DEPARTMENT OF PHILOSOPHY,
LINGUISTICS AND THEORY OF SCIENCE

ARE LANGUAGE-AND-VISION TRANSFORMERS SENSITIVE TO DISCOURSE?

The case study of ViLBERT

Ekaterina Voloshina

Master's Thesis:	15 credits
Programme:	Master's Programme in Language Technology
Level:	Advanced level
Semester and year:	Spring, 2023
Supervisor:	Simon Dobnik and Nikolai Ilinykh
Examiner:	Eleni Gregoromichelaki
Keywords:	transformers, discourse, multimodality, evaluation, grounding, word representations

Abstract

Language-and-vision models have shown good performance on image-caption matching, caption generation and other tasks. At the same time, they are known for generating pragmatically incorrect captions that are not grounded in the image. To explore the reasons behind such behaviour, evaluation is required.

In this thesis, we explore to what extent language-and-vision models are sensitive to different discourse structures, such as short or long discourse structures and to what extent factors, such as modality, the similarity of captions or images, affect models' predictions. We introduce new tasks to evaluate language-and-vision models on discourse understanding. We test the ViLBERT model on these tasks and show that the model can distinguish different situations but it is not sensitive to differences within one narrative structure. We also show that performance depends on the way the tasks are constructed, for example, what modality remains unchanged in non-matching pairs or how similar non-matching pairs are to original pairs.

Acknowledgements

I could not fully express my gratefulness to my supervisors, Simon Dobnik and Nikolai Ilinykh, for their support throughout the entire process of writing this thesis, for discussing every step and guiding my sometimes chaotic ideas to more fruitful directions. You have been of a great example of academic work in process, and I have never been that excited about working on a project before.

To the members of CLASP research group for organising all the seminars and making me feel welcome there, despite being a student. You all have proved that although there are so many more questions to answer and problems to solve, research is a collaborative process, and the right answers could be only born in discussions.

To my friends back home who support me no matter what decisions I make and how much I annoy them talking about linguistics all the time. Thank you for being my light in dark times for so many years.

To my friends in Gothenburg for making my time here one of the greatest experiences in my life. I have learnt a lot from all of you from insights in linguistics to how to use a laser cutter for making popcorn.

To Jon for always being there for me.

Special thanks to Anton, Katya and Johnny for providing helpful comments on the thesis.

Last but not the least, to my family and especially to my mom for supporting me and encouraging my curiousity. It is the curiousity that brought me here and made it all possible.

P.S. This is the preface

Contents

1 Introduction

Large language models (LLMs) have shown significant performance on different tasks related to Natural Language Processing (NLP) due to their self-attention mechanism (Vaswani et al., 2017). As the results on different benchmarks show (Rajpurkar et al., 2018; Wang et al., 2019; Gehrmann et al., 2022), LLMs perform well on tasks such as question answering, common sense reasoning, or text generation. However, these models are often considered to be 'black boxes', as due to the complexity of computations, it is hard to explain which features of input data affect their predictions. Recently, the stream of papers on *interpretability* of large language models started examining such models for the presence of linguistic knowledge (Rogers et al., 2020).

To explain models' predictions, it is essential to evaluate such models on tasks different from downstream applications and estimate if they have acquired necessary skills, including language skills, during the pre-training process (Liu et al., 2019; Belinkov, 2022; Elazar et al., 2021). Such evaluation tasks can take different forms, for example, language modelling (Warstadt et al., 2020) or classification problems (Conneau et al., 2018). Moreover, model performance on such linguistic tasks is believed to correlate with downstream applications (Saphra, 2021). From a practical point of view, such evaluation identifies a lack of knowledge in pre-trained models and allows us to improve performance by introducing modifications to architecture or datasets. Besides that, such evaluation techniques highlight the differences in how computational algorithms acquire different types of linguistic data (Clark & Lappin, 2010).

When it comes to the evaluation of LLMs on different linguistically motivated tasks, most of the research is focused on morphology and syntax (Conneau et al., 2018; Warstadt et al., 2020; Taktasheva et al., 2021; Stańczak et al., 2022; Maudslay & Cotterell, 2021; Lasri et al., 2022) while *discourse* is left behind. We understand discourse as any type of meaning that is created in the situation of communication and connected to social and cultural contexts. Discourse can be **short** (on a level of one item, such as an utterance or an image) or **long** (on a level of several items). Discourse can be understood as a combination of three different components:

- **Textual**: textual discourse is a level of language that operates with linguistic relations and dependencies that exist between linguistic units across words in a single utterance. Co-reference could serve as an example of such relations.
- **Visual**: images capture only part of the reality and, therefore, they are focused on specific actors and events among all that are visually available. This is similar to the phenomenon of *centering* described by (Grosz et al., 1995). In textual discourse, some actors are placed in the centre of the narrative and discourse coherence is built around references to the centered object.
- **Situation-level**: situation-level discourse operates on a level of bigger structures, i.e. how separate utterances are united into a coherent and cohesive narrative. Understanding situation-level discourse requires world knowledge and awareness of the social context and unites textual and visual discourse, since it operates not only on the level of text or images but also on the level of social situation. As Fairclough (1992) notices, discourse is grounded in the social and cultural background of speakers, the so-called *common ground* (Clark, 1996).

Several researchers have shown (Ettinger, 2020; Nie et al., 2019) that language models are known to struggle with textual discourse. As an example, language models struggle to generate pragmatically coherent sentences, as they do not take previous context into account. This issue might indicate that models are insensitive to discourse relations and are not able to capture them. For example, Ettinger (2020) provides

Figure 1: Pragmatically incorrect captions. The examples are taken from (Lake et al., 2017)

the outputs of a BERT model on a pragmatics dataset (see Example 1, models' top predictions are given in italics):

(1) The snow had piled up on the drive so high that they couldn't get the car out. When Albert woke up, his father handed him a *note/letter/gun/blanket/newspaper*.

Although most of the predictions are correct in the context of one sentence, they become incoherent when a broader context is taken into account. The models seem to lack contextual knowledge, as language data is not sufficient (Bender & Koller, 2020). However, a multi-modal setup gives models more data to acquire world knowledge or context about a situation, as language-and-vision models have an advantage of visual information that they process simultaneously with language data. Language models cannot capture situation-level discourse; we need to use models that can generalise over additional modalities beyond textual to be able to capture narrative relations.

As for language-and-vision models, they have shown good performance on tasks that involve visual grounding (Bugliarello et al., 2022). The research on discourse in multi-modal models, however, has just begun, and they have seldom been tested on discourse knowledge. As Ilinykh & Dobnik (2022c) claim, language-and-vision transformers are biased towards more frequent nouns and lack contextual knowledge while predicting image-specific descriptions. Consequently, although they generate grammatically and semantically correct captions (for example see Figure 1), they fail to capture pragmatics, common reasoning and understand basic physics and relations.

The reason why models fail to properly describe images is because they struggle to capture situation-level discourse. Discourse is meant to explain why some phrases are used in some situations but are not acceptable in others where their synonymous sentences are used. Taking an example from the textual modality, the following sentences could describe the same situation:

(2) a. One student came to the class.

b. It is not true that no student came to the class.

However, they differ in the way they could be used:

(3) a. One student came to the class. He answered all the questions.

b. *It is not true that no student came to the class. He answered all the questions.

This could be explained by different factors, such as narrative structure, context of a situation or discourse relations. Discourse relations could be expressed implicitly (with the phenomenon of *bridging anaphora*

Figure 2: **Description-in-isolation.** A black frisbee is sitting on top of a roof. A man playing soccer outside of a white house with a red door. The boy is throwing a soccer ball by the red door. A soccer ball is over a roof by a frisbee in a rain gutter. Two balls and a frisbee are on top of a roof.
Story-in-sequence. *A discus* got stuck up on the roof. Why not try getting it down with *a soccer ball*? Up *the soccer ball* goes. It didn't work so we tried *a volley ball*. Now *the discus, soccer ball,* and *volleyball* are all stuck on the roof.

(Clark, 1977), see Example 4) or explicitly (for example, with discourse connectors, see Example 5):

(4) John is going to France. He likes opera.

In the example above, discourse relations are implicit. Since a speaker puts together these sentences, they might be connected, i.e. going to an opera is a reason to go to France. This could be expressed explicitly:

(5) John is going to France **because** he likes opera.

In this work, our **primary goal** is to evaluate how multi-modal models can understand situations across all the three components of discourse mentioned earlier. As has been shown (Zhang et al., 2022), such models acquire more knowledge from combining textual and visual modalities than language models that lack non-linguistic knowledge.

We evaluate a model's sensitivity to three components of discourse. As for the data, we use Visual Storytelling dataset (VIST) (Huang et al., 2016), as it includes all necessary types of data required for our experiments. Specifically, VIST consists of stories based on sequences of pictures. It includes two different levels of annotation: *descriptions-in-isolation* and *stories-in-sequence* (see Figure 2). While *descriptions-in-isolation* were collected following standard instructions of MS-COCO (Lin et al., 2014) (for example, annotators have to name all important parts), for *stories-in-sequence* crowd workers were asked to write a story about a sequence of events (see Figure 2). *Stories-in-sequence* do not tend to name objects on images but rather connect an image to its previous context.

