Is BERT Blind? Exploring the Effect of Vision-and-Language Pretraining on Visual Language Understanding

Morris Alper*, Michael Fiman*, Hadar Averbuch-Elor

Tel Aviv University

Alex gave Riley a present   Alex gave Riley an ultimatum   the ocean is [*]-colored   a corn chip is [*]-shaped

How concrete are the words *present* and *ultimatum*?   **What word should be inserted in the blank?**

Figure 1. In this paper, we propose a suite of *visual language understanding* tasks for probing the visual reasoning capabilities of text encoder models. While we consider text-only tasks (i.e., processing only the textual descriptions above without associated imagery), we argue that they require visual imagination to complete and can thus benefit from vision-and-language pretraining. For instance, consider the words *present* and *ultimatum*. A simple online query (that considers only freely-available images) roughly yields a coherent set of images for the more concrete word (namely *present*), while the latter cannot be uniquely depicted. Likewise, selecting the most natural color or shape descriptors in cloze contexts as shown in the two examples on the right requires implicit knowledge of the appearance of the referent under consideration (*ocean* and *corn chip* respectively).

Abstract

Most humans use visual imagination to understand and reason about language, but models such as BERT reason about language using knowledge acquired during text-only pretraining. In this work, we investigate whether vision-and-language pretraining can improve performance on text-only tasks that involve implicit visual reasoning, focusing primarily on zero-shot probing methods. We propose a suite of visual language understanding (VLU) tasks for probing the visual reasoning abilities of text encoder models, as well as various non-visual natural language understanding (NLU) tasks for comparison. We also contribute a novel zero-shot knowledge probing method, Stroop probing, for applying models such as CLIP to text-only tasks without needing a prediction head such as the masked language modelling head of models like BERT. We show that SOTA multimodally trained text encoders outperform unimodally trained text encoders on the VLU tasks while being underperformed by them on the NLU tasks, lending new context to previously mixed results regarding the NLU capabilities of multimodal models. We conclude that exposure to images during pretraining affords inherent visual reasoning knowl-

edge that is reflected in language-only tasks that require implicit visual reasoning. Our findings bear importance in the broader context of multimodal learning, providing principled guidelines for the choice of text encoders used in such contexts[1].

1. Introduction

Humans are multimodal learners. We communicate with each other about things that we have experienced and knowledge we have gained using our senses—most commonly including sight as well as hearing, touch, smell, and taste. Our communication channel is limited to a single modality—spoken language, signed language, or text—but a reader or listener is expected to use his or her imagination to visualize and reason about the content being described. In general, language is used to describe scenes, events, and images; the words used to describe these are used to conjure up a visual impression in the listener. Therefore, it is natural to consider the types of visual reasoning used in understanding language, and to ask how well we can currently

---

*These authors contributed equally to this work

[1]Our code will be made available at https://isbertblind.github.io/

model them with computational methods.

Consider, for instance, the questions in Figure 1. Concreteness is typically correlated with how well a concept can be visually imagined. For example, a concrete word such as *present* often has a unique visual representation. In addition, common associations such as *ocean→blue* (color) and *corn chip→triangle* (shape) reflect properties of an imagined visual representation of the item in question. These properties may be difficult to infer from text alone without prior knowledge gained from visual input; for instance, a number of studies have investigated the partial ability of blind English speakers to predict color associations and how it differs from the intuition of sighted speakers[2] [45, 57, 58, 61, 73].

There has been a wealth of recent research vision-and-language (V&L) tasks involving both text and image data, and the use of vision-language pretraining (VLP) to create models that are able to reason jointly about both of these modalities together [11, 12, 31, 39]. Notable in this regard is CLIP [52], consisting of paired text and image encoders jointly trained on a contrastive objective, that learns to align text and image embeddings in a shared semantic space. On the other hand, text encoder models such as BERT [15] learn to reason about text in a unimodal vacuum, with knowledge derived from pretraining tasks that only involve textual data.

Prior work has investigated the performance of multimodally trained text encoders on various natural language understanding (NLU) tasks with mixed results, sometimes finding that they are outperformed by unimodal models [24] and at other times suggesting improved performance [78]. However, these works fine-tune the models under consideration on NLU tasks before evaluation, making it difficult to disentangle the effects of multimodal pretraining and fine-tuning configuration on the observed performance. Additionally, these works do not address the distinction between NLU tasks requiring implicit visual reasoning and ones that are purely non-visual. We refer to natural language inference involving implicit visual reasoning as *visual language understanding* (VLU) and propose a suite of VLU tasks that may be used to evaluate visual reasoning capabilities of pretrained text encoders, focusing primarily on zero-shot methods.

We compare multimodally trained text encoders such as that of CLIP to BERT and other unimodally trained text encoders, evaluating their performance on our suite of VLU tasks. We evaluate these models in without modifying their internal weights in order to probe their knowledge obtained during pretraining. A key design aspect of these tests is the probing method used to evaluate knowledge. Previ-

ous work has probed the knowledge of BERT and similar models using a masked language modelling (MLM) paradigm [48, 55], but this cannot be directly applied to CLIP since it was not pretrained with MLM. We therefore propose a new zero-shot probing method that we term *Stroop probing*. This is based on the psychological Stroop effect [43] (described in Section 3.2), which suggests that salient items should have a stronger interference effect on the representation of their context.

Strikingly, we find that the multimodally trained text encoders under consideration outperform unimodally trained text encoders on VLU tasks, both when comparing to much larger encoders as well as ones of comparable size. We also compare these models on baseline NLU tasks that do not involve visual reasoning and find that models such as CLIP underperform on these tasks, demonstrating that they do not have a global advantage on NLU tasks. We conclude that exposure to images during pretraining improves performance on text-only tasks that require visual reasoning. Furthermore, our findings isolate the effect of the text component of multimodal models for tasks such as text to image generation, providing principled guidelines for understanding the knowledge that such models inject into downstream vision tasks.

2. Related Work

Building models to create text embeddings for a large range of language tasks has been broadly explored over the past several years. In our work we compare two types of transformer-based models which encode these types of embeddings: those trained on text-only data (unimodally trained), and those exposed to both text and image data during training (multimodally trained).

Since the introduction of the self-attention-based transformer architecture by Vaswani *et al.* [74], transformers have become the predominant architecture for tasks involving textual data. Devlin et al. [15] introduce BERT, a transformer model encoding contextual information for each token in an input sequence in a bidirectional manner. They suggest a self-supervised pretraining method to compute contextual feature representations of textual data, via masked language modelling and next sentence prediction objectives. This pretrained model can then be applied to other downstream tasks by end-to-end fine-tuning. Subsequently, various other text encoder transformers have been proposed, such as RoBERTa [41], DistilBERT [56], and ERNIE [68, 69]. While these models differ on some architectural details and on precise pretraining objectives, they all share the basic transformer architecture and the use of denoising pretraining objectives. In particular, they are all trained on unimodal data, meaning that they are only exposed to text during training.

In contrast to unimodally trained text encoders, V&L

---

[2]This phenomenon is illustrated in this interview with Tommy Edison, a congenitally blind man, in which he describes his understanding and frequent confusion regarding color associations.

models have been exposed to both text and image data during training. These models are typically used for tasks that require joint reasoning on text and images, such as visual question answering, grounding referring expressions, and vision-language retrieval [11, 18]. Fusion encoder models such as LXMERT [70], UNITER [12], ViLT [31], and ALBEF [39] output a fused representation of text and image data, while dual encoder models like CLIP [52] and ALIGN [26] consist of dual text and image encoder models that are jointly trained to produce embeddings in a shared semantic space. FLAVA, a vision-and-language model introduced recently by Singh et al. [65], also includes dual text and image encoders, and is trained on both multimodal objectives involving the alignment of images and text, as well as unimodal objectives. In this work we focus on the text encoder component of dual encoder models, since it may be used after V&L pretraining for text-only tasks.

Various works have explored the use of multimodal learning to benefit text understanding. Most related to our work is the very recent study of Zhang et al. [83] which investigates the use of unimodal and multimodal models for understanding visual commonsense in text. Their analysis follows a line of related work investigating the contribution of multimodal learning to visual commonsense knowledge in text, since such knowledge is typically not written explicitly in text but is abundantly present in visual information [35, 40, 75]. Unlike Zhang et al. [83] who only evaluate CLIP with an added set of task-specific learned weights, we are able to probe CLIP and other similar models in the strictly zero-shot setting via our novel Stroop probing method. This allows for directly evaluating properties learned by the models, independent of differences that result, for instance, from specific training configurations. In addition, we also study performance on both visual and non-visual NLU tasks in order to provide a controlled benchmark.

