Probing Conceptual Understanding of Large Visual-Language Models

Madeline Schiappa[1] Raiyaan Abdullah[1*] Shehreen Azad[1]
Jared Claypoole[2] Michael Cogswell[2] Ajay Divakaran[2]
Yogesh Rawat[1]

[1]Center for Research in Computer Vision, University of Central Florida
[2]SRI International

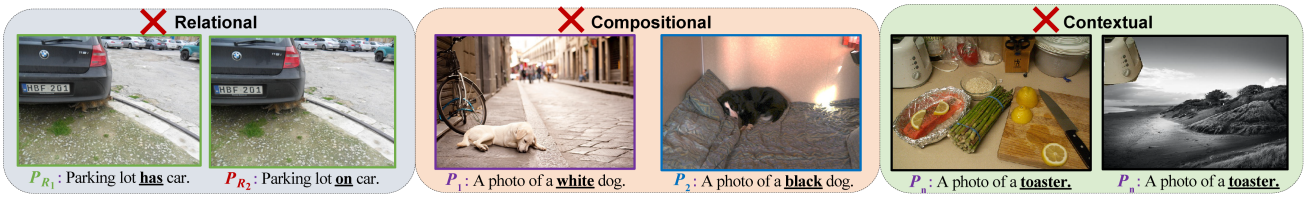

Figure 1. **Conceptual understanding of an existing V+L model.** Here, CLIP failure to understand relational, compositional and contextual reasoning is shown. This benchmark presents three datasets to evaluate V+L models on relational, compositional, and contextual understanding. They utilize image-text matching tasks with predicate, object/subject, compositions, or background swaps.

Abstract

In recent years large visual-language (V+L) models have achieved great success in various downstream tasks. However, it is not well studied whether these models have a conceptual grasp of the visual content. In this work we focus on conceptual understanding of these large V+L models. To facilitate this study, we propose novel benchmarking datasets for probing three different aspects of content understanding, 1) relations, 2) composition, and 3) context. Our probes are grounded in cognitive science and help determine if a V+L model can, for example, determine if snow garnished with a man is implausible, or if it can identify beach furniture by knowing it is located on a beach. We experimented with many recent state-of-the-art V+L models and observe that these models mostly fail to demonstrate a conceptual understanding. This study reveals several interesting insights such as that cross-attention helps learning conceptual understanding, and that CNNs are better with texture and patterns, while Transformers are better at color and shape. We further utilize some of these insights and investigate a simple finetuning technique that rewards the three conceptual understanding measures with promising initial results. The proposed benchmarks will drive the community to delve deeper into conceptual understanding and foster advancements in the capabilities of large V+L models. The code and dataset is available at: https://tinyurl.com/vlm-robustness

1. Introduction

Humans navigate the world by learning an "understanding" of how it works. Understanding may be defined as the underlying organization of all concepts, including objects, situations, events, and more [8, 27]. They are organized in our brains as *conceptual maps*, which encode structured, relational information [13]. *Conceptual maps* highlight major objects and actions in a system and the causal relations between them. While deep learning models have impressive performance in a variety of tasks, it is still unclear if their impressive performance is due to learnt *conceptual maps*.
Large visual-language (V+L) models are recently and greatly successful deep learning models that learn representations of image and text in a shared space. These representations are useful for downstream tasks like image classification, visual-question answering, image retrieval and more [1, 31, 32, 43, 47, 53]. However, for use in real-world applications, it is also vital that models "understand" rather than memorize to perform on more general tasks [29]. While large-language models have been shown to have a moder-

*Corresponding Author: raiyaanabdullah@gmail.com

ate amount of "theory of mind," as measured by conceptual consistency [35], V+L models have not been investigated in a similar way using real-world examples. This is partly because images are more challenging, as shown by preliminary studies [5, 10, 42]. With this in mind, we focus on probing models on their conceptual maps.

We develop a benchmark by combining insights from well-known tests such as the Peabody Picture test, semantic analysis underpinning knowledge bases such as ConceptNet [39], and comprehension in elementary school education [35] to identify three key areas for probing: relations, composition, and context (Figure 1). Our benchmark could be seen as a computational instantiation of visual comprehension testing along three important fundamental skills. These skills form a compact set of necessary, but not sufficient, prerequisites for key tasks such as concept transfer, analysis, evaluation, and generation. They thus provide us a basis for probing comprehension of large V+L models.

We propose three benchmark datasets, *Probe-R*, *Probe-C*, and *Probe-B*. Probe-R looks at model understanding of possible object relations by comparing an image to a correct prompt and an incorrect prompt where the predicate is swapped with an unlikely relation. Probe-C looks at model understanding of possible compositional relations by comparing two images and two prompts where either the composition is swapped with an antonym or the object is swapped. Finally, Probe-B looks at model understanding of objects and their relationships to their surroundings by removing background and observing the change in performance.

We experimented with several state-of-the-art V+L models and provide several interesting insights regarding these models. For compositional understanding, we observe that (1) models struggle with compositionality, and (2) CNN based backbones may be better at recognizing texture and patterns while ViT backbones are better with color and shape. For relational understanding, we observe that (1) both modality specific attention and co-attention in parallel improve relational understanding, and (2) Predicate swapping that violates expectations surfaces the lack of an underlying conceptual model. For contextual understanding we observe that (1) models tend to not use context in order to recognize most objects, again indicating a lack of an underlying conceptual model. We further utilize these findings and develop a simple finetuning approach based on selective negatives paradigm and observe improvement on our understanding-related probes.

In summary, we make the following contributions:

- We study the capability of existing large V+L models for complex visual perception focusing on relational, compositional, and contextual understanding.
- We propose three benchmark datasets: Probe-R, Probe-C, and Probe-B focusing on subject-object relations, composition-object relations, and background-object relations.
- We perform extensive evaluation of existing models and provide new insights about their capabilities.
- We present a simple approach, based on prompting that rewards compositionality and preservation of relations between objects, which yields a more robust performance on complex visual perception tasks.

