Taller, Stronger, Sharper: Probing Comparative Reasoning Abilities of Vision-Language Models

Prateek Agarwal Lakshita Bhargava Deep Chakraborty Kartik Choudhary
{prateekagarw, lbhargava, dchakraborty, kartikchoudh}@umass.edu

1 Problem Statement

Visual reasoning capabilities are largely grounded in the language we use every day. Studies have shown that linguistic differences shape the way we think (Boroditsky, 2011), and affect our conceptions of time, space, and other visual attributes (Boroditsky, 2001, 2000; Winawer et al., 2007). In this work, we assess the role comparative adjectives play in building representations of the visual world. Humans use words like *bigger, brighter, shorter, deeper, etc.* several times a day to compare entities or states. We *hypothesize* that such comparisons are a key component to assessing the performance of models that perform visio-linguistic reasoning, and may offer a way to build richer representations for downstream tasks.

Vision Language (VL) Models (Radford et al., 2021; Tan and Bansal, 2019; Chen et al., 2020) are a relatively new class of methods based on transformers (Dosovitskiy et al., 2020; Vaswani et al., 2017), that are capable of joint visio-linguistic reasoning to solve tasks such as visual question answering (VQA), image to text (image captioning) and text to image (using natural language descriptions), optical character recognition (OCR), scene graph generation, etc. While these models have been largely successful, they still struggle with elementary reasoning from compositional language prompts (Diwan et al., 2022; Yuksekgonul et al., 2023; Lewis et al., 2023). Datasets like *Winoground* (Thrush et al., 2022) and *CLEVR* (Johnson et al., 2017) have been proposed as diagnostic toolkits to assess the performance of these models on such tasks. Specifically, in *Winoground*, given two images and two captions that contain the same words but in a different order, the task is to pair them correctly. The authors show that state-of-the-art VL models seldom do better than random chance on such instances.

Following such works, we offer a way to diagnose the performance of VL models using a new dataset of images that we collect. Each of these images has two objects that can be compared using at least one comparative adjective. The model is then faced with a series of tasks that involve pairing the noted adjective(s) to the image in a way that constrains the model to make comparisons between the objects present in the image, through natural language sentences. An example of (image, adjective) pair is shown in Fig. 2. Although the English language has over 4,800 adjectives, less than 500 of them accept the '-er' form, and only about a fifth of such adjectives are appropriate for visual comparisons. We identify and focus on this subset exclusively for the rest of this work.

We assess: (a) the ability of VL models to associate the correct comparative adjectives given an image, (b) the ability of VL models to retrieve images that can be described using a given comparative adjective, (c) performance gains for VL models using AI-powered prompt-engineered sentences containing comparative adjectives, (d) effect of adding object annotations alongside adjectives, and (e) failure modes that offer insight into desirable features that must be encoded by VL models in the future. Our dataset is also a first of its kind and we hope it'll serve as a benchmark for comparative-reasoning capabilities. Code is available here [1].

2 Proposed vs. Accomplished

- ~~Collect dataset of images containing two objects that can be compared using a sentence with a comparative adjective~~: We had originally proposed to use an AI like DALL-E 2 (Ramesh et al., 2022) to generate these images, but that didn't work well in our initial

[1] https://github.com/deepc94/685-project

exploration, so we resorted to search-based images.

- ~~Design an extensive set of experiments that can test the model's comparative reasoning abilities~~
- ~~Determine failure modes of the model based on our experiments~~
- ~~Improve the zero-shot performance of the model using prompt engineering and additional annotations~~: We added this after our original proposal based on the feedback we received.
- *Propose a method to fine-tune the VL models representations for better performance on downstream tasks*: We ran out of time for this, but hopefully our conclusions can point readers to future directions.

3 Related work

Our project is inspired by previous work in cognitive science that shows how discriminative abilities in visual stimuli develop as a result of the language we speak. For example, Winawer et al. (2007) show that the obligatory distinction between light blue ("goluboy") and dark blue ("siniy") in the Russian language results in Russian speakers being faster at color discrimination tasks due to a distinct categorization advantage that is not otherwise accessible to English speakers. This leads us to believe that the ability of visio-linguistic models to make fine distinctions is an important property to be studied.

Vision Language (VL) models (Li et al., 2022; Radford et al., 2021; Tan and Bansal, 2019; Chen et al., 2020) in the NLP and machine learning literature, therefore, stand out as obvious candidates for running such tests. CLIP (Radford et al., 2021) is a Contrastive Learning-based Language-Image Pre-training method that leverages large-scale training on web data. It uses a matching-loss which contrastively maximizes the cosine-similarity between the vision and text encodings of a image-text pair. By replacing text-prediction with image-text matching, CLIP allows for zero-shot predictions on a variety of downstream tasks, and achieves competitive performance. However, the authors did not comprehensively test the visual-reasoning and compositional-reasoning abilities of the model.

