Diagnosing the Compositional Knowledge of Vision Language Models from a Game-Theoretic View

Jin Wang [1] **Shichao Dong** [4] **Yapeng Zhu** [4] **Kelu Yao** [3] **Weidong Zhao** [3] **Chao Li** [3] **Ping Luo** [2 1]

Abstract

Compositional reasoning capabilities are usually considered as fundamental skills to characterize human perception. Recent studies show that current Vision Language Models (VLMs) surprisingly lack sufficient knowledge with respect to such capabilities. To this end, we propose to thoroughly diagnose the composition representations encoded by VLMs, systematically revealing the potential cause for this weakness. Specifically, we propose evaluation methods from a novel game-theoretic view to assess the vulnerability of VLMs on different aspects of compositional understanding, *e.g.*, relations and attributes. Extensive experimental results demonstrate and validate several insights to understand the incapabilities of VLMs on compositional reasoning, which provide useful and reliable guidance for future studies. The deliverables will be updated here.

1. Introduction

Recently, Vision Language Models (VLMs) (Radford et al., 2021; Jia et al., 2021; Li et al., 2022b; Singh et al., 2022; Goel et al., 2022; Yao et al., 2021) have made remarkable strides, which significantly advance a wide range of unimodal and multimodal applications, *e.g.*, object detection (Gu et al., 2021; Du et al., 2022; Zang et al., 2022), semantic segmentation (Zhou et al., 2022; Li et al., 2022a; Xu et al., 2022) and text-to-image generation (Rombach et al., 2022; Ramesh et al., 2022). However, recent studies have unveiled a surprising weakness of state-of-the-art VLMs: they struggle with compositional reasoning capabilities (Yuksekgonul et al., 2022; Thrush et al., 2022), such as object relations and object attributes. For instance, BLIP (Li et al., 2022b) failed to correctly comprehend the subtle differences between "the horse is eating the grass" and "the grass is eating the horse" (Yuksekgonul et al., 2022). Given the fundamental status of compositionality in human intelligence (Cresswell, 1973), this lack of compositional knowledge has hindered the further development of vision language models.

Previous studies on the compositionality of VLMs mainly focused on two perspectives. Some studies proposed to evaluate the compositional reasoning capabilities of VLMs in a black-box probing manner (Yuksekgonul et al., 2022; Thrush et al., 2022; Ma et al., 2023; Hsieh et al., 2023; Zhao et al., 2022). They usually measured the accuracy performance on whether VLMs correctly retrieved the matching text for a given image between two captions with minimal changes. Other studies proposed to improve the compositionality of VLMs in an empirical manner by introducing the supervision of scene graphs (Herzig et al., 2023; Huang et al., 2023b) or curated hard-negative samples (Yuksekgonul et al., 2022; Doveh et al., 2023). However, there still lack in-depth analyses to thoroughly diagnose the internal compositional representations of VLMs, which can help us understand the essential cause of this weakness and provide reliable guidance for future studies.

Therefore, in this paper, we propose to take a further step and conduct detailed analyses on the potential causes of VLMs' poor compositional reasoning capabilities. Since a VLM usually contains an image encoder and a text encoder as a whole, we propose to comprehensively evaluate VLMs by firstly focusing on the compositional knowledge of each unimodal encoder *separately* and then the multimodal compositional knowledge *jointly*. Under this scheme, we expect to answer the following questions.

• *Question 1.* **Does the text encoder of a VLM understand texts compositionally?**

• *Question 2.* **Does the image encoder of a VLM understand images compositionally?**

• *Question 3.* **Do the text encoder and the image encoder of a VLM have mutually-matching knowledge on compositionality?**

In this way, such a disentangled representation dissection scheme can help us obtain a more meticulous and fine-

[1]Department of Computer Science, The University of Hong Kong, Hong Kong [2] Shanghai AI Laboratory, China [3] Zhejiang Laboratory, Hangzhou, China [4] Baidu Inc, Beijing, China . Correspondence to: Ping Luo <pluo@cs.hku.hk>, Chao Li <lichao@zhejianglab.com>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