Table 1. **Comparison of ours with various recent works** probing relational, attribute, and context understanding of models.

Approach	Relational	Compositional	Contextual
VL-CheckList [52]	✓	✓	×
ARO [50]	✓	✓	×
SVLC [12]	✓	✓	×
ControlledImCaps [19]	✓	✓	×
CREPE [26]	✓	✓	×
SugarCREPE [15]	✓	✓	×
Ours	✓	✓	✓

2. Related Works

Several works have probed models to understand what models are learning [10, 12, 15, 18, 19, 26, 28, 42, 49, 50, 52]. Table 1 shows a comparison of our proposed benchmark against several other existing works that probe the different understanding property of V+L models. From the table, it is evident that none of the existing works probe the contextual understanding of V+L models, which our proposed dataset does. Moreover, our proposed benchmark has more images and is evaluated on more models than most of the existing methods, which will be discussed later. Even though SVLC [12] is a large-scale benchmark in terms of both size of dataset and number of evaluation models than our proposed benchmark, this is not sufficient for contextual understanding of the V+L models. An extension of Winoground [10] showed that models perform worse than humans because it requires both compositional understanding and commonsense reasoning. Without disentangling the individual skills required to perform well, it limits insights to why/how they are failing and where to improve. In this work, we generate a benchmark that isolates components of understanding into compositional, relational, and context, allowing for more detailed insights.

3. Benchmark and Evaluation Metrics

We evaluate three discrete concepts: object-relations, compositionality, and background context. We have generated three datasets: Probe-R, Probe-C, and Probe-B. A summary of each dataset is shown in Table 2 and an overview in Figure 2. *Prompting* is typically done in downstream image classification by forming sentences with each class name in the prompt, such as "a photo of a dog." The one with the highest similarity to the visual features is the predicted class [31, 38]. These benchmarks heavily rely on "prompting"

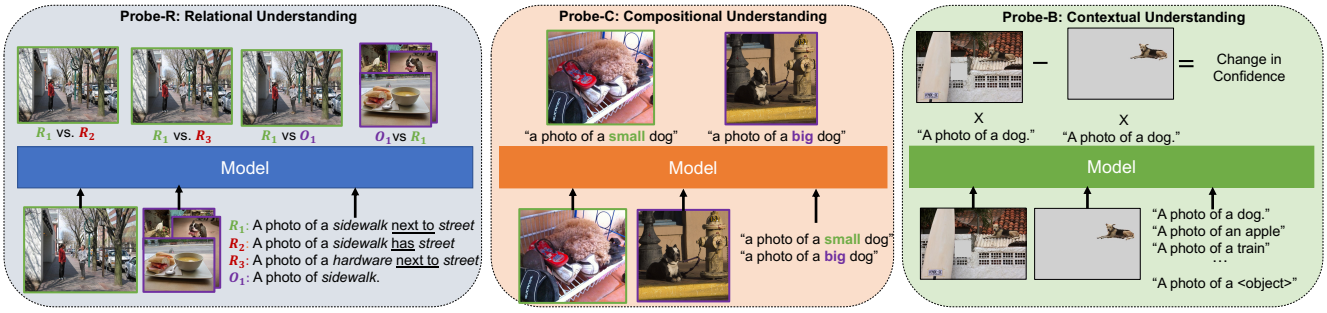

Figure 2. **Overview of proposed benchmarks**. *Probe-R* swaps the real subject or relation with an unlikely one and swaps a set of subject-only images to a subject-only prompt and the ground-truth relation prompt. *Probe-C* asks the model to match two images and two prompts, swapping object or composition. *Probe-B* compares object recognition performance before and after swapping out context from background and other surrounding objects.

the model by changing text input as well as image input in Probe-B.

3.1. Dataset

Probe-R: Relational Understanding To generate a dataset that can be used to probe for relational understanding, we collected samples from the Visual Genome [21] dataset. These samples are used to probe whether models have learned consistent concepts of objects and their potential relationships to each other. For each group, we have four prompts $P \in \{R_1, R_2, R_3, O_1\}$, one anchor image X_{R_1} and 10 images $\mathbf{X}_{O_1}$ with the subject present and no other objects found in the anchor image. For each X_{R_1}, the ground truth relation $R_1 = \langle s_1, r_1, o_1 \rangle$ is compared to a swap of subject $R_3 = \langle \overline{s}_1, r_1, o_1 \rangle$ or predicate $R_2 = \langle s_1, \overline{r}_1, o_1 \rangle$. We sample $\overline{s}_1$ uniformly from subjects that do not occur in the dataset with r_1 and o_1 and similarly $\overline{r}_1$ is sampled uniformly from relations that do not occur with s_1 and o_1. This swapping of unlikely subjects and predicates allows us to test whether V+L models have learned consistent conceptual models of what object relations are possible in a system by comparing existing ones to unlikely ones. The final comparison is subject-only images X_{O_1} to P_{R_1} and a prompt with only the subject P_{O_1}.

Probe-C: Compositional Understanding To generate a dataset that can be used to probe for compositional understanding, we collected samples from the MS COCO Captions dataset [24]. These samples are used to probe whether models have learned an understanding of object attributes and their relationships to each other. For each group, we have two images x_1 and x_2 and two prompts p_1 and p_2. This dataset has two splits, one where the compositions are swapped in the prompts and the other where objects are swapped. When swapping compositions, antonyms were manually mapped to each attribute to ensure that the attribute is not present in the image. For example, if there is a "small dog" in an image, the comparison could be "a large dog." When swapping objects, the images must have the same composition but different objects.

