MSc Artificial Intelligence
Master Thesis

Exploring Basic Language Abilities of Multimodal Pretrained Transformers

by
Xinyi Chen
13218042

August 4, 2022

48 EC
1 NOV, 2021 - 1 JULY, 2022

Supervisors:
Dr Sandro Pezzelle
Dr Raquel Fernández

Examiner:
Dr Sandro Pezzelle

Second reader:
Dr Iacer Calixto

Universiteit van Amsterdam

Contents

Abstract

Vision and language(V&L) transformers have achieved impressive results on a variety of tasks that rely on the pretrain-and-finetune method. The notable performance seems to suggest that they are very good at understanding visually-grounded language. However, little is known about whether and to what extent this is the case. Our work aims to fill in this gap by proposing a novel benchmark task called **BLA** (Benchmarks for **B**asic **L**anguage **A**bilities), which minimizes the requirement for the reasoning abilities of the models and focuses on linguistic knowledge. The benchmark contains tasks for three language phenomena - active-passive voices, coordination and relative clause. The test contains one image, two captions correctly describing the image and two distractors highly similar to the captions but do not match the image contents. We evaluate the state-of-the-art V&L pretrained models in a zero-shot setting to match the correct captions to the image, and find that, surprisingly, none of them do much better than chance. An extensive analysis is performed to shed light on the shortcomings of current models and provide insights into how future work can drive further progress in the field.

Acknowledgements

First and foremost, I would like to thank my supervisors Sandro Pezzelle and Raquel Fernández for all the guidance and encouragement during the thesis period. This thesis is the result of their constant support and freedom to explore various challenging problems. Working on this thesis is a challenging but fruitful journey, from which I learned research skills and critical thinking that would better prepare me for future research career. I also would like to thank them for providing very detailed and insightful feedback for the thesis writing.

Many thanks to Ece Takmaz, who provides many inspirations for the task designs and experiments as well as feedback on the thesis. I enjoyed every conversation I had with her, which ranged from research ideas to personal growth. I am also very grateful for the opportunity to work in the Dialogue Modeling Group, where everyone is very friendly and open to sharing their knowledge. I had a wonderful time spent with the group members and joining the group activities. Finally, I gratefully recognize the help of Dylan O'Sullivan and Jie Li, who spent a lot of time for reviewing and proofreading this thesis.

Chapter 1

Introduction

In recent years, transformer-based pretrained Language Models (LMs), such as BERT (Devlin et al., 2019) and GPT-3 (Brown et al., 2020), have seen widespread use in Natural Language Processing (NLP). By training on large volumes of text data, LMs learn universal language representations and transfer this knowledge to downstream tasks, topping leaderboards on various NLP benchmarks. Following the success of language models, the community has proposed various transformer vision-and-language (V&L) models, such as LXMERT (Tan and Bansal, 2019), ViLBERT (Lu et al., 2019), UNITER (Chen et al., 2020), which combine representations from image and text modalities and reaches the state-of-the-art (SOTA) in various tasks (Lu et al., 2019; Tan and Bansal, 2019; Li et al., 2019; Chen et al., 2020; Li et al., 2020a; Su et al., 2019). While the model performance is impressive, it is important to understand why the models perform so well, what information they learn from pretraining and what their limitations are.

A few benchmarks and probing tasks have been proposed to better understand V&L transformer models. Hendricks and Nematzadeh (2021) suggests that models are insensitive to linguistic distinctions of verb-argument structure. Parcalabescu et al. (2022) finds that pretrained transformer models can identify named objects and their presence in images well but struggle to ground their interdependence and relationships in visual scenes when forced to respect linguistic indicators. Thrush et al. (2022) probes V&L models on visio-linguistic compositional reasoning, and finds models struggle in all the tasks. In most of the previous work, language and reasoning abilities are not distinguished. For example, the ability to count, which is connected to reasoning, is involved in performing the task designed to evaluate the model's linguistic capability for understanding negation (Dobreva and Keller, 2021). The same V&L model is required to perform both language and reasoning tasks at the same time. However,However, neuroscientific evidence suggests that these two capabilities may be separate processes for humans, as humans perform reasoning and language tasks in different brain regions (Fedorenko and Varley, 2016). Therefore, it would be fairer to evaluate a model's reasoning abilities and language abilities separately.

In this light, we propose a novel benchmark for pretrained V&L models, which aims to evaluate **B**asic **L**anguage **A**bilities (henceforth, we refer to it as BLA) that do not involve the reasoning process. The benchmark explores whether models have a general understanding of sentence-level semantics and process more advanced linguistic knowledge on specific language phenomena. The language phenomena that we are interested in are active-passive voices, coordination and relative clause. These phenomena were selected from the Coloring Book tests (Pinto and Zuckerman, 2019), which evaluate children's linguistic knowledge with a wide range of grammatical topics. The active-passive task can test whether one can comprehend two sentences that have the same meaning but are phrased differently. The coordination and relative clause tasks assess one's ability in understanding how different parts of a sentence

are composed together into a whole. The ability to complete these tasks demonstrates good language comprehension. Pinto and Zuckerman (2019) shows that children as young as six years old are capable of solving these tasks. While understanding these phenomena is trivial for humans, we would like to investigate whether the powerful pretrained V&L models can perform well on these tasks.

The BLA benchmark is divided into three tasks, each task relates to one specific language phenomenon and uses the same structure for each test set, where one image is paired with four sentences. The four sentences are highly similar but two are captions that correctly describe the image while the others are distractors that either swap the participants of an action (in Active-Passive task) or mismatch person entities and attributes (in Coordination and Relative Clause tasks) to make them incorrect. In order to force the models to utilize linguistic knowledge, we carefully construct all the sentences to only contain objects existing in the image. Therefore, models cannot succeed in the tasks by simply performing object recognition but have to leverage more fine-grained language abilities to understand "who does what to whom" or "who has what attributes" in the two modalities.

With this newly proposed benchmark, we evaluate recent state-of-the-art pretrained V&L transformers CLIP, LXMERT and ViLBERT in a zero-shot setting. This means that we do not train or finetune the models on our evaluation tasks but directly apply them to our tasks by leveraging the existing prediction heads. Surprisingly, the models perform close to or even worse than the chance level. We conduct extensive analysis to better understand the capabilities and limitations of the V&L models. The main contributions of this thesis work are:

- We propose a framework to tease apart, as much as possible, linguistic vs. reasoning abilities of models and design the evaluation tasks more similar to the tests for humans.
- We introduce a novel benchmark BLA, aiming to explore the basic language abilities of pretrained V&L models. The benchmark evaluates the models on their understanding of active-passive voices, coordination and sentences embedding a relative clause, three linguistic phenomena that are commonly used in English.
- We perform experiments on current SOTA V&L models, which suggests that the model struggle with all the benchmark tasks. We conduct further analysis to better understand why the models might fail in these tasks. Based on that, we provide some insights on how to improve the model's visually-grounded language abilities for future work.

1.1 Overview of Chapters

The remainder of this thesis is organized into eight chapters. Chapter 2 introduces the background knowledge on pretrained vision-and-language models and common downstream tasks used to evaluate them. Chapter 3 covers an overview of recent research works aiming at investigating what the state-of-the-art V&L models learn from pretraining. In particular, we focus on related work that uses minimally-edited counterfactual examples to understand specific abilities of V&L models in a zero-shot setting. In Chapter 4, we introduce the tasks of the BLA benchmark and the methods taken to construct the benchmark as well as some statistics on the constructed dataset. Chapter 5 introduces the experimental setup for evaluating model performance and human performance on the BLA benchmark. Chapter 6, we demonstrate the results of model performance and human performance on the tasks mentioned in the previous chapter. In Chapter 7, we conducted quantitative and qualitative analyses to better understand the performance of the models. Chapter 8 summarizes the findings of this thesis and discusses potential opportunities for future work.

Chapter 2

Background

Vision-and-language models jointly process information from the two modalities and encode them into the same semantic space. Following the successful applications of pretrained models in computer vision (Simonyan and Zisserman, 2014; He et al., 2016) and natural language processing (Devlin et al., 2019) tasks, recent research adopts the pretrain-and-finetune paradigm for multimodal models, which pretrain the models on a large amount of data with self-supervision tasks and then finetune on the downstream V&L tasks. Three of the most common V&L tasks are Image Captioning (generating a caption for a given image), Visual Question Answering (VQA) (for a given image-question pair, answering the question based on the image), and Image Retrieval (for a visual or textual query, finding the semantically closet target from another modality). These tasks are difficult because they require the model to understand visual concepts, language semantics and most importantly, to relate fine-grained elements of the two modalities.

Major V&L model architectures can be divided into two categories: single- and dual-stream models. Single-stream models concatenate image and text features and jointly encode the resulting sequence with one single encoder, while dual-stream models encode the two modalities with two separate encoders and optionally use additional layers to fuse them into cross-modality features. Models are usually pretrained on tasks including but not limited to masked language modeling, masked region modeling and image-text matching to learn general multimodal representations.

In the following sections, we introduce the three state-of-the-art pretrained models that are evaluated on our BLA datasets. The reasons for selecting these models are described in the Experimental Setup (Section 5.2).

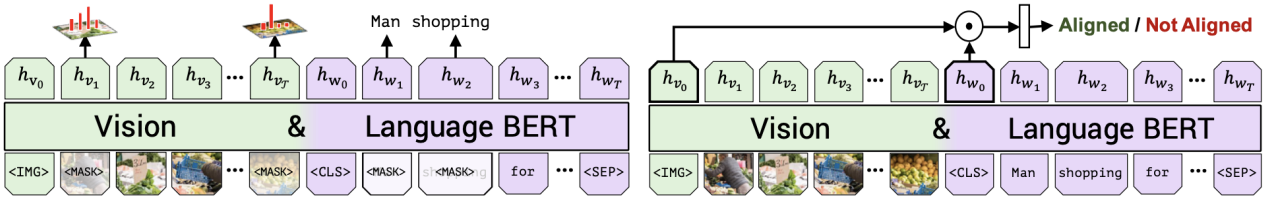
2.1 ViLBERT

ViLBERT (Vision-and-Language BERT) (Lu et al., 2019) extends the BERT architecture to a multi-modal dual stream model, which processes visual and textual inputs with two parallel BERT-style models operating over image regions and text segments. The information exchange between the two modalities is enabled by co-attentional transformer layers.

ViLBERT is pretrained on Conceptual Captions (Sharma et al., 2018), a dataset consisting of 3.3 million images with weakly-associated descriptive captions automatically collected from alt-text enabled images on the web. Similar to the text representation of BERT, the image is represented as visual "tokens", which are visual features of bounding boxes extracted by Faster R-CNN (Ren et al., 2015), an object detection network pretrained on Visual Genome (Krishna et al., 2017). The position encodings of the 'visual' tokens are represented by the spatial location of the bounding boxes. In analogy to the BERT, pretraining tasks of ViLBERT are masked multi-modal modeling and multi-modal alignment prediction. The masked multi-modal

modeling task (Figure 2.1 (a)) asks the mode to reconstruct the masked image region categories or words given the remaining inputs. In the multi-modal alignment task, the model has been presented with an image-text pair and is asked to predict whether the image and text are aligned. Through pretraining, the model learned a semantically meaningful alignment between vision and language.

After fine-tuning on downstream tasks, ViLBERT outperforms tasks-specific state-of-the-art models on four vision-and-language tasks - visual question answering (Antol et al., 2015), visual commonsense reasoning (Zellers et al., 2019), grounding referring expressions (Kazemzadeh et al., 2014), and caption-based image retrieval (Young et al., 2014).



(a) Masked multi-modal learning (b) Multi-modal alignment prediction

Figure 2.1: Pre-training procedure of ViLBERT by Lu et al. (2019). The model is trained under two training tasks to learn visual grounding. In masked multi-modal learning, the model must reconstruct image region categories or words for masked inputs given the observed inputs. In multi-modal alignment prediction, the model must predict whether or not the caption describes the image content.

2.2 LXMERT

LXMERT (Learning Cross-Modality Encoder Representations from Transformers) (Tan and Bansal, 2019) is another dual-stream pretrained transformer designed to learn vision-and-language connections. It consists of three encoders (an object-relationship encoder, a language encoder, and a cross-modality encoder) and uses five pretraining tasks to learn cross-modality representations. The model is pretrained on image-text pairs from five captioning or image question answering datasets: MS COCO (Lin et al., 2014), Visual Genome (Krishna et al., 2017), VQA v2.0 (Antol et al., 2015), GQA balanced version (Hudson and Manning, 2019), and VG-QA (Zhu et al., 2016).

The pretraining procedure is shown in Figure 2.2. The inputs to the model are an image and its related sentence (a caption or a question). Each sentence is split into words by the WordPiece tokenizer (Wu et al., 2016) in BERT (Kenton and Toutanova, 2019). Each word w_i and its index i (w_i's absolute position in the sentence) are projected to word embedding and index embedding, which add to the index-aware word embeddings and input to the language encoder. The image is represented by the features of detected objects. The objects are detected by Faster R-CNN and each object is represented by its position features (bounding box coordinates) and its 2048-dimensional region-of-interest (RoI) feature. The position-aware object-level embedding is fed into the object-relationship encoder.

The top branch of Figure 2.2 is the vision task, masked object prediction, where objects are masked randomly and the model is asked to predict the properties of these masked objects. The task is divided into two sub-tasks **RoI-Feature Regression** (regresses the object RoI feature with L2 loss) and **Detected Label Classification** (learns the labels of masked objects with cross-entropy loss). The model can learn to infer the masked object from the vision side, which helps learn the object relationships, as well as from the language side, which helps with the cross-modality alignment. The bottom branch of Figure 2.2 is the language task,

Masked Cross-modality Language Modeling, where words are randomly masked and the model is asked to predict these masked words based on the non-masked words and the vision modality. The middle-rightmost part of Figure 2.2 contains two cross-modality tasks. The **Cross-Modality Matching** task is similar to 'Next Sentence Prediction' in BERT (Kenton and Toutanova, 2019), where a sentence is replaced with a mismatched sentence with some probability and the model learns to predict whether an image and a sentence match each other. When the sentence is the question about the image, the **Image Question Answering** task learns to predict the answer to these image-related questions.

By learning from these V&L pretraining tasks, LXMERT achieves state-of-the-art performance on the evaluation sets of VQA (Antol et al., 2015) and GQA (Hudson and Manning, 2019). It also outperforms previous models on Natural Language for Visual Reasoning for Real (NLVR2) (Suhr et al., 2019) after finetuning.

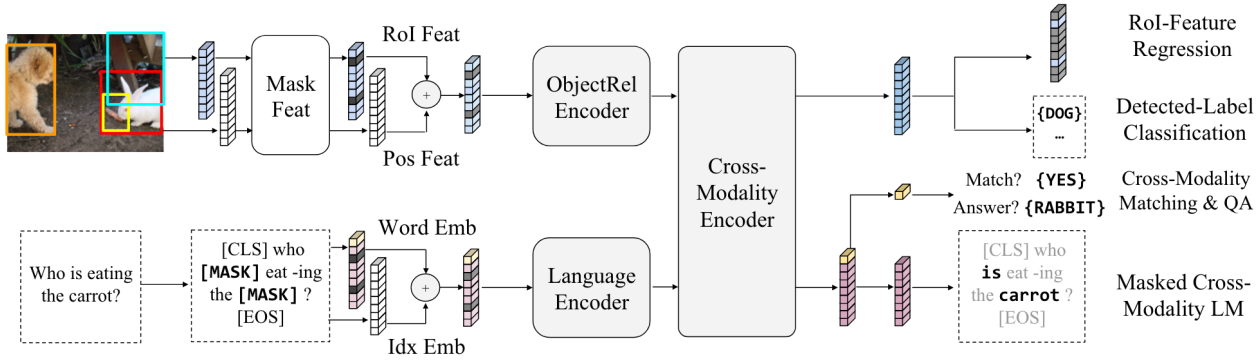


Figure 2.2: Pre-training procedure of LXMERT by (Tan and Bansal, 2019). The object RoI features and word tokens are masked. Five pre-training tasks learn the feature representations based on these masked inputs. Special tokens are in brackets and classification labels are in braces.

