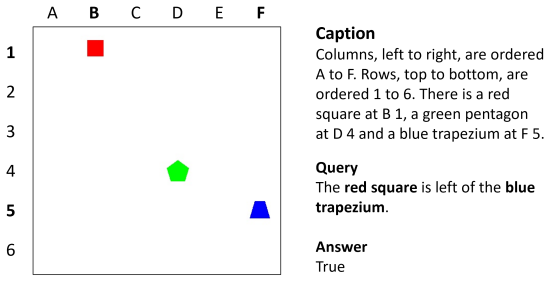
TRAVLR: Now You See It, Now You Don't! Evaluating Cross-Modal Transfer of Visio-Linguistic Reasoning

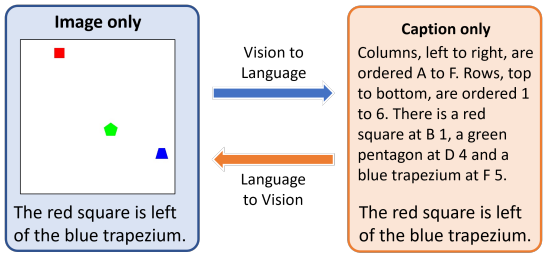
Keng Ji Chow[♮λ*] **Samson Tan**[♮§*] **Min-Yen Kan**[♮]
[♮]Department of Computer Science, National University of Singapore
[λ]Department of English Language and Literature, National University of Singapore
[§]Salesforce Research Asia
kengjichow@u.nus.edu
{samson.tmr, kanmy}@comp.nus.edu.sg

Abstract

Numerous visio-linguistic (V+L) representation learning methods have been developed, yet existing datasets do not evaluate the extent to which they represent visual and linguistic concepts in a unified space. Inspired by the crosslingual transfer and psycholinguistics literature, we propose a novel evaluation setting for V+L models: *zero-shot cross-modal transfer*. Existing V+L benchmarks also often report global accuracy scores on the entire dataset, rendering it difficult to pinpoint the specific reasoning tasks that models fail and succeed at. To address this issue and enable the evaluation of cross-modal transfer, we present TRAVLR, a synthetic dataset comprising four V+L reasoning tasks. Each example encodes the scene bimodally such that either modality can be dropped during training/testing with no loss of relevant information. TRAVLR's training and testing distributions are also constrained along task-relevant dimensions, enabling the evaluation of out-of-distribution generalisation. We evaluate four state-of-the-art V+L models and find that although they perform well on the test set from the same modality, all models fail to transfer cross-modally and have limited success accommodating the addition or deletion of one modality. In alignment with prior work, we also find these models to require large amounts of data to learn simple spatial relationships. We release TRAVLR as an open challenge for the research community.[1]



(a) A complete example for spatiality task.



(b) Possible directions of cross-modal transfer.

Figure 1: An example from TRAVLR (a). Both image and caption fully represent the scene; either can be dropped during training/testing, enabling the evaluation of cross-modal transfer ability (b).

1 Introduction

Research in psycholinguistics has found that human processing of spatial words activates brain regions associated with the visual system (Tang et al., 2021), suggesting the latter's involvement in processing linguistic input. It is therefore reasonable to expect multimodal neural models to resemble humans in this respect. Following its recent success in the text domain (Devlin et al., 2019), the pretraining–fine-tuning paradigm has been applied to the vision and text modalities to create unified visio-linguistic (V+L) representations. Just as pre-trained multilingual models have been shown capable of zero-shot cross-lingual transfer on various NLP tasks (Conneau et al., 2020), we may expect *true* V+L models to be capable of generalising to a modality not seen during fine-tuning.

However, current approaches of benchmarking V+L models often involve reporting global accuracy scores on the entire dataset, rendering the specific sources of success and failure difficult to diagnose (Ribeiro et al., 2020; Goel et al., 2021). For instance, Visual Question Answering (VQA, Goyal et al. 2017) tasks may allow models to exploit dataset bias (Dancette et al., 2021), or may reduce to object recognition problems which do not evaluate the models' ability to perform more complex tasks beyond aligning words or phrases in the text to a portion of the image (Hudson and

*Equal contribution
[1]Code and dataset to be released shortly.

Manning, 2019; Acharya et al., 2019), which does not require knowledge of syntactic structure or the ability to reason over several objects in a scene (Bernardi and Pezzelle, 2021). This concern is pertinent given that pretraining tasks often primarily involving masking either the textual or image modality.

Datasets such as NLVR2 (Suhr et al., 2019) address this limitation, but do not allow for fine-grained evaluation along specific dimensions (Tan et al., 2021). CLEVR (Johnson et al., 2017) and SHAPEWORLD (Kuhnle and Copestake, 2017) enable targeted evaluations of a V+L model's reasoning abilities but only encode the scene unimodally, as images. Additionally, their test examples may still be in the training distribution with respect to task-relevant dimensions, making it difficult to draw conclusions about generalisation ability.

We thus propose TRAVLR, a synthetic dataset comprising four V+L reasoning tasks: spatiality, cardinality, quantifiers, and numerical comparison. Unlike SHAPEWORLD, we control the train/test split such that examples in the out-of-distribution (OOD) test set are OOD *with respect to task-relevant dimensions*. We focus on tasks involving spatial and numerical reasoning, which require reasoning over multiple objects and have been shown to be challenging for V+L models (Johnson et al., 2017; Parcalabescu et al., 2020).

Inspired by the word/picture sentence verification task from psycholinguistics (Goolkasian, 1996), we further propose the cross-modal transfer setting, where the model is trained on input from one modality and tested on input from another. By representing the scene bimodally as both an image and a caption (Figure 1), TRAVLR is the first V+L dataset to support such an evaluation setting, to our knowledge. Being able to transfer cross-modally in a zero-/few-shot manner will improve data efficiency in applications where diverse image data is more difficult to obtain than written descriptions.

We use TRAVLR to evaluate the minimum amount of data and training steps required for various V+L models to learn simple reasoning tasks, in addition to comparing their final performance. We show that existing models often require unreasonably large amounts of data and training steps to learn simple tasks. We argue that our dataset serves as a basic sanity check for the abstract reasoning capabilities of models, and is complementary to datasets such as GQA (Hudson and Manning, 2019) that evaluate real-world object recognition and compositional reasoning abilities. Finally, we find current pretrained V+L models to be generally unsuccessful at learning to perform a task from one modality alone, and thus pose this as an open challenge for future V+L models.