Moreover, this dataset allows us to permute data in several ways, across stories and within one story. We therefore adapt a standard image-caption matching task to our experiments, but we create different perturbations, i. e. non-matching pairs that we call **distractors**[1].

Hence, we formulate our research questions as follows:

I Can a language-and-vision model perform *visual grounding of descriptive captions* when non-matching images or captions are taken from different situations?

II Can a language-and-vision model perform *story visual grounding*, i. e. ground captions that are parts of a narrative but there is a mismatch between images or captions and the situation?

III Can a language-and-vision model understand the narrative structure, i. e. the model can say if two parts of the same story form a coherent whole?

[1] We define the word *distractor* in a different way than the way it is used in the field of psycholinguistics.

To address these research questions, we conduct three experiments (Table 1). For the first two experiments, we take captions and images from different albums and regard them as examples of short discourse isolated from a larger narrative. The last experiment aims to access whether the models are sensitive to differences inside a narrative: in other words, this task is meant to shed light on models' sensitivity to fine-grained pragmatic differences, such as distinguishing discourse units within one narrative structure.

Captions	**Short discourse**	**Long discourse**
Descriptions-in-isolation	Experiment I	-
Stories-in-sequence	Experiment II	Experiment III

Table 1: The summarisation of the experiments by different parameters.
Short discourse operates on the level of one item, such as an utterance or an image. Long discourse functions on the level of several items.
Based on the annotation of the VIST dataset, we take two levels of annotations: *descriptions-in-isolation* were collected with the standard procedure of image-caption annotation, *stories-in-sequence* are the part of a narrative.

The rest of the thesis is structured as following: Section 2 covers background necessary for further understanding of the research questions and gives an overview of previous work on this topic; in Section 3, we describe the methods used in the research, including the dataset, models, and an experimental setup; Section 4 and 5 include the description and discussion of results respectively.

2 Background

This chapter explains the theoretical background relating to grounded language-and-vision models and discourse structure. It starts with an explanation of transformers and the concept of self-attention (Section 2.1), then proceeds to descriptions of word embeddings and BERT-like models (Section 2.2). The final subsection (Section 2.3) dives deeper into language-and-vision models and their knowledge of discourse.

2.1 Transformers

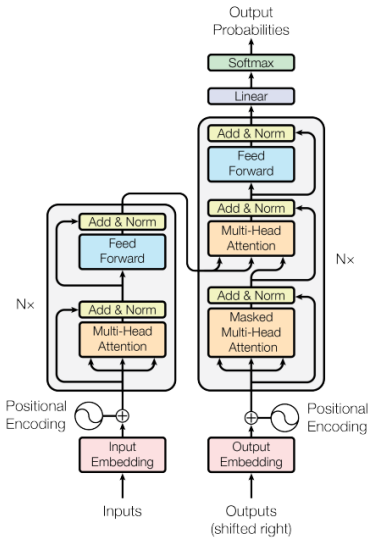


Figure 3: Transformer architecture from (Vaswani et al., 2017). A model takes a sequence as an input, then the data goes through an encoder stack, and is transformed into hidden representations. These representations get through the decoder stack to produce another sequence as an output.

The transformer architecture (Vaswani et al., 2017) consists of two blocks: an encoder and a decoder (Figure 3). The encoder generates vector representations of input sequences that are later transformed to output by the decoder block. Originally, transformers were introduced for the machine translation task: in this task, the encoder will take sentences in a source language, for example English, and transform them to vectors. The decoder will generate sentences in a target language (e.g. French) from these representations (Vaswani et al., 2017). In the original architecture, the encoder includes six layers where one layer consists of a multi-head attention sub-layer and a linear sub-layer. The decoder also consists of six layers, but each layer includes sub-layers: multi-head attention over a vector representation from an encoder stack. Both the encoder and the decoder include encoding of a token's position in a sentence.

The main benefit of this architecture comes from multi-head attention mechanism. As an input, an attention layer has three vector projections: query (Q), key (K), and value (V). The concept of keys, values and queries is taken from information retrieval, where a system matches queries with keys of search results and presents candidates, i.e. values. When predicting a word, a query represents a current token and it is matched with a set of keys to find the best candidates for a next word. Keys and values in transformers are linear transformations of a vector representation of a token to be predicted, and queries are linear transformations of a current token, except for self-attention, where queries are also taken from the same token. The formula of the attention mechanism where d stands for dimension is given below:

$$Attention(Q, K, V) = softmax(\frac{\mathbf{QK}^T}{\sqrt{d_k}})\mathbf{V}. \quad (1)$$

The output of this layer is the weighted sum of the resulting values. However, to make models generalise better and encode more information, several attention heads are calculated simultaneously. The multi-head attention mechanism combines different representations at different positions by concatenation:

$$MultiHead(Q, K, V) = Concat(head_1, ..., head_h)W^O \quad (2)$$
$$where\ head_i = Attention(QW_i{}^Q, KW_i{}^K, VW_i{}^V), \quad (3)$$

where W are parameter matrices for linear projections of Q, K, and V.

Attention is implemented in three different ways in transformer models. First, in encoder-decoder layers, queries come from a previous decoder layer while keys and values are taken from an encoder block. Second, the encoder stack contains self-attention layers. In these layers all queries, keys and values come from the previous layer of the encoder. Each position has access to all positions from a previous layer. The idea of a self-attention layer is to reflect the context by replacing each element by weighted average of the other elements in the sequence. Third, in the decoder, a multi-head attention sub-layer masks forthcoming context with zeros so the model cannot use advantage of seeing future tokens while generating text.

The models based on the transformer architecture could be of different types: sequence-to-sequence (seq2seq), such as GPT-3 (Brown et al., 2020), encoder-only, such as BERT (Devlin et al., 2018), and decoder-only, such as GPT-2 (Radford et al., 2019). Since a decoder stack is responsible for generation of a new sequence, the seq2seq and decoder-only models are used for text generation, while encoder-only models are made for classification tasks. However, encoder-only models, such as BERT, are used to extract word representations which can be used for different types of text classification, e.g. sentiment analysis. Since our experiments are formulated as a binary classification task, we focus on BERT-like models.

2.2 Language models

To be used in different computational applications, language should be represented in a meaningful way. The most popular way to get such language representations is to extract them from neural networks that are able to learn from large amount of data. Neural networks iteratively learn meanings of words based on the context the words occur in. This idea goes back to distributional semantics that postulates that meaning of a word can be represented as the set of its co-occurrences with other words. Since neural networks can work with large amount of data, they can produce generalised word representations learnt from the data. The word embeddings extracted from hidden layers of models contain different type of information, including data on word morphology, syntactic relations, and world knowledge (Rogers et al., 2020; Conneau et al., 2018).

There are several ways of creating word embeddings, such as Word2Vec (Mikolov et al., 2013) or GloVE (Pennington et al., 2014). However, the most popular model for word representations is BERT. Devlin et al. (2018) introduced an architecture based on Transformer encoders with bidirectional self-attention heads. The novelty of BERT comes from itstwo pre-training objectives: Masked Language Modelling (MLM) and

Next Sentence Prediction (NSP).

The first task is designed to make models learn contextualised meanings of tokens. In this task, the data generator chooses 15% of tokens to predict. For 80% of these examples, it masks a chosen word with a special token [MASK], for 10% the mask is a random word, and in the remaining 10% the token remains unchanged. A model takes a sentence with a masked token as an input and fills in the gap with a correct token. The second task is meant to make a model learn relations between two sentences. The training data included 50% pairs of sentences where one sentence goes after another in an original text and 50% pairs of random sentences. The model's aim is to predict if the second sentence is truly the next sentence which follows the first one. The changes in the architecture and training approach led BERT to achieve the best results at that time on the GLUE benchmark (Wang et al., 2018), which includes several tasks for language models' evaluation, including discourse-based tasks, such as coreference resolution and question answering.

BERT and similar models have been tested on discourse-based tasks. However, most of the previous work on discourse evaluation of BERT and BERT-like models was focused on short discourse structures: Nie et al. (2019) evaluate models on explicit discourse relations expressed with conjunctions. Chen et al. (2019) propose the benchmark for model evaluation on different discourse tasks, such as prediction of implicit discourse relations based on the standard of annotation used in the Penn Discourse Treebank (Prasad et al., 2008), discourse coherence, and others. Araujo et al. (2021) attempt to improve the results on discourse tasks from DiscoEval by changing the pre-training objective of BERT-like models.

2.3 Language-and-vision models

In our work, we use a language-and-vision model based on the BERT architecture. Language-and-vision encoders can be divided into two groups: uni- and dual-stream models. The former models' input is a concatenation of visual and language embeddings (Zhou et al., 2020; Su et al., 2019). The latter models include two different stacks for processing visual and language features. These representations are then fed to a cross-modal stack that includes intra-modal and inter-modal layers alternating one after another (Figure 4). Each stream first computes their own keys, values, and queries and then passes it to another modality stream. Examples of such models are LXMERT (Tan & Bansal, 2019) and ViLBERT (Lu et al., 2019).

The difference between these two types of multi-modal models (illustrated with Figure 4) lies in the way they merge visual and textual information. Single-stream models encode both modalities at the same time, which may lead to learning better understanding of the interaction between two modalities. However, since dual-stream encoders represent each modality separately first, these separate representations can be more enriched with information about each modality. Dual-stream encoders are supposed to learn how to link learned representations of texts and images while single-stream encoders learn a shared representation of an image-caption pair.