2.3 CLIP

CLIP (Contrastive Language-Image Pre-Training) (Radford et al., 2021) is a dual stream V&L model that is trained through contrastive learning to maximize scores for aligned image-text pairs. The text encoder uses a modified transformer architecture proposed by Radford et al.. Different from the previous two models, CLIP uses a Byte-Pair Encoding tokenizer (Sennrich et al., 2016) trained from scratch. The image encoder has two optional architectures, a Vision Transformer (ViT) (Dosovitskiy et al., 2020) or a ResNet-50 (He et al., 2016). We chose the Vision Transformer as an image encoder in our experiments.

Figure 2.3 (1) shows the pretraining process. The pretraining dataset contains around 400 million image-text pairs sourced from the Internet, a strategy similar to the web-scale training approach used for unimodal models such as GPT-3 (Brown et al., 2020). The text encoder and image encoder are trained from scratch without any initialization. N image-text pairs are fed to the model as a training batch. The N texts and images are converted to text embeddings and image embeddings by the text encoder and the image encoder respectively. CLIP is trained to maximize the cosine similarity of the image and text embeddings of the real pairs while minimizing the cosine similarity of the embeddings of the incorrect pairings. This enables the model to learn which of the $N \times N$ possible (image, text) pairings are real. Training on text in addition to images could allow for zero-shot classification via providing downstream labels as text. Compared to fully supervised baselines, the model achieves competitive performances on various computer vision tasks like image classification using a zero-shot or few-shot setting.

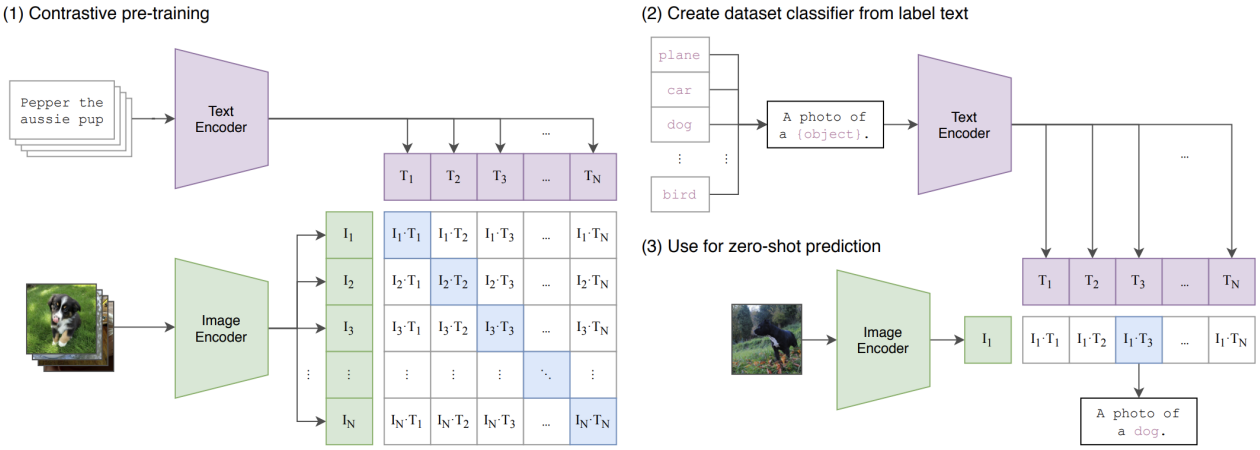


Figure 2.3: Summary of CLIP approach by Radford et al. (2021). While standard image models jointly train an image feature extractor and a linear classifier to predict some label, CLIP jointly trains an image encoder and a text encoder to predict the correct pairings of a batch of (image, text) training examples. At test time the learned text encoder synthesizes a zero-shot linear classifier by embedding the names or descriptions of the target dataset's classes.

Chapter 3

Related Work

The previous chapter introduces various pretrained V&L transformers and their impressive performance on a wide range of V&L tasks using the pretrain-and-finetuned paradigm. However, the performance on the downstream tasks is less informative in telling what models learn from pretraining because performance after finetuning depends on the size of finetuning data and other experimental setups (Yogatama et al., 2019). The research field of V&L has just started to investigate what models learn from pretraining, what makes them perform so well in the downstream tasks and what their limitations are. One approach is probing V&L models by examining the learned attention weights. Li et al. (2020b) analyzed attention heads in the model and showed that words in the text are correctly mapped to image regions that correspond to them. Cao et al. (2020) probed pre-trained vision-and-language models and reported similar observations. But these methods still use tasks that focus on general vision-and-language alignment abilities (i.e. Visual Relation Detection) to probe the models, lacking a more fine-grained understanding of what abilities the models obtained from pretraining.

Another trend is using counterfactual and minimally-edited examples (foils) for few- or zero-shot evaluation to understand V&L models on specific linguistic and/or reasoning abilities. Our research goes in this direction. In the following sections, we introduce datasets and tasks similar to our benchmark, which utilize foils to the V&L model's grounding capabilities on specific aspects.

3.1 FOIL

FOIL (Find One mismatch between Image and Language caption) (Shekhar et al., 2017) is proposed to understand the interaction of the visual and textual modalities. The dataset contains 297,268 datapoints and 97,847 images. Each datapoint is formed by one real-world image from MSCOCO, one original caption that matches the image and one "foil" caption that is highly similar to the original ones but contains one single mistake (e.g. changing the original word "bird" to "dog"). The foil examples are constructed automatically by replacing one noun in the original caption with an incorrect but similar word (nouns belong to the same supercategory of MSCOCO labels).

The dataset contains three tasks: caption classification between correct and foil, foil word detection and foil word correction (See Figure 3.1). Since V&L models are usually pretrained on the image-text alignment task, it is possible to evaluate models in a zero-shot setting to measure whether an image-sentence pair match. Hessel et al. (2021) proposed a reference-free metric to measure image-text similarities and use this metric to perform the caption classification task by comparing the similarities of the image and the correct captions vs. the foil. The model makes correct predictions if the correct caption receives a higher similarity score than the foil caption. Our experiment uses the same setting and compares the image-text similarities of the

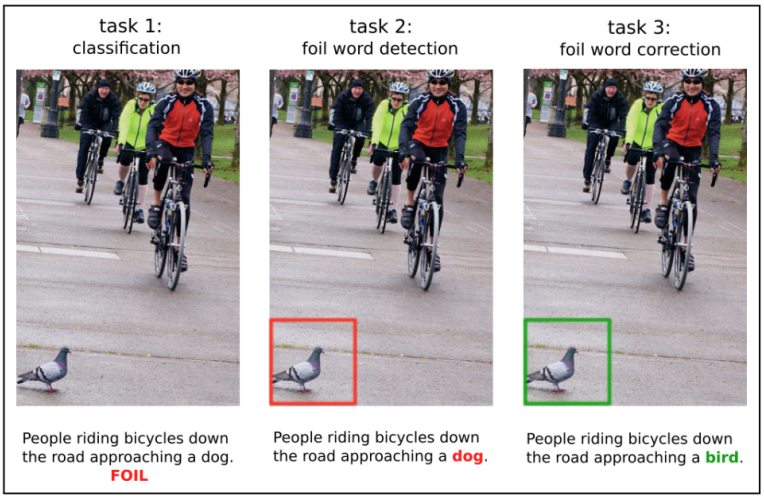


Figure 3.1: Tasks and examples of FOIL dataset by Shekhar et al. (2017). The task 1 requires to predict whether the caption is correct or foil. The task 2 requires the model to determine where the error is if it is foil. The task 3 requires to predict which word will be used to correct the foil one.

correct captions and distractors. Applying CLIP with the CLIP-Score metric achieves an 87.2 accuracy on the FOIL caption classification task.

3.2 SVO-Probes

Hendricks and Nematzadeh (2021) proposed a benchmark to examine pretrained V&L models on understanding **s**ubject, **v**erb, **o**bject triplets with minimally-edited counterfactual examples. For each sentence, two images are given: one positive image that matches the sentence and one controlled negative image that does not correspond to a specific aspect of the sentence regarding the subject, verb or object (Figure 3.2). The dataset contains 421 verbs and over 48,000 image-sentence pairs.

The benchmark is used to evaluate a few architectural variations of image–language transformer models pretrained on Conceptual Captions or MSCOCO. Similar to the FOIL setup, the model is asked to identify whether a sentence and an image match each other in a zero-shot classification task. Their results suggest that the pretrained V&L models are insensitive to verb-argument structures.

3.3 Winoground

After finishing the construction of the BLA benchmark, the preprint of Winoground (Thrush et al., 2022) came out, which is proposed to evaluate the visio-linguistic compositional reasoning abilities of pretrained transformer models. Each datapoint contains two images and two captions, where both captions contain exactly the same set of words but are ordered in a way that each describes primarily one of the images. Our Active-Passive task also constructs the distractor captions (regarding sentences using the same active or passive voice) in such a way.

The task requires models to correctly match the image-sentence pairs. The dataset uses 70 linguistic tags generated by human annotators, which is used for the swaps that differentiate caption pairs. This collection of tags is divided into three broad categories: objects, relations, and swaps involving both relations and objects. Object swaps reorder elements such as noun phrases that typically refer to real-world objects. Relation swaps reorder elements like verbs,

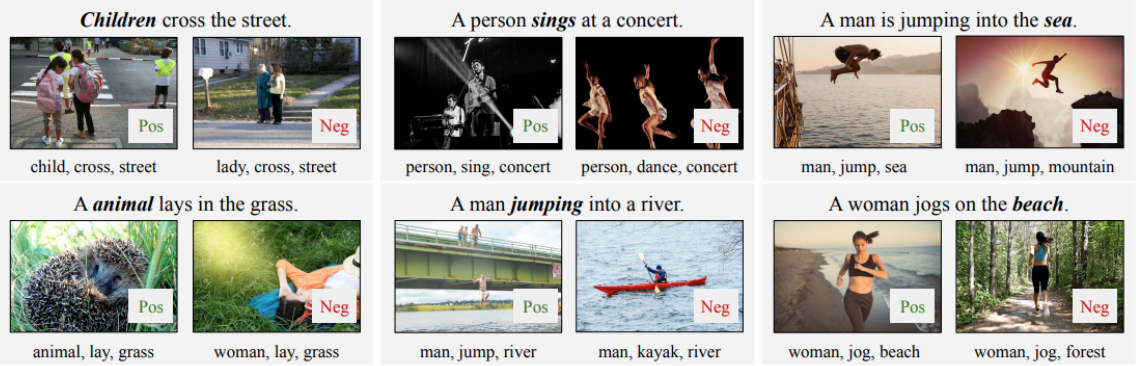
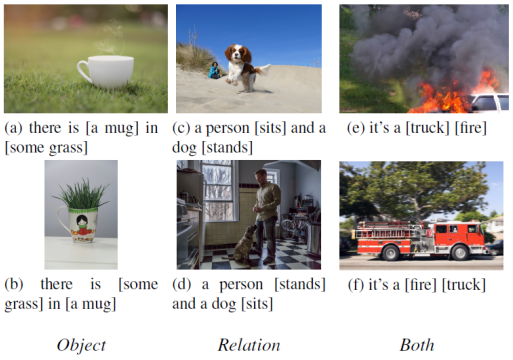


Figure 3.2: Examples of SVO Probe by Hendricks and Nematzadeh (2021). Images on the left and right show positive and negative image examples for each sentence. Below each image is the subject, verb, object triplet corresponding to the image.

adjectives, prepositions, and/or adverbs, which typically take nouns referring to objects as semantic arguments. Swaps of relations and objects can involve two separate swaps or a single swap that changes parts of speech (for example,"it's a [fire] [truck]" vs. "it's a [truck] [fire]"). Figure 3.4a shows examples of each broad tag group.

This task is used to evaluate a variety of state-of-the-art V&L models. Their results demonstrated that models perform no better than chance level across all tests, which indicated the visio-linguistic compositional reasoning capabilities of these models fall dramatically short of what they had hoped.



(a) linguistic tags

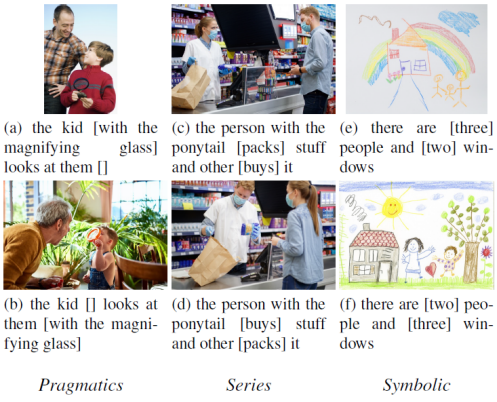


Figure 3.3: visual tags

(a) Examples of Winoground for linguistic tags and visual tags by Thrush et al. (2022). The visual examples are additionally tagged with the *Relation* tag, and 1, 2, and 1 main predicates from left to right. The linguistic examples are additionally tagged with 2, 1, and 1 main predicates from left to right.

3.4 VALSE

Our proposed benchmark is most similar to the VALSE (Vision And Language Structured Evaluation), which tests general-purpose pretrained V&L models for their visio-linguistic grounding capabilities on specific linguistic phenomena. With the same goal of understanding pretrained V&L models' linguistic abilities, our benchmark focuses on different phenomena that are eminently linguistic without involving reasoning.

The VALSE benchmark contains a wide range of linguistic phenomena from morphosyntax to semantics in the visual modality (Bernardi and Pezzelle, 2021). A model is asked to determine whether a caption is a real or foil one in relation to a given image. A foil caption is constructed from the correct caption by altering a word or phrase that realizes a specific linguistic phenomenon. The tasks and examples are shown in Figure 3.5. Our Active-Passive task is similar to the VALSE action(actant swap) one, which requires models to determine the participants of an action. But our task is more challenging as it involves the active-passive voice phenomenon.

The VALSE benchmark was constructed using automatic and human validation. After constructing the foil captions using an automatic pipeline, they ensured the high quality of the generated captions using diverse automatic metrics, which includes applying the grammatical filter GRUEN (Zhu and Bhat, 2020a). We also use GRUEN to check the grammatical correctness of our constructed sentences.

Multiple V&L models are evaluated on the VALSE benchmark. While the tasks centered on individual objects are almost solved by the best V&L model ViLBERT 12-in-1 (Lu et al., 2020). However, the other tasks can not be reliably solved.

	pieces	**existence**	**plurality**	**counting**	**relations**	**actions**	**coreference**
Data collection & metadata	instruments	*existential quantifiers*	*semantic number*	*balanced, adversarial, small numbers*	*prepositions*	*replacement, actant swap*	*standard, clean*
	#examples†	505	851	2,459	535	1,633	812
	foil generation method	*nothing ↔ something*	NP replacement (`sg2pl`; `pl2sg`) & quantifier insertion	numeral replacement	SpanBERT prediction	action replacement, actant swap	*yes ↔ no*
	MLM	✗	✗	✗	✓	✓	✗
	GRUEN	✗	✓	✗	✓	✗	✗
	NLI	✗	✓	✗	✓	✗	✗
	src. dataset	Visual7W	MSCOCO	Visual7W	MSCOCO	SWiG	VisDial v1.0
	image src.	MSCOCO	MSCOCO	MSCOCO	MSCOCO	SituNet	MSCOCO
Example data	caption (blue) / foil (orange)	*There are no animals / animals shown.*	*A small copper vase with some flowers / exactly one flower in it.*	*There are four / six zebras.*	*A cat plays with a pocket knife on / underneath a table.*	*A man / woman shouts at a woman / man.*	*Buffalos walk along grass. Are they in a zoo? No / Yes.*
	image						

Figure 3.5: Tasks and examples of VALSE dataset by Parcalabescu et al. (2022)

Chapter 4

Dataset Construction

The existing research tried to probe and benchmark current state-of-the-art V&L models in a wide range of tasks. But previous work tends to intertwine language and reasoning abilities in their evaluation tasks, but language and reasoning go through different processes in human brains (Fedorenko and Varley, 2016). In this work, we aim at testing the basic language abilities of models by proposing a new benchmark that teases apart, as much as possible, language and reasoning tasks. The BLA benchmark focuses on the model's linguistic knowledge on three commonly used linguistic phenomena and the general understanding of sentence-level semantics.

