \in Q' \text{ is correct} \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

The consensus score is the ratio between the number of subsets where all answers are correct and the total number of subsets of size k. Where, C_k is the number of subsets of size k within a group of rephrasings (1 main question and 3 rephrasings).

	Question	Answer
OQ	How many lamps are in this picture?	2
RQ	Can you tell me how many lamps are in this picture?	2
	This picture has how many lamps in it?	2
	How many lamps do you see in this picture?	2

Table 4: **Example questions from VQA-Rephrasing** OQ: Original Question; RQ: Rephrased Question

4. Discussion

Here we will go into depth with analysing and interpreting the results produced by the experiments detailed in Section 3. We will follow the pattern from Section 3 and go into depth with each of the data sets we have evaluated in this study.

	Out-of-distribution				In-domain
Model↓/Data→	C	S	R	I	VQA v2
LXMERT	49.51	46.61	**68.64**	**76.90**	**71.96**
BAN	**51.63**	**52.39**	61.43	70.05	66.27

Table 5: **Overall accuracy on each data set.** C: Compose, S: Supplement, R: Rephrasings, I: Introspect.

4.1. Overall accuracy

First we look into the results concerning the overall accuracy of the two models (LXMERT and BAN) on each of the evaluation data sets and the original VQA v2 given in in Table 5. What is clear is that LXMERT does overall have an advantage over BAN and pre-training helps when evaluated on the VQA v2 data set itself. Strikingly, BAN however does perform better on the Compose and Supplement data sets. This suggests that LXMERT does not cope very well with multiple questions at once. Table 6 confirms that this is the case. While LXMERT works very well on single boolean operations, it is consistently outperformed by BAN on questions with multiple boolean operations.

4.2. Boolean Compositions

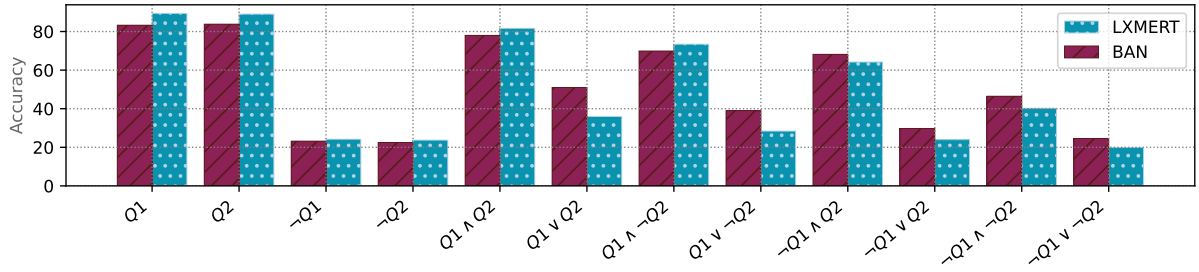
To better understand the diverging performance of the models over several boolean compositions, we investigate the per-tag performance for both models across compositions with single versus multiple operators. The results can be seen in Figure 3, while detailed results can be found in the appendix, Tables 10 and 11.

Model	Compose		Supplement	
	Single	Multi	Single	Multi
LXMERT	**57.33**	41.68	**55.32**	43.35
BAN	56.95	**46.30**	54.41	**51.63**

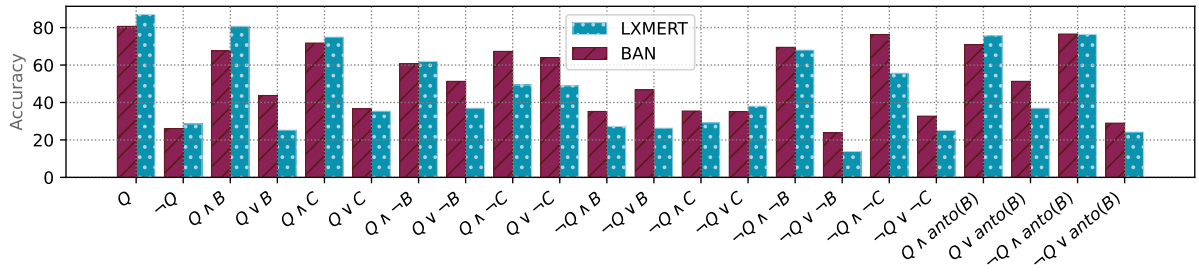
Table 6: **Accuracy for single vs multiple boolean operations.** Included are results for both VQA-Compose and VQA-Supplement.