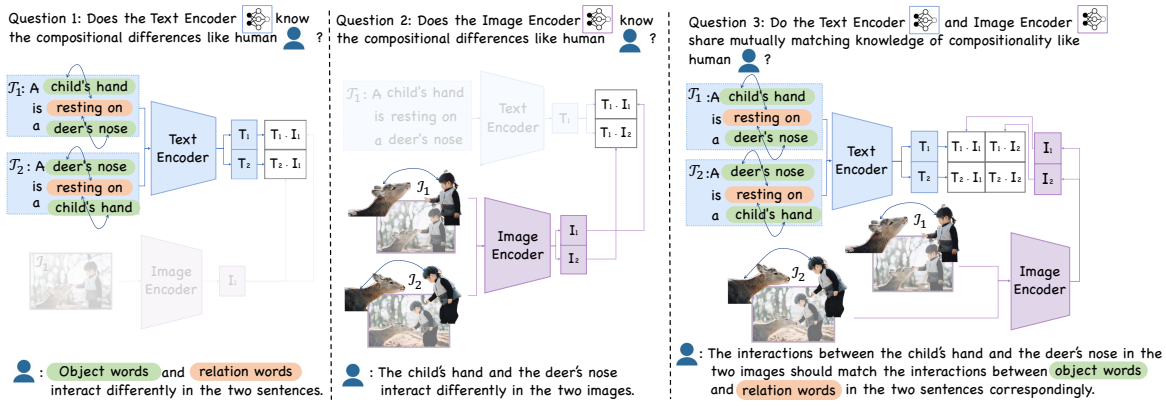

Figure 1. Diagnosing the compositional reasoning capabilities of Vision Language Models (VLMs). In this paper, we systematically analyze the potential causes for the poor compositional performance of VLMs from each unimodal *separately* and then multimodal *jointly*. In this way, three insights are obtained and validated correspondingly. Please see Section 1 for detailed elaborations.

grained understanding on VLMs' poor compositional reasoning capabilities.

To answer previous questions, we propose to evaluate the compositional representations inside VLMs from a novel game-theoretic view. Specifically, we propose several metrics based on the Harsanyi dividend (Harsanyi, 1963) to assess the sensitivities of VLMs to the changes of different compositionality aspects, *e.g.*, relations and attributes. The Harsanyi dividend was originally proposed in game theory to measure the interactions between different players, which makes itself a natural metric to dissect the compositional knowledge of DNNs. Besides, the Harsanyi dividend is related to the Shapley value (Shapley, 1953), theoretically satisfying *the efficiency, linearity, dummy, symmetry axioms*, which further ensures the trustworthiness of the interpretations for DNNs (Ren et al., 2022; 2023; Li & Zhang, 2023).

In this way, we conduct extensive evaluations on five state-of-the-art VLMs (Radford et al., 2021; Li et al., 2022b; Zeng et al., 2022; Yuksekgonul et al., 2022; Singh et al., 2022) with four widely-used datasets (Yuksekgonul et al., 2022; Zhao et al., 2022; Hsieh et al., 2023; Wang et al., 2023), obtaining and validating several fine-grained insights on the internal representations of VLMs *w.r.t.* the compositional reasoning capabilities.

• *Insight 1.* **It is to our surprise that text encoders of VLMs show excellent compositional reasoning capabilities, able to recognize the dominant compositional differences between input texts like human understanding.**

• *Insight 2.* **Image encoders of VLMs demonstrate compositional reasoning capabilities to some extent, which are relatively weaker than the corresponding text encoders, partially resulting in the poor compositional performance of VLMs.**

• *Insight 3.* **Although text encoders and image encoders show certain compositional reasoning capabilities individually, they do not share mutually-matching compositional knowledge, which also partially accounts for the poor compositional abilities of VLMs[1].**

These insights provide a detailed understanding on the potential causes of VLMs' poor compositional knowledge, which can provide beneficial and reliable instructions for future studies. For instance, to bring more significant performance gain on the compositional reasoning tasks, it may be more effective to design stronger image encoders instead of text encoders for VLMs.

The contributions of our paper are summarized as follows. 1) We conduct a systematical analysis to diagnose the internal representations of VLMs, progressively revealing the potential causes for their weakness in compositional reasoning capabilities. 2) We propose several metrics from a novel game-theoretic view to assess the vulnerability of VLMs on different aspects of compositional understanding, *e.g.*, relations and attributes. 3) Experimental results on various state-of-the-art VLMs and datasets demonstrate and validate several insights on the compositional reasoning capabilities of VLMs, which can help instruct future studies for more effective improvements of VLMs.

2. Related Work

Vision Language Models. In recent years, Vision Language Models (VLMs) (Chen et al., 2020; Li et al., 2019a;

[1]Compared to previous studies (Herzig et al., 2023; Huang et al., 2023b; Doveh et al., 2023) which mainly provided intuitive understanding in this aspect, we presented detailed and in-depth analyses from a novel game-theoretic view to further validate this assumption in this paper.