4.1 Task Setup

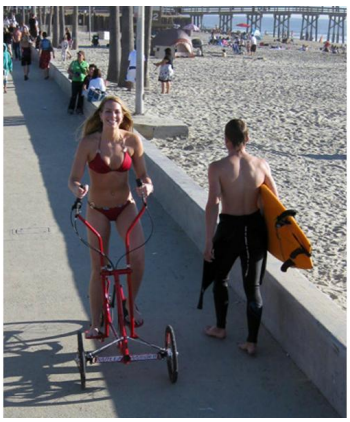
We follow similar task setups of previous work (Shekhar et al., 2017; Thrush et al., 2022; Parcalabescu et al., 2021) and design four alternative sentences for one image to test the specific abilities of the models. The four sentences contain two correct captions that correctly describe the given image and two distractor sentences that are highly similar to the target sentences but contains some mistakes in relation to the image. Instead of directly predicting whether a sentence is correct, the model can assess how much the four sentences are aligned to the given image, where the correct captions should have higher alignment scores than the distractors. Since most V&L models have been pretrained on image-text alignment tasks, this setting allows for the zero-shot evaluation, in which the pretrained models can directly perform the tasks without further training or finetuning.

The four sentences are designed to evaluate three specific language phenomena. In general, we avoid the use of reasoning abilities to finish our tasks by discarding descriptions related to reasoning tasks like numbers or positions in the construction process. To force the models to respect sentence-level semantics instead of relying on object recognition, we design the distractors that only contain objects that are present in the image. Models must understand the language phenomena to make completely correct predictions on each caption set. We describe how to design the caption sets to evaluate these language phenomena in the following paragraphs.

Active-Passive voices Passive voice is commonly used in English and understanding its difference from the active voice is crucial in processing the performer and the receiver of an action. As shown in Figure 4.1a, the two correct captions (True Active: *TA*, True Passive: *TP*) have different word orders but express the same meaning that the "woman" is the performer of the action "feed" and the "man" is the receiver. While the two sentences express the same meaning, they are phrased differently. The distractors switch the performer and receiver of the action in the original captions which change the sentence semantics while keeping the sentence syntactics relatively similar. The distractors (False Active: *FA*, False Passive: *FP*) of the

TA: the woman feeds the man
TP: the man is fed by the woman
FA: the man feeds the woman
FP: the woman is fed by the man

(a) Active-Passive



TP1: the man wears a wetsuit and carries a surfboard
TP2: the woman wears a red bikini and rides a red bike
FP1: the man wears a wetsuit and rides a red bike
FP2: the woman carries a surfboard and wears a red bikini

(b) Coordination

TP1: the man who holds a stuffed bear wears a gray polo
TP2: the man who wears a belt holds a cow
FP1: the man who holds a stuffed bear wears a belt
FP2: the man who wears a gray polo holds a cow

(c) Relative Clause

Figure 4.1: Examples of the BLA benchmark tasks

given example change the original meaning to "the man feeds the woman" by switching the participants, which contradicts to the image. To perform well in this task, models are required to understand the general sentence semantics of "who does what to whom". This can be assessed with the sentence pairs *TA-FA* and *TP-FP*. Further, the model must recognize that sentences are semantically equivalent even if they are phrased differently with different voices.

Coordination Coordination is a syntactic phenomenon in which two or more elements (conjuncts) are linked together. A coordinator (coordinating conjunction) often appears between the conjuncts. The most frequently used coordinators are "and" and "or", but they express different semantic meanings. The conjunctive coordinator "and" adds two conjuncts together, while the disjunctive coordinator "or" expresses an idea of choice or alternative between the two conjuncts. Distinguishing the different use of the coordinators and their meanings requires some extent of logical reasoning ability. Therefore, in this task, we only investigate whether models can understand coordination when conjunctive coordinator "and" is used to connect the conjuncts.

The conjuncts used in this task are attributes. Models are required to understand the two clauses are coordinated attributes that belong to a specific person entity in the given image. Using the one image with four sentences setup, the two correct captions describe two unique attributes of two different entities. As shown in Figure , the correct caption (True Person 1: *TP1*) describes that Person 1 (the "man") has two attributes ("wears a wetsuit" and "carries a surfboard") that Person 2 (the "woman") does not. These attributes are unique for the specific person entity. Another correct caption (True Person 2: *TP2*) describes such unique attributes of another person (the "woman"). The distractors (False Person 1: *FP1*, False Person 2: *FP2*) contain mismatched attributes by switching one of the unique attributes in each correct caption. For example, in the given case, the attributes "rides a red bike" and "carries a surfboard" are switched in the sentences describing the two person entities, which makes the descriptions incorrect in relation to the image. To perform this task, models need to understand the sentence semantics on "who has what attributes" and the linguistic phenomenon, which is that the subject of the sentence is connected to the parallel conjuncts.

Relative Clause A relative clause is used to modify a noun or noun phrase, which uses some grammatical device to indicate that one of the arguments within the relative clause has the same

referent as that noun or noun phrase. For example, "We met the woman who owns this hotel". In this sentence, "the woman" functions as the head, and "who owns this hotel" is the relative clause modifying the head noun phrase. The relative clause is one of the major grammatical constructions in English learning. Language learners who are capable of understanding and using relative clauses can develop a good perception of different sentence structures (Yi, 2017).

The Relative Clause task uses the design of matching person entities and their attributes, which is another version of the Coordination task. But we use a different and possibly more challenging syntactic structure. The person entity functions as the head of the clause, while one attribute becomes the relative clause that modifies the head. The example is shown in Figure 4.1c. Performing this task not only requires a basic understanding of "who has what coordinated features" but also requires comprehending the more complex sentence structures.

4.2 Data Sources

Directly constructing the sentences for our tasks is a better option, because sentences using the specific language phenomena are less common in existing V&L datasets and the construction of distractors is challenging without the annotation of the subject, object and verb. Visual Genome (VG) (Krishna et al., 2017) is used to construct our datasets because all objects, attributes and relationships in each image are annotated. This provides more flexibility in sentence construction and enables an automatic pipeline. The dataset contains images and fields of *relationships*, *attributes*, *objects*, *image data* and so on. An example of Visual Genome image annotations and fields are shown in Figure 4.2 and Figure 4.3.

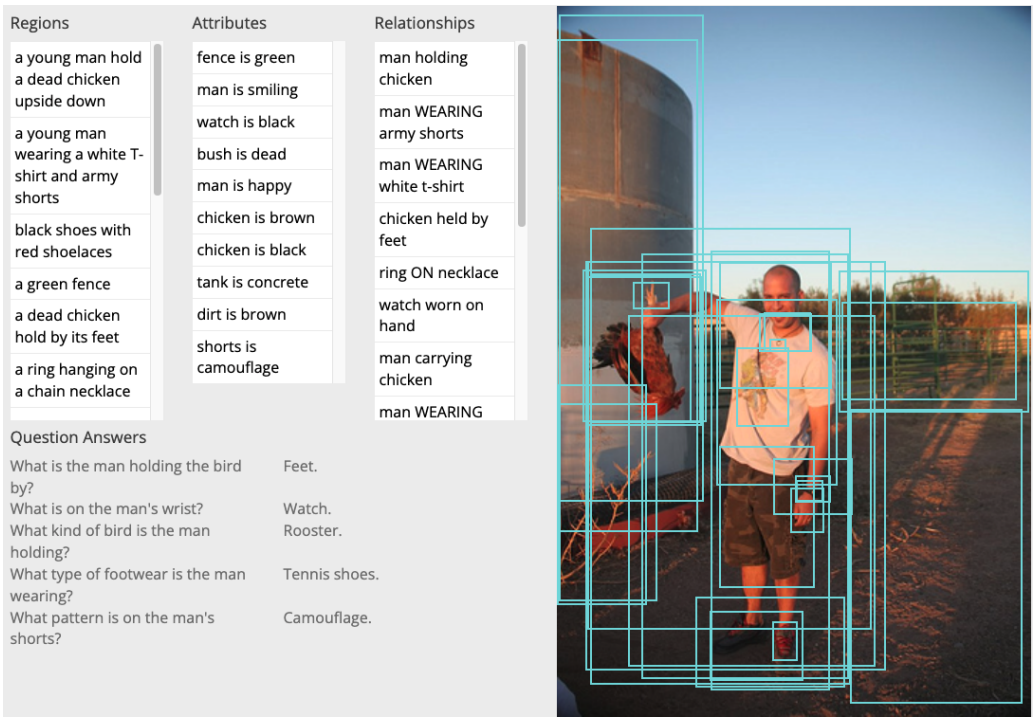
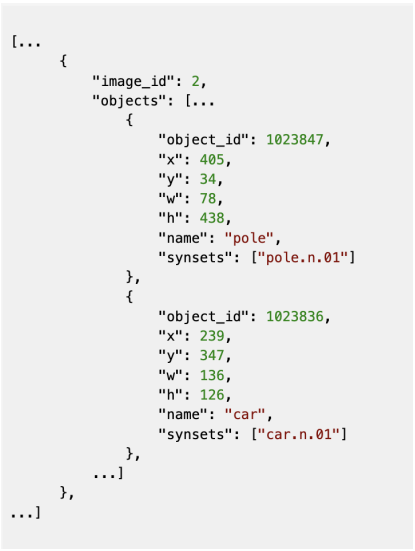


Figure 4.2: One example of Visual Genome image annotations. The bounding boxes, objects, attributes and their relationships are annotated in the VG dataset.



(a) Example for *objects*

(b) Example for *relationships*

(c) Example for *attributes*

(d) Example for *image data*

Figure 4.3: Examples of the data fields in Visual Genome

4.3 Dataset Construction

4.3.1 Active-Passive

In this task, we first construct the *TA* sentence with the selected image and relationship and then transform it into a passive voice for *TP* by changing the voice of the verb and the order of subject and object. The distractors are constructed based on the correct captions by switching the subject and object. The number of valid caption sets depends on how many valid *TA* sentences we can construct using the source dataset.

When designing the task, we aimed to make the models select the correct captions based on the image and the understanding of specific language phenomena. Therefore, we design the task in a way that the intended meaning cannot be inferred from the words alone and the switch of subjects and objects would not lead to counterfactuals in the real world. Mahowald et al. (2022) suggests that both humans and neural models can identify the agent of an action after shuffling the subject, verb and object in a sentence. This is possible because word meanings strongly constrain who is doing what to whom. For example, in the sentence "A knife is held by a woman", it is readily inferable that "woman" is the performer of the action "hold" without understanding the passive voice. And the distractor sentence "A woman is held by a knife" can be easily excluded by models without considering the image since such sentences would not occur in the training data. To avoid models taking these shortcuts, we construct the dataset with images of human interaction descriptions.

Since verbs that cannot take a direct object cannot be passive, the predicate between the subject and object must be a transitive verb. So we construct a transitive word list using transitive words from the Transitive Verbs List in English [1], and then manually add more verbs and verb phrases that can be transformed into passive voice after examining predicates that

[1] https://englishvocabs.com/transitive-verbs/184-transitive-verbs-list-in-english/

describe human interactions in Visual Genome. The following steps are used to construct the Active-Passive dataset.

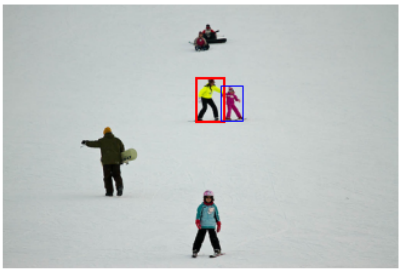
Select In this step, we select the relationships that can be used to construct the *TA* sentences. Our targets are relationships where both the subject and object are human entities and the predicate is a transitive verb. The selected relationships must meet the two requirements. To determine whether a word refers to a person entity, we use WordNet (Miller, 1995), where the nouns related to humans are categorized in a specific lexicographer set "noun.person". We check the "object" and "subject" annotations in *relationships* (an example in Figure 4.3b) and search the lexicographer file name of their "synsets" on WordNet. In this way, we can determine whether the word refers to a person entity. We use the transitive word list to check predicates. But some preprocessing steps are required to remove noisy words. For example, some parts of the noun phrases in "object" might occur in the predicate annotations like "holding a" instead of "holding". In handling such cases, we add rules to remove some common noisy words (e.g. removing the articles) in the predicates before comparing them with the word list. Also, we convert the verbs used in passive voice to simple present tense for string comparison and switch the "object" and "subject" in those cases.

Image filter In the previous step, we make some rules to select relationships that satisfy our requirements. However, since the *relationships* contain relationships of every item in the image, some small items in the background may be selected, which is difficult to recognize for both humans and models. Therefore, relationships that contain imperceptible items are discarded. We manually select a threshold for the minimum ratio size of an item in an image. The item area is calculated with the height and width of the item bounding box from *objects* (Figure 4.3a), and the image size is obtained from *image data* (Figure 4.3d). We found that when the ratio between the item area and the image area is greater than 1%, the item is usually clear enough for humans to recognize. So we discard relationships whose object or subject item ratio is smaller than 1%. Some examples whose object or subject ratio is close to 1% are shown in Figure 4.4.

Construct captions The select relationships can be used to construct a simple sentence that is formed by subject, predicate and object. But in some cases, more information is needed to avoid semantic ambiguity. In some cases, the subject and the object can be annotated with the same noun but with different object ids, which refer to two different objects in the image. For example, both the subject and object are annotated as "man" but they refer to two different person entities in the image. To distinguish the two entities, we use the object id to look for different attributes of the persons from the *attributes* (Figure 4.3c), and replace the subject and object with "distinct attribute + subject/object name" (when the attribute contains only one word) or "subject/object name + distinct attribute" (when the attribute contains more than one word). The predicate is transformed into the simple present tense and agrees with the subject in a number using Python's Pattern library. The template for the sentence construction is shown in Table 4.1. To ensure grammar correctness, a definite article "the" is added before the subject and the object. We chose a definite instead of an indefinite article because of the following reasons: (i) It's easy to use in the template without considering whether the subject/object is singular or whether the word following begins with a vowel; (ii) A definite article before subject/object is more commonly used in other image caption datasets like MSCOCO. (iii) To make it more consistent with the other two constructed datasets, where a definite article is also used before the subject and object.

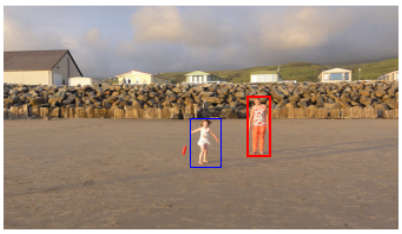
(a) Discarded example
man: 0.9%, skater: 11.5%

(b) Discarded example
woman: 6.9%, **baby: 0.8%**

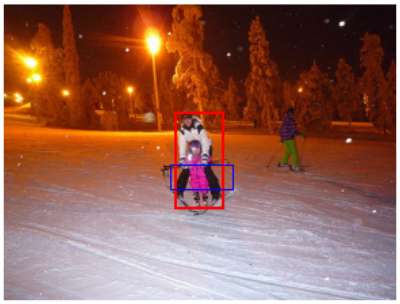


(c) Discarded example
man: 1.5%, **girl: 0.8%**

(d) Saved example
man: 1.1%, woman: 2%



(e) Saved example
woman: 1.5%, girl: 1.7%



(f) Saved example
woman: 3.9%, child: 1.3%

Figure 4.4: Example images with item size ratios, where the subject is labeled in red and the object are labeled in blue. The relationship will be discarded if any of the item ratios is smaller than the threshold of 1% (**in bold**).

	Voice	Template	Example
TA	Active	the *subject predicate(active)* the *object*	the man holds the woman
TP	Passive	the *object* is/are *predicate(passive)* by the *subject*	the woman is held by the man
FA	Active	the *object predicate(active)* the *subject*	the woman holds the man
FP	Passive	the *subject* is/are *predicate(passive)* by the *object*	the man is held by the woman

Table 4.1: Template for caption construction in Active-Passive dataset. **TA.** True Active. **TP.** True Passive. **FA.** False Active. **FP.** False Passive.