Another difference from BERT concerns pre-training objectives: multi-modal models have language, visual and cross-modal objectives. The language pre-training objective is the same as one in BERT – masked language modelling. It allows models to learn semantic representations as it has to predict a word in a given context. A similar idea was also adapted to image regions as either an object classification or feature regression task. With this pre-training objective, the model is supposed to learn an image representation with all the objects in the picture.

To learn how two modalities are linked, models are trained on an image-text matching task: a model has to predict if a given sequence of tokens describes whether a given image (a visual input) and a language input do not match. This objective helps a model to ground a textual knowledge in an image and vice

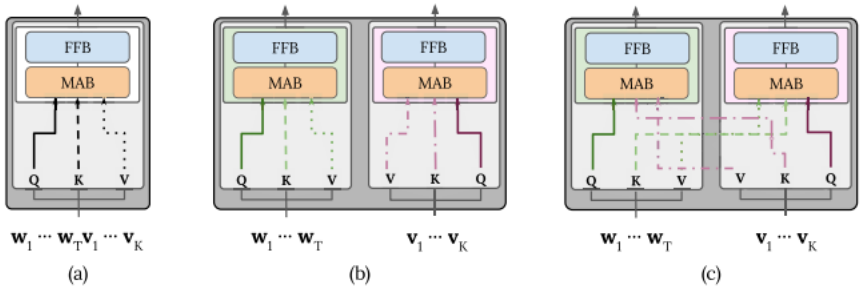
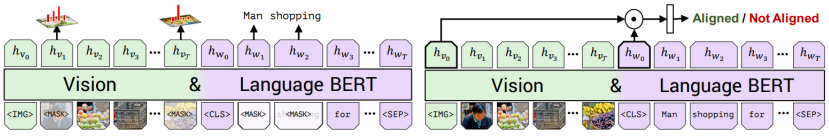


Figure 4: Examples (a) single-stream, (b) dual-stream intra-modal, and (c) dual-stream inter-modal Transformer layers. The single-stream layer takes concatenated representations of text and images and calculates joint keys, values, and queries. The dual-stream intra-modal calculates keys, values and queries separately for textual and visual vectors. The dual-stream inter-modal layer first calculates keys, values and queries for each modality separately and then passes it to another modality. Taken from Bugliarello et al. (2021)

versa. As a model learns how to link language information with non-language data, we assume that it learns discourse knowledge, therefore, later in this work we focus on this task. The pre-training objectives of language-and-vision transformers are illustrated on Figure 5.



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

Figure 5: The pre-training objectives of ViLBERT. The first image illustrates masked language modelling and region prediction (textual and visual objectives). The second image illustrates the cross-modal pre-training objective where the model has to predict if a model and a text are matching or not. Taken from Lu et al. (2019).

Moreover, language-and-vision models learn to *ground* semantic knowledge. *Grounding* can be loosely defined as linking concepts to context, such as knowledge bases, images, or discourse (Chandu et al., 2021). Frank et al. (2021) show that multi-modal models learn how to ground language data with images but not how to ground images in texts. Ilinykh & Dobnik (2022a) provide similar results that the models ground both semantic and syntactic relations. Parcalabescu et al. (2021) argue that models can ground objects but struggle with interdependence relations.

As for discourse research, little has been done on evaluating models' knowledge of discourse. Most of the work on discourse evaluation of language-and-vision models focuses on downstream tasks, such as Visual Question Answering (VQA) and Visual Coreference. Bernardi & Pezzelle (2021) provide an overview of VQA systems and possible problems with solutions taking into account reasoning, language ambiguity etc. Several works address the problem of discourse-coherent image generation. Takmaz et al. (2020) introduce a generation mechanism that produces captions grounded not only in the visual context but also in the common ground established before. Alikhani et al. (2020) improve the quality of generated captions by feeding models with additional information on the types of connections between two clauses. Ilinykh & Dobnik (2022b) show how different decoding strategies for image captioning reflect the discourse structure in comparison to reference captions.

In this work, we propose a method to evaluate discourse directly without training any additional layers or classifiers, as we adapt one of the pre-training objectives for our evaluation purposes.

3 Data and Models

3.1 Visual Storytelling dataset

The images and the captions used in the experiments below were taken from the Visual Storytelling Dataset (VIST) (Huang et al., 2016). This section describes the creation procedures as they are described by the authors of the dataset.

We choose this dataset for our experiments because it is built around stories and contains several types of annotation. The main idea of this dataset is to collect stories on the basis of albums uploaded to Flickr. Stories exemplify narrative structures and the sentences are linked with discourse relations. An example of a story taken from the dataset is illustrated in Figure 2.

VIST is an improved version of the Sequential Images Narrative Dataset (SIND) and it includes 10,117 Flickr albums with 210,819 images. The dataset annotation includes three different levels: *descriptions-in-isolation, descriptions-in-sequence*, and *stories-in-sequence*. The levels differ in the way they were originally annotated by crowd workers.

The dataset creation included several steps:

1. **Extraction of events suitable for storytelling task**: the authors of the dataset generated a list of events that were suitable for storytelling. The authors assumed that such events included some type of possession, e.g. "John's birthday party". Therefore, the authors took 5-grams of image titles and extracted possessive patterns with Stanford CoreNLP (Manning et al., 2014). Then they took heads of possessive phrases and kept only the heads that could be described as EVENT in WordNet3.0. The list of extracted events was used later for data collection.

2. **Data collection**: the authors gathered albums that included events from the mentioned list through FlickrAPI[2]. An album is a sequence of images that were uploaded by one author and taken within a 48-hour span. The albums should include from 10 to 50 images. The images came from the same event but could present different scenes, therefore, the images might not be directly connected in terms of visual discourse [3].

3. **Annotation of stories-in-sequence**: the dataset was annotated by crowd-workers through Amazon's Mechanical Turk[4]. A worker selected at least 5 images, arranged the order of images and wrote a story about these images. Then, another worker got the chosen image sequence and they had to write their own story based on these images. Workers wrote one sentence under one picture. For other details, see Huang et al. (2016).

4. **Annotation of descriptions-in-isolation and descriptions-in-sequence**: these captions were also collected through Amazon's Mechanical Turk. For these tasks, workers followed the instructions of image captioning tasks from MS COCO (Lin et al., 2014). The difference between the two tasks lies in the fact that while annotating *descriptions-in-sequence*, workers could see the entire image sequence, this is unlike *descriptions-in-isolation* where workers got to see only one image at a time.

5. **Post-processing**: this stage included tokenisation of all descriptions with CoreNLP toolkit (Manning et al., 2014) and replacement of all names with tokens MALE or FEMALE and all entities with entity

[2] https://www.flickr.com/services/api/
[3] All albums are covered with a CC-license.
[4] https://www.mturk.com/

types (such as *location*). The data was split into train, validation, and test samples in proportions of 80%/10%/10%.

In the tasks constructed on the basis of this dataset, we use *descriptions-in-isolation* and *stories-in-sequence* for our experiments. The *descriptions-in-sequence* data was lost (Frank Ferraro, p.c.).

The main limitation of the dataset that affects our work is the bias in stories' topics. The dataset topic distribution is skewed towards several major topics. The authors of the dataset report the statistics for 15 main topics: the first topic on the list — *beach* — includes 684 albums, while the 15th most popular topic (*father's day*) contains 221 albums, which is three times less than the most popular topic. As we are interested in how models can distinguish different situations, the bias in stories' topics could make our results less representative.

As has been noted before (Pyatkin et al., 2023), the construction of a task will affect the collected annotations. In the setup of *stories-in-sequence*, there is a difference in the way stories were collected: the first annotator chose images from a given album and wrote a story, while other annotators had to write stories on the basis of already existing sequences. These two ways of collecting data reflect two different sources of discourse: while for storytelling a worker had to come up with a story first and then choose images, for re-telling and tasks such as *descriptions-in-isolation*, a worker captured a situation from the visual context and then described it with words. Workers might find it difficult to write a coherent story if they do not see a connection between given images. Moreover, as Ilinykh et al. (2018) notice, the descriptions produced by crowd workers differ on a given task. For example, the workers can produce noisy, incoherent captions when they are asked to describe an image in comparison with a data collection setup where image descriptions are produced in a more natural way as a part of a task. Unfortunately, there is no way to differentiate between the two types of stories and we cannot be sure which types of stories were used in our tasks.

The last concern in connection with this dataset is the lack of reporting how the workers were recruited. This might affect collected annotations in two ways. First, since the authors do not mention if the workers had to be native speakers of English, it might affect the quality of annotations. Second, the socio-cultural information is important due to ethical considerations, as it is known that discourse depends on social factors, such as gender and age, let alone cultural background. We consider these questions to be out of scope of our research, since the original dataset description does not provide enough information.

3.2 Experimental setup and hypotheses

As mentioned in Section 1, we run three experiments based on levels of annotation presented in VIST and concepts of short or long discourse. As a reminder, short discourse operates on the level of one item (an utterance or an image) and long discourse operates on the level of several items:

I experiment with descriptive captions taken from different stories (*descriptions-in-isolation*, short discourse);

II experiment with in-story captions taken from different stories (*stories-in-sequence*, short discourse);

III experiment with *stories-in-sequence* taken from the same story (long discourse).