Other works have investigated the use of multimodal learning for NLU in various contexts. Bruni et al. [7] propose an architecture for integrating text and image-based distributional information to improve performance on tasks where meaning is grounded in perception. Kiela and Bottou [29] show that integrating features extracted from images using CNN with skip-gram representation vectors improves performance on semantic similarity datasets. Lazaridou et al. [36] train visual representations extracted using CNNs together with skip-gram embeddings to integrate visual and textual information performing well in both semantic and vision tasks. Kiela et al. [30] train a sentence embedding model using grounded information extracted from image features by attempting to predict the image features. These embeddings improve performance on various NLP tasks in comparison to text only embeddings. They show that using this method for a dataset consisting mainly of abstract words is likely to less benefit from grounding information. Shi et al. [62] show that a syntactic parser can benefit from seeing images during training; however, it was later shown that the model mostly relies on noun concreteness (which we also elaborate on in our work) rather than more complex syntactic reasoning [34]. The use of images for PCFG induction is also investigated by Jin & Schuler [27].

Along with the rise in visibility of jointly trained V&L transformer models, a number of works have explored the use of these models for text-only tasks, with mixed results. Using terms coined by Sileo et al. [64], these can be broadly split into *associative grounding* and *transfer grounding* approaches. Associative grounding uses retrieval methods to associate particular images with related texts; Kiros et al. [33] and Tan & Bansal [71] show that associative grounding methods may improve performance on various text-only NLU benchmark tasks. Transfer grounding applies V&L models directly to text-only input, disregarding the vision component of the model during inference. Wang et al. [78] apply this to weakly-supervised V&L models to outperform BERT on various text-only tasks from the GLUE benchmark. On the other hand, Iki & Aizawa [24] find that V&L-pretrained text encoders have similar or inferior results on NLU tasks including tasks from GLUE. Likewise, Cao et al. [9] find that although visually-aligned text encoders perform well on semantic linguistic probing tasks, BERT still outperforms them.

As discussed above, some prior works suggest that multimodal pretraining aids text understanding while other works show that it can lead to degradation. Our work provides new context for these seemingly contradictory results, allowing them to be reassessed in the new context of visual vs. non-visual natural language understanding.

3. Experimental Setup

3.1. Models Used

For evaluating unimodally trained text encoders, we use BERT [15], RoBERTa [41], DistilBERT and DistilRoBERTa [56], which are all trained with text-only MLM objectives. We also include results for Sentence-BERT (SBERT) [54], since its output embeddings are trained to have meaningful cosine similarity scores and thus bear more similarity to other models evaluated with Stroop proving. Results on multimodally trained text encoders are reported for CLIP [52] and FLAVA [65]; for these models we use only the text encoder with pretrained weights and discard the other subcomponents. Our tests include checkpoints from both OpenAI and the OpenCLIP open-source implementation of CLIP [13, 25]. Details of the checkpoints used for each model are listed in the supplementary material.

The text encoders of the multimodally trained models range in size from 63M (CLIP) to 109M (FLAVA) param-

eters. We compare to both comparably small unimodally trained text encoders such as DistilBERT (66M parameters) as well as much larger text encoders such as BERT-large (340M). See the supplementary material for an exhaustive list of sizes of the models under consideration.

We use each model with frozen pretrained weights. Our subsequent tests probe the contents of the feature vectors extracted by these models. For MLM probing, we also use the model's MLM head for prediction. In cases where MLM can be used we have found it to outperform Stroop probing; in such cases we report results for MLM probing here and for Stroop probing in the supplementary material.

3.2. Probing Methods

In order to probe the inherent knowledge of our models, we use the knowledge probing methods described below. The probing methods that follow are strictly zero-shot; in the supplementary material we analyze the use of linear classifiers trained on our models' frozen embeddings ("linear probing").

Masked language modelling (MLM). BERT and our other unimodally trained models were all pretrained with MLM objectives and hence can be used for zero-shot prediction of tokens in a masked context. Given a text including a [MASK] token and a set of k possible completions $C = \{c_1, c_2, \cdots, c_k\}$, a MLM assigns probabilities $p_1, \cdots, p_k$ to each corresponding token. We use $\arg\max_i p_i$ as the model's prediction. Previous works have found that BERT and other MLM can be probed for innate knowledge with this method [48, 55].

Stroop probing (SP). We propose another zero-shot probing method to extract knowledge from models based on the pooled embeddings that they extract. Consider a masked text t_m and possible completions $c \in C$, and let t_c be the text with c inserted in the mask location. Given a text encoder M, we calculate pooled embeddings $v_m = M(t_m)$ and $v_c = M(t_c)$ and unit-normalize them to $\hat{v}_m = v_m/\|v_m\|$ and $\hat{v}_c = v_c/\|v_c\|$. Stroop probing considers the cosine similarity scores $s_c := \hat{v}_m \cdot \hat{v}_c$. These can be used either directly for regression (as in the concreteness task below), or for categorical prediction by selecting $c^* = \arg\max_c s_c$.

The intuition behind Stroop probing is that items which are more surprising, incongruous, or salient in the given context may have a stronger interference effect on the encoding of the surrounding text. This is analogous to the *Stroop effect* in human psychology. When presented with congruent and incongruent stimuli such as color words printed in the same or differing colors (e.g. "red" printed in blue), readers take significantly longer on average to read the incongruent stimuli, a phenomenon known as the *Stroop*

effect [43][3]. We use Stroop probing for multiple tasks, including predicting color associations, as described below.

3.3. Prompts Used

For each task, we test the probing methods above on a wide variety of prompts in order to show the robustness of the described phenomena. In our results below we report the maximum metric value for each model over all of the prompts, since this represents a rough bound on our ability to extract intrinsic knowledge from the models under consideration. A full list of prompts used for each task and an analysis of model performance across prompts are provided in the supplementary material.

In some cases our prompt contains an empty slot, which we indicate below as [*]. Some models under consideration have a dedicated mask token, but for those such as CLIP that do not, we insert a fixed token in this slot, detailed further in the supplementary material.

4. VLU Tasks

We present three VLU tasks to probe the ability of our models to understand language with implied visual context: concreteness prediction (Section 4.1), color association prediction (Section 4.2) and shape association prediction (Section 4.3). Note that each of these tasks is performed on text alone, but requires visual reasoning to complete.

4.1. Concreteness Prediction

Task description. Words and phrases can be roughly classified as either *concrete* or *abstract*. A concrete concept is something that can be pointed to or directly sensed, while abstract concepts refer to things that cannot be easily visualized [60]. This can be conceptualized on a scale, ranging from the most abstract to the most concrete. Psychological research suggests that concrete words are easier for humans to understand and remember than abstract words [60]. Similarly, it has been shown that concreteness correlates with the learnability of visual concepts for machine learning models [20], and that MLM pretraining of V&L models may be improved by preferentially masking concrete words [4].

Because concreteness is a property of text that is tightly coupled with the visual domain, we consider concreteness prediction to be a VLU task, requiring some knowledge of the visual content of language to complete. We note that this task has been addressed in various previous works [10, 20, 22, 23, 51]. In contrast to these approaches, our unsupervised concreteness estimation procedure evaluates the concreteness of a word or phrase in a given textual context, rather than being limited to a fixed set of lexical items or discrete categories in a dataset.

---

[3]For example, try saying these colors out loud (not the printed words): Green, Red, Blue, Purple, Red, Purple.

Experimental details. We probe our models for the concreteness of words in context by using a cloze task paradigm with Stroop probing. For example, using the prompt $t_m =$ *"I see the [*]"* and testing word $\langle w \rangle$, we insert the word into the prompt to obtain $t_w =$ *"I see the $\langle w \rangle$"*, and use cosine similarity score s_w between embeddings of t_m and t_w as the regression output. All prompts used are listed in the supplementary material.

We test our approach on the dataset introduced by Brysbaert et al. [8]. This dataset contains 39,954 English unigrams and bigrams along with human-labelled concreteness scores on a scale from 1 (abstract) to 5 (concrete), averaged over annotators. We only use the unigram nouns from this list, totaling 14,562 items. Note that unlike prior concreteness prediction techniques that train supervised models on this dataset [10], we perform zero-shot prediction on this task with no supervised training, using the dataset for testing only.

Also note that we do not report results for DistilBERT or DistilRoBERTa since the checkpoints used do not contain a trained pooling layer, which is required for Stroop probing.