Compose As is evident from Figure 3a both models struggle with compositions that use multiple operators. However, it seems to be much worse for LXMERT that drops $\sim$ 16 points in performance from single (57.33) to multi (41.68) on VQA-Compose. Comparing this to BAN, that only drops $\sim$ 10 points on the same test set. This same pattern is also evident when investigating the specific per-tag performance. Here the difference in performance from the original questions, $Q1$ and $Q2$, to the most complex composition, $\neg Q1 \vee \neg Q2$, is larger still with a difference of $\sim$ 69 for LXMERT, compared to that of BAN at $\sim$ 59.

While the performance of both models does drop as the complexity of the compositions increases, some of the operations are harder than others. The operation that seems to be the hardest is the $\vee$ (OR) operation, and the easiest seem to be the $\wedge$ (AND). What is striking is that BAN is better than LXMERT on questions involving $\vee$ (OR), as seen in Figure 3a. As is evident from the answer distributions in the Compose data set, it is severely skewed towards 'no' (see Table 12 in the appendix). Answers involving a single negation or $\vee$ (OR) are instead skewed towards 'yes'. We observe that both LXMERT and BAN have a preference to output a distribution which prefers to output 'no' over 'yes', and this is even more more pronounced for LXMERT. Consequently, we observe low results on OR questions. It is difficult to give a clear reason why the models are favoring the "no"-answer, as we found that the answer distribution in the training data is balanced (yes: 49.50% no: 50.50%). However, it does show that both models are brittle and particularly LXMERT is affected by shifts in answer distributions. Another interesting point from the VQA-Compose analysis and the evaluation is the composition of negated questions. It is worth noting that the questions in the Compose and Supplement can appear unnatural, a phenomena visible in the questions shown in Table 2. This is down to the data creation method which chooses to put a 'not' or 'no' either before a preposition, verb, or noun phrase at random. This data creation fact for negations could also play a factor for both models when trying to predict the correct answer.



(a) Accuracy per composition for LXMERT and BAN in the Compose data set



(b) Accuracy per composition for LXMERT and BAN in the Supplement data set

Figure 3: **Per composition accuracy for LXMERT and BAN.** The exact accuracy scores can be found in the Appendix in tables 10 and 11.

Supplement The patterns just discussed for the Compose data set are similar on the Supplement data set (Figure 3b). This is despite the new compositions add an extra layer of difficulty to the task. In the Supplement data set, the creators generate the compositions based on one question from the original data, and subsequently generate a question from a random object, that might or might not be in the image. This adds the extra task for the model to figure out if the object actually is in the image prior to answering the question. Both models are still good at the $\wedge$ operator and bad at the $\vee$ operator. Interestingly, for the questions where they use an antonym for the random object, both models are still able to answer the questions with high accuracy, as long as it is the $\wedge$ operator joining the two questions. As discussed for the Compose set the answer distribution has a big impact on the performance of the models.

4.3. Visual Reasoning

Next we investigate the performance of the models on the VQA-Introspect data set (Tables 7 and 8).

Model	Reasoning	Perception
LXMERT	86.14	75.79
BAN	81.50	68.42

Table 7: **Accuracy on the VQA-Introspect data set.**

The overall accuracy is given in Table 7, while Table 8 shows the performance across the four quadrants. Both models have higher accuracy on the reasoning questions compared to the perception questions. This is opposite to what would be expected of these models, where the hypothesis is that the models would have an easier time answering perception questions rather than reasoning questions. However, this one score does not guarantee that the models are reasoning about the answer in a consistent manner, as they might actually not be processing the question.

To see how consistent the models are across a reasoning and a perception question, we use the four quadrants shown in Table 8. As is quite evident, LXMERT does outperform BAN across all four quadrants. However, it does still answer many reasoning (main) questions correct without answering the perception (sub) question correctly, which is a sign of inconsistency. Using the two first quadrants (Main + Sub and Only Main) we find a consistency score (Selvaraju et al., 2020) for LXMERT of 77.85% versus 70.81% for BAN. These scores can be compared to the 71.73% consistency score presented in (Selvaraju et al., 2020), where they test the Pythia (Jiang et al., 2018) model on the data, in a similar way as we do (zero-shot). Both our models perform better than their Pythia model on the 'Main + Sub' quadrant, with LXMERT doing significantly better. Reviewing their results we can see that both LXMERT and BAN do a much better job at lowering the score of the last two quadrants. Yet, LXMERT has comparable results in the 'Only Main' to that of Pythia, while BAN does have worse results. It is really good at the main questions but not so good at the sub questions. When comparing the consistency of BAN and Pythia, it can be seen that Pythia is more consistent at 71.73%.

