# VLUE: A Multi-Task Benchmark for Evaluating Vision-Language Models

**Wangchunshu Zhou** [*1] **Yan Zeng** [*1] **Shizhe Diao** [*2] **Xinsong Zhang** [*1]

## Abstract

Recent advances in vision-language pre-training (VLP) have demonstrated impressive performance in a range of vision-language (VL) tasks. However, there exist several challenges for measuring the community's progress in building general multi-modal intelligence. First, most of the downstream VL datasets are annotated using raw images that are already seen during pre-training, which may result in an overestimation of current VLP models' generalization ability. Second, recent VLP work mainly focuses on absolute performance but overlooks the efficiency-performance trade-off, which is also an important indicator for measuring progress.

To this end, we introduce the Vision-Language Understanding Evaluation (VLUE) benchmark, a multi-task multi-dimension benchmark for evaluating the generalization capabilities and the efficiency-performance trade-off ("Pareto SOTA") of VLP models. We demonstrate that there is a sizable generalization gap for all VLP models when testing on out-of-distribution test sets annotated on images from a more diverse distribution that spreads across cultures. Moreover, we find that measuring the efficiency-performance trade-off of VLP models leads to complementary insights for several design choices of VLP. We release the VLUE benchmark[1] to promote research on building vision-language models that generalize well to more diverse images and concepts unseen during pre-training, and are practical in terms of efficiency-performance trade-off.

## 1. Introduction

Building systems that can demonstrate their visual understanding by generating or responding to natural language in the context of images has been a long-standing goal in the field of artificial intelligence and cognitive science (Boden, 2008). The approaches and corresponding tasks have come to be referred to under the common banner of 'vision-and-language' (Lu et al., 2019). Vision-Language Pre-training (VLP), the paradigm of pre-training on large-scale parallel image-text pairs and then fine-tuning on vision-language (VL) tasks, has achieved state-of-the-art performance on a wide range of VL tasks. It has transformed the landscape of vision-and-language research and is considered to be a critical step for building general multi-modal intelligence. Over the last two years, a great number of research studies (Lu et al., 2019; Tan & Bansal, 2019; Chen et al., 2020; Li et al., 2020a; 2019; 2020b; Cho et al., 2021; Zhang et al., 2021; Li et al., 2021a; Jia et al., 2021; Zeng et al., 2021; Wang et al., 2021a;b; Diao et al., 2022) have been conducted in the field of vision-language pre-training. Nevertheless, it becomes increasingly complicated to track the *real* progress that the vision-language community is actually making because of two major problems in the common practice for evaluating and reporting new studies in the field. We elaborate them as follows.

First, most datasets of downstream VL tasks used for evaluating VLP models are annotated with images collected from the COCO (Lin et al., 2014) or Visual Genome (VG) (Krishna et al., 2017) dataset[2]. The problem is that most (if not all) VLP models are pre-trained on image-text pairs from COCO and VG datasets. Therefore, VLP models have already seen the images in the downstream datasets and their captions before fine-tuning and evaluating on them. As such, fine-tuning and evaluating VLP models on these datasets only measures their transferability in the setting where at least a part of data distribution (i.e., image distribution) remains the same while the label distribution shifts. This is a special case of transfer learning and is unlikely to be met in practice. Consequently, evaluating on these "in-domain"

[*]Equal contribution [1]ByteDance AI Lab [2]The Hong Kong University of Science and Technology. Correspondence to: Wangchunshu Zhou <wcszhou@outlook.com>.

*Proceedings of the $39^{th}$ International Conference on Machine Learning*, Baltimore, Maryland, USA, PMLR 162, 2022. 

[1]The benchmark is publicly available at https://vlue-benchmark.github.io. The data and codes used for training baseline models are available at https://github.com/MichaelZhouwang/VLUE.

[2]A few studies exclude the images used in some downstream datasets from their pre-training data. However, this does not fully resolve this issue because some datasets are not annotated within a specific subset. Also, even the images are not seen during pre-training, they are still from the same distribution.

datasets will lead to a biased and probably overestimated transfer and generalization ability of VLP models.

Second, most recent studies on VLP mainly focus on the absolute performance improvement but ignore the efficiency-performance trade-off, which is also very important for the application of VLP models in real-world applications. This issue is likely to be more severe because recently super-sized VLP models are pushing the state-of-the-art (SOTA) of many VL tasks to a new level, making it impossible for most researchers with moderate computation resources to reach results exceed or comparable with SOTA. This phenomenon is common in the field of natural language processing (NLP) and most studies are instead pursuing improvement on other dimensions such as the efficiency-performance trade-off (Liu et al., 2021b; Xu et al., 2021e; Jiao et al., 2020; Xu et al., 2020b; Zhou et al., 2021e;d; Xu & McAuley, 2022a;b; Zhou et al., 2021c; Xu et al., 2021a;b). We refer to the goal of these studies as "Pareto SOTA" following (Liu et al., 2021b), which means that there is no other model currently better than it on all the dimensions of interest such as performance and efficiency. Therefore, we believe it is necessary to measure and report performance-efficiency trade-off of VLP models to promote and facilitate research in the field.

In addition, the evaluation protocol used in recent VLP studies are not consistent enough and different studies report results on a different set of tasks, datasets, or experimental settings. Consequently, it is complicated for researchers to compare their methods to existing ones and for the VLP community to track progress. This is because the lack of a standard evaluation benchmark like GLUE (Wang et al., 2019b) and SuperGLUE (Wang et al., 2019a) for natural language understanding research and XTREME (Hu et al., 2020) and XGLUE (Liang et al., 2020) for multi-lingual generalization of pre-trained models.

To address these problems and promote research on truly generalizable and practical VLP, we introduce the **V**ision-**L**anguage **U**nderstanding **E**valuation (**VLUE**) benchmark. VLUE is the first multi-task benchmark focusing on vision-language understanding that covers a set of fundamental VL tasks including image-text retrieval, visual question answering, visual reasoning, and visual grounding, and maintains a leaderboard tracking the performance of representative studies and new methods on VLP. More importantly, VLUE includes a newly annotated private out-of-distribution (OOD) test set for each representative VL task. In contrast to standard datasets for these tasks that are annotated on COCO/VG images, our private OOD test sets are annotated on images from the MaRVL (Liu et al., 2021a) dataset where images are manually collected across cultures by native speakers from different countries. This ensures that the image distribution in our OOD test sets differs from that of COCO/VG images. Moreover, we carefully control the annotation protocol for our OOD test sets to be identical to the original in-domain datasets. As such, the label distribution in our OOD test sets is roughly the same as the original test set but the image distribution differs. This enables us to better measure the *true* generalization and transferability of VLP models. In addition, we also encourage researchers to measure and compare the efficiency-performance trade-off when reporting new studies in the field of VLP. To facilitate that, we measure the efficiency-performance trade-off of representative VLP models in VLUE to track a Pareto SOTA landscape for VLP research. In general, in contrast to conventional benchmarks that only capture the single performance metric, VLUE is a multi-dimension benchmark that takes multiple dimensions including performance, generalization ability, and efficiency into account. Hopefully, this will promote research on VLP models that are environmentally friendly and practical for real-world applications.

We evaluate a range of representative VLP models on VLUE to facilitate future research and analyze their generalization ability and efficiency-performance trade-off with respect to several key design choices. We find that there is a sizable generalization gap for all VLP models when evaluating on new examples annotated with images from in-the-wild distribution. Also, compared to focusing on a single dimension (i.e., absolute performance), measuring the generalization ability of different models can lead to complementary and even controversial conclusions. We also find that models with similar performance may result in completely different positions in the Pareto front measuring the efficiency-performance trade-off of VLP models, which also demonstrates the necessity of a multi-dimension benchmark for evaluating VLP models.