Caption filter The captions are generated automatically following the previous steps, but the language quality can not be guaranteed with the rule-based automatic pipeline. In addition, due to repeated annotations in *relationships*, multiple same or similar captions can be constructed for the same image (e.g. "a woman holds a baby" and "a lady holds a child"). And if an image contains more than one human interaction relationship, several different captions are generated for the same image. However, to avoid the bias that some images are easier to process than others, only one caption set for each unique image should be used for evaluation. Our solution to alleviate the above problems is by using GRUEN pretrained model (Zhu and Bhat, 2020b) to discard caption sets that contain ungrammatical sentences. And if multiple caption sets are of good language quality, we only keep the one with the highest GRUEN score for each image. A sentence is considered ungrammatical if its GRUEN score is lower than 0.7. The paper (Parcalabescu et al., 2021) also uses GRUEN to discard ungrammatical sentences but with a threshold of 0.8, but we decide to use 0.7 after examining the sentences and their GRUEN scores. The sentences discarded are shown in Table **??**. Interestingly, we found that some sentences that are grammatically correct but contradict commonsense also get low GRUEN scores (e.g. "The baby holds the man"). This may be because the GRUEN model also considers sentence likelihood. And since such counterfactual sentences seldom exist in the texts used to train the

GRUEN model, its sentence likelihood is low. But we consider this might be the same case for other pretrained models, so we discard such sentences.

Sentences	GRUEN scores
the adult is held hand of by the child	0.172
the guy records the man	0.4
the here man is watched by the man	0.448
the child holds the man	0.572
the mother is held by the girl	0.586
the woman is taken picture of by the man	0.616
the little girl holds the man	0.672
the baby holds the man	0.68
the lady dresses the man	0.698

Table 4.2: Example sentences whose GRUEN scores are lower than 0.7 (Active-Passive task)

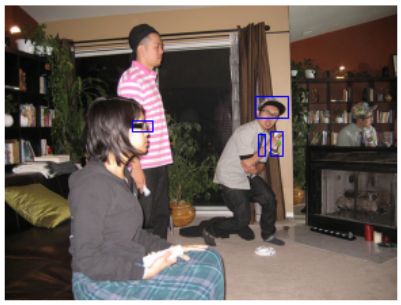
4.3.2 Coordination

Select Images are selected if at least two person entities (p_1, p_2) exist and are annotated as subjects in the *relationships*. And each person entity is annotated with at least two relationships, which suggests that the person has more than two attributes. The subjects are determined as different persons if (i) they have different object id and (ii) their bounding boxes' overlap size is less than 85%. The 85% is an empirical threshold we select after checking multiple examples where the different object ids refer to the same person entity. Step (ii) is necessary because some subjects are annotated with different object ids but refer to the same person entity in the image.

Image filter Similar to the Active-Passive dataset construction, we filter images where items in the relationship are too small to recognize. Since the subject in the sentence is a person entity and the objects are attributes, we use different size ratio thresholds for the items in the two categories. According to the observation of the data, we use 0.1% as the minimum ratio for person entities, and 0.05% for attributes. Some examples where the size ratio is around the attribute threshold are shown in Figure 4.5, which motivated our choice for the threshold values.

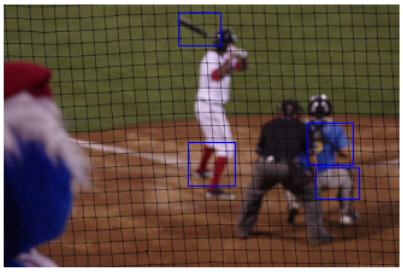
Extract unique attributes As described in the task setup in Section 4.1, we need to find two unique attributes for each person entity. Therefore, we remove the similar ones between the two entities in this step. We first construct the attribute phrases for each person. We select relationships whose subject belongs to one of our selected person entities from the *relationships*. With the *attributes*, the attribute phrase is constructed with "predicate + article of the object + attribute of the object + object" (e.g. wears a blue t-shirt). The predicate is transformed to be in the simple present tense and to match with the object. Due to repeated annotations, a person might have multiple same attributes that have the same or similar annotations. These attributes are removed with simple string comparison and Word2Vec (Church, 2017) similarity. Word2vec can only be applied to one single word. Since the different attributes usually vary on the last noun ("wears t-shirt" vs. "wears a shirt"), we only compute the similarity between the last words of the attributes rather than the average word similarity to avoid the influence of using the same predicates. When the similarity is greater than 0.7, the two attributes are considered similar and one of them will be removed. But the Word2Vec method cannot

(a) Discarded example
phone: 0.14%, watch: 0.22%
watch: 0.12%, earring: 0.05%



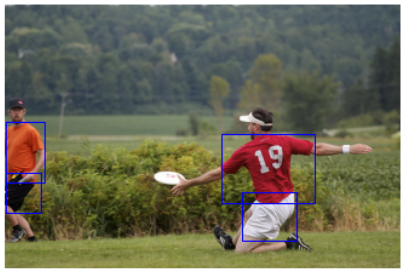
(b) Discarded example
cap: 0.59%, glasses: 0.19%
tie: 0.14%, remote: 0.25%

(c) Discarded example
bag: 0.98%, suitcase: 0.2%
sweater: 0.36%, pants: 0.82%



(d) Saved example
bat: 1.31%, socks: 1.96%
shirt: 1.83%, pants: 1.35%

(e) Saved example
shirt: 8.71%, shorts: 2.43%
shorts: 1.8%, frisbee: 1.15%



(f) Saved example
shorts: 2.53%, jersey: 6.11%
shirt: 2.31%, pants: 1.38%

Figure 4.5: Example images with attribute item size ratios. The relationship will be discarded if any of the attribute item ratios is smaller than the threshold of 0.05% (the first line).

be used to select unique attributes between two persons, because it fails to distinguish some adjectives that are commonly used to refer to different features in the VG, such as different colors. Therefore, we apply a pretrained Sentence-BERT model (Reimers and Gurevych, 2019) to compare the attributes of two persons. To reduce the computational complexity, only the first six attributes of each person are used for computation. The Sentence-BERT similarity threshold is 0.8, which is to say, if the similarity between two attributes is greater than 0.8, they are treated as similar attributes and will be eliminated. In addition, some rules are defined manually to remove potentially similar attributes between persons but are not annotated in the *relationships*. These attributes contain body parts, such as "has legs", and "has a lip", which are removed using a list of body part words.

Construct With the above steps, each selected image now has at least two persons and at least two unique attributes for each person entity. To increase the quality of the constructed cases, different combinations of two person entities (p_1, p_2) and their two unique attributes $(f_{1p_1}, f_{2p_1}, f_{1p_2}, f_{2p_2})$ are chosen to construct each caption set. We ensure that the correct caption only contains two distinct attributes of the person entity that the subject refers to, while the distractor must contain one attribute of the subject and another attribute that belongs to the other human entity. We construct at most three test sets of captions for each image to reduce the computation time. The captions are constructed with the template defined in Table 4.3.

Filter Similar to the filter step for the Active-Passive task, we filter constructed sentences whose GRUEN score is lower than 0.7. Some examples of the discarded cases using GRUEN

	Person	Template	Example
TP1	P1	the p_1 f_{1p_1} and f_{2p_1}	the man wears a white shirt and holds a controller
TP2	P2	the p_2 f_{1p_2} and f_{2p_2}	the woman wears a blue shirt and sits on a sofa
FP1	P1	the p_1 f_{1p_1} and f_{2p_2} / f_{1p_2}	the man wears a white shirt and sits on a sofa
FP2	P2	the p_2 f_{2p_1} and f_{1p_2}/ f_{2p_2}	the woman holds a controller and wears a blue shirt

Table 4.3: Template for caption construction in Coordination dataset. **TP1.** True Person 1. **TP2.** True Person 2. **FP1.** False Person 1. **FP2.** False Person 2.

are shown in Table 4.4. For images with more than one qualified caption sets, we only keep the one with the highest GRUEN score.

Sentences	GRUEN scores
the girl wears a tie dye skirt and looks at all types books	0.37
the woman wears a black and has blonde hair	0.450
the man has brown hair and watches a man	0.581
the woman watches a man and has blonde hair	0.629
the man boards an in station train and wears jeans	0.654
the man has a tan and has foot on a blue surfboard	0.692

Table 4.4: Example sentences whose GRUEN scores are lower than 0.7 (Coordination)

4.3.3 Relative Clause

As described in Section 4.1, we use the same steps of Coordination dataset construction to extract unique human attributes. During the sentence construction step, we only change the template for sentence generation to adapt to the Relative Clause task. With different combinations of two person entities (p_1, p_2) and their two unique attributes $(f_{1p_1}, f_{2p_1}, f_{1p_2}, f_{2p_2})$, we construct at most three caption sets using the template (Table 4.5). We apply the same rule for constructing the correct captions and distractors as the Coordination task. GRUEN is applied to discard low language quality caption sets and eventually, only one caption set is kept for each image if multiple sets are preserved after the caption filter step.

	Person	Template	Example
TP1	P1	the p_1 who f_{1p_1} f_{2p_1}	the man who wears a white shirt holds a controller
TP2	P2	the p_2 who f_{1p_2} f_{2p_2}	the woman who wears a blue shirt sits on a sofa
FP1	P1	the p_1 who f_{1p_1} f_{2p_2} / f_{1p_2}	the man who wears a white shirt sits on a sofa
FP2	P2	the p_2 who f_{2p_1} f_{1p_2}/ f_{2p_2}	the woman who holds a controller wears a blue shirt

Table 4.5: Template for caption construction in Relative Clause dataset. **TP1.** True Person 1. **TP2.** True Person 2. **FP1.** False Person 1. **FP2.** False Person 2. The template uses two different persons (p_1, p_2) in the image as the subject and two unique attributes of each persons $(f_{1p_1}, f_{2p_1}, f_{1p_2}, f_{2p_2})$

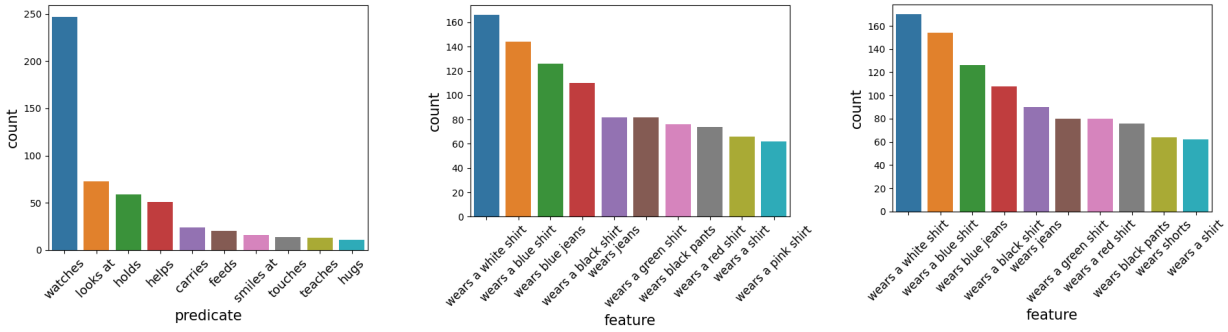
4.4 Statistics

Taken the steps mentioned above, we construct three task datasets: Active-Passive voices, Coordination and Relative Clause. We report their dataset sizes, vocabulary sizes, average

sentence length, average GRUEN scores and the number of unique predicates (for the Active-Passive task) and the number of distinctive attributes (for Coordination and Relative Clause tasks) in Table 4.6. Using the same construction process, the difference in the dataset size between Coordination and Relative Clause datasets is caused by the GRUEN score filter, where more examples in Relative Clause were discarded. For Active-Passive tasks, each caption set (one image with four sentences) is constructed with a predicate. We calculate the number of caption sets that use the unique predicates and show the ten most frequent predicates in Figure 4.6a. For Coordination and Relative Clause tasks, each caption set are constructed with four attributes. We calculate how many caption sets are constructed with the distinctive attributes and present the ten most frequent ones in Figure 4.6b and Figure4.6c.

Dataset	Size	Vocabulary size	Average sentence length	Average GRUEN score	Num. of unique predicates / distinctive attributes
Active-Passive	613	165	6.33	0.86	43
Coordination	800	827	10.46	0.84	1670
Relative Clause	789	807	10.46	0.85	1628

Table 4.6: Statistics for the BLA benchmark tasks. The number of unique predicates is for the Active-Passive dataset and the number of distinctive attributies is for the Coordination and Relative Clause tasks.



(a) Active-Passive (b) Coordination (c) Relative Clause

Figure 4.6: Top 10 predicates for the Active-Passive dataset (4.6a), Top 10 attributes for the Coordination (4.6b) and Relative Clause (4.6c) dataset. The frequency is the number of caption sets that are constructed with the specific predicate/attribute.

During dataset construction, we ensure that two correct captions and two distractors in one caption set contain the same exact words, avoiding potential bias in their sentence vocabulary use. For the visual features, as the two nouns used to construct the Active-Passive cases are always the same, they should not have potential bias regarding to the object sizes. For the Coordination and Relative Clause, the nouns for person entities and attributes are not completely the same for the four sentences in one caption set. We report the average size ratios of objects that the human entity and the attributes refer to4.7. The reported metrics are very close for the four captions, except that the size difference between Person 1 and Person 2 objects is relatively big. This is because the person entities that are more obvious (larger) usually appear first in the relationships. When analyzing the model performance, we use this feature to assess whether models show bias towards sentences that mention larger objects.

	Avarage length	Average person entity size	Average human property size
TP1	10.43	0.27	0.095
TP2	10.47	0.16	0.087
FP1	10.55	0.27	0.091
FP2	10.4	0.16	0.091

(a) Coordination

	Average length	Average person entity size (%)	Average attribute size (%)
TP1	10.44	0.27	0.097
TP2	10.46	0.16	0.086
FP1	10.56	0.27	0.092
FP2	10.39	0.16	0.09

(b) Relative Clause

Table 4.7: Statistics of the two correct captions and two distractors of the same caption set on the Coordination and Relative Clause datasets. The average "size" refer to the average of object ratio sizes for all TP1, TP2, FP1 or FP2 sentences in one dataset. Each sentence contains two attributes, so we take the average of the two for each sentence.

Chapter 5

Experimental Setup

The BLA dataset is designed to investigate the extent to which the pretrained models learn to ground specific linguistic phenomena in a zero-shot setting. Similar to previous V&L Benchmarks (Parcalabescu et al., 2022; Thrush et al., 2022), our tasks require the model to predict image-sentence matching scores between the image and the correct captions vs. the distractors. A model that performs well on these tasks is supposed to predict higher scores for the image-caption pairs than the image-distractor pairs. Each pair score is independent, which means that only one image-sentence pair is visible to the model when it computes the alignment scores.

5.1 Benchmark Tasks

Our BLA tasks require models to understand sentence semantics. But before directly moving on to high-level sentence semantics, we verify whether models have good knowledge of word-level semantics using the FOIL dataset (Shekhar et al., 2017). The dataset contains one image, one correct caption and one "foil" caption (distractor), which is highly similar to the correct caption but contains one single mistake ("foil" word). A good model should rank the correct caption higher than the "foil" caption. The FOIL task is very similar to our task and serves as a good tool to examine models on their noun understanding. Hessel et al. (2021) reported that CLIP ranks 87.2% correct captions higher than the "foil" ones on this task with their newly proposed method CLIPmbox-Scores. Parcalabescu et al. (2021) evaluates the models CLIP, ViLBERT, LXMERT and others on 934 FOIL samples that have been verified by human annotators, while our work follows the settings in Hessel et al. (2021) and uses the entire dataset.