Attempts have been made in the literature to test the elementary reasoning abilities of VL models using diagnostic datasets. Two such datasets are *CLEVR* (Johnson et al., 2017) and *Winoground* (Thrush et al., 2022). *CLEVR* is a synthetic dataset of simple 3D shapes and automatically generated complex questions that require a VL model to have abilities such as counting, comparing, logical reasoning, and short-term memory to answer correctly. *Winoground* is a dataset consisting of (2 images, 2 captions) pairs, where the captions differ only in the order of words, but the images differ substantially based on the caption. VL models have been shown to have significant issues answering questions correctly or pairing images to correct captions in these scenarios that require compositional reasoning abilities respectively. Our dataset is similar in construction the latter, except we strictly use images containing 2 prominent objects, and a sentence a valid comparative adjective as ground truth.

Lastly, works like (Diwan et al., 2022; Yuksekgonul et al., 2023; Lewis et al., 2023) test the compositional reasoning abilities of models like CLIP through a variety of experiments, and find that CLIP performs at near chance-level. (Yuksekgonul et al., 2023) find that CLIP and other VL models' text encoders behave like a bag-of-words, namely in that, given an image, some models do not significantly prefer the semantically-meaningful ground-truth sentence/caption over a shuffled version of that sentence. They test various models' relational and attributive understanding, as well as their sensitivity to order, and find them to be surprisingly poor. They show that this deficiency could be due to the image-text matching contrastive loss used for training, on which the models continue to perform well even with composition and order information removed.

The main difference in our approach from prior investigations is that we aim to test the model's *comparative*, instead of *compositional*, reasoning ability. An early work by (Gupta and Davis, 2008) uses comparative adjectives, but over a limited set (9 words) in a relational setting, whereas we test on 39 pairs of adjectives.

4 Dataset

Our dataset is comprised of 257 images that were manually curated to feature two prominent objects connected by an adjective word, which can be used

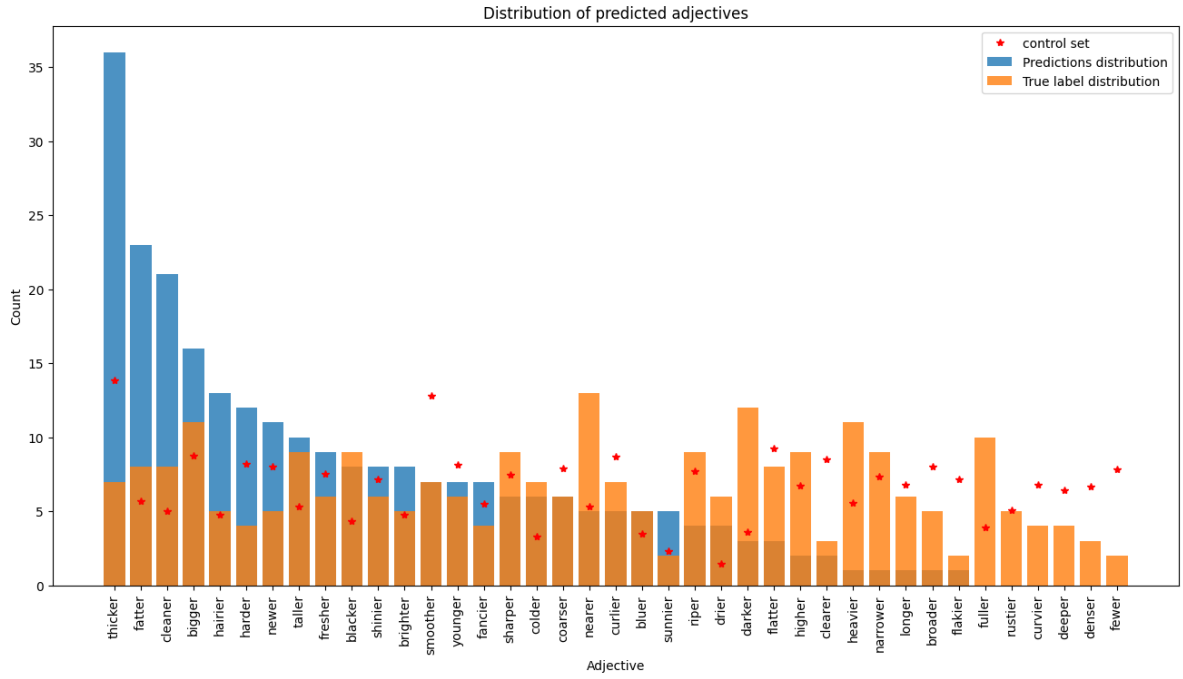

Figure 1: Distribution of adjectives in our dataset

Figure 2: Example image from the dataset. Adjective: riper, Antonym: rawer, Negative Adjective: cleaner

Figure 3: Example images from the control set, these images always have 2 identical objects in different poses or orientations.

to compare the objects in the image. The images are labeled with an associated adjective, an antonym of the adjective, and a negative adjective. The negative adjective, while typically associated with the objects in the image, is not applicable to the image itself. As shown in Fig. 2, the image has two avocados with one of them 'riper' than the other. Conversely, the other one can be called 'rawer'. While some fruits are 'cleaner' than others, these avocados cannot be compared using 'cleaner'.

We also gathered an additional set of 50 images (see Fig. 3) that contain two identical objects that would work as a control set to identify any biases in the model toward particular adjectives. See Section 5.3 for details of experiments using the control set.