The first experiment is closest to an image-caption matching task used as a pre-training objective. In the image-caption matching task, the model receives one sentence and one image as an input and it has to decide

if the image and text are matching or not. When used as a pre-training objective, non-matching captions are assigned randomly. However, we construct our experiments so non-matching captions or images will be chosen not only randomly but also based on their similarity to original items. We believe that distractors will be similar to original items because they name or picture the same objects but in different situations.

Hence, in the first experiment, we test the model on the image-caption matching task under five different conditions (see Figure 6). Besides a random assignment, we use similarity scores. For every item we first change the caption or the image to one of the most similar captions or images respectively and then we repeat the same procedure to assign the most dissimilar caption or image, as illustrated on Figure 6 (examples of constructed pairs could be found in Appendix A). To interpret the model's performance, we need a standard of comparison, hence, we run our experiments on non-matching pairs that have low similarity to a matching pair. For each condition we create a separate dataset that consists of both matching and non-matching pairs.

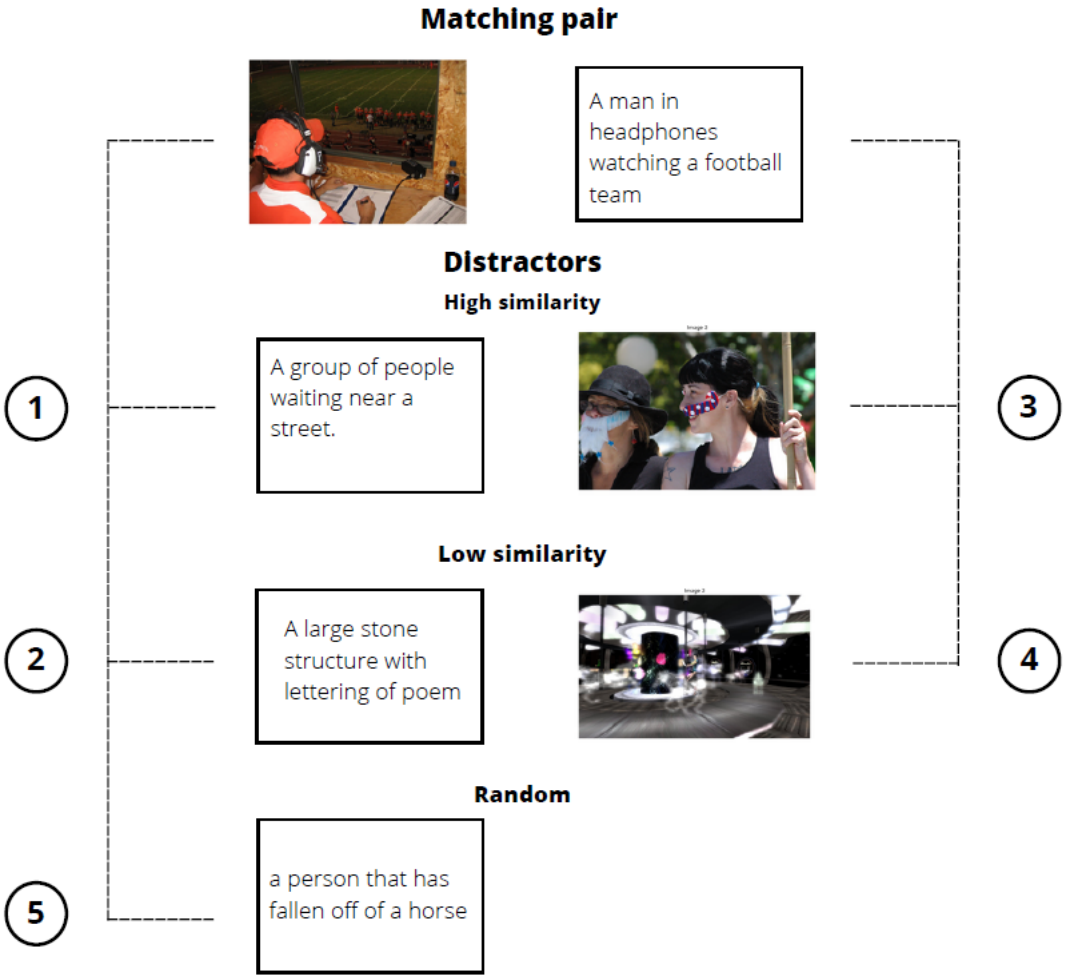


Figure 6: The dataset construction based on different similarities scores and modalities. The 5th condition repeats the procedure of assigning random captions in the pre-training objective of ViLBERT.

The random condition reproduces the setup of the pre-training objective: a model has to predict if a given image and caption are matched or not. In the pre-training task, 50% were kept untouched and 50% captions were assigned randomly. In our task, we test a model on each item (an image or a caption) twice: with its original caption or image and with its non-matching version (see Figure 7 that illustrates the pipeline). In that way, we make datasets balanced and this setup allows us to compare scores on each pair across different setups.

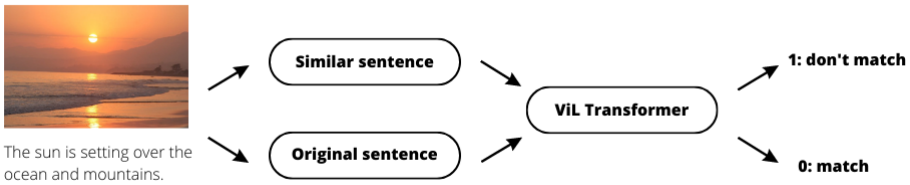


Figure 7: Evaluation pipeline for a high similarity textual setup. Original images are paired with non-matching captions and these pairs are shuffled in with original pairs and are given to the transformer which is supposed to predict label 1 for non-matching pairs and label 0 for matching labels.

First, we run the model on five datasets formed from *descriptions-in-isolation* (Experiment I) and then we repeat the same setup with more abstract, discourse-enriched captions, *stories-in-sequence* (Experiment II). As Experiment III, we randomly shuffle images and captions within one story. All the experiments are summarised in Table 2.

Experiments	**Conditions**	**Captions**	**Discourse structure**
Experiment I	$< i, (c, c_r) >$ $< i, (c, c_s) >$ $< i, (c, c_d) >$ $< c, (i, i_s) >$ $< c, (i, i_d) >$	Descriptions-in-isolation	Short discourse
Experiment II	$< i, (c, c_r) >$ $< i, (c, c_s) >$ $< i, (c, c_d) >$ $< c, (i, i_s) >$ $< c, (i, i_d) >$	Stories-in-sequence	Short discourse
Experiment III	$< i, (c, c_i) >$ $< c, (i, i_i) >$	Stories-in-sequence	Long discourse

Table 2: Summarization of experiments. i stands for image, c stands for captions, x_r stands for random items, x_s stands for high similarity items, x_d stands for low similarity items, x_i stands for items shuffled within one story (internal).
Experiments differ by factors such as the level of annotation (hypothesis C) and granularity of discourse structure (D). The first two experiments include pairs with different similarity scores (hypothesis A) and all experiments include examples where textual modality is replaced and where visual modality is changed (hypothesis B)

Our hypotheses are built around several factors that we predict will affect the model's performance:

A the degree of similarity between a matching item and a distractor, i.e. if the model will show different results on high and low similarity distractors in Experiments I and II;

B the modality that remains unchanged, i.e. if the results on visual and textual distractors will be significantly different;

C the level of caption descriptiveness, i.e. if the model will show different results on *descriptions-in-isolation* and *stories-in-sequence*;

D granularity of discourse structure, i.e. if the model performs differently on captions taken from different stories as opposed to captions taken from the same story.

3.3 Constructing datasets for experiments

As mentioned before, we probe a language-and-vision model on an image-caption matching task. To construct our non-matching pairs, we first extract textual and visual representations and then calculate similarity between them.

We calculate pairwise scores between items with cosine similarity (see Equation 5 where *v* stands for a vector):

$$\cos = \frac{\mathbf{v}_i \cdot \mathbf{v}_j}{||\mathbf{v}_i|| \cdot ||\mathbf{v}_j||}, \tag{4}$$

We then take the upper quartile from the distribution of similarity scores as a threshold for high similarity items and the lower quartile as a threshold for low similarity items. For each item, we then construct lists of items with high and low similarity and pair an original item with an item of another modality randomly chosen from these lists. This way, we make our data more diverse, as two items that are more similar to each other than any other item are assigned with two different items, and the pairs of items do not repeat.

However, the dataset can be biased and this will affect the similarity scores needed for construction of tasks for the first two experiments. As seen from Figure 8, the similarity scores between images or texts vary between 0.6 and 0.9. Since these scores are not interpretable, i.e. we do not know what a zero similarity would mean, we compare with the distribution in the dataset of Conceptual Captions that language-and-vision models are trained on. The distribution of similarities over captions is similar in the two datasets, although image similarities are closer to each other in the VIST dataset. In other words, they are less diverse. This might lead to the model struggling to distinguish similar images, as they are more similar than in the pre-training data.