2 Related Work

V+L tasks and datasets. The Visual Question Answering (VQA) task involves answering a question about an image, and is a complex task as it requires an ability to process input in both visual and textual modalities (Antol et al., 2015). A known issue with VQA datasets is the presence of real-world language priors and statistical biases in the training and testing distribution (Kervadec et al., 2021; Agrawal et al., 2018; Kafle et al., 2019). This was a problem with the original VQA dataset that Goyal et al. (2017) addresses in VQA v2.0 by balancing each query with pairs of images. However, Dancette et al. (2021) show that VQA v2.0 still contains both unimodal and multimodal biases that models can exploit. Furthermore, many questions in VQA use non-compositional language that do not require abilities beyond object recognition. Bernardi and Pezzelle (2021) argue that more complex reasoning tasks should involve reasoning about relationships between several objects in the image.

NLVR attempts to address the lack of compositionality in VQA using synthetically generated images of abstract 2D shapes accompanied by human-written English sentences to be judged true or false (Suhr et al., 2017). NLVR2 (Suhr et al., 2019) and SNLI-VE (Xie et al., 2019) also involve truth-value/entailment judgement tasks, and use photographs instead of synthetic images. Both lack detailed annotations of the specific semantic phenomena evaluated by each example. GQA improves over VQA by focusing on compositional questions that require reasoning over multiple objects and contains detailed annotations (Hudson and Manning, 2019), but still suffers from statistical imbalances and the lack of an out-of-distribution test set (Kervadec et al., 2021).

Other synthetic datasets focusing on reasoning include CLEVR (Johnson et al., 2017) and SHAPEWORLD (Kuhnle and Copestake, 2017). CLEVR is a fully synthetic 3D dataset and contains the annotations necessary to analyse model performance on specific tasks along various di-

mensions. SHAPEWORLD is a dataset targeting linguistic phenomena such as spatial relationships and quantifiers. gSCAN (Ruis et al., 2020) focuses on generalisation of commands within a 2D grid-world with objects, including various tasks such as novel composition of object properties, novel movement direction and novel adverbs.

V+L models. Pretrained V+L models differ in their architecture and pretraining methods. VL-BERT (Su et al., 2019), UNITER (Chen et al., 2020) and VisualBERT (Li et al., 2020a) are single-stream models with a single Transformer while ViLBERT (Lu et al., 2019), LXMERT (Tan and Bansal, 2019), and ALBEF (Li et al., 2021) are dual-stream models which encode image and textual inputs separately before fusing them. All models use a combination of masked language modelling and image-text matching objectives for pretraining, with LXMERT additionally pretraining on VQA and ALBEF using a contrastive loss to align the image and language representations. UNITER, VisualBERT, and LXMERT use a frozen Faster R-CNN (Ren et al., 2015) to extract region-based features from the image while ALBEF directly encodes the image with a Vision Transformer (Dosovitskiy et al., 2020).

Cross-modal transfer. Prior work has found models trained on multimodal data to perform better on unimodal downstream tasks than models trained only on one modality. Zadeh et al. (2020) found models trained on multimodal input to perform better than text-only models on three NLP tasks, while Testoni et al. (2019) showed that models trained on textual, visual, and auditory input were better at a quantification task than models trained only on a single modality. Using a task involving queries about typical colours of objects, Norlund et al. (2021) found that BERT trained on linguistic and visual features outperforms BERT trained on language data filtered for mentions of colour. Frank et al. (2021) investigated the cross-modal alignment of pretrained V+L models with an ablative method based on masked-modelling.

Summary. The datasets commonly used to evaluate V+L models such as VQA and NLVR2 lack fine-grained interpretability, due to the lack of annotations for semantic phenomena involved in each example. Additionally, multiple semantic phenomena co-occur within a single training example, making it difficult to control the training distribution and assess the generalisation abilities of models. In contrast, we show that task-specific investigation of the key reasoning capabilities of models can help to compare the data efficiency, performance and limitations of different models.

Existing V+L datasets also only present the scene in the visual modality and cannot be used to evaluate a V+L model's ability to generalise across modalities (cross-modal transfer). By encoding the underlying scene in both visual and textual modalities, we can evaluate cross-modal transfer by training on one and evaluating on the other.

Existing synthetic datasets (e.g., CLEVR and SHAPEWORLD) often fail to split the training and testing distributions along a dimension relevant to the specific task, because they generate captions based on randomly generated images. Our approach exploits the benefits of a synthetic dataset by strictly controlling the training and evaluation distributions to test the generalisation abilities of V+L models and avoid statistical biases from language priors and non-uniform distributions.

3 TRAVLR: Cross-Modal Transfer of Visio-Linguistic Reasoning

Psycholinguistic studies have demonstrated the effect of input modality on the performance of humans on truth-value judgement tasks. Goolkasian (1996)'s word/picture-sentence verification task found human subjects to exhibit faster reaction times and fewer errors when asked to provide truth value judgements on images as opposed to words, even when both encode the same underlying concept. We similarly ask if pretrained visio-linguistic models also exhibit asymmetries in accuracy and amount of required fine-tuning data when the input modality is varied.

There is also evidence that human infants learn abstract rules better when presented with bimodal cues such as visual shapes and speech sounds, compared to when information is presented in a single modality (Frank et al., 2009; Flom and Bahrick, 2007). We similarly ask if presenting the context in both visual and textual modalities improves performance for V+L models.

To answer these questions, we construct TRAVLR, a synthetic dataset comprising four visio-linguistic reasoning tasks. These tasks were previously identified to be challenging for text-only models (Lin and Su, 2021; Dua et al., 2019; Ravichander et al., 2019). TRAVLR aims to eval-

uate the extent to which pretrained V+L models already encode or are able to learn these four relations between entities present in the input scene. We first describe the general task format before elaborating on the cross-modal transfer problem.