Evaluation metrics. We report absolute values of Pearson, Spearman, and Kendall correlations between the predicted concreteness and ground truth scores ($|\rho|$, $|r_s|$, and $|\tau|$ respectively).

4.2. Color Association Prediction

Task description. Some concepts are highly associated with particular colors—for example, the word *banana* is highly associated with the color yellow, while a word like *child* does not have a strong color association. These color associations have been widely studied in experimental psychology and neuroscience [3,5]. We propose a task of *color association prediction* – given a noun (or noun phrase) $\langle w \rangle$, identify the color with which $\langle w \rangle$ is normally associated.

Experimental details. To probe our models for color associations, we use the MLM and SP methods described above. In particular, we conceive of this task as categorical prediction over a set of basic color words C. For example, using the prompt *"A picture of a [*] $\langle w \rangle$"* where $\langle w \rangle$ is the item being tested, our probing methods search for the most suitable color to place in the [*] slot. All prompts tested are listed in the supplementary material. For MLM probing, we predict the color $c \in C$ with the highest predicted probability in the [*] slot of the prompt. For SP, we predict $c^* = \arg\max s_c$ using similarity scores as defined above.

To test this method on our chosen text encoders, we use two datasets. The Color Terms Dataset (CTD) [6] provides a list of 52 concrete words and their color. The Natural-Color Dataset (NCD) of fruit images [2] is a colorization task containing images of 20 types of fruits and vegetables paired with colors. We use the provided list of fruits and colors as a fixed set of words with strong color associations, discarding the image data. For the latter, we filter objects with the color label *purple* as this label contains multiple WordPiece tokens and thus is not directly comparable with MLM probing for models such as BERT. This results in 15 unique fruits and vegetables. For each model, we calculate color predictions using the probing methods described above out of the set: {*red, orange, yellow, green, blue, black, white, grey, brown*}.

Evaluation metrics. We report categorical accuracy of predictions on the CTD and NCD datasets (acc_{CTD} and acc_{NCD}) relative to the ground truth color labels.

4.3. Shape Association Prediction

Task description. Another salient visual feature of language is the association between concrete nouns or noun phrases and particular shapes. For example, the nouns *wheel* and *compass* have a circular association, while *pyramid* and *corn chip* have a triangular association. Shape associations have been studied in the psychological literature in contexts such as child language acquisition [80] and semantic memory representation [76]. Building on this line of research, we propose the task of *shape association prediction* – given a noun (or noun phrase) $\langle w \rangle$, identify the basic shape that is most associated with $\langle w \rangle$. Because the space of possible shapes is complex and difficult to categorize unambiguously, we restrict $\langle w \rangle$ under consideration to nouns associated with a few basic shapes, as described below.

Experimental details. We construct the **ShapeIt** benchmark for shape associations[4]. This contains 109 items total, each consisting of a noun or noun phrase along with the basic shape most associated with it from the set {$rectangle, circle, triangle$}. The benchmark was constructed by performing a user study requiring users to choose a shape associated with a given word, and selecting for only those words which were consistently classified by the users. Data collection methods used in constructing this benchmark are detailed in the supplementary material, along with further analysis of its contents. Probing methods used for this task are equivalent to the color association prediction task. Prompts used for probing include *"A [*] shaped $\langle w \rangle$"* where $\langle w \rangle$ is the shape associated word; the full list of prompts used is detailed in the supplementary material. We use both shape nouns (e.g. *circle*) and associated adjectives (e.g. *circular*) and report the highest accuracy achieved between these two settings.

Evaluation metric. We report categorical accuracy of predictions (acc) relative to the ground truth shape labels.

_____

[4]available in our code repository

5. Non-visual NLU Tasks

We also present three non-visual NLU tasks to serve as a baseline comparison for our models:

5.1. Factual Knowledge Probing

Task description. It has been observed that language models have an emergent *knowledge base* property, in which they may be probed for factual knowledge about the world [48, 55]. Various works on probing BERT and other language models for commonsense world knowledge have found that they show an impressive ability to memorize knowledge, although they may be deficient in applying this knowledge to reasoning about the physical world [16]. In this task, we probe our models for fine-grained factual knowledge via a cloze test, where an empty slot must be filled in with a word. We test on factual knowledge about geographical locations since this requires factual knowledge that does not explicitly rely on visual reasoning.

Experimental details. For this task, we use the Comparative Question Completion dataset introduced by [82]. This consists of questions in which one of a pair of coordinated elements is masked; the target is the masked phrase. Specifically, we use the *cities* dataset which masks the names of geopolitical entities such as cities and countries. Example sentences from the dataset include: *which country has more part time jobs new zealand or [*]?* (the correct answer being *australia*) and *which is older saudi arabia or [*]?* (the correct answer being *persia*). The original dataset has 1,187 questions with 447 unique locations as answers. In order to fit the general method of masking tasks, we filter masked phrases with more than one token (e.g. *the west coast*) similar to the protocol presented in the original paper. As this results in an extremely limited set of candidates for MLM models such as RoBERTa that use Byte Pair Encoding tokenization, we restrict the MLM models under comparison to BERT and DistilBERT. The filtered dataset contains 825 questions with 216 unique locations.

We treat this task as a categorical classification task, choosing only from the set of unique locations given in the dataset per sample, and evaluating how often the correct target is chosen. We use MLM probing and Stroop probing for categorical prediction as described above. Similarly to our other tasks, the intuition is that more surprising completions should have a larger interference effect on the text's encoding, if the relevant information is encoded in the embedding.

Evaluation metrics. We report recall at one and five ($R@1$, $R@5$), measuring how often the ground truth answer is found among the model's top one or five predicted candidates.

5.2. Language Proficiency Probing

Task description. In order to evaluate our models' intrinsic knowledge of general language usage, we consider the task of reconstructing English text in order to produce natural-sounding language. Multiple-choice cloze tests are commonly used in language assessment tasks for students to evaluate their proficiency [1,67,72]. Similarly, a model with a good grasp of English language usage should be able to fill in missing words in cloze contexts to produce fluent English. This requires grammatical and semantic knowledge, but in general, it is not directly related to visual reasoning when applied to arbitrary masked contexts. As noted by Trace [72], cloze tasks may evaluate global reading comprehension or local contextual information in the cloze context; we focus on the latter case and refer to this task applied to our models as *language proficiency probing*.

Experimental details. To evaluate language proficiency, we use the Children's Book Test (CBT) cloze dataset provided by Meta research [21], consisting of book passages with accompanying masked sentences and possible mask completions. We discard the book passages and only consider the sentences and completions, to focus on the task of reconstructing well-formed text. Completions are grouped by part of speech (POS); we use the noun (N), verb (V), and preposition (P) groups and discard the named entity groups since the latter require long-distance context to predict while N, V, and P can often be inferred from local sentential context. In total, each of the N, V, and P groups contains 2,500 sentences with 10 possible completions each. We filter out long sentences since our multimodally trained models have shorter expected input lengths. After filtering we are left with 1,588 noun, 1,747 verb, and 2,382 preposition completion sentences. In addition, we only use sentences that have a one-word token answer for all tokenizers. For example, one sentence from the V group is *I [*] not a fellow; I am a young lady!* and the set of possible completions is {*am, born, find, picking, pricked, said, sat, seems, streamed, thinking*}. We use MLM and Stroop probing to evaluate our models on this data.

Evaluation metrics. We report categorical accuracy per POS group (acc_V, acc_N, and acc_P), measuring how often the ground truth answer is selected in each of these groups.

5.3. Sentiment Analysis

Task description. *Sentiment analysis* refers to the task of predicting speaker emotion or affect, a well-studied problem in natural language processing [28, 38, 49]. We focus on sentiment analysis in text as a subset of text classification. Since text describing the same visual scene may have a positive or negative sentiment (*This cake is delicious* vs.

This cake tastes bad), we consider this task to be a non-visual NLU task.

Experimental details. For this task, we use the IMDB movie review dataset consisting of 50K movie reviews with binary sentiment labels (positive/negative) [42]. In order to provide a fairer comparison between models rather than biasing towards models trained on longer texts, we use only a single random sentence from each review in the IMDB dataset. In addition, we filter long sentences which are too long for multimodal encoders leaving 42,567 examples. Using only a single sentence makes this task more challenging since the randomly chosen sentence is not guaranteed to contain sufficient context for understanding the review's sentiment, but we find that significantly better than random performance is achievable, as seen in the results section. We also differ from the more common learned sentiment analysis paradigm by using strictly zero-shot prediction via engineered prompts. For example, one prompt used is *"sentiment expressed for the movie is [*]. $\langle s \rangle$"*, where $\langle s \rangle$ indicates the sentence chosen from the initial review, and [*] may be filled with one of {*good, bad*}. We apply MLM and Stroop probing for binary prediction, and report categorical accuracy achieved for each model.