Finally, we created a HuggingFace dataset object for our final dataset, which can be accessed from the HuggingFace Hub (https://huggingface.co/datasets/kartik727/Test_Dataset). Each data point in this dataset has the image, and its three attributes: 'adjective', 'antonym', and 'negative'.

4.1 Data Collection

Each member collected roughly 75 images from various sources such as internet search (Getty Images, Shutterstock), AI image generating tools (DALL-E 2, Bing Image Creator), and illustrations. Images were collected in such a way that the adjectives would have similar counts. The collected images were pooled and duplicates or very similar images were removed. Some of the labels were changed to an alternate adjective that was




Figure 4: Prompt to DALL-E 2: "There are two glasses. The glass on the left is fuller than the one on the right."


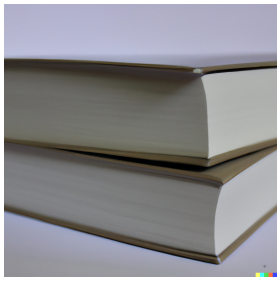

Figure 5: Prompt to DALL-E 2: "There are two books. The one on the left is thicker than the one on the right."

also applicable to the image to ensure uniformity in the final distribution of adjectives (see Fig. 1 for the distribution).

4.2 Data annotation

We selected 74 '-er' form comparative adjectives that are appropriate for visual comparisons. These include 35 antonym pairs and 4 adjectives that do not have commonly used antonyms. We chose one of the adjectives from each pair along with the 4 adjectives without antonyms to create the final list of 39 adjectives used to label our images. These adjectives (called 'left-side' adjectives) are mapped to their antonyms (called 'right-side' adjectives) or "None" for the 4 adjectives without antonyms. This mapping is used for experiments detailed in Section 5.

We initially aimed to generate images using DALL-E 2 by feeding in query sentences that have three main components: an object, a preposition, and an adjective. For e.g., "There are two apples. The apple on the left is bigger than the one on the right." After trying several templates to construct query sentences, including single- and multi-part queries, we came to the conclusion that current publicly available image-generating models like DALL-E 2 cannot reliably create images with multiple objects related by an adjective supplied in the prompt (see Fig. 4 and Fig. 5), which was also the conclusion of (Conwell and Ullman, 2022). Due to the service giving access to a limited number of free image generations and a severely rate-limited API, we were only able to gather a small number of images that fit our criteria before we abandoned this idea and pivoted to finding the rest of the images from the internet.

Each member then annotated their images with one of the 39 'left-side' adjectives based on which adjective was used to download the image, the antonym of the adjective, and a negative adjective.

4.2.1 Inter-Annotator Agreement

For every image, all 4 annotators were instructed to list their top-3 adjectives in descending order of preference. The measured agreement over top-1 i.e. the fraction of times all annotators agreed on a single top adjective (out of 39 possible adjectives) was 45.52%. Further insight into the agreement can be gained from 5.2.

4.2.2 Data Aggregation

Since each annotator found different images which corresponded to adjectives from the set, we simply aggregated the image-adjective pairs as our ground-truth dataset.

5 Our evaluation setup

Given a dataset of images and labels $\mathbb{D} = \{(I, c)\}$ where I is an image containing 2 objects and c is a comparative adjective, our goal is to pair the images I to their correct gold label c in the presence of negative label(s) C' that are constructed in different ways. We use the

`clip-vit-base-patch32` variant of CLIP (Radford et al., 2021) as our scoring function f that accepts images and text, computes $512d$ image and text embeddings $\mathbf{z}_I$ and $\mathbf{z}_c$ respectively, and computes the similarity score as $f(I, c) = {\mathbf{z}_I}^T \mathbf{z}_c$. Finally, the scoring metric is as follows:

$$s(I, c, C') = \begin{cases} 1, & \text{if } f(I, c) > f(I, c') \; \forall c' \in C' \\ 0, & \text{otherwise} \end{cases}$$

where c is the ground truth or gold label adjective assigned to the image, and C' is the set of all constructions of negative adjective labels. We compute the score over the top-K predictions in our experiments.

5.1 Label construction and baselines

CLIP is a model consisting of an image encoder and a text encoder that produces representations in a joint semantic space. The text encoder can accept prompts in natural language to produce zero-shot classifiers, thus making it a suitable candidate for our evaluation setup. For instance, it can accept an image containing an object, say a `cat`, and a list of texts containing the names of all classes in the dataset [`dog`, `cat`, `...`, `airplane`] and produce a probability distribution over classes. We use it in a similar fashion except that our images contain 2 objects each, and the list of classes is a list of adjectives [`brighter`, `bigger`, `...`, `shinier`] instead that can be used to compare the objects seen in the image. However, Radford et al. (2021) observe that they can get a performance boost of 1-2% by using the class label as a caption instead, such as `'A photo of a {label}'`. Therefore, we follow the same approach and covert our labels c into a prompt $p(c)$ such as `'The objects in this image can be compared using the adjective "{c}"'` where c is a comparative adjective of the '-er' form. Our experiments explore augmenting p with different information to analyze CLIP's behaviour.