In sum, our contributions are the following: (i) We release a vision-language benchmark consisting of 4 representative VL tasks, each equipped with a private test set annotated on images from wild distribution; (ii) We provide an online platform and leaderboard for the evaluation and comparison of VLP models and provide a set of strong baselines, which we evaluate across all tasks, and release code to facilitate adoption; (iii) We evaluate the efficiency-performance trade-off of representative VLP models and build a Pareto SOTA landscape for current VLP research; (iv) We provide an extensive analysis of the generalization ability and the efficiency-performance trade-off of representative VLP models.

## 2. Related Work

The *pre-training then fine-tuning* paradigm (Peters et al., 2018; Radford et al., 2018; Devlin et al., 2019; Liu et al., 2019; Radford et al., 2019; Raffel et al., 2019; Lewis et al., 2020; Diao et al., 2020; Brown et al., 2020; Dong et al.,

2019; Zhou et al., 2021a;b; Fu et al., 2022) has substantially advanced the state-of-the-art of natural language processing on a wide range of NLP tasks (Wang et al., 2019b;a; Warstadt et al., 2019; Socher et al., 2013; Dolan & Brockett, 2005; Agirre et al., 2007; Williams et al., 2018; Rajpurkar et al., 2016; Dagan et al., 2006; Lin et al., 2020; Xu et al., 2020a; Diao et al., 2021; Xu et al., 2021c). The success motivates researchers to adopt this paradigm to improve vision-language tasks. The idea of vision-language pre-training is to pre-train a general-purpose vision-language model on large-scale parallel image-text pairs and then fine-tune on downstream vision-language tasks.

The approaches of vision-language pre-training can be categorized according to two major criteria. The first dimension is how the visual input is represented in the vision-language model. They are typically two approaches. Most existing methods (Tan & Bansal, 2019; Lu et al., 2019; Li et al., 2019; 2020a; Chen et al., 2020; Li et al., 2020b; Gan et al., 2020; Li et al., 2021c) represent an image by dozens of object-centric features of regions of interest which are identified and extracted by object detection. They either utilize pre-trained object detectors (Ren et al., 2015; Anderson et al., 2018) or conduct object detection on-the-fly in the pre-training process (Su et al., 2020; Xu et al., 2021d). More recently, a number of work (Kamath et al., 2021; Yang et al., 2021) explored integrating detection into an end-to-end pre-training procedure. The other approaches take raw image pixels as vision input, and extract overall image features with convolutional network (Huang et al., 2020; 2021) or vision transformer (Kim et al., 2021; Li et al., 2021a).

The second dimension is how the vision and language modalities interact with each other in the vision language model, which leads to two categories: early fusion models and late fusion models. Late fusion models such as CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021) encode images and texts separately with a dual encoder architecture, and use cosine similarity or a linear projection layer to model the cross-modality interaction. The dual encoder design is effective for retrieval tasks. However, late fusion makes the interaction cross modalities too simple to handle tasks that require complex reasoning, such as visual reasoning and visual question answering. The early fusion models (Lu et al., 2019; Tan & Bansal, 2019; Li et al., 2019; Cho et al., 2021; Zhang et al., 2021; Li et al., 2021a; Zeng et al., 2021; Wang et al., 2021a;b) instead use a deep fusion encoder with cross-modal attention to improve cross-modal interaction, leading to improved performance for vision-language understanding tasks.

In addition, Cao et al. (2020) developed VALUE (short for vision-and-language understanding evaluation), a suite of probing tasks aiming to understand the inner workings of VLP models. The tasks included in VLUE are instead realistic vision and language tasks and have real-world applications. Therefore, VLUE can better serve as a standard benchmark for the evaluation of VLP models. Moreover, Li et al. (2021b) developed another VALUE benchmark where the "V" stands for video. It focuses on video-and-language tasks whereas VLUE focuses on image-and-language tasks. Recently, Su et al. (2021) introduced the GEM benchmark. GEM is a multimodal benchmark that focuses on both image-language tasks and video-language tasks. Different from VLUE, GEM focuses on multilingual multimodal models and only considers the image-text retrieval task and image captioning task.

## 3. VLUE

VLUE is a multi-task multi-dimension vision-language benchmark with the goal of providing an accessible platform and a standard practice for the evaluation of VLP models. In this section, we first describe the tasks and datasets included in the VLUE benchmark. We then introduce two additional dimensions for evaluating new models: generalization ability and efficiency performance trade-off. We describe our private out-of-distribution (OOD) test set that is annotated on images from wild distribution. Finally, we introduce the idea and protocol for evaluating the efficiency-performance trade-off of VLP models.

### 3.1. Tasks

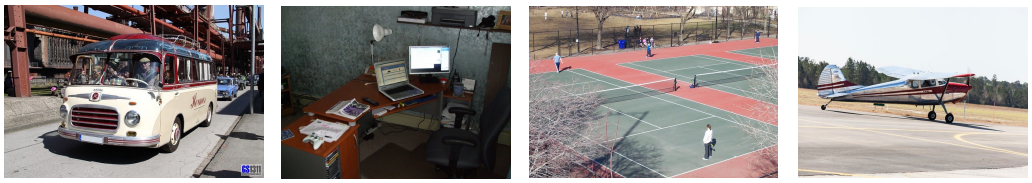
VLUE consists of five fundamental tasks requiring vision-language understanding and reasoning. We give an overview of all tasks and the corresponding datasets in Table 1, and describe the details as follows:

**Image-Text Retrieval** There are two subtasks: text retrieval (TR), where images are queries and texts are targets, and image retrieval (IR), where texts are queries and images are targets. The performance for both subtasks is measured by R@K (recall with top k predictions). This task requires VLP models to align the semantic space of vision and language modalities so that two views of a scene are represented similarly in the vector space. Previous studies generally evaluate on MSCOCO (Lin et al., 2014) and Flicker30K (Plummer et al., 2015). We opt to MSCOCO because the performance of recent VLP models is saturating on Flicker30K (e.g., ALBEF (Li et al., 2021a) achieves 100% R@10 for TR and 98.9% R@10 for IR).

**Visual Reasoning** The natural language visual reasoning task is a binary classification task that takes a pair of images and a natural language statement as input and judges if the statement is true for the image pair. The task requires understanding complex compositional language in visual context. Multiple datasets including NLVR (Suhr et al.,

| Task | Dataset | Image Domain | \|Train\| | \|Dev\| | \|Test\| | \|OOD Test\| | Metric |
|---|---|---|---|---|---|---|---|
| Image-Text Retrieval | MSCOCO | COCO | 566,747 | 25,010 | 25,010 | 27,796 | R@1 |
| Image Captioning | MSCOCO | COCO | 566,747 | 25,010 | 25,010 | 27,796 | BLEU/CIDER |
| Visual Grounding | RefCOCO+ | COCO | 120,191 | 10,758 | 10,615 | 1,313 | Accuracy |
| Visual Reasoning | NLVR2 | Google Images[3] | 86,373 | 6,982 | 6,967 | 5,662 | Accuracy |
| Visual Question Answering | VQA 2.0 | COCO | 443,757 | 214,354 | 447,793 | 11,942 | Accuracy |

Table 1: Characteristics of the datasets in VLUE.



(a) Images from COCO dataset

(b) Images from MaRVL dataset

Figure 1: Images random sampled from the COCO dataset (top) and the MaRVL dataset (bottom).

2017) and CLEVR (Johnson et al., 2017) are collected to test VL models' reasoning ability. We follow the common practice of previous studies on VLP and use the NLVR2 dataset (Suhr et al., 2019) for the visual reasoning task because it uses real photographs in contrast to the other datasets which use synthetic and unrealistic images. Per-example accuracy is used for evaluation.