Evaluation metric. We report categorical accuracy of predictions (acc) relative to the ground truth sentiment labels.

6. Results and Discussion

Results for the tasks described above are provided in Table 1. For tasks with multiple prompts the listed metrics are the maximum over prompts, providing a rough upper bound on each model's ability to perform the task in question. Further analysis of performance by prompt, as well as SP results for models shown here with MLM, are provided in the supplementary materials.

As seen in these results, multimodally trained models consistently outperform unimodally trained models on VLU tasks, including both comparably sized and much larger text encoders, while generally underperforming them on non-visual NLU tasks. This is further illustrated by qualitative analysis of the results in various tasks.

Figure 2 shows the results of concreteness prediction for CLIP and BERT. Nouns predicted as most concrete by CLIP, for example *bench* and *chalk*, that can be clearly visualized, while nouns predicted as least concrete (i.e., abstract) such as *story* and *name*, do not have a clear visual representation. In comparison, BERT's predictions are significantly noisier, with nouns such as *seed* and *jelly* predicted as abstract.

Figures 3 and 4 shows color and shape association predictions of BERT-base and CLIP on samples from the relevant datasets. Without having access to the associated images, the CLIP text encoder usually predicts the correct

snowman	sink	seed	friend
liar	bench	jelly	story
lettuce	chalk	cash	name
mailman	splinter	lightning	thanks
couch	pinecone	pudding	fun
BERT	CLIP	BERT	CLIP
Most *concrete*		**Most *abstract***	

Figure 2. Basic nouns selected as most (and least) concrete using BERT-base and CLIP, according the method described in Section 4.1. As these illustrate, concreteness can be reasonably predicted from CLIP text embeddings, whereas this knowledge is not readily accessible for the unimodally trained text encoders.

BERT	green	green	red	red	white
CLIP	green	orange	yellow	red	brown

Figure 3. **Examples of color prediction results on NCD dataset**. Depicted above are examples of results for predicting colors from the NCD dataset using MLM for BERT-base and Stroop probing for CLIP. We emphasize that the model only receives as input the name of the fruit or vegetable without the given image.

Item	pill	saucer	prism	notebook
BERT	triangle	rectangle	triangle	circle
CLIP	circle	circle	triangle	rectangle

Figure 4. **Examples of shape prediction results on ShapeIt benchmark**. Depicted above are examples of results for predicting shapes from the ShapeIt benchmark using MLM for BERT-base and Stroop probing for CLIP. Images are for illustration only, but during probing the model only receives the name of the item.

matching between the given item and its correct shape or color, while BERT fails in most cases. Our results suggest that these associations are more consistently encoded by multimodally trained encoders. Furthermore, qualitative analysis of the misclassifications of CLIP, OpenCLIP and FLAVA on color association prediction reveals that these are mostly due to ambiguities in the dataset itself; see the supplementary materials for details.

Performance on non-visual NLU tasks, shown on the right side of Table 1, demonstrates that our results are not an artifact of our probing methodology providing a global advantage to multimodally trained models, nor are these models uniformly better at language-related tasks. We also see

	Concreteness			Color		Shape	Knowledge		Proficiency			Sent.
Metric	$\lvert\rho\rvert$	$\lvert r_s\rvert$	$\lvert\tau\rvert$	acc_{CTD}	acc_{NCD}	acc	R@1	R@5	acc_{V}	acc_{N}	acc_{P}	acc
Unimodal												
BERT-base	0.414	0.416	0.283	0.353	0.400	0.559	0.198	0.522	0.898	0.753	0.893	0.618
BERT-large	0.348	0.355	0.239	0.490	0.467	0.587	**0.231**	**0.541**	**0.914**	**0.779**	0.905	0.625
DistilBERT	–	–	–	0.333	0.400	0.587	0.148	0.479	0.864	0.709	0.814	0.637
RoBERTa-base	0.433	0.404	0.275	0.431	0.333	0.431	–	–	0.877	0.718	0.881	0.666
RoBERTa-large	0.345	0.374	0.253	0.471	0.400	0.431	–	–	0.898	0.765	**0.916**	**0.703**
DistilRoBERTa	–	–	–	0.411	0.333	0.431	–	–	0.804	0.664	0.756	0.635
ERNIE	0.461	0.496	0.338	0.196	0.333	0.449	0.001	0.006	0.064	0.051	0.086	0.582
ERNIE-large	0.358	0.353	0.233	0.216	0.267	0.458	0.006	0.022	0.209	0.241	0.280	0.674
SBERT	0.338	0.337	0.228	0.198	0.067	0.513	0.013	0.141	0.237	0.158	0.126	0.554
V&L												
CLIP	0.603	0.624	0.437	0.843	**0.800**	0.798	0.009	0.118	0.134	0.133	0.126	0.560
OpenCLIP	**0.634**	0.643	0.432	**0.941**	**0.800**	**0.853**	0.009	0.121	0.211	0.135	0.123	0.560
FLAVA	0.608	**0.665**	**0.449**	0.882	**0.800**	0.798	0.020	0.138	0.116	0.118	0.139	0.519

Table 1. **Results on VLU (left) and non-visual NLU (right) tasks:** concreteness prediction, color and shape association prediction, factual knowledge probing, language proficiency probing, and sentiment analysis (Sent.) respectively. For tasks other than concreteness prediction, MLM probing is used for models supporting it (BERT, DistilBERT, RoBERTa, DistilRoBERTa); SP is used elsewhere. The definition of each metric is defined in the relevant task definition in Sections 4-5. DistilBERT and DistilRoBERTa do not have concreteness results due to the pooling layer issue mentioned in Section 4.1, and RoBERTa and DistilRoBERTa do not have results for factual knowledge probing due to the tokenization issue mentioned in Section 5.1. As these results show, V&L models yield superior performance on visual tasks, while underperforming unimodally trained models on non-visual NLU tasks.

that the non-visual tasks are highly solveable, with BERT-large and RoBERTa-large achieving high performance on all tasks despite the challenging zero-shot regime and limited information in the task inputs (ambiguity in cloze contexts for factual probing, lack of textual textual context for proficiency probing and randomly-chosen sentences for sentiment analysis). Despite this, the multimodally trained models show near-random performance.

We note a direct connection to the original Stroop effect in the field of human psychology. Follow-up works to the first Stroop effect demonstration have found it to apply to various types of stimuli, such as color-congruent and incongruent objects (e.g. a yellow banana vs. a purple banana) [46]. Our results, also including color congruence of objects, strengthen the motivation for using Stroop probing applied to tasks involving visual congruence or saliency.

We also note a connection between our results and the *reporting bias* effect, in which commonsense properties are less likely to be explicitly stated than incongruent properties (e.g. *a (yellow) banana* vs. *a blue banana*). Reporting bias in text has been studied in the context of color associations [47] and in more general contexts [19, 63]. As the multimodally trained models under consideration were trained on paired image-caption data, the distribution of text in image captions differs somewhat from the text used for training models such as BERT. In the supplementary material, we provide an analysis of reporting bias in the LAION dataset [59], the training data for the OpenCLIP model in-

cluded in our tests. These results provide evidence that the improvement in performance seen from V&L training cannot primarily be attributed to a lack of reporting bias in image caption texts, and emphasizes the significance of the visual modality in these models' language understanding.

7. Conclusion

We propose a suite of visual language understanding tasks along with non-visual natural language understanding tasks to probe the effect of V&L pretraining on such reasoning capabilities of text encoder models. We introduce Stroop probing as a zero-shot knowledge proving method for evaluating the innate knowledge of text encoders. We also show that exposure to V&L data in pretraining improves the performance of text encoder models on VLU tasks, even though they may underperform unimodally trained text encoders on non-visual NLU tasks. Beyond text-only tasks, these results bear importance in the broader context of multimodal learning, in which the isolated contribution of text encoders has previously been underexplored. Our findings suggest that multimodal pretraining has a significant effect on the knowledge represented by the text encoder component of multimodal models, facilitating in establishing best practices for the design and training of text encoders used in such contexts.