Probe-B: Contextual Understanding To generate a dataset that probes for model understanding on objects and their relationship to contextual cues found in an image's background, we collected samples from MS COCO [24] consisting of 80 objects. These samples are used to probe model reliance on background cues and reliance on co-occurrence between objects. For each group, there is an unmodified image x_0, an image with a random patch on the background $\tilde{x_0}$, a modified image where the background is removed $\tilde{x_1}$ and 80 or fewer prompts. We have two splits in this data, the first removing the background but keeping all objects Probe-B$_{MR}$ and the other removing both background and all other objects Probe-B$_R$. Probe-B$_R$ aims to probe models on whether they use conceptual maps on object co-occurrence to improve recognition. Probe-B$_{MR}$ aims to probe models on whether they have conceptual maps related to what group of objects are likely to be in what scenery or possible physical relations to each other. Poor performance on these tasks would indicate model use of such conceptual mappings, while good performance means they are focusing on object recognition only. We experiment with four fillers: black, gray, gaussian noise, or a random scene. Random scenery was collected from the Indoor Scenes Dataset [30] and the Kaggle Landscape dataset [34]. These images were manually filtered to ensure none of the 80 MS COCO classes were present. For single objects, images were only kept if the size of the object was between a threshold where the object was not too large and not too small relative to the image size.

3.2. Evaluation Metrics

We use different evaluation metrics for each of the three datasets, but with a focus on *change in model confidence*. This allows us to relate to the psychological paradigm "violation-of-expectation" (VoE) [2, 29]. If V+Ls are learning conceptual models, then a violation of those models should be easily recognized and confidence should remain

Table 2. **A summary of proposed benchmark datasets.** *Tasks* include Image-Text matching (ITM), multi-label object recognition (MLR), and object recognition (R). *Groups* refer to the group of images/text for each comparison being made. Under *attributes*, we list the dataset properties, where fillers are the types of replacements we use when removing background pixels.

Dataset	Task	Description	Source	Images	Group Description	Groups	Attributes
Probe-R	ITM	Predicate/ Object Swapping	Visual Genome	99,960	1 image, 10 pos. images, 4 prompts	99,960	2,456 Relations, 6,006 Objects
Probe-C	ITM	Composition Swapping Object Swapping	MS COCO	40,681 59,205	2 images, 2 prompts	79,925 375,607	114 Compositions, 2,462 Objects
Probe-B	MLR	Background Removal	MS COCO	31,745	3 images, 80 prompts	31,745	4 fillers, 80 objects
	R	Background+Object Removal		1,484	3 images, <80 prompts	9,375	4 fillers, 76 objects

high when choosing between the correct prompt and the prompt that is violating expectation. For Probe-R, by probing models with data that is intended to violate expectation, we expect the confidence to remain high. For Probe-C, by probing models with paired opposites, we also expect the confidence to be high. For Probe-B, by removing visual information or replacing it with a violation of the original information, we expect the model to become confused and therefore the confidence to be low.

Probe-R: We evaluate Probe-R using the mean confidence $\mu(c)$ and mean accuracy (acc) over all groups using equation 1, 2, and 3. We compare one image x to two prompts p_1 and p_2.

$$(c_1, c_2) = \sigma(f(x, p_1), f(x, p_2)) \tag{1}$$

$$\mu(c) = \frac{1}{N}\sum_{i=1}^{N} c_1^i \tag{2}$$

$$\text{acc} = \begin{cases} 1 & \text{if } c_1 > c_2 \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

Logit scores from model f are converted to softmax σ predictions to measure the confidence c_i of prompt p_i. Here N denotes total number of images.

Probe-C: To measure image and text matching between two images, x_1 and x_2, and two prompts, p_1 and p_2, using logit output from a model f, we adopt metrics from [42] measuring a text score (t), an image score (i), and a group score (g). t measures the accuracy of the model selecting the correct prompt for a given image by equation 4, 5, and 6.

$$t(p_1, x_1, p_2, x_2) = \begin{cases} 1 & \text{if } f(p_1, x_1) > f(p_2, x_1) \\ & \text{and } f(p_2, x_2) > f(p_1, x_2), \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

$$i(p_1, x_1, p_2, x_2) = \begin{cases} 1 & \text{if } f(p_1, x_1) > f(p_1, x_2) \\ & \text{and } f(p_2, x_2) > f(p_2, x_1), \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

$$g(p_1, x_1, p_2, x_2) = \begin{cases} 1 & \text{if } t(p_1, x_1, p_2, x_2) \\ & \text{and } i(p_1, x_1, p_2, x_2), \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

Probe-B: We evaluate model reliance on either the co-occurrence of objects or background cues. Both tasks compare to both an original image x_0 and the original image with an added patch of the respective filler $\tilde{x_0}$ to take into account general robustness. $\tilde{x_1}$ will have either the background removed and replaced with a filler or have the background and all other objects replaced. The fillers are one of: "black," "gray," "noise," or a random "scene" that does not have objects. The metrics we use for comparisons are the mean average precision (mAP) for multi-object recognition precision, relative robustness γ^r measuring the relative drop/increase in performance (equation 7, and mean change in mAP $\mu(\triangle(c))$ (equation 8) for the objects. γ^r and *mAP* evaluates how much the models rely on background context to accurately describe the scenario. We collect the similarity between the image x_n and for each object o placed in a prompt $p_o \in \mathbf{p}$. This results in a set of similarity scores for each object prompt which is used to calculate the score of model's change in confidence $\triangle c$.

$$\gamma^r = 1 - \frac{h(x, \mathbf{p}) - h(\tilde{x}, \mathbf{p})}{h(x, \mathbf{p})} \tag{7}$$

$$\triangle \text{c}(x, \tilde{x}) = \frac{1}{o}\sum_{i=1}^{o} f(x, p_o) - f(\tilde{x}, p_o) \tag{8}$$

4. Benchmark Results

Here we go through the models we are evaluating in this benchmark and then report the results of those models on the proposed datasets Probe-R, Probe-C, and Probe-B.