**Visual Grounding** It is another fundamental VL task that aims to locate an object instance from an image with a natural language referring expression. As introduced by Yu et al. (2016), there exist three variants of the visual grounding task: RefCOCO, RefCOCO+, and RefCOCOg. We opt to the RefCOCO+ subtask, which adds the constraint that no location words are included in the referring expression, because it is of medium difficulty and used in several existing VLP studies (Li et al., 2021a; Zeng et al., 2021). Per-example accuracy is used for evaluation.

**Visual Question Answering** It requires the model to take as input an image and a free-form, open-ended, natural language question about the image and produces or selects a natural language answer as the output. Compared to the image-text retrieval tasks and the binary reasoning task, VQA requires a more detailed understanding of the image and complex reasoning (Antol et al., 2015). We select the VQA v2.0 dataset (Goyal et al., 2017) which balanced the popular VQA dataset (Antol et al., 2015) so that the language bias is reduced and the understanding of visual clues is more important. Following the VQA 2.0 dataset, we use the publicly released VQA evaluation script in the VLUE benchmark.

**Image Captioning** It is a long standing and challenging problem in vision and language research. It requires a model to take an image as input and generate a natural language sentence (i.e., caption) describing the image. We follow the practice of most previous work and use the COCO Caption dataset (Chen et al., 2015) for training and evaluation. We use the widely used BLUE (Papineni et al., 2002) and CIDEr (Vedantam et al., 2015) score for evaluation.

### 3.2. OOD Test Set

We argue that the most important problem in the current evaluation protocol of vision-language models is the "in-distribution bias" of the datasets used for downstream tasks. As shown in Table 1, the prevalent datasets for 3 out of 4 representative VL tasks are annotated with images collected from the COCO dataset. The only exception is NLVR2, which also relies on synsets from ImageNet to form queries

for the Google Image search engine. On the other hand, almost all VLP models are pre-trained using COCO image-text pairs and many of them initialize the vision encoder with ImageNet pre-training. As such, VLP models have already seen the images (or similar images) in the downstream datasets and their corresponding captions after pre-training. Therefore, fine-tuning on these VL datasets only measures the transferability of VLP models in a special "in-distribution" scenario where at least a part of data distribution (i.e., image distribution) remains the same when transferring to a different label distribution. This will likely result in an overestimation of the true generalization ability of VLP models.

To address this problem, we collect a suite of test sets annotated with images from "in-the-wild" distribution (Koh et al., 2021) for each of the four representative downstream VL tasks. It is very challenging to find a suitable image source that is sufficiently different from the COCO and ImageNet image distribution while suitable for serving as examples for the vision-language tasks.

Fortunately, Liu et al. (2021a) recently release the MaRVL dataset which provides images from a diverse distribution across cultures and geographical location. In MaRVL, the languages and language-specific concepts are carefully selected by Liu et al. (2021a) together with members of a community of native speakers to better mitigate the limitation of ImageNet/COCO concepts and better represent a true worldwide distribution. As such, the MaRVL images significantly differ from the ImageNet/COCO hierarchy which is created in English language culture. We present four randomly selected images from the COCO dataset and MaRVL dataset in Figure 1. We can see that images in the COCO dataset clearly differ from those in the MaRVL dataset in terms of both concepts and styles. Therefore, we believe that MaRVL images can serve as a good testbed for measuring the true generalization ability of VLP models.

While MaRVL provides a great source of images for measuring the true transferability of VLP models, it is only annotated on the visual reasoning task with questions written in native languages instead of English. We extend the annotations to support all VL tasks included in VLUE by crowdsourcing. We hire crowdsource workers that are of good English proficiency from a crowdsourcing platform. The workers are asked to accomplish four annotation tasks to provide test sets for the four downstream VL tasks. We describe them as follows:

**Caption Annotation** The workers are asked to write a sentence describing the scene following the instructions from the original COCO caption dataset. To be specific, the sentence is required to have at least 8 words, describe all and only important parts of the image, and does not start with "there is". The images and annotated captions are used for the image-text retrieval task.

**Statement Annotation** The natural language statements provided by MaRVL for the visual reasoning task are written in native languages. However, most VLP models are trained with English image-text pairs only. Therefore, we need to translate these statements into English. In the original work of (Liu et al., 2021a), the statements are translated to English by neural machine translation in the Google Cloud API. However, we find that the translation quality of google translation API is not satisfactory for the low-resource native languages in MaRVL (e.g., Indonesian, Swahili, Tamil). To ensure the performance gap between the original in-distribution test set and our OOD test set only comes from the distribution shift instead of translation quality issues, we ask annotators to translate the original statements into English. The workers are given the input image pair and a Chinese translation of the statement generated by Google Translation API, which may of low quality but mostly faithful. They are asked to write an English version of the statement that is both fluent faithful to the original statement.

**VQA Annotation** In VQA annotation, the workers are asked to write an open-ended natural language question and the corresponding answer given an image. Inspired by the practice of Goyal et al. (2017) for reducing the language bias of VQA datasets and making visual clues matters more, we show a pair of images[4] of the same concept to the workers and ask them to write a question for which the answers for the two images are different. To ensure the performance gap between the original test set and our annotated OOD test set comes from different image distribution instead of answer distribution, we ask workers to write questions following the question type distribution in the original VQA test set. Also, most VLP models consider the VQA task as a classification problem and only produce labels from a pre-defined answer list consisting of 3129 most frequent answers. Therefore, we ask the workers to verify that the answer for the questions are in the answer list so that all examples in the OOD test set is answerable by the model fine-tuned on the VQA 2.0 train set.

**Visual Grounding Annotation** In this task, two groups of crowdsource workers are hired. Workers from the first group are asked to decide if there exist multiple instances of an object and (if so) draw a bounding box of a randomly selected instance. Workers from the other group are asked to write a referring expression for the selected instance following the RefCOCO+ instructions.

For all annotation tasks, a training session is conducted before annotation so that the workers are carefully trained to ensure they understand the annotation protocol as well as the vision-language tasks the annotations will be used

[4]We re-use the image pairs in the MaRVL dataset

for. We provide the English translation of concept names in native languages for each example to help workers better understand the scene and write annotations. Workers are allowed to skip an instance if they find it too hard or not appropriate for the task. The number of instances in the OOD test set for each VL task is presented in Table 1.

### 3.3. Efficiency-Performance Trade-off

Another limitation of the current practice for VLP evaluation is that the efficiency-performance trade-off is often neglected. Similarly to the trend in the field of pre-training for NLP, VLP models are growing larger and larger in size, which makes it hard, if not impossible, for researchers from most institutes to reach the state-of-the-art in terms of absolute performance. Consequently, more works will instead pursue improvements on other dimensions such as the efficiency-performance trade-off. Therefore, we include the efficiency-performance trade-off as another factor to consider in the VLUE benchmark.

Measuring and comparing the efficiency-performance trade-off of various VLP models is not straightforward. In VLUE, we follow the practice of ELUE (Liu et al., 2021b) and measure a Pareto front of the efficiency-performance trade-off of existing VLP models. In this way, we can easily determine if a new model achieves a Pareto SOTA by seeing if it appears outside of the current Pareto front, i.e., there's no existing model that outperforms the new model on both the performance and efficiency dimensions.

The performance can be easily measured by averaging task-specific metrics across a pre-defined set of tasks. For efficiency, there are three common choices to measure the efficiency of a pre-trained model in the field of NLP: number of parameters (Lan et al., 2020; Sanh et al., 2019), FLOPs (Jiao et al., 2020), and actual inference time (Schwartz et al., 2020; Zhou et al., 2020). In VLUE, we opt to the actual inference time following the Long Range Arena benchmark (Tay et al., 2021) for efficient transformers because the latency is more critical for most application scenarios while models with the same number of parameters or FLOPs can lead to very different latency because of their difference in network architectures and parametrizations designs (e.g., going deeper or wider).