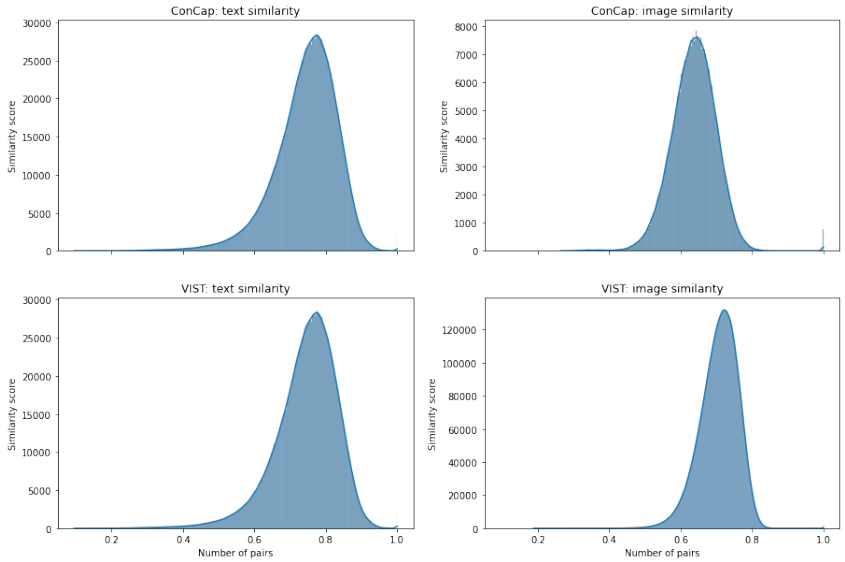


Figure 8: The distribution of similarity scores across Conceptual Captions (ConCap) and Visual Storytelling dataset (VIST) images and captions

As for implementation, we take separate representations for modalities, as we only focus on one modality at the time and find close candidates to either captions or images. To get representations for texts, we

use the HuggingFace implementation of BERT (Devlin et al., 2018) [5]. For image representations we use `ResNet101` (He et al., 2016), a deep residual network based on a CNN architecture. We choose these models since most language-and-vision transformers are based on image representations from ResNet101 and text representations from BERT. Moreover, there are other reasons to choose BERT for text representations: first, it is a standard way to get contextualised word representations learnt from a large pre-training corpus and, second, it is used to initialise the ViLBERT word embeddings.

As for image embeddings, preprocessing steps for language-and-vision transformers include feature extraction FasterRNN on the basis of ResNet101. Therefore, we use the same model for the dataset construction. Although this is a standard way to get image embeddings in computer vision, there are other ways to get visual representations, such as VGGNet or CLIP. These models could give better representations through more layers as in the case of VGGNet or more complex, transformer-based architectures like CLIP. However, we leave experiments with different ways of encoding images for future work.

3.4 Models

We work with ViLBERT (Vision-and-Language BERT) (Lu et al., 2019), the dual-stream multi-modal model based on a BERT-base configuration. We take ViLBERT as an example of the dual-stream architecture because in this setup the model first learns representations of visual and textual modalities separately and then it has to learn cross-modal grounding, in the process of which it is supposed to acquire discourse structures (Figure 4). We believe that this model will have better results on discourse tasks but we leave experiments with other models for future work. The ViLBERT model was trained on the Conceptual Captions dataset that we compared our dataset to in Section 3.3.

We choose the dual-stream transformer, as we believe that it will show better results on cross-modal tasks: intra-layers of this model explicitly pass keys, queries, and values of one modality to the stack of another modality. We believe that this feature of the model's architecture allows it to better acquire the relations between textual and visual modalities, i.e. it can acquire situation-level discourse (recall Section 1). Although ViLBERT is not the only dual-stream transformer, we prefer this model to LXMERT, since the authors of the latter model Tan & Bansal (2019) note that LXMERT shows better results on downstream tasks, e.g. visual question answering, while we are interested in image-caption matching tasks. ViLBERT has been included in previous works on evaluation of multi-modal transformers, such as (Parcalabescu et al., 2021; Bugliarello et al., 2022).

As for implementation, we test ViLBERT taken from VOLTA (Bugliarello et al., 2021), the framework that provides the code base for several transformer-based language-and-vision encoders and allows working with custom datasets. Before passing data to a multi-modal model, we need to pre-process images and captions. As an input, the model takes a set of features: masked sentences, token ids, visual features, image location, masked images with regions of interest and their object labels. Textual features come from the BERT tokenizer[6] that returns a sequence of token ids. To extract image features, we use the Caffe VG Faster R-CNN implementation (Anderson et al., 2018)[7]. The model extracts 36 proposal boxes with their features of dimension 2048 and object labels. We do not mask any tokens or regions as we focus only on one output head of ViLBERT that predicts if an image and a text match.

[5] https://huggingface.co/bert-base-uncased
[6] `https://huggingface.co/bert-base-cased`
[7] `https://github.com/airsplay/py-bottom-up-attention`

4 Results

The experiments include three parts: image-matching tasks with *descriptions-in-isolation* under different conditions, the same setup with *stories-in-sequence*, and image-matching task with captions and images within one story (Section 3). We use *accuracy* as a metric because all our experiments are binary classification tasks and the datasets are balanced, since they include every item of the original dataset with an original caption or image and a replaced caption or image. For binary classification tasks, *accuracy* is proportional to *precision* and *recall*:

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

4.1 Experiment I: evaluation of short discourse on descriptions-in-isolation captions

Changed modality	**High Similarity**	**Low similarity**
Textual	0.92	0.95
Visual	0.94	0.95

Table 3: The overall performance on image-caption matching task with *descriptions-in-isolation* based on accuracy on a balanced dataset. The accuracy of the baseline on random distractors is **0.94**.

In this experiment (Section 3), we test the model on images with descriptive, out-of-story captions permutated under five conditions based on similarity between matching and non-matching pairs and modality (recall Figure 6).

First, we look at the overall performance of the model on this task (Table 3). The model shows high performance on *descriptions-in-isolation* under all five conditions showing results greater than 0.9. As for differences on high and low similarity distractors, although the results are similar, the model shows the best quality on random distractors and distractors with low similarity, possibly because such items are easier to distinguish. The results on high similarity visual distractors are slightly worse, however, the model still can predict a correct label in the majority of cases. Regarding modality, the results on high similarity distractors are slightly worse for textual distractors than visual ones, which indicates the sensitivity of the model to changes in text and, possibly, this result might mean that the model is simply relying on text much more than on vision. This is a well-known problem with multi-modal architectures - the model grounds text in images better than images in text (Ilinykh et al., 2022). In general, we cannot see a large difference in the results of the experiment under different conditions.

To dive deeper, we can look at the confidence scores of the model. We calculate the model's confidence in the following way: we take the prediction scores of a correct answer (i.e. label 0 for matching pairs and label 1 for non-matching pairs, recall Figure 7) and interpret them as the model's confidence in this answer. The question that we are interested in is if the confidence scores will reflect any difference between high and low similarity distractors or between textual or visual distractors. As can be seen on Figure 9, the model is in general very confident in its predictions, as most of the scores are close to 1. There is almost no difference between the different graphs (conditions we tested the model on) except for the random distractors. The model is less confident on non-matching labels on this condition than in the other four conditions.

Although the model is confident in all classes, there might be a correlation between similarity scores and

the model's confidence score. However, as Figure 10 illustrates, there is no correlation and the model is confident in predictions even on distractors no matter how similar they are to original pairs. However, the model never makes mistakes on pairs with very low similarity distractors.

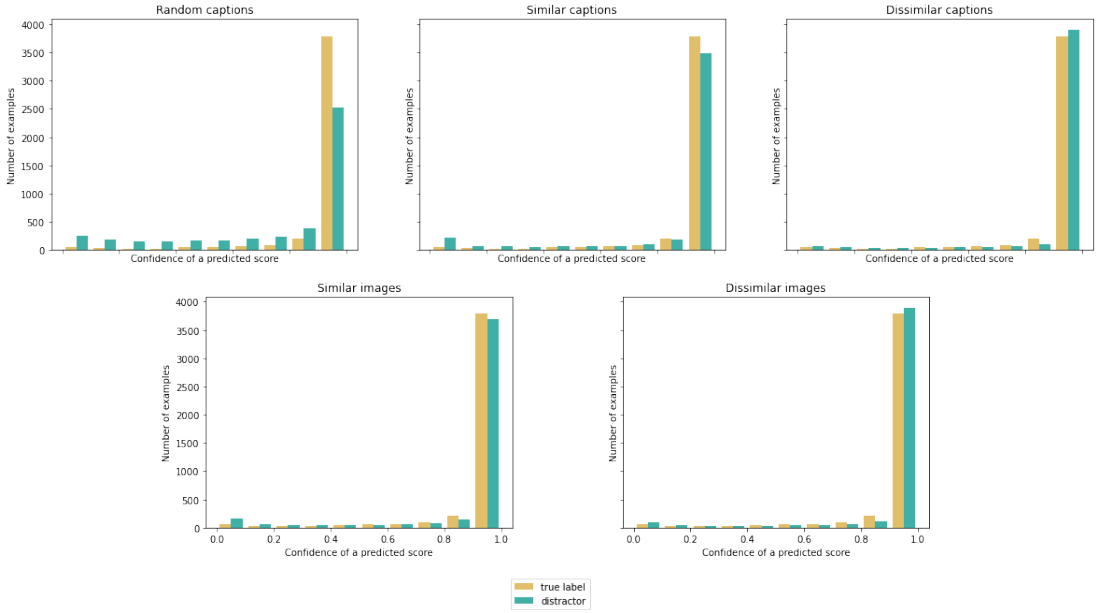


Figure 9: The cross-plot of confidence of the model in predicting correct answers in the image-caption matching task with *descriptions-in-isolation* by labels. Both ground truth examples and distractors are presented on the graph.