Models We perform our experiments on ten recently developed and publicly available models: CLIP [31], FLAVA [38], ViLT [20], BridgeTower [47], BLIP [22], BLIP2 [23], OTTER [45], ALIGN [17], MetaCLIP [46], and SigLIP [51]. **CLIP** [31] is a dual-stream, modality specific model that has a visual and text encoder of equal length and limited modality interaction. It uses a contrastive loss between text-image pairs as its only multimodal signal. **FLAVA** [38] is also a dual-stream encoder with an additional multimodal encoder that takes the ViT based [11] single-stream encoders, merges them, and co-attends. It performs unimodal training for single-stream encoders followed by multimodal training on a global contrastive loss, a masked multimodal modeling task (MMM), and an image-text matching (ITM) loss. **ViLT** [20] is a single-stream transformer that uses co-attention between modalities. It concatenates

word embeddings and linear projections of image patches as input to a pre-trained ViT [11]. It trains using an ITM loss, a masked language modeling (MLM) loss, and a word-patch alignment loss. **Bridgetower** [47] uses a dual-stream encoder with a multimodal encoder that incorporates the single-stream encoders at multiple layers using cross-attention based "bridge layers." It uses a pre-trained ViT from CLIP as visual encoder, RoBERTa [25] as text encoder, and is trained with MLM and ITM losses. **BLIP** [22] utilizes a mixture of encoder-decoder, and can operate in three functionalities: unimodal encoder, image-grounded text encoder, and image-grounded text decoder. **BLIP2** [23] uses a querying transformer that's at first trained in vision-language representation learning stage then vision-to language generative learning stage. It is a trainable module bridging the gap between the frozen image encoder and LLM. **OTTER** [45] improves upon CLIP by using online entropic optimal transport to efficiently learn image-text pairs. **ALIGN** [17] is a dual-encoder which uses EfficientNet as image encoder and BERT as text encoder trained on a noisy dataset over one-billion image-text pairs. **MetaCLIP** [46] follows CLIP by constructing metadata and carefully curating image-text pairs to imitate their dataset and training procedure. **SigLIP** [51] improves upon CLIP by introducing a pairwise Sigmoid loss instead of standard contrastive learning.

4.1. Relational Evaluation

Models become confused when predicate is swapped, but more confident when object is swapped: The overall results for the relation evaluation benchmark are shown in Figure 3 (left) where it shows each model's accuracy and mean confidence $\mu(c)$ for matching the prompt to the anchor image X_{R_1}. When comparing an image to a correct prompt and an incorrect prompt where the relation/predicate is swapped with one not likely nor present in the image, the model's $\mu(c)$ for the correct prompt compared to incorrect is very low. This may indicated that selected models become "confused" when the relation is switched, even if it is a highly unlikely relation to even exist between the two objects. When swapping objects, the object that is swapped $\overline{s}$ is one that is highly unlikely, making this task simple if the model has a consistent understanding of what relationships are possible. Model confidence is higher when the object is swapped versus when the predicate is swapped. This may indicate that models are less confused when the task is specific to object recognition, focusing more on objects rather than the relationships between them. This may additionally indicate they are not understanding prompts as a "whole" but rather parts to a whole. To visualize the differences between models, we plot some of their feature space in Figure 3 (right). We see very different structures for BridgeTower and ViLT which heavily rely on cross-attention and image-text matching (ITM) when compared to FLAVA and CLIP.
Summary: BridgeTower and ViLT's performance indicates that co-attention is a method that can improve relational understanding. (1) This would indicate that both modality specific attention and co-attention simultaneously improves relational understanding. (2) When the predicate is swapped to something that violates expectation, the drop in confidence, regardless of accuracy, indicates that their performance may not be due to an underlying conceptual map. (3) When the subject is swapped, all models show better performance compared to predicate swapping, indicating they are focusing on objects less-so than their relations to each other.

4.2. Compositional Evaluation

Modality-specific attention and co-attention simultaneously greatly improves attribute-object relation understanding: Overall results for evaluating model understanding of composition-object relationships are shown in Figure 4 (left) with additional results in the Supplementary. We show the image, group, and object scores for when the object (Obj.) is switched and for when the composition (Comp.) is switched. When presented with two images and two captions where the composition is the same but the objects are different, all models other than BridgeTower and BLIP2 perform on average double the performance versus when the composition is switched. This discrepancy indicates typically models are *relying more on object recognition* when compositions are involved. BridgeTower and BLIP2's high performance indicates further support that a combination of modality-specific attention and cross-attention in parallel improves the learning of underlying concepts.
Models stronger with more physical attributes like "materials" compared to visual-related like "color": To better understand model failures when keeping the objects the same but swapping an attribute with its antonym, we categorized each attribute into 11 categories with results shown in Figure 4 (middle). Attribute details are presented in the supplementary. All models struggle with "visibility" related compositions. The best performance was within the "material" and "pattern" category.
Transformers and CNNs differ on which attributes they understand best: To compare backbone models, we average the image scores over CLIP backbone architectures in Figure 4 (right). Some noticeable patterns are that the CNN backbone models are better with "material," "pattern," and "texture" related compositions while ViT's are better at "color" and "shape." This finding aligns with the findings of [14] where they found ImageNet trained CNNs are biased towards texture.
Summary: (1) Models struggle with compositionality but are better with those most associated with objects such as "materials." (2) CNN based backbones may be better at