In sum, VLUE is different from previous popular benchmarks because it takes not only the absolute performance but also the generalization ability and the efficiency-performance trade-off into consideration. We believe this will enable the vision and language community to better tell whether we are making genuine progress or overfitting to entrenched datasets, and guide new research in the field to focus on building more generalizable and efficient VLP models instead of overly focusing on the absolute numbers.

## 4. Experiments

### 4.1. Training and Evaluation Setup

We conduct experiments to benchmark the generalization ability and efficiency-performance trade-off of representative VLP models. For each model, we fine-tune the released pre-trained checkpoint on the VLUE tasks with the hyperparameters provided in the paper. We only consider tasks for which the original paper reported results. This is done for two reasons. First, fine-tuning on different tasks involves various design choices for which we are not able to optimize. Second, some of the models are not suitable for specific kind of tasks (e.g., encoder-decoder models are not suitable for image-text retrieval). We successfully reproduced the performance reported in the original work for all compared models and all tasks.

After fine-tuning, we evaluate the performance of fine-tuned models on the corresponding OOD test sets in the zero-shot fashion. In addition to the absolute performance, we also record the actual inference time of different models in a controlled setting where the hardware environment is fixed for all models.

### 4.2. Baselines

We evaluate a number of representative VLP models that effectively learn cross-modal representations and achieve competitive results on many VL tasks. We briefly describe the evaluated models as follows:

**ViLBERT** (Lu et al., 2019; 2020) ViLBERT is among the first VLP models. It is a two-stream transformer-based model with image representations obtained from object detectors and models cross-modal interactions via co-Attentional transformer layers. It is pre-trained on the Conceptual Captions dataset (Sharma et al., 2018) with the masked multi-modal modeling objective and the multi-modal alignment prediction objective.

**LXMERT** (Tan & Bansal, 2019) LXMERT is also among the first VLP models. The model architecture is similar to that of ViLBERT with a few slightly different designs. LXMERT does not include the multi-modal alignment prediction objective but added the RoI feature regression objective. LXMERT is pre-trained on captions of COCO and VG datasets and also the train/dev set of VQA 2.0, GQA (Hudson & Manning, 2019), and VG-QA (Zhu et al., 2016).

**UNITER** (Chen et al., 2020) UNITER is a single stream transformer-based model similar to VisualBERT (Li et al., 2019) and Unicoder-VL (Li et al., 2020a), which concatenates image representations with text representations to form a single sequence and use transformer layers for representation learning. It is pre-trained with the combination of previous objectives with a modification of applying masking

| Tasks | Metrics | ViLBERT 274M/3.3M | LXMERT 240M/180K | UNITER 300M/4M | VL-T5 224M/180K | ALBEF 210M/14M | X-VLM 216M/4M | X-VLM 216M/16M | METER 352M/4M |
|---|---|---|---|---|---|---|---|---|---|
| TR | R@1 | - | - | 65.68 | - | 77.60 | 80.40 | 81.20 | 76.16 |
| | R@5 | - | - | 88.56 | - | 94.30 | 95.50 | 95.60 | 93.16 |
| | R@10 | - | - | 93.76 | - | 97.20 | 98.20 | 98.20 | 96.82 |
| | R@1-OOD | - | - | 36.32 | - | 64.11 | 64.11 | 67.39 | 62.11 |
| | R@5-OOD | - | - | 63.81 | - | 88.40 | 88.62 | 89.95 | 87.47 |
| | R@10-OOD | - | - | 75.13 | - | 93.96 | 94.46 | 95.23 | 92.23 |
| IR | R@1 | 42.51* | - | 52.93 | - | 60.70 | 63.10 | 63.40 | 57.08 |
| | R@5 | 71.18* | - | 79.93 | - | 84.30 | 85.70 | 85.80 | 82.66 |
| | R@10 | 80.76* | - | 87.95 | - | 90.50 | 91.60 | 91.50 | 90.07 |
| | R@1-OOD | 27.51 | - | 29.73 | - | 50.08 | 51.53 | 53.88 | 45.66 |
| | R@5-OOD | 53.07 | - | 56.00 | - | 77.72 | 79.00 | 81.25 | 75.12 |
| | R@10-OOD | 63.87 | - | 66.93 | - | 85.68 | 86.71 | 88.38 | 85.28 |
| NLVR$^2$ | Dev | 77.40* | 74.90 | 79.12 | 73.11* | 82.55 | 84.16 | 84.41 | 82.33 |
| | Test-P | 78.03* | 76.20 | 79.98 | 73.6 | 83.14 | 84.21 | 84.76 | 83.05 |
| | Test-OOD | 66.53 | 65.24 | 66.60 | 73.84 | 73.17 | 74.16 | 73.02 | 73.47 |
| VQA | Test-dev | 72.70 | 69.90 | 73.82 | 69.80* | 75.84 | 78.07 | 78.22 | 77.68 |
| | Test-std | - | 72.50 | 74.02 | 70.30 | 76.04 | 78.09 | 78.37 | 77.64 |
| | Test-OOD | 48.38 | 46.18 | 47.79 | 46.31 | 51.52 | 51.55 | 52.57 | 53.76 |
| Visual Grounding | Val$^d$ | 74.49* | - | 75.31 | 67.72* | 58.46$\star$ | 80.17 | 84.51 | - |
| | Test-A$^d$ | 79.18* | - | 81.30 | 75.23* | 65.89$\star$ | 86.36 | 89.00 | - |
| | Test-B$^d$ | 66.70* | - | 65.58 | 57.86* | 46.25$\star$ | 71.00 | 76.91 | - |
| | Test-OOD$^d$ | 54.91 | - | 36.86 | 27.89 | 24.30$\star$ | 55.00 | 59.32 | - |

Table 2: Results of representative VLP models on VLUE. X-OOD denotes the results on our private OOD test sets. IR denotes Image Retrieval and TR denotes Text Retrieval. For each compared model, we report the number of parameters and the number of images used for pre-training under the model name. Results with $*$ are our reproduced numbers which are not reported in the original paper. Results with $\star$ are in the weakly supervised setting.

to only one modality at one time, and a novel word-region alignment objective. UNITER is pre-trained on the combination of four caption datasets including COCO captions, VG dense captions, Conceptual Captions, and SBU captions (Ordonez et al., 2011).

**VL-T5** (Cho et al., 2021) Different from the above encoder-only models, VL-T5 is based on the sequence-to-sequence framework (Sutskever et al., 2014) and transforms the pre-training objectives and downstream tasks into a unified text generation framework following T5 (Raffel et al., 2019). VL-T5 is composed of a single stream transformer encoder with visual embeddings obtained from an object detector and an autoregressive transformer decoder. It is pre-trained on the same set of data as in LXMERT with a combination of multimodal LM, VQA, Image-text matching, Visual Grounding, and captioning as pre-training tasks.

**ALBEF** (Li et al., 2021a) ALBEF is a two stream transformer-based encoder model. In contrast to the previous models, ALBEF does not require bounding box annotations, thus alleviating the use of an object detector. Instead, it obtains visual embeddings with a visual transformer (Dosovitskiy et al., 2021). It is pre-trained with the multi-modal language modeling objective, the image-text matching objective, and a new image-text contrastive loss inspired by MoCo (He et al., 2020). ALBEF also proposed a momentum distillation method to improve pre-training on noisy image-text pairs. It is pre-trained on COCO, VG, SBU, Conceptual Captions, and an additional dataset, named CC-12M (Changpinyo et al., 2021).