Given a scene with objects, $S = \{o_1, ..., o_n\}$, where each object can be represented as a tuple $<$ *colour, shape, position* $>$, and a textual query q involving some relation $r(o_1, ..., o_i)$ between two or more objects in S, each task involves learning a function $y = f(S, q)$ where $y \in \{true, false\}$. This is essentially a binary classification task. For instance, in the spatiality task, the relation r could be *left* or *right*, which compares the positions of two objects. In the numerical comparison task, the noun phrases in the query refer to subsets of objects, while the relations (e.g., *more*) compare the cardinality of two sets of objects. Successfully assigning a truth value to the query thus involves reasoning over several objects (Bernardi and Pezzelle, 2021).

However, a model can never have direct access to the underlying representation scene in reality and must operate on visual or textual forms. Depending on the modality under evaluation, S may be presented in the form of an image or a textual description. In prior work such as VQA, S is presented as an image. In TRAVLR, S is represented bimodally as an $<$ *image, caption* $>$ pair.

Each example consists of an image, an accompanying caption, and a query. Images include abstract objects arranged in a grid, where each object has two properties: colour and shape. In our experiments, we draw from 5 possible colours (red, blue, green, yellow, orange) and 7 possible shapes (square, circle, triangle, star, hexagon, octagon, pentagon), giving 35 unique objects in total. Each caption fully describes the image with the coordinates of each object (e.g., "There is a red circle at A 1, a blue square at B 2..."). A description of the coordinate system, e.g., "Columns, left to right, are ordered A to F. Rows, top to bottom, are ordered 1 to 6." is prepended to the caption. The caption and query are separated by the `[SEP]` token when presented to the models. Removing the caption reduces our tasks to VQA-like tasks.

3.1 Reasoning Tasks

When generating the examples for each task, we constrain the training distribution along a dimension relevant to the specific task. For instance, in generating the training and out-of-distribution (OOD) test sets for the spatial relationship task, we ensure that the positions of the queried objects do not overlap between the training and test sets *along the relevant axis* (e.g., the horizontal axis for horizontal relations *left*/*right*). This differs from the approach adopted by SHAPEWORLD, which randomly generates images which are subsequently fed to a module responsible for generating query statements and assigning a truth value based on the corresponding scene. Consequently, the distribution of the images in SHAPEWORLD cannot be directly constrained depending on the specific task, and may lead to statistical bias in the distribution of queries. Furthermore, SHAPEWORLD does not enforce task-specific train/test splits. We next explain how we construct the train/test splits.

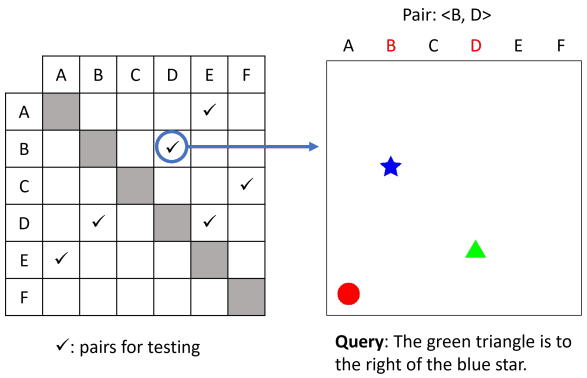


Figure 2: An example of OOD test set construction. In a left/right relationship reasoning task, the relevant dimension is the column ID. Specific ID pairs (✓) are held out to form this test distribution.

Spatiality. The spatiality task involves queries of the form "The `[object1]` is `[relationship]` the `[object2]`" (e.g., "The red circle is right of the blue triangle"), where the possible relationships are *to the left of*, *to the right of*, *above*, *below*. For horizontal relationships (left/right), the train and test sets are split based on the pair $<$`column(object1), column(object2)`$>$ (Figure 2), while for vertical relationships (above/below), the train and test sets are split based on the pair $<$`row(object1),row(object2)`$>$. This tests the model's ability to generalise its understanding of spatial relationships along the relevant dimension, as opposed to memorising fixed positions.

All	< [attr1] ∩ [attr2], [attr2] \ [attr1] >
Not all	< [attr1] ∩ [attr2], [attr1] \ [attr2] >
No	< [attr1] \ [attr2], [attr2] \ [attr1] >
Some	< [attr1] \ [attr2], [attr1] ∩ [attr2] >
Only	< [attr1] ∩ [attr2], [attr1] \ [attr2] >
Not only	< [attr1] ∩ [attr2], [attr2] \ [attr1] >

Table 1: Pairs for each quantifier.

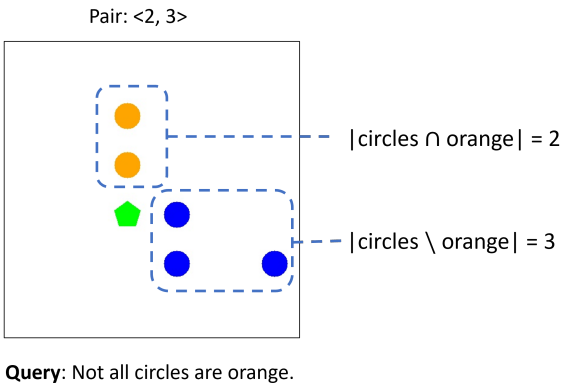


Figure 3: Example instance for *not all* quantifier with pair $< 2, 3 >$.

Cardinality. The cardinality task involves queries of the form "There are [number] [shape/colour] objects." (e.g., "There are 3 circle objects"). The train and test sets are split by the <number, shape/colour> pair occurring in the input image/caption. For instance, instances containing 2 circles and 3 triangles could occur in the training distribution, while instances containing 3 circles occur only in the OOD test distribution.

Quantifiers. This task involves queries of the form "[quantifier] the [attr1] objects are [attr2] objects.", where the quantifiers include *all*, *some*, *only* and their negated counterparts *not all*, *none* and *not only*. The train-test split is performed based on the pair $< a, b >$, which varies based on the quantifier, as given in Table 1. For instance, for the relationship *not all*, a is the number of objects which fulfil both [attr1] and [attr2], and b is the number of objects which fulfil [attr1] but not [attr2]. In the example in Figure 3, the pair is $< 2, 3 >$.