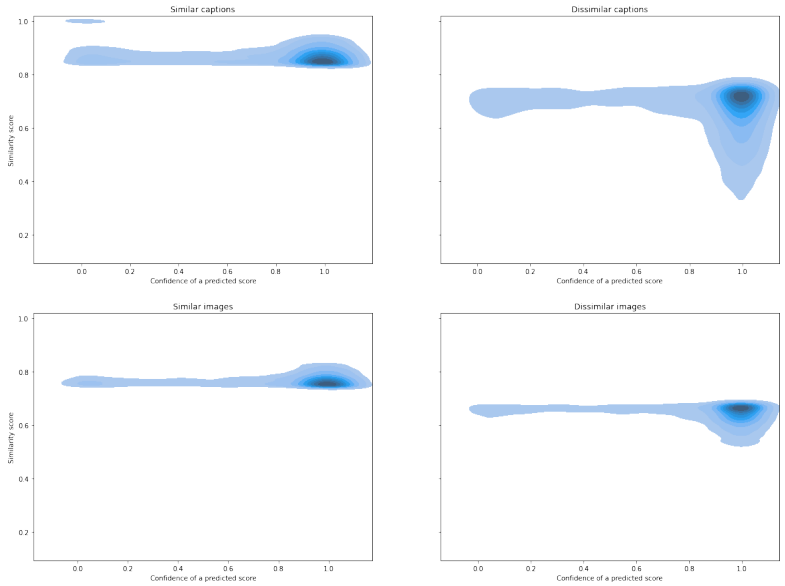


Figure 10: The correlation of similarity scores and the model's confidence in the correct answers in the image-caption matching tasks with *descriptions-in-isolation*. Only distractors are presented on this graph.

As we have seen so far, the model does not make mistakes in distinguishing situations and it is very confident in predicting correct answers. However, we can compare the model's predictions under different conditions. In that way, we can investigate the significance of different factors, such as similarity scores and modality, for the model's predictions. To achieve that, we run *Student's t-test* on prediction scores for non-matching pairs. The test shows if two paired groups have significantly different means. The null hypothesis is that two distributions have the same means. If the test shows p-value less than 0.05, we reject the null hypothesis and say that two distributions are significantly different. The results of the tests are summarized in Table 4 where we postulate p-values. As seen from the table, the p-values are extremely small for every comparison,

except for the modality on low similarity distractors. Hence, the difference between high and low similarity distractors is a significant factor for the model's performance. When the task is challenging, such as with high similarity distractors, the difference in the modality is also essential. In other words, similarity scores affect the model's performance as well as the modality affects the results when the distractors are similar to the original items. However, the performance on low similarity distractors does not vary between modalities.

Group 1	Group 2	Descriptions-in-isolation
$< i, (c, c_s) >$	$< i, (c, c_d) >$	<0.001
$< c, (i, i_s) >$	$< c, (i, i_d) >$	<0.001
$< i, (c, c_s) >$	$< c, (i, i_s) >$	<0.001
$< i, (c, c_d) >$	$< c, (i, i_d) >$	0.8

Table 4: The p-values reported by *Student's t-test* on comparison of the model's predictions under different conditions after testing the model on *descriptions-in-isolation*.
The first two rows compare results on the same original items in two different setups: in the first row textual distractors with a different degree of similarity are compared, in the second row visual distractors with different degree of similarity are compared.
In the last rows the distractors with the same degree of similarity are compared. In the third row, high similarity distractors are tested to see if there is any significant difference due to the modality. In the last row, the low similarity distractors are compared.

4.2 Experiment II: evaluation of short discourse on stories-in-sequence captions

Changed modality	High Similarity	Low similarity
Textual	0.78	0.85
Visual	0.81	0.85

Table 5: Overall performance on image-caption matching task with *stories-in-sequence* based on accuracy on a balanced dataset. The results on random distractors are **0.82**

We implement the same experimental setup for *stories-in-sequence* to test short discourse structures on the captions that originally were part of a narrative. As seen from Table 5, the results are worse than on *descriptions-in-isolation*. As for similarity, the model performs slightly better on low similarity distractors than on high similarity distractors. Regarding modality, the scores are different if the distractors are of high similarity.

We can take a closer look at the model's confidence. As seen from Figure 11, the model gets more confused on distractors, especially on high similarity distractors, as they are more challenging for the model. However, there is no difference in the performance when different modalities are replaced. As for the possible correlation between similarity scores and the model's predictions, there is still no linear correlation between similarity scores and the model's confidence, as illustrated on Figure 12.

Now we repeat the *t-test* for comparison of results obtained under different conditions (see Table 6). As in the previous experiment, we see that the scores on groups by similarity differ significantly, while modality is important only if the task is already challenging, such as with high similarity distractors.

4.3 Experiment III: evaluation of long discourse on stories-in-sequence captions

The last task includes image-caption matching of sentences from the *stories-in-sequence* annotation layer when they are randomly shuffled within one story. The results can be seen in Table 7. Unlike previous experiments, the model's performance is higher on textual distractors than on visual distractors. The possible

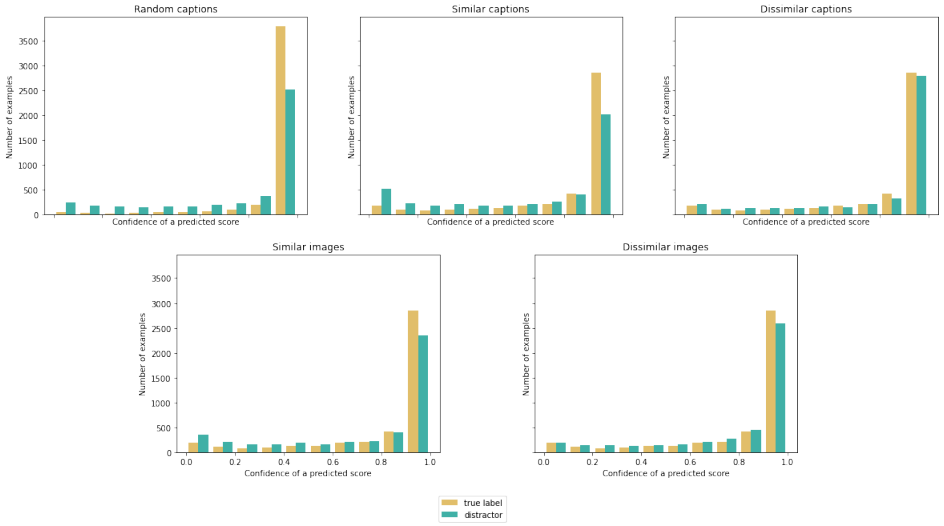


Figure 11: Confidence of the model on the image-caption matching task with *stories-in-sequence* by labels. Both ground truth examples and distractors are presented on the graph.

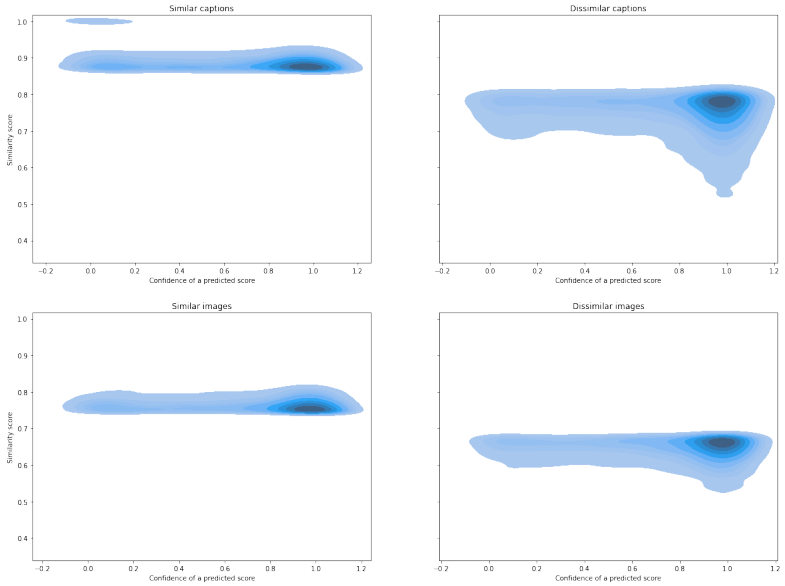


Figure 12: Cross-plot of similarity scores and the model's confidence on image-caption matching tasks with *stories-in-sequence*. Only distractors are presented on this graph.

Group 1	Group 2	Stories-in-sequence
$< i, (c, c_s) >$	$< i, (c, c_d) >$	<0.001
$< c, (i, i_s) >$	$< c, (i, i_d) >$	<0.001
$< i, (c, c_s) >$	$< c, (i, i_s) >$	<0.001
$< i, (c, c_d) >$	$< c, (i, i_d) >$	0.08

Table 6: The results of *t-test* on the model's scores under different conditions on non-matching pairs of *stories-in-sequence*

reason behind this behaviour is that images within one story are more similar than captions. Another reason might be the way *stories-in-sequence* were collected (see Section 3). Some workers were not the ones who actually combined pictures in stories and were only asked to write a story about a pre-chosen sequence of images. This might result in a lower quality of annotations, as different crowdworkers would have different perspectives and understanding of images and sequences.