**X-VLM** (Zeng et al., 2021) X-VLM proposes to learn multi-grained vision language alignments in pre-training. Unlike methods relying on object-centric features, e.g. UNITER and VL-T5, X-VLM relieves the need of object detection and is able to directly leverage the learned multi-grained alignments in downstream tasks. X-VLM is optimized by: 1) locating visual concepts in the image given associated texts by a combination of box regression loss and intersection over union loss; 2) in the meantime aligning the texts with the visual concepts by a contrastive loss, a matching loss, and a masked language modeling loss, where the alignments are in multi-granularity. The pre-training dataset of X-VLM is the same as UNITER.

**METER** (Dou et al., 2021) METER investigates on the transformer-based model designs extensively and proposes a well-designed VLP architecture, METER-CLIP-ViT. It applied CLIP-ViT (Radford et al., 2021) as the image encoder and RoBERTa (Liu et al., 2019) as the text encoder. On top of each encoder, there is a 6-layer transformer to model

the cross-modal interaction based on a cross-attention block. The pre-training objectives are masked language modeling and image-text matching only. The pre-training datasets are the same as UNITER.

### 4.3. Results

**Generalization Gap** We present the main result of VLUE in Table 2. We find that there exists a sizable generalization gap for all models between the original in-domain tests and our OOD test sets. The state-of-the-art VLP models that seemingly succeed several VL tasks including visual reasoning and VQA (with 80% accuracy) still struggle when generalizing to examples from a more diverse distribution. Specifically, the R@1 for the best performing model on the image-text retrieval task drops from 80.4 to 64.1 for text retrieval and from 63.1 to 51.5 for image retrieval. Similarly, for VL reasoning tasks, the best performance drops from 84.2 to 74.2 for NLVR and from 78.1 to 51.7 for VQA. This suggests that the performance of current VLP models is probably overestimated because of the aforementioned in-distribution bias.

**Generalization vs. Performance** In addition, we find that compared to reporting the results in a single dimension (i.e., absolute performance), including the OOD generalization performance into account sometimes leads to complementary and even controversial conclusions. For example, UNITER seems to significantly outperform ViLBERT on both visual reasoning and VQA tasks with an absolute improvement of 1.9 and 1.1 points on the original test sets. However, ViLBERT performs comparably with UNITER on the OOD test set of NLVR2 and even outperforms UNITER by 0.7 points on the OOD test set of VQA. Moreover, VL-T5 performs 9.5 points worse than ALBEF on the in-domain test set of NLVR2 while outperforming ALBEF by 0.7 points on the OOD test set. Similar trends can also be observed when comparing ALBEF, X-VLM, and METER on VQA, and comparing ViLBERT and UNITER on Visual Grounding. These observations demonstrate the necessity of evaluating on the OOD test sets when comparing different models.

**Efficiency-Performance Trade-off** We then present the Pareto front in terms of the efficiency-performance trade-off of VLP models in Figure 2. The performance of considered VLP models is measured by the mean accuracy on NLVR2 and VQA test sets. We measure the efficiency of VLP models by their average inference time on these test sets. Note that for models requiring object detection to extract features from the raw images, we include the object detection time in the total inference time. We fix the hardware environment to 1 Nvidia Tesla V100 GPU and the batch size to 1 to simulate real application scenarios[5]. We can see that models with similar performance may have completely different positions in the Pareto front of efficiency-performance trade-off. For example, VL-T5 performs only marginally worse than ALBEF, but requires an average inference time 2.4 times longer than ALBEF. Therefore, ALBEF should be considered to significantly outperform VL-T5 in the dimension of efficiency-performance trade-off. This further demonstrates the necessity of a multi-dimension benchmark like VLUE for more throughout comparisons.

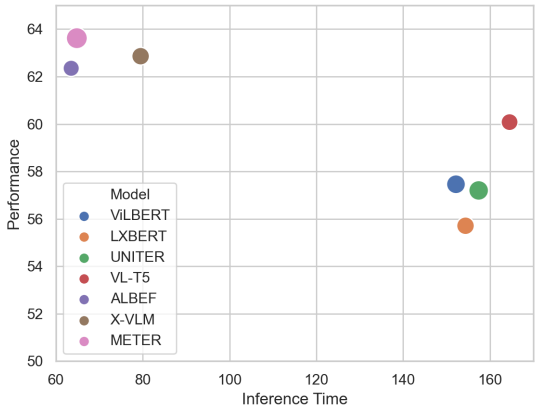



Figure 2: Pareto front of the efficiency-performance trade-off across seven models.

| Models | BLEU4 | CIDER | BLEU4-OOD | CIDER-OOD |
|---|---|---|---|---|
| **VL-T5** | 34.2 | 113.7 | 10.9 | 36.6 |
| **X-VLM** | 39.9 | 134.0 | 17.4 | 67.4 |
| **OFA** | 41.7 | 140.7 | 18.7 | 71.2 |

Table 3: Results on Image Captioning.

**Results on Image Captioning** Most aforementioned models are encoder-only models and thus not suitable for image captioning. Therefore, we also include OFA (Wang et al., 2022), the state-of-the-art pre-traiend VLM for image captioning. We present the results on image captioning in Table 3. We can see that the performance of all evaluated model drops significantly, which is in line with the previous results on vision-language understanding tasks. Moreover, we find that VL-T5, which is based on object detection for visual representation, suffers from a larger performance drop. We conjecture this is because object tags and embeddings from

[5]The actual inference time of different models depends on hardware. We provide the code for measuring actual inference time to facilitate researchers compare efficiency of different compared models in their own envirnment. We also provide the actual inference time of popular vision-language models in our setting in the Appendix of our paper for reference.

a pre-trained object detector plays an important role in its success on the COCO caption dataset, and the object detector may fail to generalize well on more diverse concepts in our OOD test set.

## 5. Analysis

We conduct a series of analyses investigating the relationship between different evaluation dimensions in VLUE and the influence of different design choices for VLP models on these dimensions.

**Performance-Generalization Correlation** We calculate the Pearson correlation coefficient $\rho$ of the models' performance on in-domain test sets and that on our OOD test sets. We obtain a relatively high correlation ($\rho = 0.75$). This suggests that models perform well on in-domain test sets also tend to succeed on the OOD test set. However, the correlation is not very high, which indicates that there may be some differences between the trend of in-domain performance and OOD performance. This also supports the need of a multi-dimensional benchmark like VLUE.

**Object Detection vs. Vision Transformer** We investigate the impact of vision feature sources on VLUE. We find that models with different vision feature sources such as object detection features or directly using vision transformers only slightly differ in terms of absolute performance. However, from the Pareto front of efficiency-performance trade-off, we can clearly find that models with vision transformers as image encoders are more practical in terms of efficiency-performance trade-off.

## 6. Conclusion and Discussion

We introduce VLUE, a multi-task multi-dimension benchmark for the evaluation of vision-language pre-trained models. VLUE is a first step towards evaluating a model not only by its absolute performance but also on other useful dimensions including the generalization ability of the model and its efficiency. We include four representative vision-language tasks/datasets in VLUE. For each task, we crowdsource an OOD test set annotated with images from a diverse distribution to measure the generalization gap. We benchmark the performance, generalization ability and efficiency-performance trade-off of 7 representative VLP models to facilitate future research. In sum, VLUE aims to promote vision-and-language research that does not solely focus on absolute performance but also takes other important factors like generalization and efficiency into account.

In addition, we would also like to highlight that while the original intended use of the collected OOD data in VLUE is to evaluate vision-language models via direct OOD generalization (i.e., fine-tune on original datasets then directly test on OOD test sets), it is also possible to fine-tune, or few-shot fine-tune on a subset of the provided OOD data for other research settings such as transfer learning and domain adaptation.