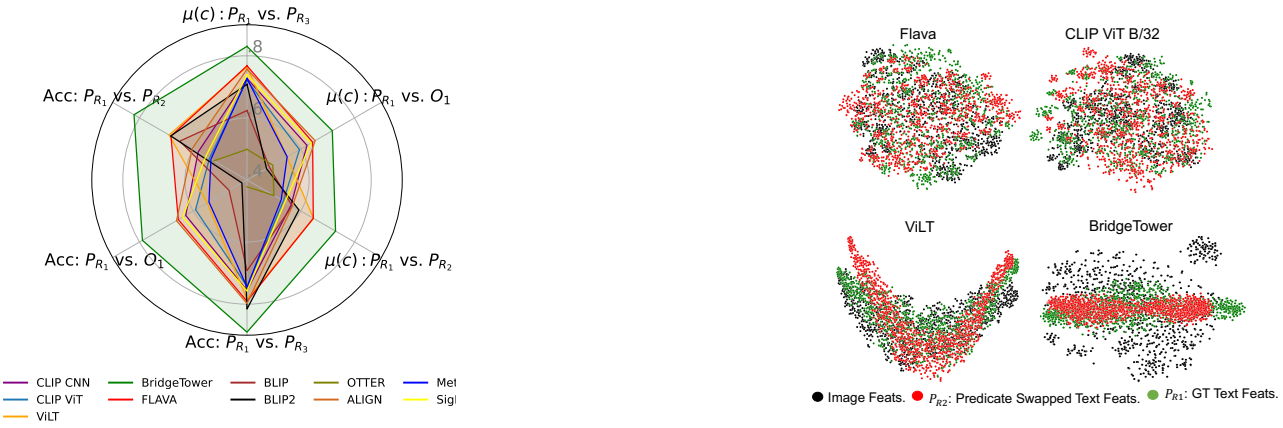

Figure 3. **Model's performance on relational understanding on Probe-R**. (left) Radar plot showing *accuracy* and *mean confidence* $\mu(c)$ of different models. Here, the anchor image X_{R_1} contains the relation $R_1 = \langle s, r, o \rangle$, image X_{O_1} contains $O_1 = \langle s \rangle$. Prompts contain either the relation P_{R_1}, $P_{R_2} = \langle s, \overline{r}, o \rangle$, $P_{R_3} = \langle \overline{s}, r, o \rangle$, or $P_{O_1} = \langle s \rangle$. (right) TSNE plot of the feature space for image features for some models where the prompt with the predicate swapped is denoted by P_{R_2} and the ground truth prompt denoted by P_{R_1}.

recognizing texture and patterns while ViT backbones with color and shape. Surprisingly, (3) these models are typically better at matching captions given the image rather than text.

4.3. Background Context Evaluation

Models ignore what the background is replaced with, indicating little use of it: Overall results for evaluating model context understanding of background-object relationships are shown in Figure 5 and 6. Figure 5 (left) shows the results averaged over filler type when only the background is removed. The most noticeable change is when comparing the ground truth image to $\tilde{x_0}$ and $\tilde{x_1}$ as expected. Overall, models are slightly less robust to when the background is replaced with either Gaussian noise or scenery. However, if models had underlying understanding of what objects belong in what context, models should be less robust to scenery. This indicates they may not have conceptual maps about objects and their relationship to context.

More co-attention may result in greater trade-off between robustness and performance: Figure 5 (right) shows the overall results averaged over model type when only the background is removed. Similar to when looking at fillers, models are typically robust to background removal, indicating little use of context. ALIGN tends to improve when the background is removed. However, ViLT, ALIGN and MetaCLIP tend to be less robust when a patch is added to the image, noticeable even more so when the robustness between $\tilde{x_0}$ and $\tilde{x_1}$ is so high. This appears to be a trade-off between robustness and performance. **Some objects benefit from the presence of others, but most are better off without:** Figure 6 shows the overall results for when the background and all other objects but one are removed, averaged over either filler (left) or model (middle). When averaging over filler, models appear to be more robust when detecting one object as opposed to multiple objects in an image. When averaging scores over models, γ^r tends to be over 1 when comparing to the background removed image $\tilde{x_1}$, indicating models improve when objects are in isolation. This case is especially prominent for ALIGN. This indicates that models may be distracted from background information rather than using it for object recognition. In order to better understand what objects models are using background with more than others, we categorize objects into sub-categories as shown in Figure 8. The object-types that models struggle with most appear to be large objects used in a common setting, such as "ovens" for appliances and "sink" for fixtures. This may indicate that there is some context used but for certain objects more than others.

Summary (1) Models tend to not use context in order to recognize multiple objects but (2) for some individual objects, models do use context. (3) These models are typically robust to a change in background where models like ViLT, ALIGN, and BridgeTower are more susceptible to a particular patch being changed. (4) When objects are placed in random scenery that violates-expectation, models still perform similarly to when the original background is there. This may indicate that overall, models are not learning conceptual maps relating objects to their context.

5. Finetuning for better conceptual understanding

Dual-stream encoders like CLIP and FLAVA allow unimodal feature representations that can be extracted and used