Numerical comparison. The numerical comparison task involves queries of the form "There are [more/fewer] [attr1] objects than [attr2] objects" (e.g., "There are more circles than squares."). The train and test sets are split by the pair $< a, b >$ where a is the number of [attr1] objects, and b is the number of [attr2] objects. Instances for which $|a - b|$ is smaller than some threshold is assigned to the training distribution, and the remaining pairs are assigned to the testing distribution. Success in this task is evidence of generalisation based on an implicit understanding of numeral scales and the transitivity of comparison i.e., $a > b$ and $b > c$ implies that $a > c$.

3.2 Cross-Modal Transfer

Humans can often reason about relationships between objects regardless of whether they are described with language or presented as an image. If pretrained V+L models have learnt a truly multimodal representation, they should similarly be able to learn a reasoning task with input from one modality and perform inference using input from the other modality with no extra training. We term this ability *zero-shot cross-modal transfer*, which may have significant implications for sample efficiency. Since annotated examples comprising diverse real-world images may be more difficult to collect compared to written descriptions, it may be desirable to be able to train multimodal models on only textual input before using them to process visual input. Furthermore, it is hoped that transfer from the visual modality can improve spatial reasoning ability even if the scene is represented as text instead of an image.

We draw an analogy to the concept of *zero-shot cross-lingual transfer* in multilingual NLP, which is often used to evaluate a multilingual model's ability to generalise to languages unseen during fine-tuning (Conneau et al., 2018). Similar to cross-modal transfer, a model is first pretrained on multiple languages before being fine-tuned on a task data from a single language. The model is then evaluated on examples from languages unseen during fine-tuning. Just as an ideal multilingual model is expected to perform well in this setting, we expect a perfectly multimodal model to perform just as well on the "unseen" modality.

Encoding the scene as both an image and a caption allows models to be trained and evaluated on a combination of three settings: i) image-only input, ii) caption-only input, and iii) both image and caption inputs. We note that the query is presented as part of the text input in each setting. In the caption-only setting, a blank white image is presented to the models. TRAVLR is, to our knowledge, the first dataset that supports the evaluation

Task	Train	Val.	InD Test	OOD Test
Spatial	15837 / 16163	4993 / 5007	5007 / 4993	9960 / 10040
Cardinality	4040 / 3960	4927 / 5073	5043 / 4957	10079 / 9921
Quantifier	4006 / 3994	5003 / 4997	5030 / 4970	10029 / 9971
Comparison	4088 / 3912	4926 / 5074	4992 / 5008	10033 / 9967

Table 2: Dataset statistics (no. of True / False)

of zero-shot cross-modal transfer.

3.3 Generating TRAVLR

We generate the dataset for each task separately. To generate each example, we select objects and determine their attributes with their values randomly sampled uniformly from the predefined distributions. The training and OOD test distributions are determined prior to the generation of both the input scene and queries based on the *pairs* explained above. We thus ensure that the *pairs* relevant to each task do not overlap between the train and OOD test sets, and also that all queries in the OOD test set cannot be found in the training set. Distractor objects irrelevant to the intended query are finally added to the scene.

For example, to generate queries for the spatial relationship task, we select two objects and their positions based on the training/testing distributions, before adding a distractor object to the scene. We then randomly select a relationship (e.g., either *left* or *right* for a horizontal relationship) for the query, which corresponds to either a `true` or `false` answer.

We also generate metadata for each example, comprising abstract representations of the input scene, the caption and the query, and crucial information about each example (e.g. the pairs). The spatiality task's training set comprises 32k examples, the training sets of the other tasks comprise 8k examples each due to differences in the amount of data required for convergence.

In- and out-of-distribution test sets. Prior work on generalisation evaluation recommended the use of in- and out-of-distribution (henceforth InD and OOD, respectively) test sets (Csordás et al., 2021). Hence, we include validation and InD test sets are randomly sampled from the training distribution (10k examples each) in addition to the OOD test set described in section 3.1 (20k examples). Table 2 summarises these statistics.

4 Experiments

Models. We perform experiments with VisualBERT, LXMERT, UNITER, and ALBEF. We use Li et al. (2020b)'s implementation of VisualBERT, LXMERT, and UNITER, and the original implementation of ALBEF. The image features of the first three models are 36 regions of interest extracted by a pretrained Faster R-CNN (Ren et al., 2015; Anderson et al., 2018), for which we use Tan and Bansal (2019)'s implementation.[2] We also use two text-only models, RoBERTa (Liu et al., 2019) and BERT (Devlin et al., 2019), as baselines in the caption-only setting.

Setting. We train models on each task for 80 epochs. Following Csordás et al. (2021)'s finding that early stopping may lead to underestimation of model performance, we do not do early stopping. Hyperparameters are fixed at a batch size of 256 and 2e-5 for ALBEF, based on the recommended parameters for fine-tuning on SNLI-VE (Xie et al., 2019), and a batch size of 32 and a learning rate of 5e-6 for VisualBERT, UNITER and LXMERT. As the hyperparameters recommended for fine-tuning on VQA on VisualBERT, UNITER and LXMERT did not lead to convergence on some tasks, we adjusted learning rates downwards which led to convergence or better performance on our dataset.

4.1 Within-Modality Results

We first discuss the results of *within-modality* testing, i.e., testing the model on the modality it was trained on (Table 3).

Spatiality. In the image-only setting, UNITER achieves the highest F_1 score, followed by LXMERT, VisualBERT, and finally ALBEF. VisualBERT requires at least 32k examples to achieve above random performance, while ALBEF completely fails to learn the task (Figure 4a). We note that 32k is a rather significant number of examples given the task's simplicity, where there are only 36 possible positions for each object. For comparison, the full VQA dataset, which aims to cover all possible tasks, consists of only 443k training examples. A potential explanation for the superior performance of UNITER and LXMERT could be that unlike the other models, spatial coordinates from the bounding boxes are explicitly encoded as features in the input to the image encoders, which they are able to directly exploit. This option is