## References

Agirre, E., Màrquez, L., and Wicentowski, R. (eds.). *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, Prague, Czech Republic, 2007. Association for Computational Linguistics. URL https://aclanthology.org/S07-1000.

Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., and Zhang, L. Bottom-up and top-down attention for image captioning and visual question answering. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pp. 6077–6086. IEEE Computer Society, 2018. doi: 10.1109/CVPR.2018.00636. URL http://openaccess.thecvf.com/content_cvpr_2018/html/Anderson_Bottom-Up_and_Top-Down_CVPR_2018_paper.html.

Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L., and Parikh, D. VQA: visual question answering. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pp. 2425–2433. IEEE Computer Society, 2015. doi: 10.1109/ICCV.2015.279. URL https://doi.org/10.1109/ICCV.2015.279.

Boden, M. A. *Mind as machine: A history of cognitive science*. Oxford University Press, 2008.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. Language models are few-shot learners. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.

Cao, J., Gan, Z., Cheng, Y., Yu, L., Chen, Y., and Liu, J. Behind the scene: Revealing the secrets of pre-trained vision-and-language models. In *ECCV (6)*, volume 12351 of *Lec-*

*ture Notes in Computer Science*, pp. 565–580. Springer, 2020.

Changpinyo, S., Sharma, P., Ding, N., and Soricut, R. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3558–3568, 2021.

Chen, X., Fang, H., Lin, T., Vedantam, R., Gupta, S., Dollár, P., and Zitnick, C. L. Microsoft COCO captions: Data collection and evaluation server. *CoRR*, abs/1504.00325, 2015.

Chen, Y.-C., Li, L., Yu, L., Kholy, A. E., Ahmed, F., Gan, Z., Cheng, Y., and Liu, J. Uniter: Universal image-text representation learning. In *ECCV*, 2020.

Cho, J., Lei, J., Tan, H., and Bansal, M. Unifying vision-and-language tasks via text generation. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 1931–1942. PMLR, 2021. URL http://proceedings.mlr.press/v139/cho21a.html.

Dagan, I., Glickman, O., and Magnini, B. The PASCAL recognising textual entailment challenge. In *Machine learning challenges. evaluating predictive uncertainty, visual object classification, and recognising tectual entailment*, pp. 177–190. Springer, 2006.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. 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)*, pp. 4171–4186, Minneapolis, Minnesota, 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

Diao, S., Bai, J., Song, Y., Zhang, T., and Wang, Y. Zen: Pre-training chinese text encoder enhanced by n-gram representations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 4729–4740, 2020.

Diao, S., Xu, R., Su, H., Jiang, Y., Song, Y., and Zhang, T. Taming pre-trained language models with n-gram representations for low-resource domain adaptation. 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)*, pp. 3336–3349, 2021.

Diao, S., Zhou, W., Zhang, X., and Wang, J. Prefix language models are unified modal learners, 2022.

Dolan, W. B. and Brockett, C. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*, 2005. URL https://aclanthology.org/I05-5002.

Dong, L., Yang, N., Wang, W., Wei, F., Liu, X., Wang, Y., Gao, J., Zhou, M., and Hon, H. Unified language model pre-training for natural language understanding and generation. In Wallach, H. M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E. B., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 13042–13054, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/c20bb2d9a50d5ac1f713f8b34d9aac5a-Abstract.html.

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. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Dou, Z., Xu, Y., Gan, Z., Wang, J., Wang, S., Wang, L., Zhu, C., Zhang, P., Yuan, L., Peng, N., Liu, Z., and Zeng, M. An empirical study of training end-to-end vision-and-language transformers. *CoRR*, abs/2111.02387, 2021.

Fu, Z., Zhou, W., Xu, J., Zhou, H., and Li, L. Contextual representation learning beyond masked language modeling. In *ACL (1)*, pp. 2701–2714. Association for Computational Linguistics, 2022.

Gan, Z., Chen, Y., Li, L., Zhu, C., Cheng, Y., and Liu, J. Large-scale adversarial training for vision-and-language representation learning. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/49562478de4c54fafd4ec46fdb297de5-Abstract.html.

Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., and Parikh, D. 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 2017, Honolulu, HI, USA, July 21-26, 2017*, pp. 6325–6334. IEEE Computer Society, 2017. doi: 10.1109/CVPR.2017.670. URL https://doi.org/10.1109/CVPR.2017.670.

He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. B. Momentum contrast for unsupervised visual representation learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pp. 9726–9735. IEEE, 2020. doi: 10.1109/CVPR42600.2020.00975. URL https://doi.org/10.1109/CVPR42600.2020.00975.

Hu, J., Ruder, S., Siddhant, A., Neubig, G., Firat, O., and Johnson, M. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 4411–4421. PMLR, 2020. URL http://proceedings.mlr.press/v119/hu20b.html.

Huang, Z., Zeng, Z., Liu, B., Fu, D., and Fu, J. Pixel-bert: Aligning image pixels with text by deep multi-modal transformers. *ArXiv preprint*, abs/2004.00849, 2020. URL https://arxiv.org/abs/2004.00849.

Huang, Z., Zeng, Z., Huang, Y., Liu, B., Fu, D., and Fu, J. Seeing out of the box: End-to-end pre-training for vision-language representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12976–12985, 2021.

Hudson, D. A. and Manning, C. D. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 6700–6709. Computer Vision Foundation / IEEE, 2019. doi: 10.1109/CVPR.2019.00686. URL http://openaccess.thecvf.com/content_CVPR_2019/html/Hudson_GQA_A_New_Dataset_for_Real-World_Visual_Reasoning_and_Compositional_CVPR_2019_paper.html.

Jia, C., Yang, Y., Xia, Y., Chen, Y., Parekh, Z., Pham, H., Le, Q. V., Sung, Y., Li, Z., and Duerig, T. Scaling up visual and vision-language representation learning with noisy text supervision. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 4904–4916. PMLR, 2021. URL http://proceedings.mlr.press/v139/jia21b.html.

Jiao, X., Yin, Y., Shang, L., Jiang, X., Chen, X., Li, L., Wang, F., and Liu, Q. TinyBERT: Distilling BERT for natural language understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 4163–4174, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.372. URL https://aclanthology.org/2020.findings-emnlp.372.

Johnson, J., Hariharan, B., van der Maaten, L., Fei-Fei, L., Zitnick, C. L., and Girshick, R. B. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pp. 1988–1997. IEEE Computer Society, 2017. doi: 10.1109/CVPR.2017.215. URL https://doi.org/10.1109/CVPR.2017.215.

Kamath, A., Singh, M., LeCun, Y., Synnaeve, G., Misra, I., and Carion, N. Mdetr-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1780–1790, 2021.

Kim, W., Son, B., and Kim, I. Vilt: Vision-and-language transformer without convolution or region supervision. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5583–5594. PMLR, 2021. URL http://proceedings.mlr.press/v139/kim21k.html.

Koh, P. W., Sagawa, S., Marklund, H., Xie, S. M., Zhang, M., Balsubramani, A., Hu, W., Yasunaga, M., Phillips, R. L., Gao, I., Lee, T., David, E., Stavness, I., Guo, W., Earnshaw, B., Haque, I., Beery, S. M., Leskovec, J., Kundaje, A., Pierson, E., Levine, S., Finn, C., and Liang, P. WILDS: A benchmark of in-the-wild distribution shifts. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5637–5664. PMLR, 2021. URL http://proceedings.mlr.press/v139/koh21a.html.

Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L., Shamma, D. A., Bernstein, M. S., and Fei-Fei, L. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *Int. J. Comput. Vis.*, 123(1): 32–73, 2017.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., and Soricut, R. ALBERT: A lite BERT for self-supervised learning of language representations. In *8th International*

*Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=H1eA7AEtvS.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7871–7880, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.703. URL https://aclanthology.org/2020.acl-main.703.

Li, G., Duan, N., Fang, Y., Gong, M., and Jiang, D. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pp. 11336–11344. AAAI Press, 2020a. URL https://aaai.org/ojs/index.php/AAAI/article/view/6795.

Li, J., Selvaraju, R. R., Gotmare, A. D., Joty, S., Xiong, C., and Hoi, S. Align before fuse: Vision and language representation learning with momentum distillation. In Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, 2021a. URL https://openreview.net/forum?id=OJLaKwiXSbx.

Li, L., Lei, J., Gan, Z., Yu, L., Chen, Y.-C., Pillai, R., Cheng, Y., Zhou, L., Wang, X. E., Wang, W. Y., et al. Value: A multi-task benchmark for video-and-language understanding evaluation. In *35th Conference on Neural Information Processing Systems (NeurIPS 2021) Track on Datasets and Benchmarks*, 2021b.

Li, L. H., Yatskar, M., Yin, D., Hsieh, C., and Chang, K. Visualbert: A simple and performant baseline for vision and language. *CoRR*, abs/1908.03557, 2019.

Li, W., Gao, C., Niu, G., Xiao, X., Liu, H., Liu, J., Wu, H., and Wang, H. UNIMO: Towards unified-modal understanding and generation via cross-modal contrastive learning. 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)*, pp. 2592–2607, Online, 2021c. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.202. URL https://aclanthology.org/2021.acl-long.202.

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. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *ECCV (30)*, volume 12375 of *Lecture Notes in Computer Science*, pp. 121–137. Springer, 2020b.

Liang, Y., Duan, N., Gong, Y., Wu, N., Guo, F., Qi, W., Gong, M., Shou, L., Jiang, D., Cao, G., Fan, X., Zhang, R., Agrawal, R., Cui, E., Wei, S., Bharti, T., Qiao, Y., Chen, J.-H., Wu, W., Liu, S., Yang, F., Campos, D., Majumder, R., and Zhou, M. XGLUE: A new benchmark datasetfor cross-lingual pre-training, understanding and generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6008–6018, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.484. URL https://aclanthology.org/2020.emnlp-main.484.

Lin, B. Y., Zhou, W., Shen, M., Zhou, P., Bhagavatula, C., Choi, Y., and Ren, X. CommonGen: A constrained text generation challenge for generative commonsense reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 1823–1840, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.165. URL https://aclanthology.org/2020.findings-emnlp.165.

Lin, T., Maire, M., Belongie, S. J., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft COCO: common objects in context. In *ECCV (5)*, volume 8693 of *Lecture Notes in Computer Science*, pp. 740–755. Springer, 2014.

Liu, F., Bugliarello, E., Ponti, E. M., Reddy, S., Collier, N., and Elliott, D. Visually grounded reasoning across languages and cultures. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 10467–10485, Online and Punta Cana, Dominican Republic, 2021a. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.818. URL https://aclanthology.org/2021.emnlp-main.818.

Liu, X., Sun, T., He, J., Wu, L., Zhang, X., Jiang, H., Cao, Z., Huang, X., and Qiu, X. Towards efficient NLP: A standard evaluation and A strong baseline. *CoRR*, abs/2110.07038, 2021b.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. Roberta: A robustly optimized bert pretraining approach. *ArXiv preprint*, abs/1907.11692, 2019. URL https://arxiv.org/abs/1907.11692.

Lu, J., Batra, D., Parikh, D., and Lee, S. Vilbert: Pretraining task-agnostic visiolinguistic representations

for vision-and-language tasks. In Wallach, H. M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E. B., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 13–23, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/c74d97b01eae257e44aa9d5bade97baf-Abstract.html.

Lu, J., Goswami, V., Rohrbach, M., Parikh, D., and Lee, S. 12-in-1: Multi-task vision and language representation learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pp. 10434–10443. IEEE, 2020. doi: 10.1109/CVPR42600.2020.01045. URL https://doi.org/10.1109/CVPR42600.2020.01045.

Ordonez, V., Kulkarni, G., and Berg, T. L. Im2text: Describing images using 1 million captioned photographs. In Shawe-Taylor, J., Zemel, R. S., Bartlett, P. L., Pereira, F. C. N., and Weinberger, K. Q. (eds.), *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011. Proceedings of a meeting held 12-14 December 2011, Granada, Spain*, pp. 1143–1151, 2011. URL https://proceedings.neurips.cc/paper/2011/hash/5dd9db5e033da9c6fb5ba83c7a7ebea9-Abstract.html.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 2227–2237, New Orleans, Louisiana, 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1202. URL https://aclanthology.org/N18-1202.

Plummer, B. A., Wang, L., Cervantes, C. M., Caicedo, J. C., Hockenmaier, J., and Lazebnik, S. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pp. 2641–2649. IEEE Computer Society, 2015. doi: 10.1109/ICCV.2015.303. URL https://doi.org/10.1109/ICCV.2015.303.

Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I. Improving language understanding by generative pre-training. 2018.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning transferable visual models from natural language supervision. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021. URL http://proceedings.mlr.press/v139/radford21a.html.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. Exploring the limits of transfer learning with a unified text-to-text transformer. *ArXiv preprint*, abs/1910.10683, 2019. URL https://arxiv.org/abs/1910.10683.

Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2383–2392, Austin, Texas, 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1264. URL https://aclanthology.org/D16-1264.

Ren, S., He, K., Girshick, R. B., and Sun, J. Faster R-CNN: towards real-time object detection with region proposal networks. In Cortes, C., Lawrence, N. D., Lee, D. D., Sugiyama, M., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pp. 91–99, 2015. URL https://proceedings.neurips.cc/paper/2015/hash/14bfa6bb14875e45bba028a21ed38046-Abstract.html.

Sanh, V., Debut, L., Chaumond, J., and Wolf, T. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108, 2019.

Schwartz, R., Stanovsky, G., Swayamdipta, S., Dodge, J., and Smith, N. A. The right tool for the job: Matching model and instance complexities. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 6640–6651, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.593. URL https://aclanthology.org/2020.acl-main.593.

Sharma, P., Ding, N., Goodman, S., and Soricut, R. 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)*, pp. 2556–2565, Melbourne, Australia, 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1238. URL https://aclanthology.org/P18-1238.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A., and Potts, C. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 1631–1642, Seattle, Washington, USA, 2013. Association for Computational Linguistics. URL https://aclanthology.org/D13-1170.

Su, L., Duan, N., Cui, E., Ji, L., Wu, C., Luo, H., Liu, Y., Zhong, M., Bharti, T., and Sacheti, A. GEM: A general evaluation benchmark for multimodal tasks. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 2594–2603, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.229. URL https://aclanthology.org/2021.findings-acl.229.

Su, W., Zhu, X., Cao, Y., Li, B., Lu, L., Wei, F., and Dai, J. VL-BERT: pre-training of generic visual-linguistic representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=SygXPaEYvH.

Suhr, A., Lewis, M., Yeh, J., and Artzi, Y. 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)*, pp. 217–223, Vancouver, Canada, 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-2034. URL https://aclanthology.org/P17-2034.

Suhr, A., Zhou, S., Zhang, A., Zhang, I., Bai, H., and Artzi, Y. A corpus for reasoning about natural language grounded in photographs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 6418–6428, Florence, Italy, 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1644. URL https://aclanthology.org/P19-1644.