References

[1] J Charles Alderson. The cloze procedure and proficiency in english as a foreign language. *TESOL quarterly*, pages 219–227, 1979. 6

[2] Saeed Anwar, Muhammad Tahir, Chongyi Li, Ajmal Mian, Fahad Shahbaz Khan, and Abdul Wahab Muzaffar. Image colorization: A survey and dataset. *arXiv preprint arXiv:2008.10774*, 2020. 5

[3] Michael M Bannert and Andreas Bartels. Decoding the yellow of a gray banana. *Current Biology*, 23(22):2268–2272, 2013. 5

[4] Yonatan Bitton, Gabriel Stanovsky, Michael Elhadad, and Roy Schwartz. Data efficient masked language modeling for vision and language. *arXiv preprint arXiv:2109.02040*, 2021. 4

[5] Inês Bramão, Alexandra Reis, Karl Magnus Petersson, and Luís Faísca. The role of color information on object recognition: A review and meta-analysis. *Acta psychologica*, 138(1):244–253, 2011. 5

[6] Elia Bruni, Gemma Boleda, Marco Baroni, and Nam-Khanh Tran. Distributional semantics in technicolor. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 136–145, 2012. 5

[7] Elia Bruni, Nam-Khanh Tran, and Marco Baroni. Multimodal distributional semantics. *Journal of artificial intelligence research*, 49:1–47, 2014. 3

[8] Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior research methods*, 46(3):904–911, 2014. 5

[9] Jize Cao, Zhe Gan, Yu Cheng, Licheng Yu, Yen-Chun Chen, and Jingjing Liu. Behind the scene: Revealing the secrets of pre-trained vision-and-language models. In *European Conference on Computer Vision*, pages 565–580. Springer, 2020. 3

[10] Jean Charbonnier and Christian Wartena. Predicting word concreteness and imagery. In *Proceedings of the 13th International Conference on Computational Semantics-Long Papers*, pages 176–187. Association for Computational Linguistics, 2019. 4, 5

[11] Feilong Chen, Duzhen Zhang, Minglun Han, Xiuyi Chen, Jing Shi, Shuang Xu, and Bo Xu. Vlp: A survey on vision-language pre-training. *arXiv preprint arXiv:2202.09061*, 2022. 2, 3

[12] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer, 2020. 2, 3

[13] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. *arXiv preprint arXiv:2212.07143*, 2022. 3

[14] Yuqing Cui, Apoorv Khandelwal, Yoav Artzi, Noah Snavely, and Hadar Averbuch-Elor. Who's waldo? linking people across text and images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1374–1384, 2021. 12

[15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 2, 3, 16

[16] Maxwell Forbes, Ari Holtzman, and Yejin Choi. Do neural language representations learn physical commonsense? *arXiv preprint arXiv:1908.02899*, 2019. 6

[17] W Nelson Francis and Henry Kucera. A standard corpus of present-day edited american english, for use with digital computers. *Brown University, Providence*, 1964. 12

[18] Zhe Gan, Linjie Li, Chunyuan Li, Lijuan Wang, Zicheng Liu, and Jianfeng Gao. Vision-language pre-training: Basics, recent advances, and future trends. *arXiv preprint arXiv:2210.09263*, 2022. 3

[19] Jonathan Gordon and Benjamin Van Durme. Reporting bias and knowledge acquisition. In *Proceedings of the 2013 workshop on Automated knowledge base construction*, pages 25–30, 2013. 8, 14

[20] Jack Hessel, David Mimno, and Lillian Lee. Quantifying the visual concreteness of words and topics in multimodal datasets. *arXiv preprint arXiv:1804.06786*, 2018. 4

[21] Felix Hill, Antoine Bordes, Sumit Chopra, and Jason Weston. The goldilocks principle: Reading children's books with explicit memory representations. *arXiv preprint arXiv:1511.02301*, 2015. 6

[22] Felix Hill and Anna Korhonen. Learning abstract concept embeddings from multi-modal data: Since you probably can't see what i mean. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 255–265, 2014. 4

[23] Felix Hill, Roi Reichart, and Anna Korhonen. Multi-modal models for concrete and abstract concept meaning. *Transactions of the Association for Computational Linguistics*, 2:285–296, 2014. 4

[24] Taichi Iki and Akiko Aizawa. Effect of visual extensions on natural language understanding in vision-and-language models. *arXiv preprint arXiv:2104.08066*, 2021. 2, 3

[25] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, July 2021. 3, 16

[26] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR, 2021. 3

[27] Lifeng Jin and William Schuler. Grounded pcfg induction with images. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, 2020. 3

[28] Dan Jurafsky and James H Martin. Speech and language processing (3rd draft ed.), 2023. Chapter 4. Naive Bayes and Sentiment Classification. 6

[29] Douwe Kiela and Léon Bottou. Learning image embeddings using convolutional neural networks for improved multi-modal semantics. In *Proceedings of the 2014 Conference on empirical methods in natural language processing (EMNLP)*, pages 36–45, 2014. 3

[30] Douwe Kiela, Alexis Conneau, Allan Jabri, and Maximilian Nickel. Learning visually grounded sentence representations. *arXiv preprint arXiv:1707.06320*, 2017. 3

[31] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pages 5583–5594. PMLR, 2021. 2, 3

[32] Yongmin Kim, Chenhui Chu, and Sadao Kurohashi. Flexible visual grounding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 285–299, 2022. 12

[33] Jamie Kiros, William Chan, and Geoffrey Hinton. Illustrative language understanding: Large-scale visual grounding with image search. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 922–933, 2018. 3

[34] Noriyuki Kojima, Hadar Averbuch-Elor, Alexander M Rush, and Yoav Artzi. What is learned in visually grounded neural syntax acquisition. *arXiv preprint arXiv:2005.01678*, 2020. 3

[35] Satwik Kottur, Ramakrishna Vedantam, José MF Moura, and Devi Parikh. Visual word2vec (vis-w2v): Learning visually grounded word embeddings using abstract scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4985–4994, 2016. 3

[36] Angeliki Lazaridou, Nghia The Pham, and Marco Baroni. Combining language and vision with a multimodal skip-gram model. *arXiv preprint arXiv:1501.02598*, 2015. 3

[37] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*, 2019. 12

[38] Jiwei Li and Eduard Hovy. Reflections on sentiment/opinion analysis. In *A practical guide to sentiment analysis*, pages 41–59. Springer, 2017. 6

[39] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021. 2, 3

[40] Xiao Lin and Devi Parikh. Don't just listen, use your imagination: Leveraging visual common sense for non-visual tasks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2984–2993, 2015. 3

[41] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. 2, 3, 16

[42] Andrew Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pages 142–150, 2011. 7

[43] Colin M MacLeod. Half a century of research on the stroop effect: an integrative review. *Psychological bulletin*, 109(2):163, 1991. 2, 4

[44] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L. Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *CVPR*, pages 11–20, 2016. 12

[45] Gloria Strauss Marmor. Age at onset of blindness and the development of the semantics of color names. *Journal of experimental child psychology*, 25(2):267–278, 1978. 2

[46] Galit Naor-Raz, Michael J Tarr, and Daniel Kersten. Is color an intrinsic property of object representation? *Perception*, 32(6):667–680, 2003. 8

[47] Cory Paik, Stéphane Aroca-Ouellette, Alessandro Roncone, and Katharina Kann. The world of an octopus: How reporting bias influences a language model's perception of color. *arXiv preprint arXiv:2110.08182*, 2021. 8, 14

[48] Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H Miller, and Sebastian Riedel. Language models as knowledge bases? *arXiv preprint arXiv:1909.01066*, 2019. 2, 4, 6

[49] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, and Rada Mihalcea. Beneath the tip of the iceberg: Current challenges and new directions in sentiment analysis research. *IEEE Transactions on Affective Computing*, 2020. 6

[50] Tingyu Qu, Tinne Tuytelaars, and Marie-Francine Moens. Weakly supervised face naming with symmetry-enhanced contrastive loss. *arXiv preprint arXiv:2210.08957*, 2022. 12

[51] Ella Rabinovich, Benjamin Sznajder, Artem Spector, Ilya Shnayderman, Ranit Aharonov, David Konopnicki, and Noam Slonim. Learning concept abstractness using weak supervision. *arXiv preprint arXiv:1809.01285*, 2018. 4

[52] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 2, 3, 16

[53] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 12

[54] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019. 3, 16

[55] Anna Rogers, Olga Kovaleva, and Anna Rumshisky. A primer in bertology: What we know about how bert works. *Transactions of the Association for Computational Linguistics*, 8:842–866, 2020. 2, 4, 6