Model	Main + Sub	Only Main	Only Sub	Neither
LXMERT	**67.60**	**19.20**	**8.20**	**5.00**
BAN	58.34	24.05	10.08	7.53
Pythia	50.05	19.73	17.40	12.83

Table 8: **Quadrants of the VQA-Introspect data set.** Higher score on Main + Sub and lower score on other quadrants is better performance. Pythia scores reported from (Selvaraju et al., 2020).

4.4. Linguistic Consistency

Lastly we evaluate the models' ability to stay consistent across linguistic variation such as rephrasing. Using the consensus score from Equation (1) we can see how well the models stack up to answer the same question rephrased ~ 3 times.

In Table 9 we can first see the accuracy for the original in-distribution VQA v2 questions (ORI) compared to their rephrasings (REP). It is evident that LXMERT is more robust when it comes to changes in the language of the question, but performance drops overall for both (~ 8 to ~ 10).

On the out-of-distribution data, LXMERT is more consistent across a larger k. This robustness is evident in the smaller difference between $k = 1$ and $k = 4$, which for LXMERT is smaller compared to the difference for BAN ($\sim$ 15 vs. $\sim$ 18). The advantage of LXMERT over BAN could arise from the two-stream approach of LXMERT. Here, LXMERT first processes and attends to each of the modalities (Vision and Language) separately before attending across the modalities. This could help it be more robust to the smaller changes in the language part, helping it understand that it is the same question even though it is phrased differently. At the same time, the language part of the model is improved by the MLM pre-training task compared to BAN. Studying these two parts and their contribution is an interesting venue for future work.

Model	K				Accuracy	
	1	2	3	4	ORI	REP
LXMERT	**75.11**	**68.20**	**63.91**	**60.69**	**80.83**	**73.18**
BAN	67.37	58.54	53.34	49.77	74.75	64.91

Table 9: **Consensus score on VQA-Rephrasings.** The reaults are presented along with the accuracy on the original questions (ORI), and the rephrasings (REP).

5. Conclusion

Besides surveying current research on VQA, pre-trained Vision+Language models and evaluation of such models, in this paper we investigate the advantage of using pre-trained V+L models such as LXMERT over previous state-of-the-art models like BAN. We found that overall LXMERT does appear more robust to certain linguistic variances confirming the benefit to pre-training. We find that both LXMERT and BAN are fairly good at answering reasoning and perception questions, with high consistency in related questions (reasoning and perception questions).

However, there is still room for improvement, especially when answering more complex questions involving logic operators. We found that the models struggle with questions involving OR compositions, especially LXMERT. It underperforms in comparison to BAN and overpredicts 'no'. Some of these problems might be down to data set imperfections, such as skewed answer distributions like we saw with VQA-Compose and VQA-Supplement, and computer-generated questions being unnatural as shown in Table 2. However, strikingly models like LXMERT do not yet generalize well over such imperfections, they have difficulty with perception questions and should use the question and image to answer rather than relying on shortcuts like the prior answer distribution.

Our experiments in this paper has focused on the linguistic advantage of using pre-trained V+L models, and are limited to the specific selected architectures. However, another interesting aspect to investigate would be the visual advantage of using these pre-trained V+L models. This could be done using similar tests as performed in this paper, but evaluating on a data set such as IV-VQA or CV-VQA (Agarwal et al., 2020) that alters the images rather than the questions. Similarly, it would be interesting to extend this analysis to further architectures.

6. Bibliographical References

Agarwal, V., Shetty, R., and Fritz, M. (2020). Towards Causal VQA: Revealing and Reducing Spurious Correlations by Invariant and Covariant Semantic Editing. pages 9690–9698.

Agrawal, A., Lu, J., Antol, S., Mitchell, M., Zitnick, C. L., Batra, D., and Parikh, D. (2016). VQA: Visual Question Answering. *arXiv:1505.00468 [cs]*, October. arXiv: 1505.00468.