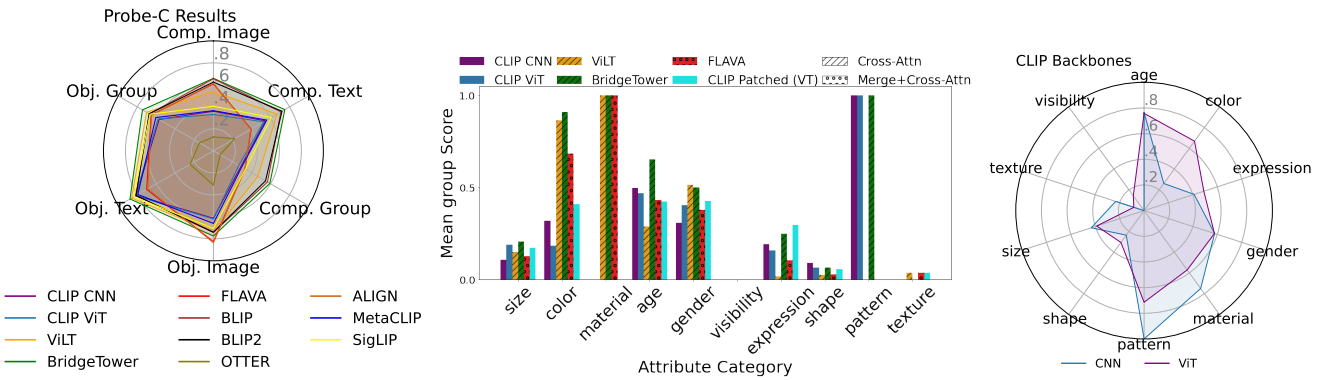

Figure 4. **Model's performance on compositional understanding on Probe-C.** (left) The overall results for Probe-C showing the image, text, and group scores for when the object is swapped (*Obj.*) or when the composition is swapped (*Comp.*). (middle) Mean group score averaged across attribute categories. (right) CLIP scores averaged over different backbones.

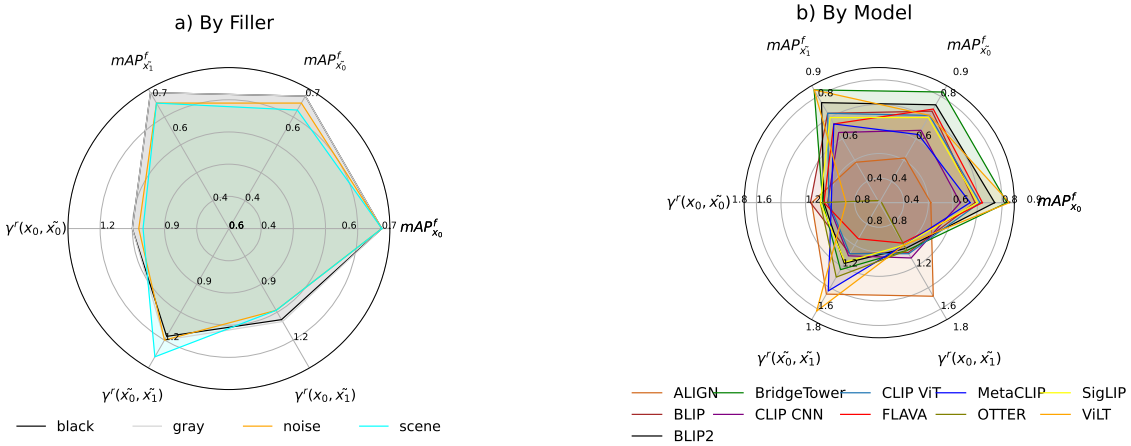

Figure 5. **Model's performance on contextual understanding on Probe-B for only background removal.** (left) Mean results for replacing background with filler and (right) for each model averaged over fillers. Comparisons between the original x_0, original+random patch $\tilde{x_0}$ and modified $\tilde{x_1}$. The metrics are *mAP* and γ^r.

for a variety of downstream tasks. Improving models that do not require paired input would provide greater value and stronger representations. To explore this idea, we finetune (FT) CLIP ViT-B/32 on a new dataset inspired by this benchmark called RelComp. The new dataset RelComp for attribute-object and object-object relations is based on MS COCO [24] and VisualGenome [21] and has no overlap between the benchmark datasets. We propose using selective negative and positive pairing based on attribute and predicate swaps and finetune using image-text matching (ITM) loss and a contrastive loss (C) [31, 38] for finetuning (see Figure 7). We linearly interpolate the original CLIP weights with our FT weights using an alpha= 0.2 to prevent "catastrophic forgetting" [16, 44]. We call this "CLIP Patched" and finetune visual-encoder only (V), text-encoder only (T), or both (VT). More details about losses, implementation, and dataset are in the Supplementary.

Overall results for our exploratory experiment are shown in Table 3. We observe drift as measured by ImageNet accuracy, even when patching. When finetuning using the visual-encoder only, the drift is less pronounced, but so is the improvement on RelComp. The largest increase is seen with FT text encoder only. This may indicate that for non-cross-attention models, text is more important for conceptual mapping. Our findings indicate that by using selective negative sampling we can enforce compositional and relational learning without extensive co-attention and computa-

Table 3. **Performance on finetuned and patched CLIP on proposed RelComp dataset.** ImageNet accuracy is shown to measure the drift from the original CLIP space. RelComp and Probe-C/R respectively report image score and mean accuracy for the correct image-to-prompt matching.

Model	ImageNet	RelComp	Probe-C	Probe-R
ViLT	–	76.00	90.78	69.00
BridgeTower	–	85.00	90.06	82.20
FLAVA	56.83	47.12	83.85	68.29
CLIP ViT B32	**63.60**	51.93	88.15	53.52
CLIP Patched (T)	57.85	**67.85**	89.49	<u>71.14</u>
CLIP Patched (V)	<u>61.45</u>	54.66	<u>89.81</u>	61.40
CLIP Patched (VT)	54.61	<u>64.27</u>	**90.30**	**71.20**