We now have a dataset $\mathbb{D} = \{(I, p(c))\}$ where c is one of 74 possible adjectives. The issue remains however that some adjectives have their antonyms in the overall set, and we should consider the model's prediction of an adjective to be correct if it chooses either the correct ground truth adjective or its antonym. This is because an image containing 2 objects, say a big cat and a small cat, can be equivalently described by either of the captions "The objects in this image can be compared using the adjective "bigger" or "The objects in this image can be compared using the adjective "smaller". We therefore create a mapping between distinct adjectives and their antonyms such as (`brighter` ⇔ `dimmer`, `sharper` ⇔ `duller`, etc.) resulting in a reduced set of 39 distinct adjectives most of which have a paired antonym, and proceed to resolve the prediction ambiguity in one of 3 ways:

1. Jointly include the adjective-antonym pair in the same caption: `The objects in this image can be compared using the adjective "{bigger}" or "{smaller}"`.

2. Average the embeddings/logits of adjective and antonym containing captions before making predictions: `The objects in this image can be compared using the adjective "{bigger}"` and `The objects in this image can be compared using the adjective "{smaller}"`

3. Remap the ground-truth antonyms of all images to one of the 39 distinct adjectives and only use those for prediction: `The objects in this image can be compared using the adjective "{bigger}"`. We finally use this mapping for our ground-truth labels and use them for all further experiments i.e. we finally evaluate over only 39 distinct adjectives.

The results from the aforementioned variations are reported in 6 and 1. Jointly including the adjective-antonym pair in the prompt performs slightly better, across all k, than either averaging the embeddings or logits; this indicates that CLIP is directly able to leverage the added information through its text encoding. The joint antonym baseline also outperforms the remap antonym baseline after k¿5, indicating that simply adding the antonym to prompt directly improves performance.

5.2 Human baseline

In order to measure human performance on the task of identifying comparative-adjectives, we construct a simple human baseline whose results

Table 1: Summary of all important results. The top category shows naive and human performance baselines, the middle shows strategies to handle antonyms of adjectives, and the last section shows the results from prompt engineering and additional object-level annotations

Experiments	Top-1 (%)	Top-3 (%)	Top-5 (%)
Random Chance Baseline	2.56	7.69	12.82
Control Set Baseline	2.72	8.56	15.56
Human Baseline	66.83	82.34	
Remap Antonym Baseline	**22.57**	39.30	46.70
Joint Antonym Baseline	22.18	**40.86**	**53.70**
Average Antonym Baseline (Embedding)	23.34	36.58	44.75
Average Antonym Baseline (Logits)	21.40	38.13	51.80
Prompt Engineering (AI prompts) 6.1	22.57	**38.13**	**46.69**
Prompt Engineering (human prompts) 6.1	**24.90**	37.74	46.69
Including Object Information 6.2	12.06	26.84	35.40

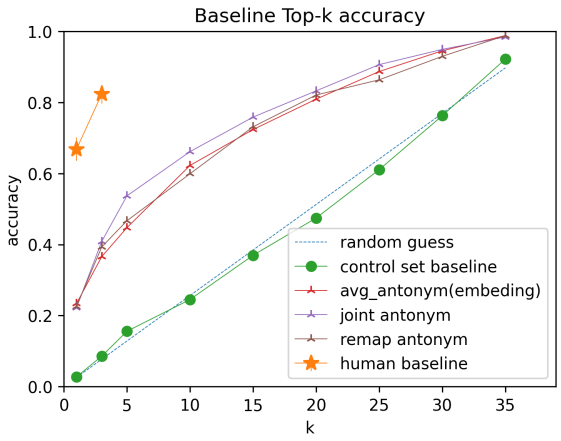

Figure 6: Top-k accuracies of CLIP over several baselines, including manual human evaluation

are reported in 1 and can be seen in 6. Specifically, for each image not including ones they gathered, each annotator was instructed to construct a top-3 list of adjectives(from 39 possible values). Then, in reference to the ground-truth adjectives, we calculated the top-1 and top-3 accuracies for each annotator (excluding their own set); the average top-1 accuracy was 66.830% (std of 3.325%) and the average top-5 accuracy was 82.341% (std of 2.647%) In comparison to all the other baselines which consisted of a simple prompt, the human baseline performed significantly better on both top-1 and top-3 accuracy, highlighting CLIP's misalignment with human judgements. Additionally, since the ground-truth is created by fellow human annotators, this baseline indirectly shows that there are high levels of annotator agreement.

5.3 Control Set Baseline

To capture the possible bias of our model towards different adjectives, we use our remapped groundtruth evaluation setup on the control dataset. Since the images consist of identical objects which cannot be compared with any adjective, we expect the model to have no preference in its predictions. For each of the 39 adjectives, we use a caption such as `'The objects in this image can be compared using the adjective "{bigger}"'`, and take the score as the average of the logits over all of the control images. Taking a softmax distribution over these adjective scores (see figure 1 for frequency of predicting a certain adjective), we find that the entropy over the distribution is 0.981, indicating that our model is very weakly biased. Using a fixed list of adjectives ordered according to decreasing scores, we create a new 'control set' baseline model and evaluate it's accuracy on the ground-truth labels as seen in 6. We can observe that this baseline performs very close to random chance and is only slightly biased. Hence, the predictions of our other models are likely due to the images themselves, and not due to the bias of the model.