Agrawal, A., Batra, D., Parikh, D., and Kembhavi, A. (2018). Don't Just Assume; Look and Answer: Overcoming Priors for Visual Question Answering. *arXiv:1712.00377 [cs]*, June. arXiv: 1712.00377.

Alsentzer, E., Murphy, J. R., Boag, W., Weng, W.-H., Jin, D., Naumann, T., and McDermott, M. B. A. (2019). Publicly Available Clinical BERT Embeddings. *arXiv:1904.03323 [cs]*, June. arXiv: 1904.03323.

Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., and Zhang, L. (2018). Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. pages 6077–6086.

Biten, A. F., Litman, R., Xie, Y., Appalaraju, S., and Manmatha, R. (2021). LaTr: Layout-Aware Transformer for Scene-Text VQA. *arXiv:2112.12494 [cs]*, December. arXiv: 2112.12494.

Cao, J., Gan, Z., Cheng, Y., Yu, L., Chen, Y.-C., and Liu, J. (2020). Behind the Scene: Revealing the Secrets of Pre-trained Vision-and-Language Models. *arXiv:2005.07310 [cs]*, July. arXiv: 2005.07310.

Chen, X., Fang, H., Lin, T.-Y., Vedantam, R., Gupta, S., Dollar, P., and Zitnick, C. L. (2015). Microsoft COCO Captions: Data Collection and Evaluation Server. *arXiv:1504.00325 [cs]*, April. arXiv: 1504.00325.

Chen, Y.-C., Li, L., Yu, L., Kholy, A. E., Ahmed, F., Gan, Z., Cheng, Y., and Liu, J. (2020). UNITER: UNiversal Image-TExt Representation Learning. *arXiv:1909.11740 [cs]*, July. arXiv: 1909.11740 version: 3.

Demner-Fushman, D., Kohli, M. D., Rosenman, M. B., Shooshan, S. E., Rodriguez, L., Antani, S., Thoma, G. R., and McDonald, C. J. (2016). Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310, March.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. (2021). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv:2010.11929 [cs]*, June. arXiv: 2010.11929.

Gan, Z., Chen, Y.-C., Li, L., Zhu, C., Cheng, Y., and Liu, J. (2020). Large-Scale Adversarial Training for Vision-and-Language Representation Learning. *arXiv:2006.06195 [cs]*, October. arXiv: 2006.06195.

Gokhale, T., Banerjee, P., Baral, C., and Yang, Y. (2020). VQA-LOL: Visual Question Answering under the Lens of Logic. *arXiv:2002.08325 [cs]*, July. arXiv: 2002.08325.

Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., and Parikh, D. (2017). Making the v in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. pages 6904–6913.

Guo, J., Zhu, C., Zhao, Y., Wang, H., Hu, Y., He, X., and Cai, D. (2020). LAMP: Label Augmented Multimodal Pretraining. *arXiv:2012.04446 [cs]*, December. arXiv: 2012.04446.

Hendricks, L. A., Mellor, J., Schneider, R., Alayrac, J.-B., and Nematzadeh, A. (2021). Decoupling the Role of Data, Attention, and Losses in Multimodal Transformers. *arXiv:2102.00529 [cs]*, January. arXiv: 2102.00529.

Huang, Z., Zeng, Z., Liu, B., Fu, D., and Fu, J. (2020). Pixel-BERT: Aligning Image Pixels with Text by Deep Multi-Modal Transformers. *arXiv:2004.00849 [cs]*, June. arXiv: 2004.00849.

Hudson, D. A. and Manning, C. D. (2019). GQA: A New Dataset for Real-World Visual Reasoning and Compositional Question Answering. pages 6700–6709.

Jiang, Y., Natarajan, V., Chen, X., Rohrbach, M., Batra, D., and Parikh, D. (2018). Pythia v0.1: the Winning Entry to the VQA Challenge 2018. *arXiv:1807.09956 [cs]*, July. arXiv: 1807.09956.

Johnson, J., Hariharan, B., van der Maaten, L., Fei-Fei, L., Lawrence Zitnick, C., and Girshick, R. (2017). CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning. pages 2901–2910.

Kim, J.-H., Jun, J., and Zhang, B.-T. (2018). Bilinear Attention Networks. *arXiv:1805.07932 [cs]*, October. arXiv: 1805.07932.