[56] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019. 2, 3, 16

10

[57] Armin Saysani, Michael C Corballis, and Paul M Corballis. Colour envisioned: Concepts of colour in the blind and sighted. *Visual cognition*, 26(5):382–392, 2018. 2

[58] Armin Saysani, Michael C Corballis, and Paul M Corballis. Seeing colour through language: colour knowledge in the blind and sighted. *Visual Cognition*, 29(1):63–71, 2021. 2

[59] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *arXiv preprint arXiv:2210.08402*, 2022. 8, 14

[60] Paula J Schwanenflugel. Why are abstract concepts hard to understand? In *The psychology of word meanings*, pages 235–262. Psychology Press, 2013. 4

[61] Roger N Shepard and Lynn A Cooper. Representation of colors in the blind, color-blind, and normally sighted. *Psychological science*, 3(2):97–104, 1992. 2

[62] Haoyue Shi, Jiayuan Mao, Kevin Gimpel, and Karen Livescu. Visually grounded neural syntax acquisition. *arXiv preprint arXiv:1906.02890*, 2019. 3

[63] Vered Shwartz and Yejin Choi. Do neural language models overcome reporting bias? In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6863–6870, 2020. 8, 14

[64] Damien Sileo. Visual grounding strategies for text-only natural language processing. *arXiv preprint arXiv:2103.13942*, 2021. 3

[65] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15638–15650, 2022. 3, 16

[66] Shane Storks, Qiaozi Gao, and Joyce Y Chai. Recent advances in natural language inference: A survey of benchmarks, resources, and approaches. *arXiv preprint arXiv:1904.01172*, 2019. 13

[67] Joseph Bartow Stubbs and G Richard Tucker. The cloze test as a measure of english proficiency. *The Modern Language Journal*, 58(5/6):239–241, 1974. 6

[68] Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, and Hua Wu. Ernie: Enhanced representation through knowledge integration. *arXiv preprint arXiv:1904.09223*, 2019. 2, 16

[69] Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Hao Tian, Hua Wu, and Haifeng Wang. Ernie 2.0: A continual pre-training framework for language understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8968–8975, 2020. 2, 16

[70] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019. 3

[71] Hao Tan and Mohit Bansal. Vokenization: Improving language understanding with contextualized, visual-grounded supervision. *arXiv preprint arXiv:2010.06775*, 2020. 3

[72] Jonathan Trace. Clozing the gap: How far do cloze items measure? *Language Testing*, 37(2):235–253, 2020. 6

[73] Jeroen van Paridon, Qiawen Liu, and Gary Lupyan. How do blind people know that blue is cold? distributional semantics encode color-adjective associations. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 43, 2021. 2

[74] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2

[75] Ramakrishna Vedantam, Xiao Lin, Tanmay Batra, C Lawrence Zitnick, and Devi Parikh. Learning common sense through visual abstraction. In *Proceedings of the IEEE international conference on computer vision*, pages 2542–2550, 2015. 3

[76] Brian N Verdine, Kelsey R Lucca, Roberta M Golinkoff, Kathryn Hirsh-Pasek, and Nora S Newcombe. The shape of things: The origin of young children's knowledge of the names and properties of geometric forms. *Journal of Cognition and Development*, 17(1):142–161, 2016. 5

[77] Mingzhe Wang, Mahmoud Azab, Noriyuki Kojima, Rada Mihalcea, and Jia Deng. Structured matching for phrase localization. In *ECCV*, pages 696–711, 2016. 12

[78] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. Simvlm: Simple visual language model pretraining with weak supervision. *arXiv preprint arXiv:2108.10904*, 2021. 2, 3

[79] Adina Williams, Nikita Nangia, and Samuel R Bowman. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426*, 2017. 12, 14

[80] Eiling Yee, Stacy Huffstetler, and Sharon L Thompson-Schill. Function follows form: activation of shape and function features during object identification. *Journal of Experimental Psychology: General*, 140(3):348, 2011. 5

[81] Wenpeng Yin, Jamaal Hay, and Dan Roth. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. *arXiv preprint arXiv:1909.00161*, 2019. 13

[82] Avishai Zagoury, Einat Minkov, Idan Szpektor, and William W Cohen. What's the best place for an ai conference, vancouver or ⎵: Why completing comparative questions is difficult. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14292–14300, 2021. 6

[83] Chenyu Zhang, Benjamin Van Durme, Zhuowan Li, and Elias Stengel-Eskin. Visual commonsense in pretrained unimodal and multimodal models. *arXiv preprint arXiv:2205.01850*, 2022. 3

Task	Metric	VisualBERT	LMXERT	BERT	CLIP
VLU					
Conc.	Pearson	0.400	0.421	0.233	0.513
	Spearman	0.412	0.370	0.238	0.495
	Kendall	0.281	0.249	0.159	0.339
NCD	Accuracy	0.467	0.400	0.267	0.823
CTD	Accuracy	0.314	0.431	0.353	0.800
NLU					
Cites	R@1	0.003	0.007	0.199	0.019
NLI	AUC	0.704	0.688	0.754	0.696

Table 2. Evaluating non-dual V&L encoders (VisualBERT and LMXERT) on several VLU and NLU tasks with BERT and CLIP added for reference.

Appendix

A. Additional Results and Comparisons

A.1. Non-dual V&L encoder models

Although non-dual (fusion) encoder models are not directly comparable to purely textual encoders such as BERT or the text encoder component of CLIP which do not fuse modalities, we consider them here for completeness We evaluate the VisualBERT and LMXERT non-dual encoder models on several tasks from our task suite by only feeding them textual input. Results are shown in Table 2. As illustrated in the table, even though these models were trained using image features together with text tokens, the models outperform BERT on visual tasks, though the gap is not as significant as with dual encoder models.

A.2. Additional tasks using linear probing

We present two additional tasks for comparing V&L and unimodal models using linear probing, one VLU and one NLU task. For both tasks, we use a linear classifier on the pooled embedding output of a model for categorical prediction. Specifically, we use a logistic regression model using the scikit-learn `linear_model.LogisticRegression` implementation. For all tasks we use the default parameters except for `max_iter` which was changed according to task requirements to allow convergence. In particular, we use parameters `penalty='l2'`, `C=1.0`, `solver='lbfgs'`.

A.2.1 Groundability classification

Task description. In paired text-image data, there is normally an implied mapping between referential expressions in the text and objects or regions in the accompanying image. The task of learning these mappings is known as *visual grounding* and is of general interest for visual semantic understanding [14, 44, 77]. In captions accompanying images, some expressions refer directly to regions in images while others give non-visual context; we refer to the former as *groundable* referents and the latter as *non-groundable*. A similar paradigm was recently proposed by Kim et al. [32] that separately considers "answerable" and "unanswerable" phrases.

We propose a *groundability classification* task, consisting of classifying referents in text as groundable or non-groundable. This is a text-only task as it uses text alone and the visual context is only implied. Since this task requires visual imagination to complete, we consider it to be a VLU task.

Experimental details. In line with previous works that consider person-centric visual grounding [14, 50], we construct a dataset of *person-centric groundability* sentences where a fixed human participant is either implied to be groundable (i.e., on-camera) or non-groundable (i.e., off-camera). The associated task consists of binary classification applied to these texts according to whether the given participant would be visible in a description of an event. Due to the lack of existing labelled data for this task, we created a synthetic dataset of sentences with a common format: *Alex [MASK]ing Riley's [MASK]*, where the first masked word is a randomly drawn verb, and the second masked word is a randomly drawn noun, and the task is to classify whether or not the second mentioned individual (i.e., Riley) is groundable. Groundability labels are estimated using zero-shot text classification with a pretrained natural language inference model. We created synthetic data for the groundability task by taking the prompt template "Alex ___ing Riley's ___", filling in various verb-noun pairs into the given slots, and filtering using a pretrained language model to select for natural-sounding samples. We then estimated ground-truth labels using zero-shot inference with a pretrained natural language inference (NLI) model.

To find verb-noun pairs, we listed all verbs and nouns in the Brown corpus of standard American English with part-of-speech labels [17]. We converted all text to lowercase and then selected the 5,000 most common verb lemmas and 1,000 most common noun lemmas in this corpus. Using all possible verb-noun combinations among these, inserted into the prompt template shown above, yielded 5M candidate phrases. From the given 5M candidates we sample randomly 200K phrases. We then calculate the total negative log-likelihood (NLL) for each candidate relative to the pretrained language model GPT2-large [53] and kept only those samples in the 20th percentile of NLL (i.e. the most likely samples), corresponding to 40,000 descriptions.