Each test set of the BLA tasks contains one image, two correct captions and two distractors (Figure 5.1). The BLA dataset evaluates a model's language abilities at different levels. For general sentence semantic understanding, the Active-Passive task accesses whether the models can understand "who does what to whom" in the image and the sentence, while the tasks Coordination and Relative Clause requires models to correctly match "who has certain attributes" in the two modalities. Further, the tasks also evaluate more fine-grained understanding on different language phenomena. The Active-Passive task requires models to understand the performer and receiver of an action when two semantically equivalent sentences are phrased in different voices. The Coordination task tests whether the models understand the use of coordination while the Relative Clause task uses a more complex sentence structure. In the BLA task, the two correct captions should be ranked higher than the distractors, which is to say that, the correct captions should be placed in the top two, while the distractors should be ranked third or fourth.

Correct: A man with glasses using a laptop computer
Foil: A man with glasses using a mouse computer

(a) Example of FOIL

TA: the man holds the baby
TP: the baby is held by the man
FA: the baby holds the man
FP: the man is held by the baby

(b) Active-Passive example of BLA

TP1: the woman wears blue dress and holds wineglass
TP2: the man wears black glasses and wears a white shirt
FP1: the woman wears blue dress and wears black glasses
FP2: the man holds wineglass and wears a white shirt

(c) Coordination example of BLA

TP1: the man who wears a red hat walks a dog
TP2: the man who wears a brown jacket wears a gray hat
FP1: the man who a red hat wears a brown jacket
FP2: the man who walks a dog wears a gray hat

(d) Relative Clause example of BLA

Figure 5.1: Examples of FOIL and BLA dataset

5.2 Models

In our experiment, we use multimodal models CLIP, LXMERT, ViLBERT and unimodal model GPT2 to perform on the evaluation tasks in a zero-shot setting.

CLIP CLIP is a state-of-the-art V&L model that achieves a high accuracy of 87.2% on FOIL (Hessel et al., 2021), a similar task to our work. And without training or finetuning on specific tasks, this model can be directly applied to the FOIL and BLA tasks by comparing the visual and textual feature similarities between correct captions and distractors. To use a comparable setup as CLIP, other V&L models are required to predict the alignment scores between each sentence and image pair independently in a zero-shot setting. Therefore, we choose the state-of-the-art dual-stream models ViLBERT and LXMERT, which contains a well-trained image-text matching (ITM) head.

The version of CLIP that we employ is ViT-B/32. The text encoder and image encoder output a 512-dimensional vector that represent the textual features and visual features independently. The image-text similarity is measured with the *CLIPScore* (Hessel et al., 2021):

$$CLIP\text{-}S(c, v) = w * max(cos(c, v), 0) \tag{5.1}$$

where we scale the cosine similarity between visual embedding v and a candidate caption with textual embedding c and set the scaling factor w to 2.5 as the original paper did. The image-sentence pair with a higher CLIPScore is ranked in a higher order.

ViLBERT and LXMERT We employ the VOLTA framework (Bugliarello et al., 2021) to evaluate ViLBERT (Lu et al., 2019) and LXMERT (Tan and Bansal, 2019) on our tasks.

The pretrained weights we use are ViLBERT (CTRL) and LXMERT (CTRL)[1], which are pretrained on the Conceptual Captions (CC) dataset (Sharma et al., 2018) using the controlled setup described in Bugliarello et al. (2021). There are differences in the pretraining details between the VOLTA implementation and the original papers. LXMERT is pretrained on CC instead of the five datasets (described in Section 2.2) using 10 epochs instead of 20. An overall [*IMG*] feature was added to denote the representation of the entire image which did not exist in the original version. The visual stream of ViLBERT uses a lower dimensional representation. Same as the original models, the image features (regions of interest) are extracted with a Faster R-CNN (Ren et al., 2015) trained on the Visual Genome dataset (Krishna et al., 2017) for both LXMERT and ViLBERT. The number of features is set to 36, the same number used in the VOLTA pretraining process.

Similar to CLIP, the models are used in a zero-shot setting. We use the image-text matching (ITM) head of the models, which outputs a logit score for image-sentence alignment prediction. A higher logit score means the image and text are more aligned and thus the pair should be ranked higher. In addition, we also explore whether using similarities between visual features and textual features of ViLBERT and LXMERT, whose settings are similar to CLIP, would also achieve a high accuracy on the task FOIL. Such settings make the three models more comparable. The similarity score is computed using the dot product of the visual embedding and token embedding.

GPT2 Multimodal models with strong language abilities may be able to solve V&L tasks by taking advantage of linguistic bias in a single modality (Goyal et al., 2017; Shekhar et al., 2019). To investigate whether our tasks suffer from such bias, we evaluate a language-only model GPT2 (Radford et al.) on the BLA benchmark. GPT2 is an autoregressive language model pretrained on a very large English corpus. Such large language models incorporate commonsense knowledge through training with a large volume of texts and may be able to detect plausibility biases in the dataset (Petroni et al., 2019; Wang et al., 2020). This can be done by leveraging the perplexity scores of sentences. As perplexity measures the likelihood of a given sentence with reference to previously encountered text, a sentence that is less plausible and contradicts commonsense should be assigned a higher perplexity score. The perplexity score of each sentence is calculated with the Hugging Face evaluate library [2]. A sentence with a low perplexity is ranked higher.

Without the visual input, we expect the unimodal model to perform at chance level in our tasks since model can only select the correct captions based on the image. If this is not the case and GPT2 is able to make correct predictions based on sentence perplexity only, the benchmark tasks can be solved by models with strong linguistic ability without replying on visual grounding.

5.3 Evaluation Metrics

To evaluate models on the BLA tasks in a zero-shot setting, it is not always possible for a model to directly predict whether a sentence is correct or wrong in relation to the image because it requires a well-trained classifier. Therefore, we use sentence-image alignment scores to indicate which sentence is more aligned to the image. The correct captions should have a higher alignment score than the distractor captions. If a correct caption is ranked first or

[1]ViLBERT (CTRL) and LXMERT (CTRL) are provided by `https://github.com/e-bug/volta/blob/main/MODELS.md`

[2]`https://github.com/huggingface/evaluate`

second, or a distractor is ranked third or fourth, the prediction for the sentence is considered CORRECT.

We use **sentence accuracy** (*sen_acc*), **set accuracy** (*set_acc*) and **set error rate** (*set_error*) to evaluate the model performance.

- **Sentence accuracy** is used in the sentence level to judge how many sentences the model makes correct predictions in the whole task. The sentence accuracy provides insights on how well a model can identify the correct or incorrect sentences. The chance level of the sentence accuracy is 50% since each sentence can be either predicted correct or incorrect.

- **Set accuracy** is used in the caption set (one image with four sentences) level, which measures how many caption sets in a task that the model makes correct predictions on. A caption set is predicted correctly only if **both** correct captions are ranked higher than the two distractors. It goes beyond guessing whether one sentence is correct but requires a model to fully understand the interactions between the image and sentences and rank all the sentences in completely correct orders. For each caption set sentences (Correct Caption 1, Correct Caption 2, Distractor 1, Distractor 2), only ranking orders (1, 2, 3, 4), (2, 1, 3, 4), (1, 2, 4, 3) and (2, 1, 4, 3) are correct. The number of all possible ranking orders for four sentences in a set is $A_4^4 = 24$. So the chance level is $4/16 = 16.67\%$.

- **Set error rate** is the opposite of the set accuracy and is also used in the caption set level. It measures how many caption sets in a task that the model makes "completely wrong" predictions, which is to say **both** distractors are ranked higher than the correct captions. A high error rate indicates that the models might have learnt some incorrect conceptions. For example, in the Active-Passive task, if the model makes "completely wrong" predictions, the model might misunderstand the agent and patient of the verb. Similar to the set accuracy, the chance level for the error rate is 16.67%.

5.4 Human Annotation

To further assess the task difficulty and the dataset quality, we collect human annotations from crowdworkers and experts via the Appen platform [3].

The crowdworkers we employed are from English-speaking countries. Like the models, annotators are shown one image-sentence pair at a time. The annotators are asked to make a binary choice on whether the sentence is correct in relation to the image. The question interface is shown in Figure 5.2. Each question is labeled by five annotators. The answer takes the majority choice of the annotators, which means at least three annotators agree on the answer. Our original plan was to collect annotations for 80 randomly selected caption sets for each language phenomenon, which are 320 questions for each language phenomenon job. But due to some unexpected problems, we ended up collecting 200 questions for the Active-Passive and Coordination tasks, and 273 for Relative Clause. Details are described in Section B. To ensure data quality, we set 20 Test Questions for each job. The Test Questions are selected from the same language phenomenon task of the BLA dataset but never appear in the 80 caption sets in the job. We provide the golden label as the answer and manually add reasons for the answer choice for each Test Question. The crowdworkers are required to take Test Questions during participating in the task and must maintain the accuracy of 70% in the Test Questions to finish their annotations. We also require the crowdworkers to spend at least 2 seconds on one question and contribute to 60 judgments at maximum. We launch 100 questions at a time as a pilot for each job and verify the answer quality before launching the rest.

[3] `https://appen.com/`

We also collect expert annotations using Appen's internal channel option. The three experts are all involved in this thesis project and are all non-native English speakers. Two annotators are experts in linguistics and one is the dataset constructor who is the expert on the tasks. Similar to the crowdworker annotation, the annotated answer to each question takes the majority choice of the annotators. We assume the answer quality of the experts are good and did not use Test Questions.

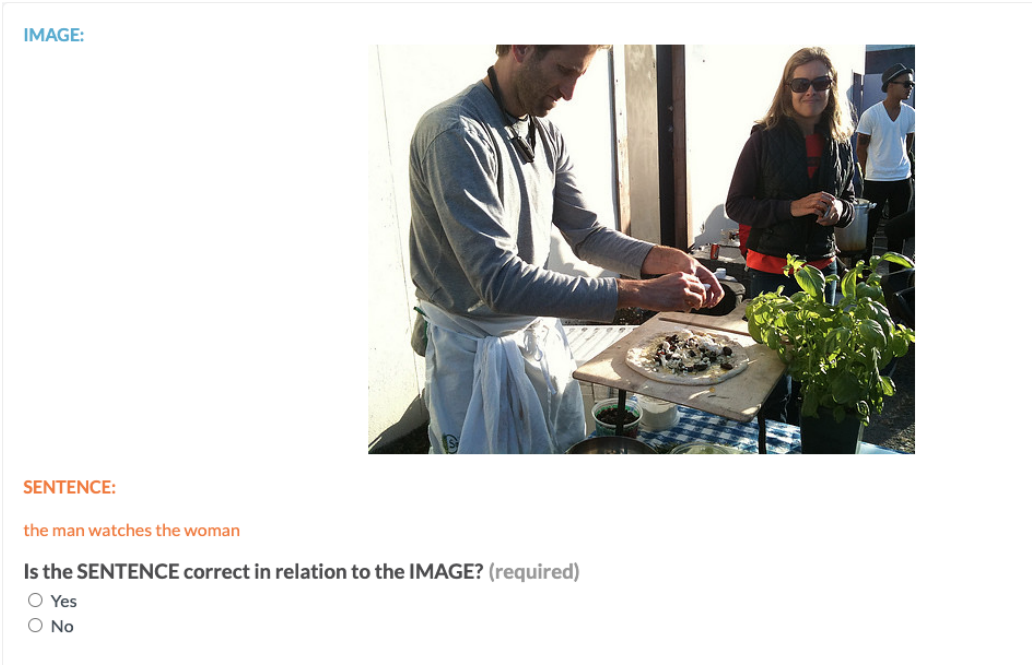


Figure 5.2: Example of the Appen question interface. The golden label of this question is "No".

Chapter 6

Results

In this chapter, we first present results on FOIL dataset as a sanity check for the model's performance in understanding word-level semantics, and human performance on BLA tasks as a ceiling for a model's performance on this benchmark. Following this, we demonstrate the zero-shot evaluation results of the unimodal and multimodal models.

6.1 Preliminary Results on FOIL

The results in Table 6.1 show that the pretrained multimodal models can understand word-level semantics and are able to identify objects and their presence in the image. CLIP achieves a high accuracy of 88.8% on FOIL, followed by ViLBERT and then LXMERT (Table 6.1). The accuracy of using dot product is significantly lower than using the Image-Text Matching (ITM) head for LXMERT, so we choose to use the ITM head predictions of ViLBERT and LXMERT for the BLA evaluation.

The CLIP and ViLBERT results are very close to the results reported in VALSE (Parcalabescu et al., 2021). The accuracy of LXMERT (ITM) is about 5% lower than that in VALSE. Two possible reasons are: (i) the LXMERT using VOLTA controlled setup was pretrained on Conceptual Captions only while LXMERT used in VALSE was pretrained on a larger dataset. (ii) we tested with the entire FOIL datasets instead of the 934 human verified samples in VALSE. But overall, our results are comparable to the results in the previous paper.

	CLIP	ViLBERT	LXMERT
dot product t,v	88.8	85.7	73.5
ITM head	//	85.4	81.3
Parcalabescu et al. (2021)	88.8	87.1	86.9

Table 6.1: Model performance on FOIL. Our experiment is conducted on the entire FOIL dataset. ITM head refers to the result using the Image-Text Matching head. dot product t,v : the dot product of visual and text features. The results obtained by Parcalabescu et al. (2021) are obtained on 934 human verified samples from FOIL.

6.2 Human Performance

As described in the Experimental Setup, the accuracy is calculated by comparing the label and the majority vote across annotators per image-sentence pair. As shown in Table 6.2, the experts performed the best on Active-Passive tasks, achieving an accuracy of 91.5% on sentence accuracy. Discussions after the annotation had been carried out indicated that the two expert

annotators, who very often agreed, partly relied on world knowledge to solve the task and achieve the accuracy of over 90%, while the task expert achieved a lower accuracy of 84% mostly based on the information in the questions. The sentence accuracy on Coordination and Relative Clause tasks is lower than the Active-Passive task but still reaches an accuracy higher than 80%. The set accuracy is relatively low. This is expected because distinguishing person entity attributes is more challenging and some samples contain attributes that are less obvious or not unique. Although the evaluation uses a small subset of the tasks, it serves as a good indication of the performance ceiling on the BLA benchmark.

The results collected from crowdworkers are very low and seem to be less reliable. The crowdworkers are more lenient to the incorrect descriptions of the image and select "yes" for most of the cases, which is mentioned in Thrush et al. (2022). Their performance in two pilots varies a lot. More details are discussed in Appendix B. Due to time and budget constraints, we stopped collecting more data and decided to use expert performance as human performance results. For this reason, we did not collect annotations for all four samples of one image (one image contains four image-sentence pairs regarding to its four sentence captions). Therefore, there are no set accuracy or set error results from the crowdworkers.

		Active-Passive	Coordination	Relative Clause
Experts	number of samples	200	200	200
	sentence accuracy	91.5	84.0	88.5
	set accuracy	82.0	54.0	64.0
	set error↓	0.02	0.0	0.0
Crowdworkers	number of samples	200	200	273
	sentence accuracy	54.5	57.5	64.8

Table 6.2: Results of human performance. The annotations were collected from experts and crowdworkers. Within each group, the annotated label takes the majority vote across annotators per item. The chance level of set accuracy and sentence accuracy are 16.7% and 50% respectively. Set error, the lower the better.

6.3 Unimodal Results

As shown in Table 6.3, the sentence accuracy of GPT2 is close to random and the set accuracy is much lower than the chance level for all tasks. This indicates that the tasks are not solvable by relying on plausibility bias or other linguistic biases in text. The error rate reaches 0% for the Active-Passive task, but this does not suggest good model performance. If the model always ranks sentences using the same active or passive voice in the top 2, there must be one correct caption. In this case, the model never makes "completely wrong" predictions for the four sentences in one caption set. This hypothesis is confirmed by the analysis in Section 7.1.1.

6.4 Multimodal Results

As shown in Table 6.3, the models struggle across the board on BLA, often performing close to or even below random chance. All multimodal models have similar results in each evaluation metric but their performance varies regarding different tasks.