Kim, W., Son, B., and Kim, I. (2021). ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 5583–5594. PMLR, July. ISSN: 2640-3498.

Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.-J., Shamma, D. A., Bernstein, M. S., and Fei-Fei, L. (2017). Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. *International Journal of Computer Vision*, 123(1):32–73, May.

Li, L. H., Yatskar, M., Yin, D., Hsieh, C.-J., and Chang, K.-W. (2019). VisualBERT: A Simple and Performant Baseline for Vision and Language. *arXiv:1908.03557 [cs]*, August. arXiv: 1908.03557.

Li, G., Duan, N., Fang, Y., Gong, M., and Jiang, D. (2020a). Unicoder-VL: A Universal Encoder for Vision and Language by Cross-Modal Pre-Training. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07):11336–11344, April. Number: 07.

Li, L., Gan, Z., and Liu, J. (2020b). A Closer Look at the Robustness of Vision-and-Language Pre-trained Models. *arXiv:2012.08673 [cs]*, December. arXiv: 2012.08673.

Li, X., Yin, X., Li, C., Zhang, P., Hu, X., Zhang, L., Wang, L., Hu, H., Dong, L., Wei, F., Choi, Y., and Gao, J. (2020c). Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks. *arXiv:2004.06165 [cs]*, July. arXiv: 2004.06165 version: 5.

Li, Y., Wang, H., and Luo, Y. (2020d). A Comparison of Pre-trained Vision-and-Language Models for Multimodal Representation Learning across Medical Images and Reports. In *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1999–2004, December.

Li, C., Yan, M., Xu, H., Luo, F., Wang, W., Bi, B., and Huang, S. (2021). SemVLP: Vision-Language Pre-training by Aligning Semantics at Multiple Levels. *arXiv:2103.07829 [cs]*, March. arXiv: 2103.07829.

Lin, T.-Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C. L., and Dollár, P. (2015). Microsoft COCO: Common Objects in Context. *arXiv:1405.0312 [cs]*, February. arXiv: 1405.0312.

Lin, J., Men, R., Yang, A., Zhou, C., Zhang, Y., Wang, P., Zhou, J., Tang, J., and Yang, H. (2021a). M6: Multi-Modality-to-Multi-Modality Multitask Mega-transformer for Unified Pretraining. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 3251–3261. Association for Computing Machinery, New York, NY, USA, August.

Lin, J., Yang, A., Zhang, Y., Liu, J., Zhou, J., and Yang, H. (2021b). InterBERT: Vision-and-Language Interaction for Multi-modal Pretraining. *arXiv:2003.13198 [cs]*, April. arXiv: 2003.13198.

Liu, H., Xu, S., Fu, J., Liu, Y., Xie, N., Wang, C.-C., Wang, B., and Sun, Y. (2021). CMA-CLIP: Cross-Modality Attention CLIP for Image-Text Classification. *arXiv:2112.03562 [cs]*, December. arXiv: 2112.03562.

Lu, J., Batra, D., Parikh, D., and Lee, S. (2019). ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. *arXiv:1908.02265 [cs]*, August. arXiv: 1908.02265.

Lu, J., Goswami, V., Rohrbach, M., Parikh, D., and Lee, S. (2020). 12-in-1: Multi-Task Vision and Language Representation Learning. pages 10437–10446.

Malinowski, M. and Fritz, M. (2014). A Multi-World Approach to Question Answering about Real-World Scenes based on Uncertain Input. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

Malinowski, M., Rohrbach, M., and Fritz, M. (2015). Ask Your Neurons: A Neural-Based Approach to Answering Questions About Images. pages 1–9.

Ordonez, V., Kulkarni, G., and Berg, T. (2011). Im2Text: Describing Images Using 1 Million Captioned Photographs. In *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc.

Parcalabescu, L., Gatt, A., Frank, A., and Calixto, I. (2020). Seeing past words: Testing the cross-modal capabilities of pretrained V&L models. *arXiv:2012.12352 [cs]*, December. arXiv: 2012.12352.

Qi, D., Su, L., Song, J., Cui, E., Bharti, T., and Sacheti, A. (2020). ImageBERT: Cross-modal Pre-training with Large-scale Weak-supervised Image-Text Data. *arXiv:2001.07966 [cs]*, January. arXiv: 2001.07966.