After generating these texts, we estimated labels using a pretrained NLI model. We used BART-large [37] fine-tuned on the MNLI dataset [79] (using the

`facebook/bart-large-mnli` checkpoint from Hugging Face model hub[5]). This model takes pairs of texts as inputs (the "premise" and "hypothesis" texts) and outputs three probabilities per pair: p_c, p_n, p_e, corresponding to probabilities of a contradictory, neutral, or entailment relation between the texts respectively. As observed by Yin et al. [81], NLI can be used for zero-shot text classification by designing premise and hypothesis prompts for the task of interest. In our case, we use the following prompts:

Premise: "This is a picture of _____."
Hypothesis: "Riley can be seen in the picture."

For each of our 40,000 texts, we insert the text in the slot given in the premise and calculate p_e with the NLI model. If $p_e > 0.5$ we assign the sample label 1 (groundable), otherwise we assign it label 0 (non-groundable). Below are several example sentences from the synthetic dataset. Examples of Riley being groundable (sample label 1):

- Alex facing Riley's figure

- Alex viewing Riley's participation

- Alex seeing Riley's enjoyment

Examples of Riley being non-groundable (sample label 0):

- Alex hiding Riley's file

- Alex announcing Riley's absence

- Alex stealing Riley's evidence

For evaluation we created a test set, containing 200 sentences judged by human evaluators to be natural sounding, half labeled as groundable and the other half as non-groundable. To provide an example, sentences such as *Alex cutting Riley's hair* or *Alex blocking Riley's shot* were labeled as groundable, whereas sentences such as *Alex painting Riley's house* or *Alex counting Riley's vote* were labeled as non-groundable.

For this binary classification task, we apply linear probing to assess our models' understanding of groundability, and report ROC-AUC scores for each model. We also provide 95% confidence intervals, calculated using bootstrap resampling with 200 bootstraps, in order to analyze the robustness of these results.

Results and discussion. Results for the groundability classification are provided in Table 3. As these results illustrate, CLIP significantly outperforms all unimodally trained text encoders on average. We observe that the score gaps are not as distinct as in the previous zero-shot tasks, as this task is a learnable task which requires training, allowing all models to learn this task to some extent. Nonetheless, CLIP's ability to surpass the unimodally trained encoders suggest that

Model	AUC (95% CI)
BERT	0.789 ± 0.0007
RoBERTa	0.799 ± 0.0005
ERNIE	0.766 ± 0.0006
CLIP	$\mathbf{0.822 \pm 0.0007}$

Table 3. **Groundability Classification Evaluation**. We report ROC-AUC with 95% bootstrap confidence intervals scores for a manually assembled test set comparing linear probing for text based encoders and V&L CLIP model. As shown above, CLIP significantly outperforms the unimodally trained models.

V&L trained text encoders have a better ability to grasp if an object is grounded or not due to additional perceptual information that is encoded during the pretraining phase. Furthermore, note that in comparison to the other VLU tasks, here the subject in question (i.e., Riley) is not directly connected to visual information and the prediction is based only on context relating to the performed action and the associated object. The improved performance on this task illustrates that V&L models can better encode higher-level perceptual reasoning.

A.2.2 Natural language inference

Task description. Natural language inference (NLI) refers to inferring the logical relation between pairs of statements, as well as more generally referring to logical inference based on text [66]. In particular, NLI commonly considers the following logical relations between sentences A and B:

- **Contradiction**: For example, A=*It is rainy outside.* is contradicted by B=*It is sunny outside.*, since they cannot be simultaneously true.

- **Neutral**: For example, A=*It is rainy outside.* is neutral with regards to B=*It is summer.*, since A neither contradicts nor entails B.

- **Entailment**: For example, A=*It is cold and rainy outside.* entails B=*It is cold outside.*, since if A is true then B must also be true.

Solving this task requires an understanding of the fine-grained semantics of language and logical reasoning. On the other hand, visual cues are not tightly related to this task and are even potentially misleading. For example, the sentences *This cup contains grape juice.* and *This cup contains wine.* are contradictory even though the scenes they describe are visually identical. Therefore, we consider this to be a non-visual NLU task.

Model	AUC $\pm$ std
BERT	0.754 ± 0.001
RoBERTa	0.777 ± 0.001
ERNIE	$\mathbf{0.787} \pm 0.001$
CLIP	0.696 ± 0.001

Table 4. **NLI Evaluation**. We report ROC-AUC scores for the NLI task using linear probing, comparing text based encoders to the V&L CLIP model. As depicted above, the V&L trained text encoder is inferior to all other text based encoders for this non-visual language understanding task.

Experimental details. For this task we use the MNLI dataset introduced by Adina et al. [79]. We remove sentence pairs with a neutral relation and treat this as a binary classification task to predict sentence pairs as contradictory or entailing. We perform 5-fold cross validation on a dataset of 261,775 pairs of sentences using 80% of samples for training and 20% for testing.

For each sentence pair, we concatenate the sentences' two pooled embeddings and apply linear probing. Note that some models such as BERT include a special [SEP] token for encoding sentence pairs as a single unit, but we encode sentences separately and concatenate their embeddings in order to have a fair comparison between all models. We report the ROC-AUC score on the MNLI test set.

Results and discussion. Results for NLI are provided in Table 4. As shown in the table, text-based models outperform CLIP by a large margin. Similar to our findings regarding linguistic acceptability classification, we see that V&L trained models are less effective in tasks that do not incorporate perceptive information, suggesting that for non-visual tasks, V&L pretraining is not necessarily beneficial.

A.3. Comparing usage of SP on text based models

In the main paper we presented results for text models using MLM probing, and for CLIP using Stroop probing (SP). To allow for a full comparison between both types of models, and to strengthen the choice of using MLM probing for text based models, we present additional results comparing SP and MLM probing for text based models. Table 5 presents results for comparing SP and MLM probing methods for BERT and RoBERTa. As illustrated, using SP with unimodally trained models results in lower performance than using MLM probing with these models. This result supports our choice of using MLM probing for text based models trained to perform MLM tasks as the preferred probing method.

A.4. Additional task results information

We provide additional detailed results for our suite tasks including the mean and standard deviation of the results over all used prompts in Table 6.

A.5. Qualitative analysis for V&L model misclassifications on color prediction

Our results for color association prediction show that V&L models outperform unimodally trained text encoders in the given setting. Additional qualitative analysis of the results show that even the reported misclassifications of V&L models such as CLIP may be explained by ambiguities in the dataset itself. For example, the noun "ash" has ground truth value "grey" in our dataset, while CLIP with SP predicts the color "black", which is arguably also correct. Table 7 presents all of the objects from both color datasets misclassified by CLIP, containing the ground truth and the predicted color. As seen there, most of these predictions may be interpreted as valid colors for the given objects.

A.6. Analysis of reporting bias in LAION

Prior works have noted that commonsense properties that can be inferred from text are less likely to be explicitly stated than incongruent properties, notably including color terms(e.g. *a (yellow) banana* vs. *a blue banana*) [19,47,63]. In particular, text in image captioning datasets such as the web-scale LAION dataset [59] (used to train OpenCLIP) might have a different incidence of reporting bias than the text used to train models such as BERT. To disentangle this from the effect of training on the visual modality, we provide an analysis of reporting bias in LAION for color associations.

We use the `laion-2B-en` subset of 2.33 billion English-language image-caption pairs in the LAION-5B dataset, and estimate reporting bias by searching for bigram pairs (c, w) where c is a basic color term[6] and w is a unigram noun from our color association datasets (CTD and NCD). The empirical probability of color c immediately preceding w is $P(c|w) = n_{(c,w)}/n_w$, where n indicates the number of instances of the given ngram, and the associated color estimates are $\hat{c}_w = \arg\max_c P(c|w)$. For these estimates, the corresponding accuracy scores on the unigrams in our datasets are $acc_{\text{CTD}} = 0.549$ and $acc_{\text{NCD}} = 0.714$, significantly below the accuracies achieved by all of the multimodally trained models under consideration on these datasets for the color prediction task. We also provide qualitative examples in Table 8 showing the effect of reporting bias for various common nouns from these datasets. These results provide evidence that multimodally trained models' strong performance on VLU tasks cannot be explained away as stemming from a lack of reporting bias in the texts used to train them.