Sutskever, I., Vinyals, O., and Le, Q. V. Sequence to sequence learning with neural networks. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D., and Weinberger, K. Q. (eds.), *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pp. 3104–3112, 2014. URL https://proceedings.neurips.cc/paper/2014/hash/a14ac55a4f27472c5d894ec1c3c743d2-Abstract.html.

Tan, H. and Bansal, M. 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)*, pp. 5100–5111, Hong Kong, China, 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1514. URL https://aclanthology.org/D19-1514.

Tay, Y., Dehghani, M., Abnar, S., Shen, Y., Bahri, D., Pham, P., Rao, J., Yang, L., Ruder, S., and Metzler, D. Long range arena : A benchmark for efficient transformers. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=qVyeW-grC2k.

Vedantam, R., Zitnick, C. L., and Parikh, D. Cider: Consensus-based image description evaluation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pp. 4566–4575. IEEE Computer Society, 2015. doi: 10.1109/CVPR.2015.7299087. URL https://doi.org/10.1109/CVPR.2015.7299087.

Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. R. Superglue: A stickier benchmark for general-purpose language understanding systems. In Wallach, H. M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E. B., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 3261–3275, 2019a. URL https://proceedings.neurips.cc/paper/2019/hash/4496bf24afe7fab6f046bf4923da8de6-Abstract.html.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. R. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019b. URL https://openreview.net/forum?id=rJ4km2R5t7.

Wang, P., Yang, A., Men, R., Lin, J., Bai, S., Li, Z., Ma, J., Zhou, C., Zhou, J., and Yang, H. Unifying

architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. *arXiv preprint arXiv:2202.03052*, 2022.

Wang, W., Bao, H., Dong, L., and Wei, F. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. *CoRR*, abs/2111.02358, 2021a.

Wang, Z., Yu, J., Yu, A. W., Dai, Z., Tsvetkov, Y., and Cao, Y. Simvlm: Simple visual language model pretraining with weak supervision. *CoRR*, abs/2108.10904, 2021b.

Warstadt, A., Singh, A., and Bowman, S. R. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641, 2019. doi: 10.1162/tacl_a_00290. URL https://aclanthology.org/Q19-1040.

Williams, A., Nangia, N., and Bowman, S. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 1112–1122, New Orleans, Louisiana, 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1101. URL https://aclanthology.org/N18-1101.

Xu, C. and McAuley, J. J. A survey on dynamic neural networks for natural language processing. *CoRR*, abs/2202.07101, 2022a.

Xu, C. and McAuley, J. J. A survey on model compression for natural language processing. *CoRR*, abs/2202.07105, 2022b.

Xu, C., Pei, J., Wu, H., Liu, Y., and Li, C. MATINF: A jointly labeled large-scale dataset for classification, question answering and summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 3586–3596, Online, 2020a. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.330. URL https://aclanthology.org/2020.acl-main.330.

Xu, C., Zhou, W., Ge, T., Wei, F., and Zhou, M. BERT-of-theseus: Compressing BERT by progressive module replacing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7859–7869, Online, 2020b. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.633. URL https://aclanthology.org/2020.emnlp-main.633.

Xu, C., Zhou, W., Ge, T., Xu, K., McAuley, J., and Wei, F. Beyond preserved accuracy: Evaluating loyalty and robustness of BERT compression. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 10653–10659, Online and Punta Cana, Dominican Republic, 2021a. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.832. URL https://aclanthology.org/2021.emnlp-main.832.

Xu, C., Zhou, W., Ge, T., Xu, K., McAuley, J., and Wei, F. Beyond preserved accuracy: Evaluating loyalty and robustness of BERT compression. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 10653–10659, Online and Punta Cana, Dominican Republic, 2021b. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.832. URL https://aclanthology.org/2021.emnlp-main.832.

Xu, C., Zhou, W., Ge, T., Xu, K., McAuley, J., and Wei, F. Blow the dog whistle: A Chinese dataset for cant understanding with common sense and world knowledge. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2139–2145, Online, 2021c. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.172. URL https://aclanthology.org/2021.naacl-main.172.

Xu, H., Yan, M., Li, C., Bi, B., Huang, S., Xiao, W., and Huang, F. E2E-VLP: End-to-end vision-language pre-training enhanced by visual learning. 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)*, pp. 503–513, Online, 2021d. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.42. URL https://aclanthology.org/2021.acl-long.42.

Xu, J., Zhou, W., Fu, Z., Zhou, H., and Li, L. A survey on green deep learning. *CoRR*, abs/2111.05193, 2021e.

Yang, Z., Gan, Z., Wang, J., Hu, X., Ahmed, F., Liu, Z., Lu, Y., and Wang, L. Crossing the format boundary of text and boxes: Towards unified vision-language modeling. *ArXiv preprint*, abs/2111.12085, 2021. URL https://arxiv.org/abs/2111.12085.

Yu, L., Poirson, P., Yang, S., Berg, A. C., and Berg, T. L. Modeling context in referring expressions. In *ECCV (2)*, volume 9906 of *Lecture Notes in Computer Science*, pp. 69–85. Springer, 2016.

Zeng, Y., Zhang, X., and Li, H. Multi-grained vision language pre-training: Aligning texts with visual concepts. *CoRR*, abs/2111.08276, 2021.

Zhang, P., Li, X., Hu, X., Yang, J., Zhang, L., Wang, L., Choi, Y., and Gao, J. Vinvl: Revisiting visual representations in vision-language models. In *CVPR*, pp. 5579–5588. Computer Vision Foundation / IEEE, 2021.

Zhou, W., Xu, C., Ge, T., McAuley, J. J., Xu, K., and Wei, F. BERT loses patience: Fast and robust inference with early exit. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/d4dd111a4fd973394238aca5c05bebe3-Abstract.html.

Zhou, W., Ge, T., Xu, C., Xu, K., and Wei, F. Improving sequence-to-sequence pre-training via sequence span rewriting. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 571–582, Online and Punta Cana, Dominican Republic, 2021a. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.45. URL https://aclanthology.org/2021.emnlp-main.45.

Zhou, W., Lee, D., Selvam, R. K., Lee, S., and Ren, X. Pre-training text-to-text transformers for concept-centric common sense. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021b. URL https://openreview.net/forum?id=3k20LAiHYL2.

Zhou, W., Xu, C., and McAuley, J. J. Meta learning for knowledge distillation. *CoRR*, abs/2106.04570, 2021c.

Zhou, X., Zhang, W., Chen, Z., Diao, S., and Zhang, T. Efficient neural network training via forward and backward propagation sparsification. *Advances in Neural Information Processing Systems*, 34:15216–15229, 2021d.

Zhou, X., Zhang, W., Xu, H., and Zhang, T. Effective sparsification of neural networks with global sparsity constraint. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3599–3608, 2021e.

Zhu, Y., Groth, O., Bernstein, M. S., and Fei-Fei, L. Visual7w: Grounded question answering in images. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pp. 4995–5004. IEEE Computer Society, 2016. doi: 10.1109/CVPR.2016.540. URL https://doi.org/10.1109/CVPR.2016.540.

## A. Actual Inference Time

We present the actual inference time of compared models in Table 3.

| MODEL | VILBERT | LXBERT | UNITER | VL-T5 | ALBEF | X-VLM | METER |
|---|---|---|---|---|---|---|---|
| PERFORMANCE | 57.46 | 55.71 | 57.20 | 60.08 | 62.35 | 62.86 | 63.62 |
| INFER. TIME (O.D. TIME) | 152.16 (121.89) | 154.36 (121.89) | 157.39 (121.89) | 164.51 (121.89) | 63.50 | 79.50 | 64.80 |

Table 4: The performance and inference time across seven models. INFER.TIME and O.D. TIME refer to the inference and object detection time cost in terms of millisecond, respectively.