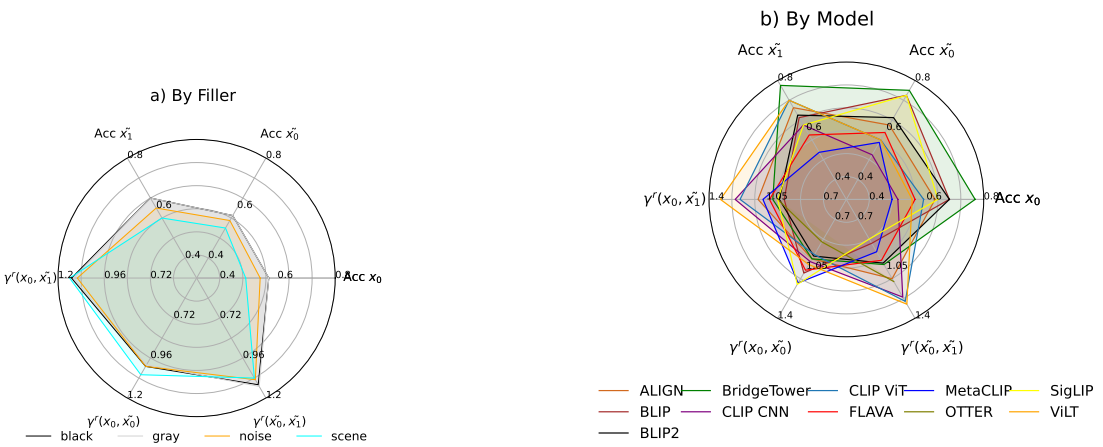

Figure 6. **Model's performance on contextual understanding on Probe-B on background and all but one object removal.**(Left): Results for when the background and all other objects are replaced with a filler $\tilde{x_1}$, compared to the original x_0, and (right) original+random patch $\tilde{x_0}$.

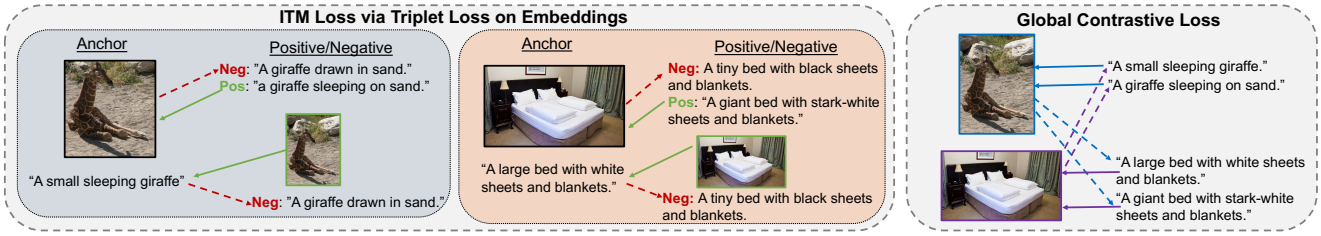

Figure 7. **Exploratory finetuning training scheme for CLIP.** Image-text matching (ITM) is used as a triplet loss whose pairings vary depending on if it is a compositional or a relational task. A contrastive loss is used to maintain general representations.

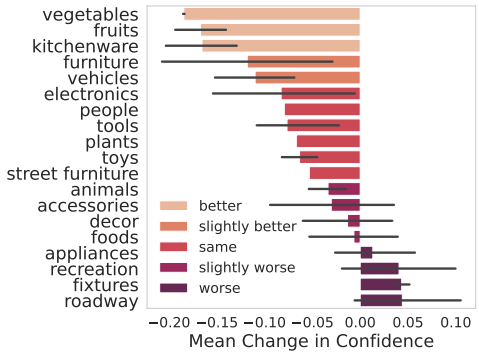

Figure 8. **Performance on contextual understanding for certain object categories.** Mean change in confidence ($\mu(\triangle c)$) from the ground truth x_0 to the modified $\tilde{x}_1$, where the background and other objects are removed.

tional complexity.

6. Conclusions

In this benchmark we evaluated large visual-language (V+L) models on relational, compositional, and contextual understanding with three new datasets: Probe-C, Probe-R, and Probe-B. For compositional understanding, we observe (1) models struggle with compositionality. (2) CNN backbones may be better at recognizing texture and patterns while ViT backbones are with color and shape. For relational understanding, we observe (1) both modality specific attention and co-attention in parallel improves relational understanding. (2) An expectation violating predicate swap surfaces the lack of a conceptual map through drop in confidence. For contextual understanding we observe (1) models mostly tend to not use context in order to recognize multiple objects. (2) When objects are placed in random scenery that violates expectation, model performance is unchanged, indicating a lack of conceptual map of context. When trying to improve CLIP, the dual-encoder with no cross-attention, by finetuning on our proposed selective negatives training paradigm on the proposed RelComp dataset, (1) we find that there is a small drop in classification performance, but (2) an improvement on Probe-R, Probe-C, and RelComp is observed, indicating an improvement in relational and compositional learning. We hope these insights will help drive future work on building V+L models that better "understand."

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch,

Katie Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *arXiv preprint arXiv:2204.14198*, 2022. 1

[2] Renee Baillargeon. Object permanence in 31/2-and 41/2-month-old infants. *Developmental psychology*, 23(5):655, 1987. 3

[3] Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.", 2009. 1

[4] Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.", 2009. 11

[5] Nicholas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramèr, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models. *arXiv preprint arXiv:2301.13188*, 2023. 2

[6] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3558–3568, 2021. 7

[7] Gal Chechik, Varun Sharma, Uri Shalit, and Samy Bengio. Large scale online learning of image similarity through ranking. *Journal of Machine Learning Research*, 11(3), 2010. 7

[8] for Social Security Administration Disability Determinations; Board on the Health of Select Populations; Institute of Medicine Committee on Psychological Testing, Including Validity Testing. Psychological testing in the service of disability determination, 2015. 1

[9] Karan Desai, Gaurav Kaul, Zubin Aysola, and Justin Johnson. Redcaps: Web-curated image-text data created by the people, for the people. *arXiv preprint arXiv:2111.11431*, 2021. 7

[10] Anuj Diwan, Layne Berry, Eunsol Choi, David Harwath, and Kyle Mahowald. Why is winoground hard? investigating failures in visuolinguistic compositionality. *arXiv preprint arXiv:2211.00768*, 2022. 2