---

[6]one of {red, orange, yellow, green, blue, black, white, grey, brown}

	Color		Shape	Knowledge		Proficiency			Sent.
Metric	acc_{CTD}	acc_{NCD}	acc	R@1	R@5	acc_{V}	acc_{N}	acc_{P}	acc
BERT-MLM	**0.353**	**0.400**	**0.559**	**0.198**	**0.522**	**0.898**	**0.753**	**0.893**	**0.618**
BERT-SP	0.137	0.067	0.412	0.000	0.005	0.048	0.038	0.013	0.596
RoBERTa-MLM	**0.431**	**0.333**	**0.431**	–	–	**0.877**	**0.718**	**0.881**	**0.666**
RoBERTa-SP	0.176	0.200	0.422	–	–	0.016	0.019	0.063	0.616

Table 5. **Comparing SP to MLM probing for text base models**. As the results show, using probing using MLM method for text based models outputs better results than using SP

	Concreteness			Color		Shape	Sent.
Metric	$|\rho|$	$|r_s|$	$|\tau|$	acc_{CTD}	acc_{NCD}	acc	acc
Unimodal							
BERT-base	0.27 ± 0.10	0.27 ± 0.09	0.18 ± 0.07	0.26 ± 0.13	0.25 ± 0.08	0.47 ± 0.08	0.56 ± 0.03
BERT-large	0.18 ± 0.13	0.26 ± 0.10	0.17 ± 0.06	0.28 ± 0.14	0.27 ± 0.15	0.51 ± 0.06	0.56 ± 0.03
DistilBERT	–	–	–	0.23 ± 0.08	0.31 ± 0.04	0.45 ± 0.09	0.56 ± 0.04
RoBERTa-base	0.30 ± 0.09	0.29 ± 0.10	0.19 ± 0.07	0.27 ± 0.10	0.27 ± 0.07	0.43 ± 0.00	0.61 ± 0.04
RoBERTa-large	0.21 ± 0.10	0.23 ± 0.11	0.16 ± 0.07	0.30 ± 0.12	0.26 ± 0.08	0.43 ± 0.00	**0.63 ± 0.06**
DistilRoBERTa	–	–	–	0.24 ± 0.12	0.25 ± 0.10	0.43 ± 0.01	0.57 ± 0.02
ERNIE	0.23 ± 0.10	0.20 ± 0.12	0.13 ± 0.08	0.10 ± 0.04	0.13 ± 0.11	0.31 ± 0.08	0.53 ± 0.02
ERNIE-large	0.23 ± 0.08	0.22 ± 0.07	0.14 ± 0.05	0.12 ± 0.06	0.07 ± 0.09	0.30 ± 0.05	0.57 ± 0.05
SBERT	0.24 ± 0.09	0.25 ± 0.09	0.17 ± 0.06	0.13 ± 0.02	0.07 ± 0.01	0.43 ± 0.05	0.53 ± 0.02
V&L							
CLIP	**0.47 ± 0.09**	0.49 ± 0.09	0.34 ± 0.07	0.67 ± 0.15	**0.70 ± 0.08**	0.69 ± 0.08	0.52 ± 0.01
OpenCLIP	0.45 ± 0.12	0.47 ± 0.12	0.32 ± 0.09	**0.77 ± 0.12**	0.66 ± 0.17	**0.79 ± 0.08**	0.53 ± 0.01
FLAVA	0.46 ± 0.10	**0.52 ± 0.10**	**0.36 ± 0.07**	0.52 ± 0.30	0.47 ± 0.22	0.68 ± 0.10	0.50 ± 0.01

Table 6. **Mean and STD Results.** Additional details of mean and standard deviations calculated across prompts, for all tasks which use multiple prompts.

Word	Ground Truth	Predicted Color
apple	green	red
ash	grey	black
cauliflower	white	brown
cello	brown	black
chalk	white	grey
foam	white	grey
garlic	white	brown
lady finger	green	red
pear	green	yellow
sea	blue	grey
sky	blue	white

Table 7. **Qualitative results for CLIP misclassified objects from the CTD and NCD datasets.** As can be seen by analyzing the misclassified objects, most mistakes can be explained by ambiguity of the data.

Word	Ground Truth	LAION
banana	yellow	green
cherry	red	black
orange	orange	red
soil	brown	red
swan	white	black
wood	brown	white

Table 8. **Reporting bias in the LAION dataset**, illustrated by unigram nouns from the CTD and NCD datasets, along with their ground truth colors and the most commonly preceding colors in LAION.

B. Additional Details

B.1. Models

Table 9 presents the different models and Hugging Face checkpoints used for comparing results on the presented tasks.

Model family	Size	Pretraining	Params	MLM head?	Checkpoint
BERT [15]	base	text	110M	Y	`bert-base-uncased`
BERT [15]	large	text	340M	Y	`bert-large-uncased`
RoBERTa [41]	base	text	124M	Y	`roberta-base`
RoBERTa [41]	large	text	355M	Y	`roberta-large`
ERNIEv2 [68, 69]	base	text	109M	Y*	`ernie-2.0-base-en`
ERNIEv2 [68, 69]	large	text	335M	Y*	`ernie-2.0-large-en`
DistilBERT [56]	base	text	66M	Y	`distilbert-base-uncased`
DistilRoBERTa [56]	base	text	82M	Y	`distilroberta-base`
SBERT [54]	–	text	23M	N	`paraphrase-MiniLM-L6-v2`
FLAVA [65]	–	text & VLP	109M	Y	`facebook/flava-full`
CLIP [52]	–	VLP	63M	N	`openai/clip-vit-base-patch32`
OpenCLIP [25]	–	VLP	352M	N	`laion/CLIP-ViT-H-14-laion2B-s32B-b79K`

Table 9. **Models table**. Note that the number of parameters listed for CLIP, OpenCLIP and FLAVA refers to their text encoder components alone. * Note: ERNIE was trained with an MLM head, but because the public checkpoints provided do not include this, we do not evaluate it with MLM probing.

B.2. Prompts used per task

We present further implementation details elaborating the list of prompts used per task.

Concreteness Prediction As explained in the main paper, we use the following prompts to probe our models for the concreteness of words in context by using a cloze task paradigm with Stroop probing. For each word tested, we insert the masked prompt and the prompt with the tested word and calculate the cosine similarity between them.

- *Alice giving the `[*]` to Bob*
- *Bob giving the `[*]` to Alice*
- *I see the `[*]`*
- *A photo of my `[*]`*
- *A close-up photo of a `[*]`*
- *A painting of the `[*]`*
- *A photo of the `[*]`*
- *A photo of a nice `[*]`*
- *A drawing of the `[*]`*

Color Association Prediction For the color association prediction, we use the following prompts. For each given object denoted as $\langle w \rangle$, we use all color options to probe for the correct color.

- *A picture of a `[*]` $\langle w \rangle$*
- *A photo of a `[*]` $\langle w \rangle$*
- *A photo of the `[*]` $\langle w \rangle$*

- *A `[*]` $\langle w \rangle$*
- *`[*]` $\langle w \rangle$*
- *The normal color of a $\langle w \rangle$ is `[*]`*
- *$\langle w \rangle$ usually has a `[*]` color*
- *$\langle w \rangle$ s have a `[*]` color*
- *What is the color of a $\langle w \rangle$? `[*]`*
- *The natural color of a $\langle w \rangle$ is `[*]`*

Shape Association Prediction For the shape association prediction, we use the following prompts. For each given object denoted as $\langle w \rangle$, we use the given shape o to probe for the correct object shape.

- *A photo of a `[*]` shaped $\langle w \rangle$*
- *A photo of a `[*]` $\langle w \rangle$*
- *A photo of the `[*]` $\langle w \rangle$*
- *A `[*]` $\langle w \rangle$*
- *`[*]` $\langle w \rangle$*
- *An image of a `[*]` $\langle w \rangle$*
- *A $\langle w \rangle$ usually has a `[*]` shape*
- *$\langle w \rangle$ s commonly have a `[*]` shape*
- *The basic shape of a $\langle w \rangle$ is `[*]`*
- *What is the shape of a $\langle w \rangle$? `[*]`*

Sentiment Analysis For sentiment analysis, we concatenate the following prompts to the given reviews and use the different options for sentiment prediction.

- *Is this review positive? [*]; Yes, No*

- *Is this a good movie? [*]; Yes, No*

- *I conclude the movie was [*]; good, bad*

- *The film was [*]; good, bad*

- *I had a [*] time; good, bad*

- *The following movie review expresses what sentiment? [*]; Positive, Negative*

- *Sentiment expressed for the movie is [*]; Positive, Negative*

- *The overall review of the film is [*]; good, bad*

- *The movie was [*]; good, bad*

- *This movie is [*]; good, bad*