[2] https://github.com/airsplay/py-bottom-up-attention

Spatiality

Train / Test	Image	Image: Caption	Image: Img. + Cap.	Caption	Caption: Caption	Caption: Img. + Cap.	Image + Caption	Image + Caption: Caption	Image + Caption: Img. + Cap.
VisualBERT	77.78 (+1.11)	46.21 (-0.18)	51.35 (-0.05)	40.79 (+0.56)	50.44 (+0.05)	50.49 (+0.00)	50.69 (+0.35)	50.50 (-0.56)	50.56 (-0.66)
UNITER	**99.63** (-0.09)	38.96 (+0.32)	**58.99** (-1.64)	45.46 (-1.69)	73.52 (-7.65)	73.57 (-7.64)	**52.08** (+0.68)	**97.66** (-2.34)	**97.58** (-2.42)
LXMERT	99.37 (-0.34)	33.44 (+0.08)	45.59 (+0.08)	33.42 (+0.12)	33.42 (+0.12)	33.42 (+0.12)	33.42 (+0.12)	33.42 (+0.12)	33.42 (+0.12)
ALBEF	48.01 (+0.08)	48.31 (+0.54)	48.14 (+0.58)	47.84 (-0.18)	**95.28** (-4.72)	**95.28** (-4.72)	39.31 (-1.10)	96.33 (-3.67)	96.34 (-3.66)

Cardinality

Train / Test	Image	Image: Caption	Image: Img. + Cap.	Caption	Caption: Caption	Caption: Img. + Cap.	Image + Caption	Image + Caption: Caption	Image + Caption: Img. + Cap.
VisualBERT	83.67 (-1.56)	46.78 (-0.85)	50.08 (-1.05)	42.34 (-1.47)	**99.82** (+0.28)	**99.77** (+0.62)	42.34 (-1.47)	**99.87** (+0.06)	**99.82** (+0.21)
UNITER	74.56 (-3.87)	47.72 (-0.13)	50.24 (-0.94)	42.34 (-1.47)	98.60 (+0.80)	98.77 (+1.04)	42.34 (-1.47)	98.82 (-0.04)	99.65 (+0.8)
LXMERT	**83.69** (-2.92)	33.28 (+0.05)	45.55 (+1.37)	33.51 (-0.01)	33.51 (-0.01)	33.51 (-0.01)	33.51 (-0.01)	33.51 (-0.01)	33.51 (-0.01)
ALBEF	59.23 (+0.39)	32.66 (-0.06)	**53.16** (-0.52)	43.91 (+0.8)	99.76 (+0.48)	99.76 (+0.48)	43.91 (-0.42)	99.57 (+0.35)	99.57 (+0.35)

Quantifiers

Train / Test	Image	Image: Caption	Image: Img. + Cap.	Caption	Caption: Caption	Caption: Img. + Cap.	Image + Caption	Image + Caption: Caption	Image + Caption: Img. + Cap.
VisualBERT	91.50 (-2.93)	34.44 (+0.26)	48.15 (-0.16)	46.89 (+0.49)	99.63 (+0.47)	99.65 (+0.51)	41.84 (+0.21)	99.48 (+0.50)	99.46 (+0.56)
UNITER	**96.93** (-0.12)	49.31 (-0.37)	54.10 (-0.85)	48.51 (-2.00)	99.40 (-0.28)	99.33 (-0.31)	44.79 (-1.35)	94.55 (+7.04)	97.30 (+4.78)
LXMERT	96.09 (-1.58)	36.63 (-1.06)	36.26 (-1.09)	48.45 (-1.24)	97.86 (+6.88)	93.99 (+7.59)	49.33 (-0.30)	34.75 (-0.07)	51.21 (+0.03)
ALBEF	60.45 (-2.15)	40.64 (-0.26)	**54.32** (-0.90)	48.45 (-0.07)	**99.97** (+0.00)	**99.97** (+0.00)	46.10 (-0.55)	**100.00** (+0.03)	**99.99** (+0.03)

Numerical Comparison

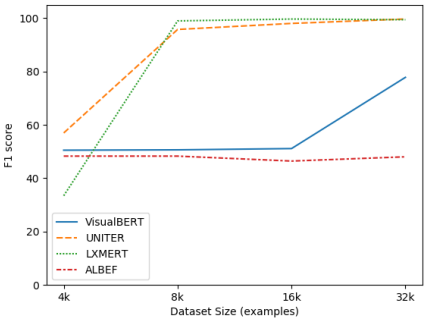
Train / Test	Image	Image: Caption	Image: Img. + Cap.	Caption	Caption: Caption	Caption: Img. + Cap.	Image + Caption	Image + Caption: Caption	Image + Caption: Img. + Cap.
VisualBERT	62.07 (-24.72)	33.83 (-2.32)	33.58 (-2.93)	49.99 (+1.23)	89.55 (-10.14)	89.49 (-10.24)	50.00 (+1.23)	81.92 (-16.77)	82.08 (-17.63)
UNITER	57.47 (-28.09)	34.18 (-1.61)	50.00 (-2.10)	47.79 (+0.06)	61.90 (-37.76)	60.59 (-39.04)	49.03 (-1.80)	61.38 (-37.92)	59.64 (-40.11)
LXMERT	**63.01** (-21.44)	46.09 (+0.98)	43.40 (-1.08)	46.50 (-2.70)	50.05 (-49.63)	50.11 (-49.52)	50.32 (+3.97)	58.75 (-22.42)	58.17 (-40.27)
ALBEF	47.28 (+4.14)	32.94 (+0.16)	32.94 (+0.16)	47.28 (-1.13)	**99.24** (-0.56)	**99.24** (-0.56)	44.98 (-4.25)	**97.46** (-2.45)	**97.44** (-2.45)

Table 3: F_1 scores on the OOD test sets for all four tasks (relative change from InD results in parentheses). Above random results are underlined; best result in each column is **bolded**.

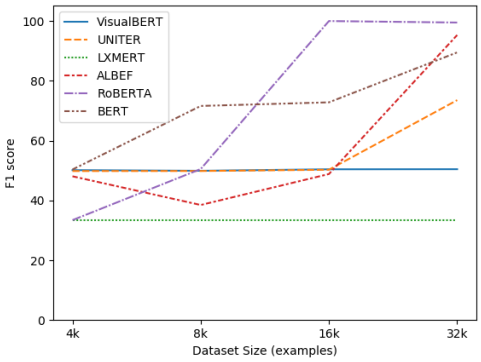
unavailable to ALBEF, which takes in the image as input directly instead of relying on a separate object detector. VisualBERT does not make use of these spatial coordinates, which may have impaired its ability to relate the positions of objects. Bugliarello et al. (2021) and Frank et al. (2021) posited this limitation of VisualBERT to be the reason for its poor performance on tasks such as RefCOCO+ and Masked Region classification, but the impact of this limitation on spatial reasoning has hitherto not been directly investigated.