6 Above and beyond the baseline

In this section, we attempt to improve the baseline performance of zero-shot CLIP as reported in the previous section, by exploiting a host of extra modalities such as prompt engineering and object-level annotations. We hoped that this would also reveal additional behavioral properties of CLIP.

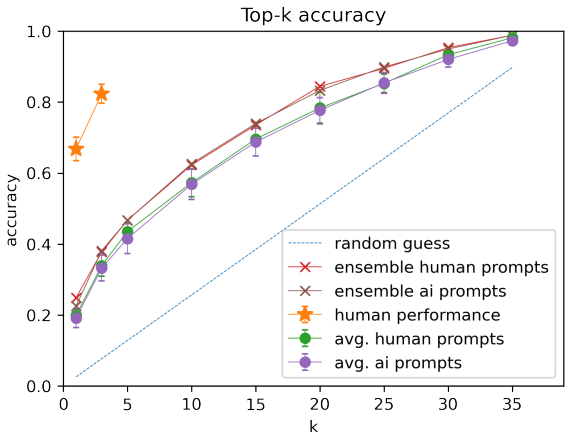

Figure 7: Top-k accuracies from different prompt engineering techniques, including human and AI generated prompts, against a human baseline.

6.1 Prompt Engineering

Radford et al. (2021) note that the performance of CLIP can be pushed by a further 3-5% by ensembling over multiple zero-shot classifiers, by combining different prompt formats for the caption. In the original paper, the authors manually design a set of 80 prompts that are averaged in the embedding space before making predictions. These prompts were designed by trial and error to improve the accuracy on lowest performing classes on ImageNet. Inspired by this, and the recent success of large language models such as ChatGPT (OpenAI, 2023) and Bard (AI, 2022), we wanted to create a large set of AI-generated prompts from reference human prompts to improve CLIP's ability to understand images that require comparisons using comparative adjectives.

The human prompts consisted of descriptive (`There are two objects in the image. They can be compared using the adjective {brightness}.`), speculative (`The objects in the image could be compared using the adjective {brightness}.`), and assertive (`{Brightness} can be used to compare the objects in the image.`) sentences, among others. These prompts were designed to test the model's understanding and comparative abilities in various contexts.

To generate AI prompts, we leveraged the capabilities of ChatGPT and Bard. For ChatGPT, we asked the model to generate 10-20 variations for each of the human prompts and excluded the ones that were repetitive. This approach aimed to explore different possible phrasings and expressions for the same underlying comparative concept. On the other hand, with Bard, we provided an explanation of our project and task, and asked the model to autonomously generate sentences for our prompt engineering task. By involving multiple AI models, we aimed to capture a broad range of prompt variations and perspectives.

Examples of prompts given to AI tools and the generated sentence variations:

1. (ChatGPT) `Generate variations of the sentence: The objects in the image could be compared using the adjective brighter`
   - A comparison of the objects in the image can be made based on their brightness.
   - The objects in the image can be distinguished based on their varying degrees of brightness.

2. (ChatGPT) `Generate variations of the sentence using a speculative tone: Brightness could be used to compare the objects in the image`
   - Could it be that the attribute of brightness is what sets the objects in the image apart and makes them comparable?
   - Perhaps the brightness of the objects in the image can be used as a means of comparison.

3. (Bard) `I want to use CLIP for a task where I give in an image with two prominent objects and ask which "-er" type adjective can be used to compare the two objects. For example, given an image with an apple and a watermelon, I will ask if we can compare the objects using the word "bigger". What kind of text prompt should I give in this case?`
   - The objects in this image are different in brightness.

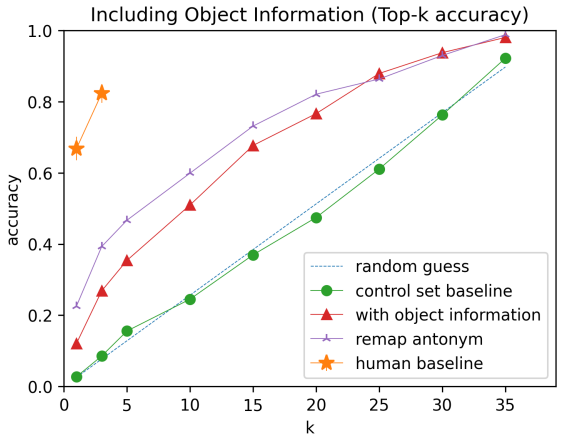

Figure 8: Top-k accuracies of CLIP when object information is included, in comparison to similar baselines

In total, we used 70 sentences generated by ChatGPT and 23 sentences generated by Bard, for a total of 93 AI generated prompts.