Metric	Random	Model	Active-Passive	Coordination	Relative Clause
sen_acc	50.0	CLIP	50.8	49.88	**50.51**
		ViLBERT	**51.22**	50.25	50.19
		LXMERT	48.12	50.06	50.0
		GPT2	50.08	**50.38**	49.94
set_acc	16.7	CLIP	6.2	1.75	**3.17**
		ViLBERT	**8.48**	1.88	1.9
		LXMERT	5.55	1.62	1.65
		GPT2	0.16	**2.5**	1.77
set_error↓	16.7	CLIP	6.04	2.0	2.15
		ViLBERT	6.04	**1.38**	**1.52**
		LXMERT	9.3	1.5	1.65
		GPT2	**0.0**	1.75	1.9

Table 6.3: Model performance on Active-Passive, Coordination and Relative Clause. We use blue bold for the best scores per metric per task, and green underline for the worst scores per metric per task. *sen_acc.* Sentence accuracy. *set_acc.* Set accuracy. *set_error.* Set error, the lower the better.

Compared to humans Multimodal results are significantly lower than human performance. We observe 70% absolute difference between humans and the best performing model ViLBERT on set accuracy in the Active-Passive task. And for the other tasks and evaluation metrics, the multimodal models have more than 40% performance gap compared to humans. Such a trend, which is also reported in Thrush et al. (2022), is not unique to our task. These results suggest that the models find it difficult to ground these specific linguistic phenomena. There is still a lot of space to improve the models.

Compared to the unimodal model Although both unimodal and multimodal models perform similarly on sentence accuracy, their performance varies on set accuracy and error rate. In Active-Passive tasks, multi-modal models outperform GPT2 significantly in set accuracy, which indicates that they process the abilities to understand and align the visual and text semantics to some extent. However, they still lack more fine-grained understanding of the two modalities as suggested by their set accuracy on the Coordination and Relative Clause tasks, where the existence of visual information does not help the multimodal models perform better than the unimodal model. The set error rates of the multimodal model on the Active-Passive task are higher than the unimodal results.

Comparison among tasks Multimodal models perform better on set accuracy in the Active-Passive tasks as suggested by the higher set accuracy. This is in line with the trend in human performance. Set error rates are also higher in the Active-Passive task. In these cases, models misunderstand the agent and patient of an action.

Chapter 7

Analysis

As presented in the previous chapter, the pretrained V&L models perform much worse than expected on the BLA benchmark. In this chapter, we perform quantitative and qualitative analyses to better understand why the models show such performance.

7.1 Quantitative Analyses

7.1.1 Bias on Active-Passive Voices

Since the active voice is more frequently used in English compared to the passive voice, we investigate whether models show a preference for active voice sentences, as they are expected to be more common in the pretraining dataset. We compare their preference within the sentence pairs True Passive vs. False Active (*TP-FA*), True Active vs. False Passive (*TA-FP*), True Active vs. True Passive (*TA-TP*), and False Active vs. False Passive (*FA-FP*) in the Active-Passive task. The results are presented in Table 7.1.

Metric	CLIP	VilBERT	LXMERT	GPT2	Random
TA-FA	42.41	47.31	47.31	58.56	50.0
TP-FP	58.08	58.89	44.54	**65.58**	50.0
TP-FA	**70.64**	54.81	54.32	**99.18**	50.0
TA-FP	**30.18**	45.35	**38.66**	**0.98**	50.0
TA-TP	**27.9**	44.21	41.92	**0.65**	50.0
FA-FP	**34.09**	51.71	43.88	**0.98**	50.0

Table 7.1: Pair comparision results on BLA Active-Passive task. Results above 60% in blue bold, and below 40% in green underline. *TA-FA.* True Active vs. False Active. *TP-FP.* True Passive vs. False Passive. *TP-FA.* True Passive vs. False Active. *TA-FP.* True Active vs. False Passive. *TA-TP.* True Active vs. True Passive. *FA-FP.* False Active vs. False Passive.

Unimodal model For *TP-FA*, the model prefers the True Passive (*TP*) sentences for 99.18% of cases. *TA-FP*, *TA-TP* and *FA-FP* also suggest that when a passive voice sentence and an active voice sentence are given, the model would predict lower perplexity scores in more than 99% of the cases. Such passive voice preference might be caused by the model's preference for longer sentences since in the same caption set, passive voice sentences are always longer than the active voice ones.

The sentence pair preference results also suggest that GPT2 prefers syntactically similar sentences. Sentences expressed in the same voice (e.g. "a man holds a baby" and "a baby holds

a man") are more syntactically similar than sentences that use different voices, but express the same meaning (e.g. "the man holds the baby" and "the baby is held by the man"). The model predicts a higher rank for passive voice sentences *TP-FP* in 98.69% of samples and predicts a higher rank for active voice sentences *TA-FA* in 0.48% of samples, so in 99.18% of the evaluation set, the top 2 predictions are sentences using the same active or passive voice. This explains why GPT2 achieves a zero set error rate in the Active-Passive task.

Multimodal models As shown in Table 7.1, to our surprise, CLIP shows a strong preference over passive voice sentences for selecting passive sentences in about 70% of cases even when the passive sentence description does not match the image. LXMERT and ViLBERT slightly prefer passive voice sentences but the tendency is not so obvious. Such passive voice sentence preference is in line with the unimodal results, so this might be related to the bias of the text encoder.

Sentences using the same voice The above results suggest V&L model shows preference for passive voice sentences. However, we would like to investigate whether this leads to the poor performance of the models. We simplify the Active-Passive task to rank within sentence pairs using the same active or passive voice, which are True Active vs. False Active (*TA-FA*) and True Passive vs. False Passive (*TP-FP*). As shown in Table 7.1, the models choose the true captions at a chance level, indicating that they fail to distinguish the semantic roles in an action that involves two participants. That is, the model struggles in understanding the general sentence semantics on "who does what to whom" even without the influence of different voices.

7.1.2 Bias on Noun Order

When comparing model preference within sentence pairs *TA-FA* and *TP-FP*, we notice that CLIP selects *TA* and *FP* sentences for almost the same frequency (*TA* for 42.41% cases and *FP* for 41.2% cases). Sentence pairs *TA-FP* and *FA-TP* have the same noun orders and subject-verb-object orders, which might suggest that the model uses specific noun order or subject-verb-object (SVO) order as cues for prediction. We further investigate this hypothesis by checking whether the ranking of the sentences can be predicted by the order of nouns or SVO orders only.

As suggested in Section 4.3.1, sentences *TA* and *FP* follow the noun order of (agent, patient) and the SVO order of (agent, verb, patient), and sentences *TP* and *FA* follow the noun order of (patient, agent) and the SVO order of (patient, verb, agent). If the model utilizes noun order or SVO order as cues to predict sentence and image alignment score, it should rank $TA > FA$ and $FP > TP$ if it predicts agent > patient and/or (agent, verb, patient) > (patient, verb, agent). But our experiment suggests that only 32% of cases follow the same ranking orders for nouns and sentences and 36% of cases follow the same ranking orders for SVO and sentences. This indicates, contrary to our hypothesis, that there is no strong relationship between the ranking orders of nouns/SVO and the ranking orders of sentences.

7.1.3 Difference Across Verbs

Previous results suggest that models struggle to understand an action's agent and patient. We further investigate whether it is caused by the difficulties in verb understanding. Therefore, we group caption sets according to their predicates (verbs) and select the verbs that contain more than 20 cases. The verbs and their frequency are "watch"(247), "look at"(73), "hold"(59), "help"(51), "carry"(24), "feed"(20). Verbs that appear less than 20 times in the dataset are

grouped as "others"(139). We also analyze whether the passive voice preference differs among verbs by comparing model preference within sentence pairs.

Model performance varies on verbs As shown in Table 7.2, the hardest verbs and easiest verbs vary on models. But overall, the most common predicate "watch" is challenging for all the models, which is not surprising given that the images of "watch" are ambiguous in person entity actions. As shown in Figure 7.1 (a), the original annotation is "the girls watching the boys", but we can only see girls standing behind the boys and watching something ahead of them. They might look at something else other than the boys. CLIP finds predicate "watch" most difficult and "feed" easiest while LXMERT performs worst in "feed" and best in "carry". ViLBERT achieves the highest sentence accuracy and set accuracy on "hold" and did badly in "watch". When sentences with the same voice are given (Table 7.3), models achieve high accuracy in some cases. CLIP achieves an accuracy of 90.0% in detecting the agent and patient of the action "feed" when only passive voice sentences are given. And it also achieves relatively high accuracy ($> 70\%$) for the verb "hold" and "help". This suggests that they possess a better understanding of the semantic roles regarding some specific actions. However, it is difficult to explain why models do not achieve such high accuracy when only active voice sentences are given. After all, these models were pretrained on the corpus where active voice sentences are more common.

Metric	Random	Model	Predicates							
			ALL	watch	look at	hold	help	carry	feed	other
sen_acc	50.0	CLIP	50.08	47.77	49.32	53.59	54.9	50.0	**55.0**	50.72
		VilBERT	51.22	48.79	52.74	**56.78**	50.98	52.08	52.5	52.16
		LXMERT	48.12	45.14	50.68	51.69	50.0	**54.17**	45.0	49.28
set_acc	16.7	CLIP	6.2	2.43	4.11	10.17	**11.76**	8.33	10.0	9.35
		VilBERT	8.48	6.07	10.96	**13.56**	7.84	8.33	5.0	10.07
		LXMERT	5.55	5.67	4.11	6.78	7.84	**16.67**	0.0	3.6
set_error↓	16.7	CLIP	6.04	6.88	5.48	3.39	1.96	8.33	**0.0**	7.91
		VilBERT	6.04	8.5	5.48	**0.0**	5.88	4.17	**0.0**	5.76
		LXMERT	9.3	15.38	**2.74**	3.39	7.84	8.33	10.0	5.04

Table 7.2: Multimodal model performance with different predicates on BLA Active-Passive task. We use blue bold for the best scores per metric, and green underline for the worst scores per metric. *sen_acc.* Sentence accuracy. *set_acc.* Set accuracy. *set_error.* Set error, the lower the better.

Passive voice sentence preference varies on verbs As shown in the results of *TP-FA*, *TA-FP*, *TA-TP* and *FA-FP* in Table 7.3, models prefer passive voice sentences for most predicates, but CLIP shows a preference for active voice sentences with the verb "hold". This indicates that sentence length should not be the only factor that leads to passive sentence preference. Another explanation could be that CLIP choose the passive sentences because they are more "surprising" as they seldom appear in the training dataset. In two popular V&L pretraining datasets Conceptual Captions and MSCOCO, the predicates and their frequency of using in passive voice are "hold" (456 / 89), "watch" (332 / 45), "feed" (296 / 32), "carry" (129 / 14), help (69 / 4), "looks at" (6 / 8). Compared to the other verbs, the passive voice of "hold" appears more frequently in the training corpus, so CLIP finds these cases less surprising and assigns lower scores to them. But further investigation is required to verify this hypothesis.

Metric	Random	Model	Predicates							
			ALL	watch	look at	hold	help	carry	feed	other
TA-FA	50.0	CLIP	42.41	34.41	38.36	50.85	43.14	54.17	50.0	51.8
		VilBERT	47.31	40.49	50.68	**62.72**	37.25	**62.5**	40.0	53.24
		LXMERT	47.31	34.82	58.9	54.24	58.82	45.83	55.0	55.4
TP-FP	50.0	CLIP	58.08	53.85	50.68	**71.19**	**70.59**	**66.67**	**90.0**	53.24
		VilBERT	58.89	59.51	54.79	**72.88**	58.82	58.33	**70.0**	52.52
		LXMERT	44.54	44.13	43.84	49.15	49.02	**66.67**	55.0	37.41
TP-FA	50.0	CLIP	**70.64**	**83.81**	56.16	47.46	**68.63**	**83.33**	50.0	**66.19**
		VilBERT	54.81	44.94	57.53	**84.75**	54.9	**70.83**	**60.0**	54.68
		LXMERT	54.32	44.94	**82.19**	**71.19**	**66.67**	54.17	55.0	44.6
TA-FP	50.0	CLIP	30.18	9.31	47.95	**61.02**	43.14	16.67	55.0	38.85
		VilBERT	45.35	43.32	42.47	35.59	43.14	41.67	50.0	55.4
		LXMERT	38.66	38.46	19.18	25.42	39.22	58.33	30.0	52.52
TA-TP	50.0	CLIP	27.9	10.53	43.84	45.76	29.41	20.83	45.0	41.01
		VilBERT	44.21	44.94	43.84	22.03	31.37	37.5	45.0	58.27
		LXMERT	41.92	43.72	26.03	37.29	33.33	45.83	35.0	52.52
FA-FP	50.0	CLIP	34.09	18.22	50.68	57.63	43.14	12.5	**70.0**	38.85
		VilBERT	51.71	**62.35**	42.47	33.9	47.06	33.33	**65.0**	48.2
		LXMERT	43.88	52.23	20.55	27.12	35.29	58.33	45.0	48.92

Table 7.3: Multimodal pair comparision results with different predicates on BLA Active-Passive task. Results above 60% in blue bold, and below 40% in green underline. *TA-FA.* True Active vs. False Active. *TP-FP.* True Passive vs. False Passive. *TP-FA.* True Passive vs. False Active. *TA-FP.* True Active vs. False Passive. *TA-TP.* True Active vs. True Passive. *FA-FP.* False Active vs. False Passive.

7.1.4 Bias towards Person Entities and Attributes

Correct vs. incorrect attributes of the same person entity To test the extent to which models understand "who has what attributes", we simplify the Coordination and Relative Clause tasks with the correct captions and distractors that describe the same person, which are True Person 1 vs. False Person 1 (*TP1-FP1*) and True Person 2 vs. False Person 2 (*TP2-FP2*). As shown in Table 7.4, models do not show a clear preference for the true caption that correctly describes the image. This suggests that models fail to match the person entity and their attributes in the visually-grounded task when all attributes exist in the image.

Person entity 1 vs. person entity 2 In pair comparison results *TP1-FP1*, *FP1-FP2* and *TP2-FP2*, CLIP shows a slight preference for the sentences in the order of *TP1* > *FP1* > *FP2* > *TP2*. This shows a preference over sentences that describe the person entity 1, which coordinates with the average person entity size of the four sentences (Table 4.7) reported in the previous chapter. LXMERT and ViLBERT also show a similar tendency. As models lack a fine-grained understanding of the semantics, they might predict the alignment scores by simply matching objects and nouns and ignoring their relationships. The larger objects have more salient attributes in the image, and thus might be predicted to have higher alignment scores.

Metric	Coordination				Relative Clause			
	CLIP	ViLBERT	LXMERT	GPT2	CLIP	ViLBERT	LXMERT	GPT2
TP1-FP1	54.37	52.75	49.38	50	53.49	52.85	49.94	45.75
TP2-FP2	47.25	50.62	49.75	54	47.53	50.19	50.44	53.87
TP2-FP1	39.12	41.5	43	48.5	38.4	42.21	42.59	48.67
TP1-FP2	59.25	58.75	56.88	53.62	**62.86**	57.92	58.68	51.96
TP1-TP2	59.5	56.88	54.75	49.88	**60.08**	57.29	56.15	48.16
FP1-FP2	57.12	56.5	54.75	53.5	57.54	55.89	55.64	52.6

Table 7.4: Pair comparision results on Coordination and Relative Clause. Results above 60% in blue bold, and below 40% in green underline. *TP1-FP1.* Person1 True, Person1, False. *TP2-FP2.* Person2 True, Person2, False. *TP2-FP1.* Person2 True, Person1, False. *TP1-FP2.* Person1 True, Person2, False. *TP1-TP2.* Person1 True, Person2, True. *FP1-FP2.* Person1 False, Person2, False.

7.2 Statistical Analysis

To explore what makes the models struggle in the BLA benchmark in a more formal way, we run statistical analyses to provide more insights into the factors that affect the model performance. Some visual and textual attributes of the BLA dataset are selected as predictors for a model's output score (the logit score of ViLBERT and LXMERT, the CLIPScore of CLIP). A higher score means the model predicts a higher alignment between the visual and text inputs. The analyses are conducted within the R statistical computing environment and the model is a Linear Model. If the p-value for a predictor is less than 0.01, we consider the predictor to have a significant relationship in predicting the model score. It is less interesting to report the specific p-values in our analysis, so we only report whether they are significant and the direction in which the predictor contributes to the prediction. The predictors used are explained below and the results are shown in Table 7.5 and Table 7.6, where "×" means the predictor is not significant, "-" and "+" represent the direction of influence by the significant predictors. The details on how to analyze statistical model results are described in Appendix A.