Although LXMERT and UNITER achieve similar F_1 scores, UNITER succeeds at learning the task with substantially less data ($\leq$4k examples) compared to all the other models while LXMERT converges in fewer epochs. For instance, LXMERT only requires 4 epochs of training on the 32k dataset to exceed 99% accuracy on the validation set, while UNITER requires 39 epochs. A possible reason for the faster convergence of LXMERT on the spatiality task is that it was additionally pretrained on a VQA task, unlike all the other models. We can conclude that LXMERT is more efficient in terms of training steps, while UNITER is more sample efficient. Johnson et al. (2017) previously found CNN and LSTM models to have trouble learning spatial relationships and often memorise absolute object positions. Our results indicate that Transformer-based models likely face similar issues.

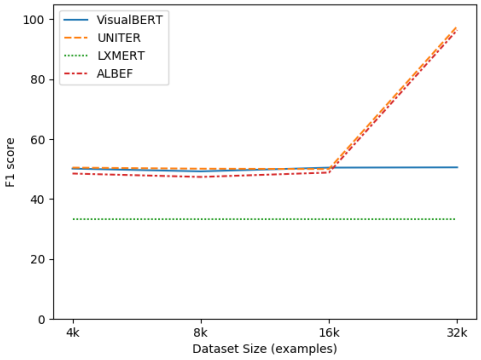
In the caption-only setting, only UNITER and ALBEF manage to achieve non-random performance. Only ALBEF achieves performance close to that of RoBERTa, which achieves an F_1 score of 99.46 on the OOD test set with 32k examples, but requires 16k examples to achieve above random performance (Figure 4b). BERT achieves an F_1 score of 89.47 on the OOD test set, outperforming all models other than ALBEF. Nevertheless, BERT requires at least 8k examples to achieve above random performance, corroborating findings by Lin and Su (2021) that BERT requires



(a) Image-only setting



(b) Caption-only setting



(c) Image+caption setting

Figure 4: Performance on spatiality task on (a) image-only, (b) caption-only and (c) image+caption input at different dataset sizes.

a significant number of examples to learn a simple natural language inference task.

While ALBEF achieves similar results in the caption-only and image+caption settings, UNITER's performance in the image+caption setting is significantly better than performance in the caption-only setting (Figure 4c). This may indicate a benefit to training UNITER on both modalities on the spatiality task.

Cardinality. The cardinality task requires less data than the spatiality task, and all models are able to achieve non-random performance in the settings where they were trained with 8k examples. In the image-only setting, LXMERT is the best performing model, followed by VisualBERT, UNITER, and finally ALBEF. Furthermore, performance on the OOD test set is poorer than performance on the InD test set for all models except ALBEF. Our results corroborate Parcalabescu et al. (2020)'s finding that current V+L models face difficulties counting objects in images.

All models are generally able to achieve close to a perfect F_1 score in the caption-only and image+caption settings, with the exception of LXMERT. It is notable that VisualBERT is the best performing model in the caption-only and image+caption settings, in contrast to its poor performance on the spatiality task. The performance of VisualBERT, UNITER and ALBEF are comparable to that of RoBERTa (OOD: 99.82; InD: 99.93) and BERT (OOD: 98.93; InD: 98.98). These results corroborate findings by Wallace et al. (2019) that numeracy is encoded in the embeddings of language-only models. We hypothesise that the poor performance of LXMERT compared to the other models is a result of not being initialised with BERT parameters prior to pretraining.

Quantifiers. All models perform well on the quantifiers task in most settings, with some exceptions. In the image-only setting, all models exceed an F_1 score of 90, except for ALBEF, which achieves an F_1 score of 60.45. Performance in the caption-only and image+caption settings are similar with the exception of LXMERT, and the best performing model is ALBEF, as in the numerical comparison task. Both RoBERTa and BERT achieve a F_1 score of 100 both the InD and OOD datasets. Good performance on the OOD dataset indicates that models are not memorising specific numbers of objects and instead use more general strategies for understanding quantifiers. This parallels psycholinguistic findings that comprehension of (non-exact) quantifiers does not correlate with counting skills in human children (Dolscheid et al., 2015).

Numerical comparison. Recall that the InD and OOD test sets for the comparison task are split based on the pair $< a, b >$ where a is the number of objects with the first attribute in the query and b

is the number of objects with the second attribute. In the main experiment, the value of $|a - b|$ in the InD test set is between 1 and 3, inclusive, and the maximum value of a and b is 9. In contrast to the simpler cardinality task, there is a significant difference between the InD and OOD settings for the numerical comparison task in across most settings, although the models still manage to achieve above random performance on the OOD test set.

In the image-only setting, performance on the InD test set is above 80 with the exception of ALBEF, which does not achieve above random performance. The performance of the other models on the OOD test set is significantly lower, between 55 to 65, indicating that all models only have a limited ability to generalise beyond the training distribution. In the caption-only setting, all models achieve close to an F_1 score of 100 on the InD test set, but do not generalise well to the OOD test set. Only ALBEF maintains a close to perfect F_1 score on the OOD test set, while VisualBERT (F_1=89.55) and UNITER (F_1=61.90) show a significant drop in performance, and LXMERT's performance is not better than random. Performance in the image+caption setting is similar to the caption-only setting, although performance on the OOD test set is poorer compared to the caption-only setting for all models, with the exception of LXMERT. Notably, the performance of ALBEF is like that of RoBERTa, which achieves similar results on OOD and InD test sets (OOD: 99.94; InD: 100), while VisualBERT and UNITER are closer to that of BERT which performs significantly more poorly on the OOD test set (OOD: 68.47; InD: 99.60).