Ren, S., He, K., Girshick, R., and Sun, J. (2015). Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Roberts, A., Raffel, C., Lee, K., Matena, M., Shazeer, N., Liu, P. J., Narang, S., Li, W., and Zhou, Y. (2019). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. Technical report, Google.

Selvaraju, R. R., Tendulkar, P., Parikh, D., Horvitz, E., Ribeiro, M., Nushi, B., and Kamar, E. (2020). SQuINTing at VQA Models: Introspecting VQA Models with Sub-Questions. *arXiv:2001.06927 [cs]*, June. arXiv: 2001.06927.

Shah, M., Chen, X., Rohrbach, M., and Parikh, D. (2019). Cycle-Consistency for Robust Visual Question Answering. pages 6649–6658.

Sharma, P., Ding, N., Goodman, S., and Soricut, R. (2018). 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, Melbourne, Australia, July. Association for Computational Linguistics.

Silberman, N., Hoiem, D., Kohli, P., and Fergus, R. (2012). Indoor Segmentation and Support Inference from RGBD Images. In Andrew Fitzgibbon, et al., editors, *Computer Vision – ECCV 2012*, Lecture Notes in Computer Science, pages 746–760, Berlin, Heidelberg. Springer.

Simonyan, K. and Zisserman, A. (2015). Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv:1409.1556 [cs]*, April. arXiv: 1409.1556.

Singh, A., Natarajan, V., Shah, M., Jiang, Y., Chen, X., Batra, D., Parikh, D., and Rohrbach, M. (2019). Towards VQA Models That Can Read. pages 8317–8326.

Su, W., Zhu, X., Cao, Y., Li, B., Lu, L., Wei, F., and Dai, J. (2020). VL-BERT: Pre-training of Generic Visual-Linguistic Representations. *arXiv:1908.08530 [cs]*, February. arXiv: 1908.08530.

Sun, C., Myers, A., Vondrick, C., Murphy, K., and Schmid, C. (2019). VideoBERT: A Joint Model for Video and Language Representation Learning. pages 7464–7473.

Tan, H. and Bansal, M. (2019). LXMERT: Learning Cross-Modality Encoder Representations from Transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5100–5111, Hong Kong, China, November. Association for Computational Linguistics.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention Is All You Need. *arXiv:1706.03762 [cs]*, December. arXiv: 1706.03762.

Wang, Z., Yu, J., Yu, A. W., Dai, Z., Tsvetkov, Y., and Cao, Y. (2021). SimVLM: Simple Visual Language Model Pretraining with Weak Supervision. *arXiv:2108.10904 [cs]*, August. arXiv: 2108.10904.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A.,

Salakhudinov, R., Zemel, R., and Bengio, Y. (2015). Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 2048–2057. PMLR, June. ISSN: 1938-7228.

Xu, H., Yan, M., Li, C., Bi, B., Huang, S., Xiao, W., and Huang, F. (2021). E2E-VLP: End-to-End Vision-Language Pre-training Enhanced by Visual Learning. *arXiv:2106.01804 [cs]*, June. arXiv: 2106.01804.

Yang, Z., He, X., Gao, J., Deng, L., and Smola, A. (2016). Stacked Attention Networks for Image Question Answering. pages 21–29.

Yu, Z., Yu, J., Xiang, C., Fan, J., and Tao, D. (2018). Beyond Bilinear: Generalized Multimodal Factorized High-Order Pooling for Visual Question Answering. *IEEE Transactions on Neural Networks and Learning Systems*, 29(12):5947–5959, December. Conference Name: IEEE Transactions on Neural Networks and Learning Systems.

Yu, F., Tang, J., Yin, W., Sun, Y., Tian, H., Wu, H., and Wang, H. (2020). ERNIE-ViL: Knowledge Enhanced Vision-Language Representations Through Scene Graph. *arXiv:2006.16934 [cs]*, July. arXiv: 2006.16934 version: 2.

Zhang, S., Jiang, T., Wang, T., Kuang, K., Zhao, Z., Zhu, J., Yu, J., Yang, H., and Wu, F. (2020). DeVLBert: Learning Deconfounded Visio-Linguistic Representations. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 4373–4382. Association for Computing Machinery, New York, NY, USA, October.