Active/passive voice In previous analyses, CLIP shows clear preferences for passive voice sentences on the Active-Passive task. This predictor is a categorical attribute specifically used in the Active-Passive task. We hypothesize that passive voice significantly affects CLIP score in a positive direction. But contrary to our hypothesis, CLIP predicts lower scores for sentences using passive voice than those using active voice. But this result is based on all sentences in the dataset, while results in Section 7.1.1 are based on sentences of the same image. Therefore, models predict higher alignment scores for active voice sentences in general, but it prefers the passive voice sentences of the same caption set in most cases.

Sentence length In the statistical analysis, we hypothesize that sentence length (the number of words in each sentence) is a significant predictor and affects scores in a positive direction. In the Active-Passive task, the hypothesis is supported by the results in predicting CLIP and LXMERT scores. When constructing the sentences, adjective words are added to distinct two nouns when the subject and object are the same words, so longer sentences provide more attributes of the nouns. It is not surprising that models predict higher alignment scores for such cases. Similar rules are applied to the Relative Clause and Coordination tasks as well, where adjective words are added to distinguish attributes in one image when the nouns are the same. The sentence length predictor is significant in predicting all model scores in the two

scores	ViLBERT	LXMERT	CLIP
len_sentence	×	+	+
voice/passive	×	×	-
gruen_score	×	-	-
person_size	+	+	+
avg_concretness	+	×	+
avg_freq	-	-	-
verb/feed	+	+	+
verb/help	+	+	+
verb/hold	+	×	-
verb/look at	-	×	-
verb/watch	+	×	×
verb/other	-	×	-

Table 7.5: Statistical analysis on Active-Passive task. Significant features have a p-value $<$ 0.01

	ViLBERT	LXMERT	CLIP
len_sen	+	+	+
gruen_score	-	-	-
person_size	+	+	+
attribute_size	+	+	+
avg_concretness	+	+	×
avg_freq	-	-	-

(a) Coordination

	ViLBERT	LXMERT	CLIP
len_sen	+	+	+
gruen_score	-	-	-
person_size	+	+	+
attribute_size	+	+	+
avg_concretness	+	+	×
avg_freq	-	-	-

(b) Relative Clause

Table 7.6: Statistical analysis result for Coordination and Relative Clause

tasks. However, in the Active-Passive task, passive sentences are always longer than the active sentences in the same caption set. Therefore, it could be the sentence length that causes the passive voice preference, but it is difficult to understand the reasons behind it.

Visual attributes We further investigate whether the size of objects mentioned in the sentence would be significant predictors. We hypothesize that larger objects are easier to recognize and more salient, so they should help models in the task. The visual attributes are *person_size* in both Active-Passive task as well as the *attribute_size* in the Coordination and Relative Clause tasks. The "size" refers to the object size ratio of the image. The *person_size* in the Active-Passive task is the average size of two person entities mentioned in the sentence. In the Coordination and Relative Clause tasks, the *person_size* is the size of the one single person entity mentioned in the sentence and the *attribute_size* is the average size of two mentioned attributes. As shown by the results, the larger object size leads to higher prediction scores for all models and all tasks.

Word concreteness Concrete words usually describe physical phenomena, relying on the five senses of sight, hearing, touch, taste, and smell. V&L models are trained to align such words to text, so we expect sentences with words that are more concrete will achieve higher alignment scores. The word concreteness ratings we use are collected by Brysbaert et al. (2014), which ranks words from abstract (1) to concrete (5). After removing stop words, we take the average of concreteness ratings for each word in the sentence and use it as a predictor (*avg_concreteness*). Except for the LXMERT, word concreteness plays a significant role in predicting scores. Higher

word concreteness leads to higher alignment scores between the two modalities.

Word frequency Words that appear less in the language (English) might contain more specific attributes that make it easier for matching text and images. For example, the word "policeman" has a lower word frequency than the word "man". It refers to the man wearing a policeman's uniform in the image and is more salient than what the word "man" refers to. To obtain the word frequency (*avg_freq*), we first remove the stop words in the sentence and compute the average logarithmic scales of word frequency for each word in the sentence. The word-frequency dictionary we used is taken from kaggle[1], which contains 333,333 most commonly-used single words on the English language web derived from the Google Web Trillion Word Corpus. For all models in the three tasks, lower word frequency leads to higher model scores.

Predicates Extending previous analysis on model performance regarding predicates (Section 7.1.3), we investigate whether the predicate verbs are significant in predicting scores in the Active-Passive task. We hypothesize that models predict lower scores for sentences with more challenging verbs. Challenging can mean that the verbs are less visually recognizable or cases with these verbs are more ambiguous. The results suggest that some predicates lead to higher scores while the others lead to lower scores or do not significantly affect it. All models predict higher scores for sentences using the verbs "feed" and "help". ViLBERT and CLIP predict lower scores for those using "look at" and other less frequent verbs. Verbs "Feed" and "Help" are more visual than the verb "look at". ViLBERT predicts higher scores for sentences with the verb "hold" while CLIP predicts lower scores for these cases. The CLIP's result is unexpected as the verb "hold" should be easy to recognize. But the verb "hold" could mean holding someone (usually a small child) in hands or holding someone with arms around. Perhaps CLIP, a model pretrained on a much larger dataset, has seen more ambiguous examples than the other model. And in this case, CLIP might be pretrained with more cases where "hold" is used in active voice, which might be related to CLIP's preference towards active voice sentences with "hold" (Section 7.1.3).

GRUEN scores GRUEN scores are used when filtering low-quality sentences in the dataset construction stage. A higher GRUEN score means the sentence is more grammatically correct. We expect models predict higher scores for high-quality sentences. However, the results suggest that lower GRUEN scores lead to higher alignment scores for all tasks and models except for ViLBERT in the Active-Passive task. This could be because after using a grammar filter, the sentences used in the benchmark are good enough in quality and should be easy to understand. Since the sentences are constructed automatically, the sentences with lower GRUEN scores might use expressions that are less common in human languages (e.g. "a shirtless man holds a man") but contain more attributes ("shirtless") that help with the alignment predictions.

7.3 Qualitative Analysis

We further dive into specific examples to analyze model performance on the BLA tasks. We mainly analyze some extreme cases where the model predictions are "all correct" (set accuracy of the image is 1) or "all wrong" (the set error rate of the image is 1).

[1] `https://www.kaggle.com/datasets/rtatman/english-word-frequency`

7.3.1 Active-Passive Voice

Extreme cases across models Among the 613 cases, there is only one case where all models make "all correct" predictions (Figure 7.1a). However, the image is very complex with many person entities appearing in the image. This case can be ambiguous and challenging. Using world knowledge, we can guess that "men" refers to two men on the right of the image and "children" refers to the children standing in line and watching them. Selecting the correct answers in this example is even very difficult for humans so models might have made this correct prediction based on some biases, for example, the object bounding box sizes of "men" and "children".

vilbert	lxmert	clip	caption
1	1	2	the children watch the men
2	2	1	the men are watched by the children
4	4	4	the men watch the children
3	3	3	the children are watched by the men

(a) all correct

vilbert	lxmert	clip	caption
4	4	4	the girl watches the boys
3	3	3	the boys are watched by the girl
1	1	2	the boys watch the girl
2	2	1	the girl is watched by the boys

(b) all wrong

vilbert	lxmert	clip	caption
3	4	3	the man observes the women
4	3	4	the women are observed by the man
1	1	1	the women observe the man
2	2	2	the man is observed by the women

(c) all wrong

vilbert	lxmert	clip	caption
4	3	4	the man carries the child
3	4	3	the child is carried by the man
1	2	2	the child carries the man
2	1	1	the man is carried by the child

(d) all wrong

Figure 7.1: Three models are "all correct" or "all wrong" on Active-Passive. For each set, the first 2 sentences are correct captions and the last 2 are distractors. The numbers under each model are the predicted ranking of the sentences. The correctly predicted ones are labeled in blue while the incorrect ones are labeled in green.

There are 3 cases where all models make completely wrong predictions (Figure 7.1 b - d). The images 7.1b and 7.1c are very challenging because the person entities' interactions are hard to identify. In the image 7.1b, it is not very clear where the girl is looking, but the girls stand behind the boys and they are watching something, so it should be more correct for "girls are watching boys" than "boys are watching girls". In the image 7.1c, the women are cutting something while the man seems to observe how they work. So it should be more likely that "the man observes the women" instead of "the women observe the man". However, the person entity's interaction (man holding a child) is very clear in the image 7.1d and we expect models

to make correct predictions. But again models might align the word "child" to the child holding a frisbee in the foreground since the size of this person entity is larger.

There aren't any cases where one model makes correct predictions on the set while the others make totally wrong predictions. So we select examples for one specific model at a time and analyze when the model makes "all correct" predictions and when it makes "all wrong" predictions. In the meanwhile, we check how the other two models perform in these cases.

Extreme cases for individual model By checking the "all correct" cases and "all wrong" cases, we found overall the "all correct" cases (Figure 7.2 (a)(b)) have more clear and less ambiguous person entities' interactions than the "all wrong" cases (Figure 7.2 (d)(e)). However, there are some exceptions (Figure 7.2 (c) and Figure 7.2 (f)) that models make "all correct" predictions when the case is challenging and models make "all wrong" predictions in caption sets that look intuitively easier than others.

vilbert	lxmert	clip	caption
2	2	3	the boy helps the lady
1	1	2	the lady is helped by the boy
3	4	4	the lady helps the boy
4	3	1	the boy is helped by the lady

(a) all correct

vilbert	lxmert	clip	caption
2	2	1	the woman carries the girl
1	3	2	the girl is carried by the woman
4	4	4	the girl carries the woman
3	1	3	the woman is carried by the girl

(b) all correct

vilbert	lxmert	clip	caption
2	1	3	the spectator watches the skateboarders
1	2	2	the skateboarders are watched by the spectator
4	3	4	the skateboarders watch the spectator
3	4	1	the spectator is watched by the skateboarders

(c) all correct

vilbert	lxmert	clip	caption
4	4	2	the man embraces the woman
3	2	3	the woman is embraced by the man
1	3	4	the woman embraces the man
2	1	1	the man is embraced by the woman

(d) all wrong

vilbert	lxmert	clip	caption
4	4	4	the spectators watch the player
3	3	1	the player is watched by the spectators
1	1	3	the player watches the spectators
2	2	2	the spectators are watched by the player

(e) all wrong

vilbert	lxmert	clip	caption
4	4	3	the woman touches the man
3	1	4	the man is touched by the woman
2	3	1	the man touches the woman
1	2	2	the woman is touched by the man

(f) all wrong

Figure 7.2: At least one model is "all correct" or "all wrong" on the Active-Passive task. For each set, the first 2 sentences are correct captions and the last 2 are distractors. The numbers under each model are the predicted ranking of the sentences. The correctly predicted ones are labeled in blue while the incorrect ones are labeled in green.

7.3.2 Coordination and Relative Clause

For both Coordination and Relative Clause tasks, none of the three models share "all correct" or "all wrong" cases, which indicates that the three models do not agree with each other in the

predictions for such extreme cases. We analyze the extreme cases for individual models.

Models perform worse on complex images By checking extreme cases, we found that models usually make "all wrong" predictions when the image scenes are more "complex". These images usually have multiple person entities in the image or a very noisy background. In the "all correct" cases, the person entities are usually very clear, which means their size is not too small and their attributes are obvious, and the background is relatively clean. This suggests that models might need better person entity attribute extractors. But there are some cases where models fail in relatively simple scenes (Figure 7.4 a). This suggests that the model still lacks the ability to understand and align fine-grained person entity attributes.

vilbert	lxmert	clip	caption
1	1	2	the man wears a brown shirt and wears glasses
2	4	3	the woman wears a purple shirt and wears a chain
3	2	4	the man wears a brown shirt and wears a chain
4	3	1	the woman wears glasses and wears a purple shirt

(a) Coordination

vilbert	lxmert	clip	caption
1	2	1	the skateboarder wears a white t-shirt and wears black pants
4	1	3	the skateboarder wears a maroon shirt and wears blue jeans
3	4	2	the skateboarder wears a white t-shirt and wears a maroon shirt
2	3	4	the skateboarder wears black pants and wears blue jeans

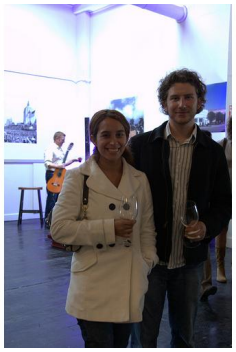
(b) Coordination

vilbert	lxmert	clip	caption
4	1	2	the man wears a gray hat and wears a yellow shirt
1	4	1	the lady wears a white shirt and wears a blue top
3	3	4	the man wears a gray hat and wears a white shirt
2	2	3	the lady wears a yellow shirt and wears a blue top

(c) Coordination

vilbert	lxmert	clip	caption
2	1	1	the man who has a grey shirt has blue shoes
1	3	4	the woman who has a brown wedge has a black jacket
4	2	2	the man who has a grey shirt has a brown wedge
3	4	3	the woman who has blue shoes has a black jacket

(d) Relative Clause



vilbert	lxmert	clip	caption
4	1	3	the man who wears a black jacket wears blue jeans
1	2	2	the woman who wears a white coat wears jeans
2	4	1	the man who wears a black jacket wears a white coat
3	3	4	the woman who wears blue jeans wears jeans

(e) Relative Clause

vilbert	lxmert	clip	caption
3	2	2	the man who sits in a chair wears a black shirt
2	3	1	the man who sits on a couch wears a hat
1	1	3	the man who sits in a chair sits on a couch
4	4	4	the man who wears a black shirt wears a hat

(f) Relative Clause

Figure 7.3: Examples of "all correct" cases on Coordination and Relative Clause. For each set, the first 2 sentences are correct captions and the last 2 are distractors. The numbers under each model are the predicted ranking of the sentences. The correctly predicted ones are labeled in blue while the incorrect ones are labeled in green.

Difference in person entity attributes sizes We notice that CLIP is more likely to make correct predictions when no person entity attribute is very small or much less obvious in the sentences. For example, CLIP makes a correct prediction in Figure 7.3 c as all the attributes that exist in the sentence are very obvious and of relatively similar sizes but it makes mistakes in 7.3 when the attribute necklace is very small compared with the other attributes. But LXMERT and ViLBERT are more acceptable for such cases and can still make correct predictions when the difference between attribute sizes is very large.

vilbert	lxmert	clip	caption
3	3	4	the girl has a purple backpack and has blue jeans
4	2	1	the girl has a green shirt and rides a walking horse
2	4	3	the girl has a purple backpack and has a green shirt
1	1	2	the girl has blue jeans and rides a walking horse

(a) Coordination

vilbert	lxmert	clip	caption
4	4	3	the man wears a brown shirt and wears a green tie
1	3	4	the man wears a blue tie and wears a grey shirt
2	1	1	the man wears a brown shirt and wears a blue tie
3	2	2	the man wears a green tie and wears a grey shirt

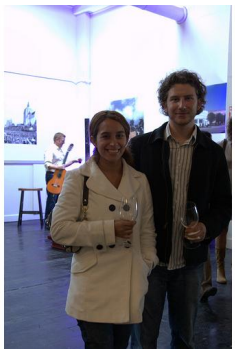
(b) Coordination

vilbert	lxmert	clip	caption
2	3	3	the man falls in grass and wears a black shirt
3	2	4	the man runs on grass and wears a pink shirt
4	4	1	the man falls in grass and wears a pink shirt
1	1	2	the man wears a black shirt and runs on grass

(c) Coordination

vilbert	lxmert	clip	caption
2	1	1	the man who has a grey shirt has blue shoes
1	3	4	the woman who has a brown wedge has a black jacket
4	2	2	the man who has a grey shirt has a brown wedge
3	4	3	the woman who has blue shoes has a black jacket

(d) Relative Clause



vilbert	lxmert	clip	caption
4	1	3	the man who wears a black jacket wears blue jeans
1	2	2	the woman who wears a white coat wears jeans
2	4	1	the man who wears a black jacket wears a white coat
3	3	4	the woman who wears blue jeans wears jeans

(e) Relative Clause

vilbert	lxmert	clip	caption
3	2	2	the man who sits in a chair wears a black shirt
2	3	1	the man who sits on a couch wears a hat
1	1	3	the man who sits in a chair sits on a couch
4	4	4	the man who wears a black shirt wears a hat

(f) Relative Clause

Figure 7.4: Examples of "all wrong" cases on Coordination and Relative Clause. For each set, the first 2 sentences are correct captions and the last 2 are distractors. The numbers under each model are the predicted ranking of the sentences. The correctly predicted ones are labeled in blue while the incorrect ones are labeled in green.