Our results suggest that models are able to generalise to unseen number pairs by constructing an implicit numeral scale, but only to a limited extent. Furthermore, unlike the cardinality and quantifiers tasks, the numerical comparison task is able to differentiate the models' understanding of the numeral scale. ALBEF performs the best on the OOD test set, followed by VisualBERT, UNITER and finally, LXMERT. As explained earlier, a possible explanation for the poorer performance of LXMERT is that it was not initialised with BERT parameters prior to pretraining.

4.2 Adding/Dropping Modalities

We now discuss the effects of either adding or dropping a modality to the input presented during testing. Understood together with the observation of a clear similarity between the results in the caption-only and image+caption settings across all models and reasoning tasks, these results reveal a bias towards the textual modality across all models. Overcoming this bias is a potential step towards modality-agnostic representations.

First, models trained in the image+caption setting at times exhibit minor drops in performance when tested in the caption-only setting. In contrast, models trained in the image+caption setting perform poorly in the image-only setting in most cases, with random or close to random performance. The only exception is UNITER on the spatiality task, which achieves slightly above random performance when the caption is dropped during testing. This indicates a clear bias towards the textual input and a tendency to be distracted by the caption across all models.

Second, models trained only on captions perform similarly when tested in the image+caption setting. In contrast, testing a model trained only on images in the image+caption setting results in a significant performance drop. This is true even for the quantifiers task, which was shown to be the easiest for all models. In most cases, the F_1 score is either close to or below random chance, although ALBEF and UNITER differ from VisualBERT and LXMERT in managing to maintain above random performance when the caption is added to the input during testing.

4.3 Cross-Modal Transfer

Despite performing well in the within-modality settings, **none** of the models succeed at performing zero-shot cross-modal transfer to an unseen modality (i.e., from image-only to the caption-only setting, and vice versa). Our results suggest that existing V+L representation learning methods have not succeeded in producing truly multimodal, or modality-agnostic, representations.

5 Discussion

Asymmetry between image and text modalities. Thus far, we have seen that performance in the caption-only setting resembles performance in the image+caption setting across all tasks. Models may be distracted by the caption to the extent that they perform more poorly in the image+caption setting than in the image-only setting. Testing a model fine-tuned on both modalities on only one

modality reveals that models often rely heavily on the caption, ignoring the image completely, to the extent that they are unable to answer questions when the caption is removed. The overall finding is hence a bias towards the textual modality. This corroborates previous findings by Cao et al. (2020) that the textual modality plays a more important role than the image for both single and dual stream models. Furthermore, we find that V+L models perform poorer than unimodal RoBERTa on various caption tasks, similar to Iki and Aizawa (2021), who show that pretraining on V+L models cause poorer performance on NLU tasks.

Comparing tasks. The spatiality task is the hardest task, requiring at least 32k examples in some cases, as opposed to the 8k examples required for the other tasks. Focusing on the image-only setting, the easiest task is the quantifiers task (models achieve F_1 scores above 90), followed by cardinality (models achieve F_1 scores below 90), and finally numerical comparison (models achieve F_1 scores below 70). In the caption-only and image+caption settings, all models apart from LXMERT achieve a close to perfect F_1 score in the cardinality and quantifiers tasks, while all models except ALBEF suffer a performance degradation on the OOD dataset.

Our results thus suggest that while most models may succeed on the quantifiers task, they succeed at counting only to a limited extent. Furthermore, while success on the cardinality task indicates an understanding of the meaning of numbers in absolute terms, the numerical comparison task is able to more clearly differentiate the models in terms of their understanding of individual numbers' relative positions on a numeral scale.

Comparing models. In general, the performance of UNITER, VisualBERT and ALBEF in the caption-only and image+caption settings is better than performance in the image-only setting. In contrast, LXMERT appears to perform better in the image-only settings compared to the caption-only settings. Although UNITER achieves slightly higher results than LXMERT on the spatiality and quantifiers tasks, LXMERT significantly outperforms UNITER on the other tasks, likely due to its having been pretrained on a VQA task.

Our findings corroborate Bugliarello et al. (2021)'s findings that differences between models cannot be clearly attributed to differences in model architecture (i.e. whether they are single or dual-stream). Since LXMERT and ALBEF are both dual-stream models, our results suggest that the pretraining method has a significant effect on the model's performance on a downstream task. The performance of ALBEF in image-only settings is poorest amongst all models across all tasks. We hypothesise that the pretrained object detector used by the other models but not ALBEF confers an advantage on the image-only setting because the embeddings presented to the models already encodes the objects directly. We further note that while ALBEF may succeed at aligning phrases in the text to a portion of the image, all our tasks involving numerical reasoning include noun phrases which refer to multiple and spatially non-contiguous objects in the image.

UNITER is the only model which succeeds on all tasks on all settings, and seems to be less susceptible to performance degradation when modalities are added or removed from the input during test. These results suggest that some component of its architecture or pretraining procedure makes it less overly biased towards one modality.

6 Conclusion

While pretrained multilingual models have been shown to demonstrate zero-shot cross-lingual transfer abilities, it is unclear whether visio-linguistic models are similarly able to perform *zero-shot cross-modal transfer* of downstream task abilities to a modality unseen during training. We hence contribute a new dataset, TRAVLR, inspired by the word/picture sentence verification task from psycholinguistics. In contrast to existing V+L reasoning datasets that only encode the scene as an image, TRAVLR enables the evaluation of cross-modal transfer ability by encoding the scene in both the visual and textual modalities, allowing either to be dropped during training or testing.

TRAVLR allows us to evaluate specific visio-linguistic reasoning skills in isolation instead of at an aggregate level, enabling finer-grained diagnosis of a model's deficiencies. We found some models to learn better from one modality than the other, and some task-setting combinations to be more challenging across the board. Our results also provide useful estimates of the amount of data required for V+L models to acquire various reasoning skills, indicating that existing models may require unreasonably large amounts of data

and training steps to learn certain types of visio-linguistic reasoning. Improving the sample efficiency and training time of V+L models in this regard is a potential direction for future research.

We further found all models to suffer from a bias towards the textual modality and are unable to perform zero-shot cross-modal transfer of reasoning capabilities despite, in some cases, achieving close to perfect performance on a test set encoded in the same modality. Developing new visio-linguistic representations that are capable of zero-shot cross-modal transfer is another direction for future research, and we pose this as a new challenge for multimodal modelling.