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021. 4, 5

[12] Sivan Doveh, Assaf Arbelle, Sivan Harary, Rameswar Panda, Roei Herzig, Eli Schwartz, Donghyun Kim, Raja Giryes, Rogerio Feris, Shimon Ullman, and Leonid Karlinsky. Teaching structured vision&language concepts to vision&language models, 2023. 2

[13] Steven M. Frankland and Joshua D. Greene. Concepts and compositionality: In search of the brain's language of thought. *Annual Review of Psychology*, 71(1):273–303, 2020. 1

[14] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*, 2018. 5

[15] Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. Sugarcrepe: Fixing hackable benchmarks for vision-language compositionality, 2023. 2

[16] Gabriel Ilharco, Mitchell Wortsman, Samir Yitzhak Gadre, Shuran Song, Hannaneh Hajishirzi, Simon Kornblith, Ali Farhadi, and Ludwig Schmidt. Patching open-vocabulary models by interpolating weights. *NeurIPs*, 2022. 7, 2, 8, 9, 10

[17] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision, 2021. 4, 5

[18] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

[19] Amita Kamath, Jack Hessel, and Kai-Wei Chang. Text encoders bottleneck compositionality in contrastive vision-language models, 2023. 2

[20] 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. 4, 7, 10

[21] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalanditis, Li-Jia Li, David A Shamma, Michael Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. 2016. 3, 7, 1, 9, 11

[22] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, 2022. 4, 5

[23] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023. 4, 5

[24] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 3, 7, 1, 9, 11

[25] 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. 5

[26] Zixian Ma, Jerry Hong, Mustafa Omer Gul, Mona Gandhi, Irena Gao, and Ranjay Krishna. Crepe: Can vision-language foundation models reason compositionally?, 2023. 2

[27] Richard E Mayer. Models for understanding. *Review of educational research*, 59(1):43–64, 1989. 1

[28] Genevieve Patterson and James Hays. Coco attributes: Attributes for people, animals, and objects. In *Computer Vision–ECCV 2016: 14th European Conference, Amster-*

dam, The Netherlands, October 11-14, 2016, Proceedings, Part VI 14, pages 85–100. Springer, 2016. 2

[29] Luis S Piloto, Ari Weinstein, Peter Battaglia, and Matthew Botvinick. Intuitive physics learning in a deep-learning model inspired by developmental psychology. *Nature human behaviour*, 6(9):1257–1267, 2022. 1, 3

[30] Ariadna Quattoni and Antonio Torralba. Recognizing indoor scenes. In *2009 IEEE conference on computer vision and pattern recognition*, pages 413–420. IEEE, 2009. 3, 1, 11

[31] 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. 1, 2, 4, 7, 10

[32] Yongming Rao, Wenliang Zhao, Guangyi Chen, Yansong Tang, Zheng Zhu, Guan Huang, Jie Zhou, and Jiwen Lu. Denseclip: Language-guided dense prediction with context-aware prompting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18082–18091, 2022. 1

[33] Arnaud ROUGETET. Kaggle landscape dataset. https : / / www . kaggle . com / datasets / arnaud58/landscape-pictures/. 11

[34] Arnaud Rougetet. Landscape pictures. https://www.kaggle.com/datasets/arnaud58/landscape-pictures, 2021. Accessed: February 16, 2023. 3, 1

[35] Pritish Sahu, Michael Cogswell, Yunye Gong, and Ajay Divakaran. Unpacking large language models with conceptual consistency. *arXiv preprint arXiv:2209.15093*, 2022. 2

[36] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021. 7

[37] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, Melbourne, Australia, 2018. Association for Computational Linguistics. 7

[38] 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. 2, 4, 7

[39] Robyn Speer, Joshua Chin, and Catherine Havasi. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the AAAI conference on artificial intelligence*, 2017. 2

[40] Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2443–2449, 2021. 7

[41] Bart Thomee, David A. Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. *Commun. ACM*, 59(2):64–73, 2016. 7

[42] Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. 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, 2022. 2, 4

[43] Zhaoqing Wang, Yu Lu, Qiang Li, Xunqiang Tao, Yandong Guo, Mingming Gong, and Tongliang Liu. Cris: Clip-driven referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11686–11695, 2022. 1

[44] Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, et al. Robust fine-tuning of zero-shot models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7959–7971, 2022. 7, 10

[45] Bichen Wu, Ruizhe Cheng, Peizhao Zhang, Tianren Gao, Peter Vajda, and Joseph E. Gonzalez. Data efficient language-supervised zero-shot recognition with optimal transport distillation, 2023. 4, 5

[46] Hu Xu, Saining Xie, Xiaoqing Ellen Tan, Po-Yao Huang, Russell Howes, Vasu Sharma, Shang-Wen Li, Gargi Ghosh, Luke Zettlemoyer, and Christoph Feichtenhofer. Demystifying clip data, 2023. 4, 5

[47] Xiao Xu, Chenfei Wu, Shachar Rosenman, Vasudev Lal, and Nan Duan. Bridge-tower: Building bridges between encoders in vision-language representation learning. *arXiv preprint arXiv:2206.08657*, 2022. 1, 4, 5, 7

[48] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014. 11

[49] Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bag-of-words models, and what to do about it? *arXiv preprint arXiv:2210.01936*, 2022. 2

[50] Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it?, 2023. 2

[51] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training, 2023. 4, 5

[52] Tiancheng Zhao, Tianqi Zhang, Mingwei Zhu, Haozhan Shen, Kyusong Lee, Xiaopeng Lu, and Jianwei Yin. Vl-checklist: Evaluating pre-trained vision-language models with objects, attributes and relations, 2023. 2

[53] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Com-*

puter Vision and Pattern Recognition, pages 16816–16825, 2022. 1