Finally, in order to improve performance using these additional prompts, we explore two averaging strategies: (i) averaging predictions obtained by using each prompt separately, and (ii) averaging all the prompts in embedding space similar to (Radford et al., 2021). The results of these experiments are in 7. The first strategy yielded a top-1 and top-5 accuracy of 19.1 ± 2.7 % and 41.5 ± 4.2 % respectively, whereas ensembling yielded a top-1 and top-5 accuracy of 22.5% and 46.7% respectively (an improvement of 3-5%, which is consistent with the paper). However neither of these approaches significantly outperformed the mean human-generated prompt accuracies (top-1: 19.9%, top-5: 43.5%) and (top-1: 24.9%, top-5: 46.7%) respectively for the two strategies. This result was really surprising to us, and might indicate that there may be better strategies for prompt engineering in comparative reasoning scenarios that are fundamentally different from object recognition scenarios. The AI generated prompt that individually yielded the highest top-1 accuracy of 25.7% was `{brighter} is an adjective that can be used to describe the objects in this image`, and that which gave the highest top-5 accuracy of 52.9% was `One of the objects in this image is {thicker} than the other.`

6.2 Including Object Information

Since the prompts in our previous experiments only jointly refer to the two objects in the image as "objects", we wanted to observe the effects of explicitly mentioning the two objects we want CLIP to reason over, results in 1 and 8. We generally expect higher performance as this inclusion is supposed to make CLIP's job easier by providing insights into human perception and recognizing human judgements. For each image in the training dataset, a human annotator noted the two objects in the scene; if there were two distinct objects, both would be noted separately, such as "cat and dog" , and jointly if they were identical, such as "apples". The adjectives were included as usual and prompts were composed such as: `'The dogs in this image can be compared using the adjective "{bigger}"'`. The inclusion of object information in a single term (as opposed to 'left dog and right dog') when applicable, and on one side of the adjective are intentional; they are mainly to avoid confounding factors, namely, the (Yuksekgonul et al., 2023) finding that CLIP is somewhat attribute and word order invariant. We find that, surprisingly, including order information significantly degrades CLIP's performance across the board in comparison to our remap antonym baseline (whose prompt simply omits object information); the model achieves only 12.06%, 26.84% and 35.40% top-1, 3, 5 accuracy respectively. This could indicate that either CLIP, in the remap antonym baseline, is not relying on the objects we intend, or, in this experiment, is not able to encode the explicit object information in its text embedding to find the intended adjective. Regardless, since downstream tasks often focus on identifying specific objects, CLIP's text encoder's behaviour and preference towards ambiguity is alarming, and needs further investigation.

7 Error analysis

Consistently through the baselines and other experiments, it can be seen that the performance of CLIP on comparative reasoning tasks such as ours is well under the human-level performance, a staggering gap of roughly 41%. Below, we try to explain why some of this gap in performance might be.

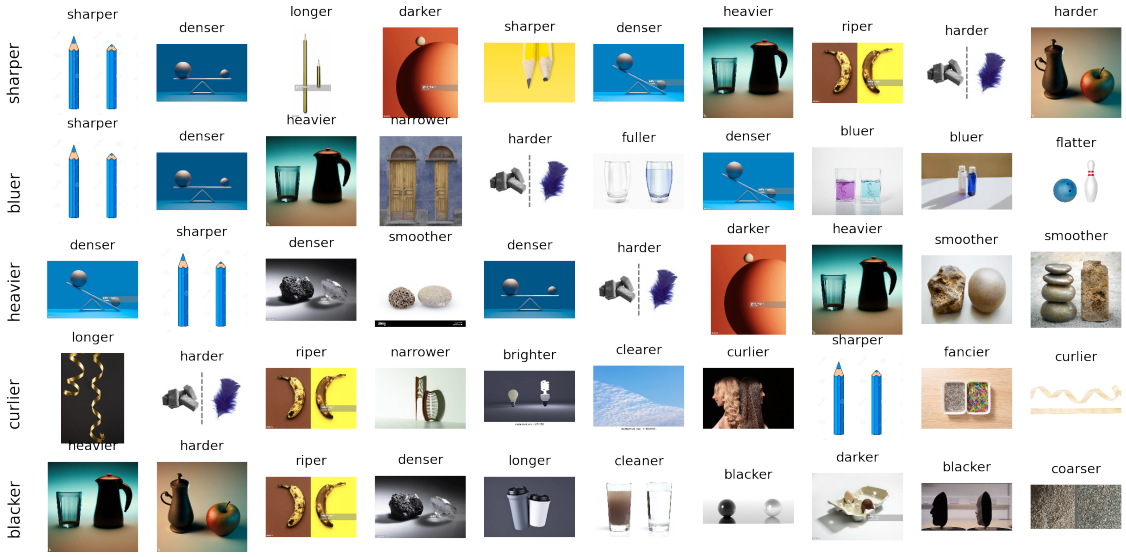

Figure 9: Top-10 most similar images (left-to-right) in embedding space for the adjectives 'sharper', 'bluer', 'heavier', 'curlier', 'blacker' (rows), using the top performing AI generated prompt "One of the objects in this image is {adjective} than the other". The ground truth adjective associated with each image is shown on top, but is not necessarily the only adjective that can be used to compare objects in that image. (best viewed by zooming in.)