References

Manoj Acharya, Kushal Kafle, and Christopher Kanan. 2019. TallyQA: Answering complex counting questions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8076–8084.

Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. 2018. 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.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*.

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

Raffaella Bernardi and Sandro Pezzelle. 2021. Linguistic issues behind visual question answering. *Language and Linguistics Compass*, 15(6):e12417.

Emanuele Bugliarello, Ryan Cotterell, Naoaki Okazaki, and Desmond Elliott. 2021. Multimodal pretraining unmasked: A meta-analysis and a unified framework of vision-and-language BERTs. *Transactions of the Association for Computational Linguistics*, 9:978–994.

Jize Cao, Zhe Gan, Yu Cheng, Licheng Yu, Yen-Chun Chen, and Jingjing Liu. 2020. Behind the scene: Revealing the secrets of pre-trained vision-and-language models. In *European Conference on Computer Vision*, pages 565–580. Springer.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. UNITER: Universal image-text representation learning. In *ECCV*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Róbert Csordás, Kazuki Irie, and Juergen Schmidhuber. 2021. The devil is in the detail: Simple tricks improve systematic generalization of transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 619–634.

Corentin Dancette, Remi Cadene, Damien Teney, and Matthieu Cord. 2021. Beyond question-based biases: Assessing multimodal shortcut learning in visual question answering. *arXiv preprint arXiv:2104.03149*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Sarah Dolscheid, Christina Winter, and Martina Penke. 2015. Counting on quantifiers: Specific links between linguistic quantifiers and number acquisition. In *EAPCogSci*.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2368–2378.

Ross Flom and Lorraine E Bahrick. 2007. The development of infant discrimination of affect in mul-

timodal and unimodal stimulation: The role of intersensory redundancy. *Developmental psychology*, 43(1):238.

Michael C Frank, Jonathan A Slemmer, Gary F Marcus, and Scott P Johnson. 2009. Information from multiple modalities helps 5-month-olds learn abstract rules. *Developmental science*, 12(4):504–509.

Stella Frank, Emanuele Bugliarello, and Desmond Elliott. 2021. Vision-and-language or vision-for-language? On cross-modal influence in multimodal transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9847–9857.

Karan Goel, Nazneen Fatema Rajani, Jesse Vig, Zachary Taschdjian, Mohit Bansal, and Christopher Ré. 2021. Robustness gym: Unifying the NLP evaluation landscape. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations*, pages 42–55.

Paula Goolkasian. 1996. Picture-word differences in a sentence verification task. *Memory & Cognition*, 24(5):584–594.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6325–6334. IEEE Computer Society.

Drew A Hudson and Christopher D Manning. 2019. 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.

Taichi Iki and Akiko Aizawa. 2021. Effect of vision-and-language extensions on natural language understanding in vision-and-language models. *arXiv preprint arXiv:2104.08066*.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. 2017. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1988–1997. IEEE.

Kushal Kafle, Robik Shrestha, and Christopher Kanan. 2019. Challenges and prospects in vision and language research. *Frontiers in Artificial Intelligence*, 2:28.

Corentin Kervadec, Grigory Antipov, Moez Baccouche, and Christian Wolf. 2021. 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.

Alexander Kuhnle and Ann Copestake. 2017. Shapeworld-a new test methodology for multimodal language understanding. *arXiv preprint arXiv:1704.04517*.

Junnan Li, Ramprasaath R Selvaraju, Akhilesh Deepak Gotmare, Shafiq Joty, Caiming Xiong, and Steven Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. In *Thirty-Fifth Conference on Neural Information Processing Systems*.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2020a. What does BERT with vision look at? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5265–5275.

Yikuan Li, Hanyin Wang, and Yuan Luo. 2020b. 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. IEEE.

Yi-Chung Lin and Keh-Yih Su. 2021. How fast can bert learn simple natural language inference? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 626–633.

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

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. ViLBERT: pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pages 13–23.

Tobias Norlund, Lovisa Hagström, and Richard Johansson. 2021. Transferring knowledge from vision to language: How to achieve it and how to measure it? In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 149–162.

Letitia Parcalabescu, Albert Gatt, Anette Frank, and Iacer Calixto. 2020. Seeing past words: Testing the cross-modal capabilities of pretrained V&L models. *arXiv preprint arXiv:2012.12352*.

Abhilasha Ravichander, Aakanksha Naik, Carolyn Rose, and Eduard Hovy. 2019. Equate: A benchmark evaluation framework for quantitative reasoning in natural language inference. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 349–361.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster R-CNN: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28:91–99.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with checklist. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912.

Laura Ruis, Jacob Andreas, Marco Baroni, Diane Bouchacourt, and Brenden M Lake. 2020. A benchmark for systematic generalization in grounded language understanding. *Advances in Neural Information Processing Systems*, 33.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2019. VL-BERT: Pre-training of generic visual-linguistic representations. In *International Conference on Learning Representations*.

Alane Suhr, Mike Lewis, James Yeh, and Yoav Artzi. 2017. A corpus of natural language for visual reasoning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 217–223.

Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. 2019. A corpus for reasoning about natural language grounded in photographs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6418–6428.

Hao Tan and Mohit Bansal. 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.

Samson Tan, Shafiq Joty, Kathy Baxter, Araz Taeihagh, Gregory A. Bennett, and Min-Yen Kan. 2021. Reliability testing for natural language processing systems. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4153–4169, Online. Association for Computational Linguistics.

Jerry Tang, Amanda LeBel, and Alexander G Huth. 2021. Cortical representations of concrete and abstract concepts in language combine visual and linguistic representations. *bioRxiv*.

Alberto Testoni, Sandro Pezzelle, and Raffaella Bernardi. 2019. Quantifiers in a multimodal world: Hallucinating vision with language and sound. In *Proceedings of the workshop on cognitive modeling and computational linguistics*, pages 105–116.

Eric Wallace, Yizhong Wang, Sujian Li, Sameer Singh, and Matt Gardner. 2019. Do NLP models know numbers? Probing numeracy in embeddings. 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 5307–5315.

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

Amir Zadeh, Paul Pu Liang, and Louis-Philippe Morency. 2020. Foundations of multimodal co-learning. *Information Fusion*, 64:188–193.