With the qualitative analysis, we found that the models can understand the sentence semantics to some extent and identify the scene of person entities' interactions and person entities with their attributes, but their performance is still below par and more work needs to be done to further investigate why they perform in such a way and how to improve them.

Chapter 8

Conclusions

In this thesis, we propose a novel benchmark BLA, which evaluates the basic language abilities that vision-and-language models obtained from pretraining. The benchmark contains three tasks that are designed to evaluate whether V&L models understand three linguistic phenomena and general sentence-level semantics on "who does what to whom" and "who has what attributes". We construct the dataset on top of VisualGenome and propose an automatic pipeline that can generate relatively high-quality datapoints. We experiment on the state-of-the-art models CLIP, ViLBERT and LXMERT and collect human annotations as a comparison. The experiment results suggest that there is still a huge performance gap between the models and humans, where models perform no better than the chance level in all our tasks. This contradicts our hypothesis that models performing well on various complex V&L tasks should have learned good visually grounded language, but our results are consistent with previous findings that models perform at around chance level in a task using similar settings as ours (Thrush et al., 2022). The following sections discuss how good the models are on basic language abilities and why this is the case. With the findings of this thesis work, we hope that our task and dataset will provide useful insights and help guide research in building better V&L models.

8.1 Model Capabilities and Limitations

Combining the findings from previous work, we further discuss how good the models are in understanding semantics on different levels.

Word-level semantics The most common categories of words are nouns, verbs and adjectives. FOIL task is almost solved by CLIP, which demonstrates the model's good understanding of nouns. Given the contrastive pretraining goal of CLIP, it is unsurprising that the model can detect what object presents in the image and what is not. The model learns good alignment between the noun words and image object attributes in pretraining. The task SVO-Probes (Hendricks and Nematzadeh, 2021) focus on verb understanding, which finds that models fail at identifying the differences in verbs in their task. They point out that models have trouble ignoring noise in the automatically-curated pretraining datasets, supported by the fact that a model can achieve better performance when training on a smaller but cleaner dataset. This is also reflected in our experiment results where models find some verbs more challenging than others. Our Coordination and Relative Clause tasks require distinguishing adjectives (e.g. "red" vs. "black") in some cases. Few existing works focus on the adjective understanding of vision-and-language models. Salin et al. (2022) finds that concepts related to object size are not well understood by current SOTA V&L models which makes them struggle with understanding size adjectives (i.e. "large" vs. "small"). However, size adjectives are seldom used to distinguish entities in our tasks.

Sentence-level semantics Instead of focusing on single words, sentence-level semantics require the model to leverage relationships between words. Our task is set up to investigate the general understanding of sentence semantics regarding "who does what to whom" and "who has what attributes". In the Active-Passive task, the pair comparison results between TA (True Active) vs. FA (False Active) and TP (True Passive) vs. FP (False Passive) suggest that models have difficulty understanding the performer and receiver of an action. This is in line with the results in the Action task in Parcalabescu et al. (2022). Recent papers Cirik et al. (2018); Akula et al. (2020) point out that transformer models are insensitive to word orders. This could explain why models have trouble with tasks requiring them to distinguish between sentences like "the man hits the woman" and "the woman hits the man". Models also struggle to match what attribute belongs to which person in the Coordination and Relative Clause tasks. The models appear to approach them as object recognition tasks, which means they simply focus on how well the words in the sentence match the objects in the image while ignoring the sentence's semantics. This indicates models still need to have a more fine-grained understanding of the interaction between the two modalities.

Insigts on better model As Sinha et al. (2021) has proposed some methods for word order sensitivity, we discuss how to make models better in understanding sentence-level semantics. Inspired by children's language learning process, Nikolaus and Fourtassi (2021) uses similar tasks as ours to evaluate vision-and-language models on their acquisition of semantic knowledge from cross-situational Learning. The models they use are trained from scratches on cartoon storybook datasets where the training set contains one image, one sentence aligned to the image and one misaligned sentence (similar setting as our task). The training loss of the model encourages aligned image-sentence pairs to have a higher similarity score than misaligned pairs. The model is proven to show good performance on the evaluation tasks. Though the scene of the dataset is relatively simple, the idea of applying contrastive learning at the sentence level ("a man holds a baby" vs. "a baby holds a man") instead of the word level ("a photo of a dog" vs."a photo of a cat") only could be a direction to further improve current models on their sentence-level semantic understanding.

8.2 Future Work and Broader Impact

We proposed a framework to evaluate the pretrained model's language abilities and reasoning abilities separately. In this thesis, we started from the linguistic tasks while future work can focus on the reasoning tasks. One idea is to use a similar setting of one sentence with four sentences in each test trial. The reasoning ability can be simple positional reasoning, which is to list objects in the image from left to right order and from right to left order. Since the model have been proven to perform well in noun understanding, this task would eminently challenge the models on their reasoning abilities.

In addition, the language benchmark can be further extended and improved. The Children's book contains tasks regarding 7 phenomena while our BLA dataset only evaluates three of them. More tasks focusing on other linguistic phenomena can be proposed to have a more thorough understanding of models' linguistic abilities. In addition, since the dataset is constructed automatically, the language is relatively simple and the attributes or predicates are less diverse. We hope future work can leverage crowdsourcing to construct more diverse and more high-quality cases for these tasks, which can be used to conduct more in-depth analyses..

Appendix A

Statistical Models

In the regression equation, for predicting an outcome variable (y) on the basis of a predictor variable (x). When the predictor variable x is a numerical number, the equation can be written as y = b0 + b1 * x, where b0 and b1 are the regression beta coefficients, representing the intercept and the slope respectively. The original predictions results are shown in the following figures. Since the coefficient values are less interesting in the analysis, we only provide the direction of the values in the results. + and - indicates the direction in which the feature contributes to the prediction. For example, the *len_sentence* contributes to the LXMERT output scores in a positive way, which means the LXMERT output score would be higher if the sentence is longer.

But for the categorical features (e.g. active or passive), the equation is y = b0 + b1 for the category it outputs. Therefore the + and - does not mean the direction to which the feature contributes to the prediction but provide a comparison between the categorical values in predicting the higher output. For example, for the passive or active features, b0 is a value greater than 0 for active, and b1 is a value less than 0 for passive. That indicates that the feature "passive" would lead to a lower model score than the "active" (y_passive ¡ y_active).

Appendix B

Collection of Human Annotations

After running the first pilot with 100 rows, we found the crowedworkers did not perform as well as we expected. We manually checked test cases where crowdworkers made mistakes and found that they could have performed better in many cases. For example, in Figure B.1, the woman is holding the red tablet but most crowdworkers might ignore this detail for some reason. Therefore, we modified the instructions of all tasks by adding examples that are more close to the real task by providing examples that require to pay attention to the details, which is explained in the judgement reasons. The two versions of instruction are shown in Figure B.2.

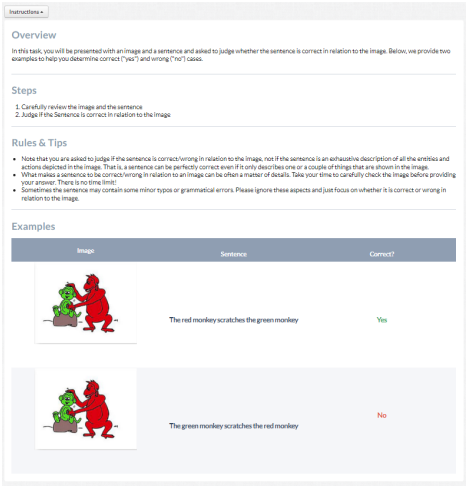
After changing the instructions, we launched the second pilot for another 100 rows. But unexpectedly, the crowdworkers performed even worse with the more detailed instructions. We found that the crowdworkers provided the answer "yes" for majority of the questions. This suggests that the annotators are more lenient to cases that the image and the sentence do not match, which is also reported in Thrush et al. (2022). In addition, the crowdworker performance vary a lot between two pilots as shown in Table B.1. Therefore, the performance of the crowdworker annotations we collected are not reliable. But due to time and budget constrains, we stopped adjusting the tasks and launching the third pilot to collect more data.

		Active-Passive	Coordination	Relative Clause
Version 1	size of cases	100	100	100
	"yes" vs. "no"	77 : 23	57 : 43	58 : 42
	accuracy	0.61	0.54	0.71
Version 2	size of cases	100	100	173
	"yes" vs. "no"	96 : 4	59 : 14	116 : 57
	accuracy	0.48	0.61	0.613
Total	size of cases	200	200	273
	"yes" vs. "no"	173 : 27	116 : 57	174 : 99
	accuracy	54.5	57.5	64.8

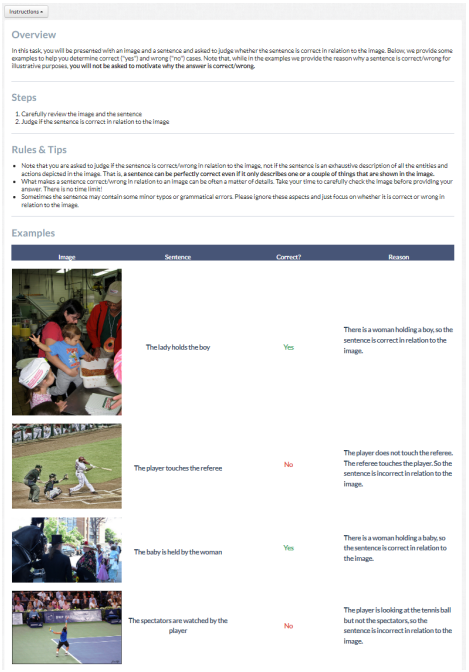
Table B.1: Crowdworker performance collected in two pilots. The "yes" and "no" represents the distribution of the majority votes in the launched questions.

sentence: the woman who holds a pen holds a red tablet
golden label: yes
size of crowdworkers selected "yes": 2
size of crowdworkers selected "no": 3

Figure B.1: Example of test case that crowdworkers made mistakes.



(a) Version 1



(b) Version 2

Figure B.2: Two versions of instructions. In the Version 1, we use the cartoon picture to explain our task, which is clear but looks very different from the real task. In the Version, we use examples from the BLA dataset that are not used in the annotation task or the Test questions. We also provide reasons explaining the reason for the answer choice.

Bibliography

Arjun Akula, Spandana Gella, Yaser Al-Onaizan, Song-chun Zhu, and Siva Reddy. 2020. Words aren't enough, their order matters: On the robustness of grounding visual referring expressions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6555–6565.

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.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior research methods*, 46(3):904–911.

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 *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX*, pages 104–120.

Kenneth Ward Church. 2017. Word2vec. *Natural Language Engineering*, 23(1):155–162.

Volkan Cirik, Louis-Philippe Morency, and Taylor Berg-Kirkpatrick. 2018. Visual referring expression recognition: What do systems actually learn? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 781–787.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

Radina Dobreva and Frank Keller. 2021. Investigating negation in pre-trained vision-and-language models. In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 350–362.

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

Evelina Fedorenko and Rosemary A. Varley. 2016. Language and thought are not the same thing: evidence from neuroimaging and neurological patients. *Annals of the New York Academy of Sciences*, 1369.

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.

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

Lisa Anne Hendricks and Aida Nematzadeh. 2021. Probing image-language transformers for verb understanding. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3635–3644.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528.

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.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. 2014. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73.

Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. 2020a. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11336–11344.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.

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

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

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32.

Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 2020. 12-in-1: Multi-task vision and language representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10437–10446.

Kyle Mahowald, Evgeniia Diachek, Edward Gibson, Evelina Fedorenko, and Richard Futrell. 2022. Grammatical cues are largely, but not completely, redundant with word meanings in natural language. *arXiv preprint arXiv:2201.12911*.

George A. Miller. 1995. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41.

Mitja Nikolaus and Abdellah Fourtassi. 2021. Evaluating the acquisition of semantic knowledge from cross-situational learning in artificial neural networks. In *Workshop on Cognitive Modeling and Computational Linguistics*, pages 200–210. Association for Computational Linguistics.

Letitia Parcalabescu, Michele Cafagna, Lilitta Muradjan, Anette Frank, Iacer Calixto, and Albert Gatt. 2021. Valse: A task-independent benchmark for vision and language models centered on linguistic phenomena. *arXiv preprint arXiv:2112.07566*.

Letitia Parcalabescu, Michele Cafagna, Lilitta Muradjan, Anette Frank, Iacer Calixto, and Albert Gatt. 2022. Valse: A task-independent benchmark for vision and language models centered on linguistic phenomena. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8253–8280.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? 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 2463–2473.

Manuela Pinto and Shalom Zuckerman. 2019. Coloring book: A new method for testing language comprehension. *Behavior research methods*, 51(6):2609–2628.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

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.

Emmanuelle Salin, Badreddine Farah, Stéphane Ayache, and Benoit Favre. 2022. Are vision-language transformers learning multimodal representations? a probing perspective. In *AAAI 2022*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *54th Annual Meeting of the Association for Computational Linguistics*, pages 1715–1725. Association for Computational Linguistics (ACL).

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

Ravi Shekhar, Sandro Pezzelle, Yauhen Klimovich, Aurelie Herbelot, Moin Nabi, Enver Sangineto, and Raffaella Bernardi. 2017. Foil it! find one mismatch between image and language caption. In *ACL*, pages 255–265. Association for Computational Linguistics.

Ravi Shekhar, Ece Takmaz, Raquel Fernández, and Raffaella Bernardi. 2019. Evaluating the representational hub of language and vision models. In *Proceedings of the 13th International Conference on Computational Semantics-Long Papers*, pages 211–222.

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Koustuv Sinha, Robin Jia, Dieuwke Hupkes, Joelle Pineau, Adina Williams, and Douwe Kiela. 2021. Masked language modeling and the distributional hypothesis: Order word matters pre-training for little. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2888–2913.

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, 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.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. 2022. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5238–5248.

Chenguang Wang, Xiao Liu, and Dawn Song. 2020. Language models are open knowledge graphs. *arXiv preprint arXiv:2010.11967*.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Dan Yi. 2017. Teaching relative clause in secondary school english classroom. *International Journal of Liberal Arts and Social Science*, 5(5):1–6.

Dani Yogatama, Cyprien de Masson d'Autume, Jerome Connor, Tomas Kocisky, Mike Chrzanowski, Lingpeng Kong, Angeliki Lazaridou, Wang Ling, Lei Yu, Chris Dyer, et al. 2019. Learning and evaluating general linguistic intelligence. *arXiv preprint arXiv:1901.11373*.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6720–6731.

Wanzheng Zhu and Suma Bhat. 2020a. Gruen for evaluating linguistic quality of generated text. In *Findings of the Association for Computational Linguistics, ACL 2020: EMNLP 2020*, pages 94–108. Association for Computational Linguistics (ACL).

Wanzheng Zhu and Suma Bhat. 2020b. GRUEN for evaluating linguistic quality of generated text. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 94–108, Online. Association for Computational Linguistics.

Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. 2016. Visual7w: Grounded question answering in images. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4995–5004. IEEE.