Changed modality	Random
Textual	0.63
Visual	0.6

Table 7: The results (accuracy) on the experiment with *stories-in-sequence* shuffled within one story.

In comparison with Experiment II where *stories-in-sequence* were used but were taken from any story, the quality drops by 0.2, which indicates that the model struggles with items taken from the same story. Since all images and texts are focused on the same objects within one story, the model does not capture causal relations that could help to distinguish items from different parts of one story.

Moreover, as seen from Figure 13, the model gets easily distracted with shuffled items and predicts true labels for them with high confidence. More than 1500 non-matching pairs out of 2500 get the null score for label 1 (non-matching image and caption), which indicates that the model predicts the opposite label (matching pair) with confidence of 1. Therefore, the model does not only predict wrong answers but it gives high confidence scores to false positives.

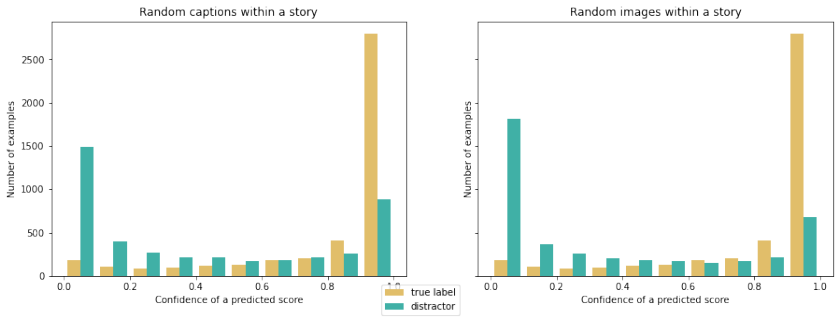


Figure 13: Confidence of the model on image-caption matching task with *stories-in-sequence* within one story by labels. The model assigns a label 0 (if it is ground-truth) or 1 (if it is a distractor). Both ground truth examples and distractors are presented on the graph.

5 Discussion

5.1 The impact of data

We have tested the model on distractors with high and low similarity to see if the model can distinguish between similar situations, in other words, whether it performs with the same quality on distractors with high similarity as on the ones with low similarity, while assuming that high similarity distractors will picture similar situations to the original situations.

In general, evaluation results show that the model was not particularly challenged by distractors with high similarity, especially for *descriptions-in-isolation* (see Section 3). Results are similarly good when presented with distractors of both high and low similarity. In other words, a model is sensitive to different situations of high or low similarity. In other words, the model distinguishes different situations well no matter if descriptions or images of the situations are close in similarity or not. However, for *stories-in-sequence*, the results on high similarity pairs were lower than on low similarity pairs. In this case, a model can distinguish between two completely different situations but is less sensitive to catch the difference in images that picture similar objects but in different situations. Moreover, there is a significant correlation in the model's performance and similarity scores, as Table 6 indicates. It is also shown in Figures 10 and 12 that there is a visible correlation between similarity scores and the model's confidence in the way that it tends to make fewer mistakes on more dissimilar items.

Overall, since the model performs well on both Experiments I and II, we could say that it captures differences in situations even under challenging conditions, such as high similarity distractors, and with more abstract annotations. If we recall the distribution of scores in the Conceptual Captions dataset it was trained on (Figure 8), the images were similar even in the training set, so the model learned how to catch fine-grained similarity differences.

If we compare results on each task with randomly shuffled pairs, we could see how well a model performs relatively to a random baseline. The results on random distractors tend to lie between results on high and low distractors. Among random distractors we expect to see both high and low distractors. Therefore, since similarity affects the model's scores, the model performs worse on random distractors than on low similarity distractors, as random pairs also include high similarity distractors. On the other hand, the model performs better than on high similarity distractors because random distractors contain low similarity distractors.

The other feature of the data we used was different levels of annotations. The first level of annotation – *descriptions-in-isolation* – has more descriptive captions, i.e. more grounded in images. For example, they name objects in images and their attributes and how they are shown on images. On the other hand, *stories-in-sequence* are more abstract in that the texts are less grounded and they focus on the situation and on events happening on an image rather than list the objects in a picture. In such captions, annotators could miss the majority of the information as this information can be retrieved from the visual modality and focus only on the part that brings novelty to a story.

As can be seen from the results, the model performs better on descriptive captions. This could happen for several reasons. First, the model was pre-trained on descriptive captions, although they come from a different dataset and the distribution of represented topics in the two datasets might be different. Second, descriptive captions are easier to ground as they tend to mention all or most of the objects presented in a picture while *stories-in-sequence* can mention only one centered object. Third, to ground long discourse captions (see Section 1), the model might need previous context, which it has not seen as it gets only one image and one caption.

5.2 The (im)balance of modalities in the model

We are interested in how the performance of the model differs when either a visual or a textual modality remains unchanged. In other words, results on textual distractors show how sensitive the model is to linguistic discourse and the results on visual distractors show to what degree the model captures the differences in visual discourse.

Frank et al. (2021) show with ablation masking studies that the model relies more on the textual modality in its predictions than on visual modality. Our results reveal the same tendency. The results on experiments for which the text was fixed and images were shuffled are better on *descriptions-in-isolation* and *stories-in-sequence*. Although the difference in accuracy is not dramatic, the model shows better results when it can base predictions on ground truth textual modality, hence, the model tends to predict false positives more often when it receives a non-matching caption than a non-matching image. In other words, it is more difficult for the model to see the mismatch between an original image and a different caption, as the model relies more on textual modality in its prediction than on a visual one, as it distinguishes better between two texts than between two images.

In Experiment III, however, the model performs better on textual distractors. Within one story, captions could be more diverse than images. In textual discourse there are more ways of making a narrative coherent, for example, using pronouns for coreference. In visual discourse, however, the same centered objects are presented on several pictures to make the narrative coherent.

In addition, as we see from Table 6, there is a significant difference in means of scores grouped by modality for distractors with high similarity. However, the modality is not important for results on low similarity distractors. Overall, similarity is a more important factor than modality, and the modality plays an essential role only on already complex tasks.

5.3 Understanding situations in the context of discourse

The model shows better performance on captions taken from different stories than on ones taken from the same story, as the model is less sensitive to more granular distinctions in a narrative and makes more mistakes on pairs shuffled within one story.

As we have shown earlier, the model is good at capturing even fine-grained differences, while it is less adept when presented with more abstract captions. However, the performance of the model on the distractors shuffled within the same story is slightly better than a random baseline, which would be 0.5 for a binary classification task. We could expect that this task would not be so challenging, since the captions might be more different within one story than in the similar captions taken from different narratives. If we recall Figure 2, the story mentions the same objects, but the sentences in the story are not interchangeable.

As mentioned before, in comparison with Experiment II, the performance in Experiment III has dropped by 0.2 in accuracy. Since Experiment II was already run on *stories-in-sequence*, the difference between experiments is in more granular discourse differences that the model cannot capture. Therefore, the model can distinguish two different narratives but it struggles with distinguishing discourse items that belong to one story. In other words, the model can differentiate different situations but it does not see the differences on the fine-grained level of discourse when all items come from the same story.

6 Conclusion

In this work, we looked at whether a language-and-vision model is sensitive to discourse structure at different levels of granularity. We focus on short and long discourse structure and adapt these terms to the task of image-caption matching which is one of the pre-training objectives of multi-modal models.

We run the experiments on ViLBERT, a language-and-vision dual-stream transformer-based model. As for the data, we use Visual Storytelling dataset where images are united into stories. The dataset has several layers of annotation including isolated captions and stories. First, we test its general image-caption matching abilities on the first type of captions under five different conditions. We constructed pairs of distractors, i. e. pairs where either an image or a caption is changed to a similar or dissimilar caption or image. We test if the model performance will be different depending on the similarity scores or modalities.

We see that the model predicts correct answers with high confidence in 90% of cases. We then take the *stories-in-sequence* annotation level for our experiments to test the model on captions included in a story that we believe are more challenging for the model.

The results drop by 0.1 in comparison with the previous experiment which shows that captions that are part of a bigger narrative are more challenging for the model that was trained on descriptive captions. However, the model shows high results on the two experiments, hence, the model has learnt to distinguish the differences in short discourse structure.

In our last experiment, we focus on *stories-in-sequence* captions shuffled within one story. The model's performance drops to an accuracy of 0.6. The model assigns wrong labels with high confidence and it cannot distinguish discourse units within one story. The model therefore struggles to capture the differences when both matching and non-matching items focus on the same objects. We believe that the model is better at distinguishing different situations as images and captions present different objects in different situations.

Moreover, we investigated the impact of similarity score values and the modality. As for similarity scores, the statistical tests show a significant difference in distributions of predicted scores for high and low similarity distractors. However, the correlation is vague and the model is confident in its predictions on any type of distractors.

The results on textual distractors are usually worse than the results on corresponding experiments with visual distractors, except for Experiment III. In other words, it is easier for the model to give a correct label to a non-matching pair where a caption was not changed than to a pair where an image was not changed, although according to the distribution of similarity scores, images are less diverse and should be harder to differentiate. As for Experiment III, we believe that images are more similar than captions, as visual discourse has fewer available tools to make a discourse coherent. However, statistical tests show that the modality is not essential when the distractors have low similarity.