7.1 Analysis using most similar images for an adjective

Figure 9 shows the top-10 most similar images (in embedding space) for 5 randomly chosen adjectives in our dataset: {'sharper', 'bluer', 'heavier', 'curlier', 'blacker'}. The first thing we notice is that the ground truth adjective of an image is not necessarily the only adjective that can be used to compare the objects in that image, and several others may apply. For instance in row 3, images $1, 3, 7$ can all potentially be described using the adjective 'heavier'. Dealing with synonymous or multiple adjectives is a challenge in such datasets. Moreover, further inspection reveals that the model in its zero-shot state might not be focusing on the actual differences between the objects in question to find the relevant adjective, but might focus on spurious information such as background color (in row 2 for 'bluer', images 2,3,4, and 7 might appear simply because of comparisons made between background and floor color, or background and foreground color), or the general description of a scene that might use a particular adjective (row 1 for 'sharper', pencils are generally associated with sharpness, row 3 for 'heavier', rocks are generally thought as heavy, etc.). Therefore, we see that models like CLIP lack the ability to compare between specific objects in the image, regardless of whether the prompt identifies the object specifically (see section 6.2 where adding object information actually hurts performance), or specifies their position (as shown in prior work (Yuksekgonul et al., 2023)).

7.2 Analysis using Negative Prompts

As described in Section 5.1, the class label adjectives c were converted into a caption using a prompt $p(c)$ = `'The objects in this image can be compared using the adjective "{c}"'`. Given a prompt $p(\cdot)$, we can also create a negation of the prompt $\hat{p}(\cdot)$ that would have the opposite meaning of the original prompt. For example, for $p(\cdot)$ shown above, we would have $\hat{p}(c)$ = `'The objects in this image cannot be compared using the adjective "{c}"'`.

We would expect the model predictions using the negative prompt $\hat{p}(\cdot)$ to be reversed when compared to the positive prompt $p(\cdot)$, and thus the performance using this prompt on the evaluation to be below random chance.

However, Fig. 10 shows that the negative prompt $\hat{p}(\cdot)$ performs just as well as the baseline prompt $p(\cdot)$ in top-k accuracy, even though $\hat{p}(\cdot)$ conveys the opposite meaning of $p(\cdot)$.

This indicates that the model does not understand the semantic meaning of the prompts very well and focuses on the adjective itself in its predictions, regardless of how the adjective is used in

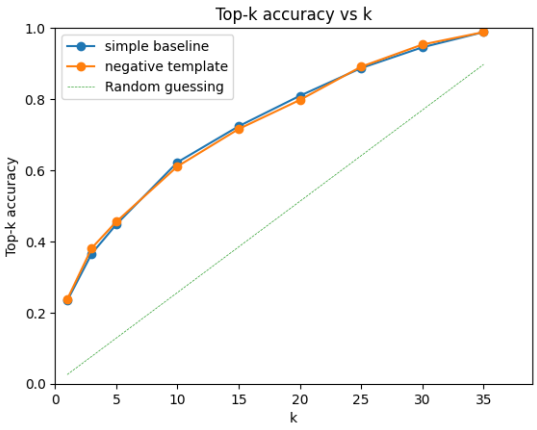

Figure 10: Top-k accuracy using the negative prompt $\hat{p}(\cdot)$ vs the baseline using the positive prompt $p(\cdot)$

Figure 11: The hands cannot be compared using the negative adjective "bigger"

the prompt.

7.3 Analysis using Negative Adjectives

In order to further investigate CLIP's performance on our task, for each image, our annotators also noted a couple negative adjectives which cannot be used to compare the two intended objects e.g. 11. We know that, at the very least, all of our negative adjectives should have a lower score (and softmax probability) when compared to our groundtruth adjectives. Therefore, over our dataset using the remap antonym prompt, we calculate the percentage of samples where atleast one negative adjective receives higher probability than the positive adjective and find this to be 49.42%; this shows that CLIP's poor performance cannot be attributed entirely to confounding, but correct, adjectives which might have crowded our Top-k, but also due to large-scale high scoring of our verified negative adjectives. To understand the extent of this misscoring, we also visualize the distribution of probability differences in 12, when a negative adjective was scored higher than a positive; we observe that most of the differences fall under 0.1 which

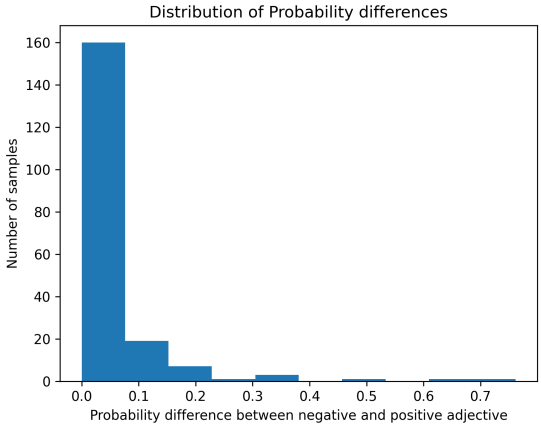

Figure 12: Distribution of Probability Differences between Negative and Positive adjectives

indicates that when our model is not behaving as expected, it is similarly confident in both our negative and positive adjectives. Since CLIP does not outright favor the negative adjective, it means that CLIP is doing some reasoning over the adjectives, however, this reasoning is not very well calibrated and sufficiently discriminative.