Zhang, P., Li, X., Hu, X., Yang, J., Zhang, L., Wang, L., Choi, Y., and Gao, J. (2021). VinVL: Making Visual Representations Matter in Vision-Language Models. *arXiv:2101.00529 [cs]*, January. arXiv: 2101.00529.

Zhou, L., Palangi, H., Zhang, L., Hu, H., Corso, J., and Gao, J. (2020). Unified Vision-Language Pre-Training for Image Captioning and VQA. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07):13041–13049, April. Number: 07.

A. Results per composition analysis

Composition	LXMERT	BAN
$Q1$	89.39	83.26
$Q2$	89.06	83.80
$\neg Q1$	24.19	23.18
$\neg Q2$	23.71	22.49
$Q1 \wedge Q2$	81.65	77.99
$Q1 \vee Q2$	35.96	50.99
$Q1 \wedge \neg Q2$	73.44	69.89
$Q1 \vee \neg Q2$	28.40	38.98
$\neg Q1 \wedge Q2$	64.19	68.16
$\neg Q1 \vee Q2$	23.95	29.75
$\neg Q1 \wedge \neg Q2$	40.24	46.45
$\neg Q1 \vee \neg Q2$	19.89	24.57

Table 10: Per composition accuracy for the VQA-Compose data set.

Composition	LXMERT	BAN
Q	87.02	80.69
$\neg Q$	28.59	26.07
$Q \wedge B$	80.72	67.53
$Q \vee B$	25.26	43.76
$Q \wedge C$	74.94	71.76
$Q \vee C$	35.39	36.62
$Q \wedge \neg B$	61.93	60.57
$Q \vee \neg B$	36.96	51.32
$Q \wedge \neg C$	49.36	67.32
$Q \vee \neg C$	48.88	64.01
$\neg Q \wedge B$	27.21	35.24
$\neg Q \vee B$	26.30	46.73
$\neg Q \wedge C$	29.15	35.44
$\neg Q \vee C$	38.04	35.17
$\neg Q \wedge \neg B$	68.02	69.49
$\neg Q \vee \neg B$	13.79	23.81
$\neg Q \wedge \neg C$	55.67	76.33
$\neg Q \vee \neg C$	25.06	32.72
$Q \wedge anto(B)$	75.69	70.91
$Q \vee anto(B)$	36.96	51.32
$\neg Q \wedge anto(B)$	76.35	76.58
$\neg Q \vee anto(B)$	24.21	28.99

Table 11: Per composition accuracies for VQA-Supplement

B. Answer distributions

Composition	Yes	No
$Q1$	36.88	63.12
$Q2$	37.07	62.93
$\neg Q1$	63.12	36.88
$\neg Q2$	62.93	37.07
$Q1 \wedge Q2$	15.54	84.46
$Q1 \vee Q2$	58.41	41.59
$Q1 \wedge \neg Q2$	21.34	78.66
$Q1 \vee \neg Q2$	78.47	21.53
$\neg Q1 \wedge Q2$	21.53	78.47
$\neg Q1 \vee Q2$	78.66	21.34
$\neg Q1 \wedge \neg Q2$	41.59	58.41
$\neg Q1 \vee \neg Q2$	84.46	15.54

Table 12: Answer distribution per tag in VQA-Compose

Composition	Yes	No
Q	49.57	50.43
$\neg Q$	50.43	49.57
$Q \wedge B$	49.57	50.43
$Q \vee B$	100.0	00.00
$Q \wedge C$	49.57	50.43
$Q \vee C$	100.0	00.00
$Q \wedge \neg B$	00.00	100.0
$Q \vee \neg B$	49.57	50.43
$Q \wedge \neg C$	00.00	100.0
$Q \vee \neg C$	49.57	50.43
$\neg Q \wedge B$	50.43	49.57
$\neg Q \vee B$	100.0	00.00
$\neg Q \wedge C$	50.43	49.57
$\neg Q \vee C$	100.0	00.00
$\neg Q \wedge \neg B$	00.00	100.0
$\neg Q \vee \neg B$	50.43	49.57
$\neg Q \wedge \neg C$	00.00	100.0
$\neg Q \vee \neg C$	50.43	49.57
$Q \wedge anto(B)$	00.00	100.0
$Q \vee anto(B)$	49.57	50.43
$\neg Q \wedge anto(B)$	00.00	100.0
$\neg Q \vee anto(B)$	50.43	49.57

Table 13: Answer distribution per tag in VQA-Supplement