In order to ensure that all our experiments are reproducible, we make our code publicly available: https://github.com/EkaterinaVoloshina/multimodal_discourse_probing.

References

Alikhani, M., Sharma, P., Li, S., Soricut, R., & Stone, M. (2020). Clue: Cross-modal coherence modeling for caption generation. *arXiv preprint arXiv:2005.00908*.

Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., & Zhang, L. (2018). Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*.

Araujo, V., Villa, A., Mendoza, M., Moens, M.-F., & Soto, A. (2021). Augmenting bert-style models with predictive coding to improve discourse-level representations. *arXiv preprint arXiv:2109.04602*.

Belinkov, Y. (2022). Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1), 207–219.

Bender, E. M. & Koller, A. (2020). Climbing towards nlu: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 5185–5198).

Bernardi, R. & Pezzelle, S. (2021). Linguistic issues behind visual question answering. *Language and Linguistics Compass*, 15(6), elnc3–12417.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33, 1877–1901.

Bugliarello, E., Cotterell, R., Okazaki, N., & Elliott, D. (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.

Bugliarello, E., Liu, F., Pfeiffer, J., Reddy, S., Elliott, D., Ponti, E. M., & Vulić, I. (2022). Iglue: A benchmark for transfer learning across modalities, tasks, and languages. In *International Conference on Machine Learning* (pp. 2370–2392).: PMLR.

Chandu, K. R., Bisk, Y., & Black, A. W. (2021). Grounding'grounding'in nlp. *arXiv preprint arXiv:2106.02192*.

Chen, M., Chu, Z., & Gimpel, K. (2019). Evaluation benchmarks and learning criteria for discourse-aware sentence representations. *arXiv preprint arXiv:1909.00142*.

Clark, A. & Lappin, S. (2010). Computational learning theory and language acquisition. *Philosophy of linguistics*, (pp. 445–475).

Clark, H. (1977). Bridging. *Thinking: Readings in Cognitive Science*.

Clark, H. H. (1996). *Using language*. Cambridge university press.

Conneau, A., Kruszewski, G., Lample, G., Barrault, L., & Baroni, M. (2018). What you can cram into a single vector: Probing sentence embeddings for linguistic properties. *arXiv preprint arXiv:1805.01070*.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Elazar, Y., Ravfogel, S., Jacovi, A., & Goldberg, Y. (2021). Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9, 160–175.

Ettinger, A. (2020). What bert is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8, 34–48.

Fairclough, N. (1992). Discourse and text: Linguistic and intertextual analysis within discourse analysis. *Discourse & society*, 3(2), 193–217.

Frank, S., Bugliarello, E., & Elliott, D. (2021). Vision-and-language or vision-for-language? on cross-modal influence in multimodal transformers. *arXiv preprint arXiv:2109.04448*.

Gehrmann, S., Bhattacharjee, A., Mahendiran, A., Wang, A., Papangelis, A., Madaan, A., McMillan-Major, A., Shvets, A., Upadhyay, A., Yao, B., et al. (2022). Gemv2: Multilingual nlg benchmarking in a single line of code. *arXiv preprint arXiv:2206.11249*.

Grosz, B. J., Joshi, A. K., & Weinstein, S. (1995). Centering: A framework for modelling the local coherence of discourse.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778).

Huang, T.-H., Ferraro, F., Mostafazadeh, N., Misra, I., Agrawal, A., Devlin, J., Girshick, R., He, X., Kohli, P., Batra, D., et al. (2016). Visual storytelling. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: Human language technologies* (pp. 1233–1239).

Ilinykh, N. & Dobnik, S. (2022a). Attention as grounding: Exploring textual and cross-modal attention on entities and relations in language-and-vision transformer. In *Findings of the Association for Computational Linguistics: ACL 2022* (pp. 4062–4073). Dublin, Ireland: Association for Computational Linguistics.

Ilinykh, N. & Dobnik, S. (2022b). Do decoding algorithms capture discourse structure in multi-modal tasks? a case study of image paragraph generation. In *Proceedings of the 2nd Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)* (pp. 480–493). Abu Dhabi, United Arab Emirates (Hybrid): Association for Computational Linguistics.

Ilinykh, N. & Dobnik, S. (2022c). Hallucinate or ground: how general or specific are object descriptions generated by a vision-and-language transformer?

Ilinykh, N., Emampoor, Y., & Dobnik, S. (2022). Look and answer the question: On the role of vision in embodied question answering. In *Proceedings of the 15th International Conference on Natural Language Generation* (pp. 236–245).

Ilinykh, N., Zarrieß, S., & Schlangen, D. (2018). The task matters: Comparing image captioning and task-based dialogical image description. In *Proceedings of the 11th International Conference on Natural Language Generation* (pp. 397–402).

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and brain sciences*, 40, e253.

Lasri, K., Pimentel, T., Lenci, A., Poibeau, T., & Cotterell, R. (2022). Probing for the usage of grammatical number. *arXiv preprint arXiv:2204.08831*.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13* (pp. 740–755).: Springer.

Liu, N. F., Gardner, M., Belinkov, Y., Peters, M. E., & Smith, N. A. (2019). Linguistic knowledge and transferability of contextual representations. *arXiv preprint arXiv:1903.08855*.

Lu, J., Batra, D., Parikh, D., & Lee, S. (2019). Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J. R., Bethard, S., & McClosky, D. (2014). The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations* (pp. 55–60).

Maudslay, R. H. & Cotterell, R. (2021). Do syntactic probes probe syntax? experiments with jabberwocky probing. *arXiv preprint arXiv:2106.02559*.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Nie, A., Bennett, E., & Goodman, N. (2019). Dissent: Learning sentence representations from explicit discourse relations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 4497–4510).

Parcalabescu, L., Cafagna, M., Muradjan, L., Frank, A., Calixto, I., & Gatt, A. (2021). Valse: A task-independent benchmark for vision and language models centered on linguistic phenomena. *arXiv preprint arXiv:2112.07566*.

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532–1543).

Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A. K., & Webber, B. L. (2008). The penn discourse treebank 2.0. In *LREC*.

Pyatkin, V., Yung, F., Scholman, M. C., Tsarfaty, R., Dagan, I., & Demberg, V. (2023). Design choices for crowdsourcing implicit discourse relations: Revealing the biases introduced by task design. *arXiv preprint arXiv:2304.00815*.

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

Rajpurkar, P., Jia, R., & Liang, P. (2018). Know what you don't know: Unanswerable questions for squad. *CoRR*, abs/1806.03822.

Rogers, A., Kovaleva, O., & Rumshisky, A. (2020). A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8, 842–866.

Saphra, N. (2021). Training dynamics of neural language models.

Stańczak, K., Ponti, E., Hennigen, L. T., Cotterell, R., & Augenstein, I. (2022). Same neurons, different languages: Probing morphosyntax in multilingual pre-trained models. *arXiv preprint arXiv:2205.02023*.

Su, W., Zhu, X., Cao, Y., Li, B., Lu, L., Wei, F., & Dai, J. (2019). Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*.

Takmaz, E., Giulianelli, M., Pezzelle, S., Sinclair, A., & Fernández, R. (2020). Refer, Reuse, Reduce: Generating Subsequent References in Visual and Conversational Contexts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 4350–4368). Online: Association for Computational Linguistics.

Taktasheva, E., Mikhailov, V., & Artemova, E. (2021). Shaking syntactic trees on the sesame street: Multilingual probing with controllable perturbations. *arXiv preprint arXiv:2109.14017*.

Tan, H. & Bansal, M. (2019). Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. (2019). Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. R. (2018). Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Warstadt, A., Parrish, A., Liu, H., Mohananey, A., Peng, W., Wang, S.-F., & Bowman, S. R. (2020). Blimp: The benchmark of linguistic minimal pairs for english. *Transactions of the Association for Computational Linguistics*, 8, 377–392.

Zhang, C., Van Durme, B., Li, Z., & Stengel-Eskin, E. (2022). Visual commonsense in pretrained unimodal and multimodal models. *arXiv preprint arXiv:2205.01850*.

Zhou, L., Palangi, H., Zhang, L., Hu, H., Corso, J., & Gao, J. (2020). Unified vision-language pre-training for image captioning and vqa. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34 (pp. 13041–13049).

A Appendix A. The examples of generated non-matching pairs

Figure 14: The example of **a random distractor**. As one item of a batch, the model gets an image and a matching caption and as another it gets the same image with a non-matchong caption
Ground-truth caption: *For St. Patrick's day, a man dresses in a leprechaun hat and huge tie covered in clover.*
Non-matching caption: *Beautiful summer day to enjoy a ride near the ocean.*

Figure 15: The example of **a high similarity textual distractor**.
Ground-truth caption: *Two hikers taking a picture in the wilderness.*
Non-matching caption: *Two people and a dog pose for a picture on a cold winter day.*

Figure 16: The example of **a low similarity textual distractor**.
Ground-truth caption: *A hidden male mimics rabbit ears behind 2 other males posing for a picture.*
Non-matching caption: *The bakery has fresh bread with different prices.*

Figure 17: The example of **a high similarity visual distractor**.
Caption: *A group of individuals are clapping at a game.*

Figure 18: The example of **a high similarity visual distractor**.
Caption: *A man is drinking a glass of liquid while the woman is wearing a birthday hat and holding a child dressed in a green outfit*