8 Contributions of group members

- Prateek: gathered data, collated and cleaned/aligned dataset, wrote and analyzed multiple baselines (remap antonym, jointly antonym, human baseline, involved with analysis on control set baseline). Did one improvement experiment(Object Information) and code for error analysis experiment (negative adjectives). Wrote corresponding sections, which was a solid amount.

- Lakshita: data collection, annotation, wrote code for prompt engineering, control baseline dataset collection, control baseline experiments, and wrote corresponding sections for the report.

- Deep: conceptualized the project and designed the overarching evaluation framework, wrote code for baseline and prompt engineering experiments, did data collection and annotation, did error analysis checking for image adjective alignment. Wrote a solid amount.

- Kartik: Data collection, cleaning, aggregation and correction. Wrote boilerplate code

for zero-shot evaluation of CLIP. Initial experiments of detecting attributes (shape, size, color). Experiments on negative prompts aggregating embeddings for synonyms.

9 Conclusion

We proposed methods to evaluate the representations learned by VL models like CLIP in comparative reasoning settings. Is the model able to correctly identify an adjective that adequately describes the two objects in question in the image? The short answer is No, there's a 40% gap in the model's performance compared to average human performance on this task. We find that models like CLIP struggle with multiple things: exactly identifying the object in question, identifying spatial relationships between objects, focusing on the relevant differences between objects, and focusing on positive and negative context around an adjective, etc. We show that prompt engineering and additional modalities like object level annotations are unable to push simple baselines, and this suggests a ground-up retraining of VL models with comparative abilities in mind. Handling synonymous adjectives was especially difficult in our task and may have been a huge confounding factor, a problem which cannot be necessarily rectified by detailed annotation due to its subjective nature. Curating images for this task was also really hard, and relying on generative AI models didn't help us too much. In future, we would like to repeat our experiments with other VL models like BLIP (Li et al., 2022), propose a fine-tuning and downstream transfer task, and have more control over confounding factors in our dataset.

10 AI Disclosure

- Did you use any AI assistance to complete this proposal? If so, please also specify what AI you used.
  - Yes, we used BARD(AI, 2022) and ChatGPT(OpenAI, 2023)

If you answered yes to the above question, please complete the following as well:

- If you used a large language model to assist you, please paste *all* of the prompts that you used below. Add a separate bullet for each prompt, and specify which part of the proposal is associated with which prompt.
  - Prompts are an integral part of 6.1 and some of them are mentioned in the report.
- **Free response:** For each section or paragraph for which you used assistance, describe your overall experience with the AI. How helpful was it? Did it just directly give you a good output, or did you have to edit it? Was its output ever obviously wrong or irrelevant? Did you use it to generate new text, check your own ideas, or rewrite text?
  - It was mild to moderately helpful. We used it to generate new text.

References

AI, G. (2022). Bard: A large language model from google ai. *arXiv preprint arXiv:2201.08237*.

Boroditsky, L. (2000). Metaphoric structuring: Understanding time through spatial metaphors. *Cognition*, 75(1):1–28.

Boroditsky, L. (2001). Does language shape thought?: Mandarin and english speakers' conceptions of time. *Cognitive psychology*, 43(1):1–22.

Boroditsky, L. (2011). How language shapes thought. *Scientific American*, 304(2):62–65.

Chen, Y.-C., Li, L., Yu, L., El Kholy, A., Ahmed, F., Gan, Z., Cheng, Y., and Liu, J. (2020). Uniter: Universal image-text representation learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX*, pages 104–120. Springer.

Conwell, C. and Ullman, T. (2022). Testing relational understanding in text-guided image generation. *arXiv preprint arXiv:2208.00005*.

Diwan, A., Berry, L., Choi, E., Harwath, D., and Mahowald, K. (2022). Why is winoground hard? investigating failures in visuolinguistic compositionality. *arXiv preprint arXiv:2211.00768*.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Gupta, A. and Davis, L. S. (2008). Beyond nouns: Exploiting prepositions and comparative adjectives for learning visual classifiers. *ECCV (1)*, 5302:16–29.

Johnson, J., Hariharan, B., Van Der Maaten, L., Fei-Fei, L., Lawrence Zitnick, C., and Girshick, R. (2017). Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2901–2910.

Lewis, M., Nayak, N. V., Yu, P., Yu, Q., Merullo, J., Bach, S. H., and Pavlick, E. (2023). Does clip bind concepts? probing compositionality in large image models.

Li, J., Li, D., Xiong, C., and Hoi, S. (2022). Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation.

OpenAI (2023). Chatgpt: Large language model. GitHub repository.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. (2021). Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. (2022). Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*.

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

Thrush, T., Jiang, R., Bartolo, M., Singh, A., Williams, A., Kiela, D., and Ross, C. (2022). Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5238–5248.

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

Winawer, J., Witthoft, N., Frank, M. C., Wu, L., Wade, A. R., and Boroditsky, L. (2007). Russian blues reveal effects of language on color discrimination. *Proceedings of the national academy of sciences*, 104(19):7780–7785.

Yuksekgonul, M., Bianchi, F., Kalluri, P., Jurafsky, D., and Zou, J. (2023). When and why vision-language models behave like bags-of-words, and what to do about it?