Tan & Bansal, 2019; Li et al., 2021a; Jia et al., 2021; Goel et al., 2022; Singh et al., 2022; Liu et al., 2023; Zhu et al., 2023) have received an increasing research focus, providing benefits for both unimodal and multimodal applications. Generally, VLMs were trained to generate correspondences between input texts and images. Their encoded representations were then evaluated on many zero-short or few-short downstream tasks, such as CLIP (Radford et al., 2021), X-VLM (Zeng et al., 2022), BLIP (Li et al., 2022b) and etc. However, recent studies showed that despite the high performance on dozens of well-established benchmarks (Yao et al., 2021; Li et al., 2021b; Gao et al., 2022), most of the VLMs surprisingly exhibited poor compositional understanding capabilities (Yuksekgonul et al., 2022; Thrush et al., 2022).

Compositionality of VLMs. Previous studies firstly designed a number of new benchmarks to comprehensively evaluate the compositional reasoning capabilities of VLMs, such as Winoground (Thrush et al., 2022), ARO (Yuksekgonul et al., 2022), VL-CheckList (Zhao et al., 2022), CREPE (Ma et al., 2023), SUGARCREPE (Hsieh et al., 2023), COLA (Ray et al., 2023), EQBEN (Wang et al., 2023), SyViC (Cascante-Bonilla et al., 2023) and SPEC (Peng et al., 2024). However, these benchmarks usually evaluated the compositionality of VLMs with the accuracy metric, testing whether the paired images and texts could be correctly retrieved among perturbed samples. Such a black-box probing scheme failed to provide further explanations for the unsatisfying performance. Besides the development of these benchmarks, other studies focused on improving the compositional performance of VLMs in an empirical manner, such as introducing the guidance of scene graphs (Herzig et al., 2023; Huang et al., 2023b) and generating curated hard-negative samples (Yuksekgonul et al., 2022; Doveh et al., 2023; Sahin et al., 2023; Momeni et al., 2023). However, there still lacks a thorough representation diagnosis on the compositional reasoning capabilities of VLMs, so as to systematically unveil the essential causes for this weakness.

Interactions of DNNs. Considerable studies have focused on quantifying the interactions among input units for diagnosing the representation of DNNs (Grabisch & Roubens, 1999; Zhang et al., 2020b; 2021a; Wang et al., 2020; Ren et al., 2021; Wang et al., 2021; Yao et al., 2023; Dong et al., 2022; Chen et al., 2023). Based on the Shapley value (Shapley, 1953) originally proposed in game theory, some studies (Grabisch & Roubens, 1999; Sundararajan et al., 2020) proposed interaction metrics to quantify the relationships among the input units, such as the Harsanyi dividend (Harsanyi, 1963). Besides, Zhang *et al.* (2021b; 2020a) further extended the interaction metric to the multi-order and multivariate interactions, which were applied to explain several phenomena of DNNs (Deng et al., 2021; Wang et al., 2020; Ren et al., 2021). In comparison, our study aims to provide detailed explanations on the poor compositional reasoning capabilities of VLMs.

3. Quantifying the compositional knowledge of VLMs with the Harsanyi dividend

In this paper, we propose several metrics based on the Harsanyi dividend (Harsanyi, 1963) to evaluate the compositional reasoning capabilities of VLMs from different aspects. To this end, we first present a brief introduction to the Harsanyi dividend for better understanding.

3.1. The Harsanyi dividend

The Harsanyi dividend was a typical metric in game theory (Harsanyi, 1963), which measures the interaction among a set of players. Specifically, given a set of players $\mathcal{N} = \{1, 2, ..., n\}$ participating in a game v, certain *rewards* can be obtained. Here, $v(\cdot)$ represents a function to map any subset of players $\mathcal{S} \subseteq \mathcal{N}$ to a real number, representing the obtained numerical *reward*. Intuitively, during such a game, each player usually does not contribute to the *reward* individually, but interacts with each other, forming different *coalitions*/*patterns* to cause casual effects on the final outcome. Mathematically, such effects can be measured by the Harsanyi dividend, which is defined as follows:

$$w(\mathcal{S}|\mathcal{N}) = \sum_{\mathcal{S}' \subseteq \mathcal{S}} (-1)^{|\mathcal{S}'|-|\mathcal{S}|} \cdot v(\mathcal{S}'). \tag{1}$$

Besides, the Harsanyi dividend also satisfies many axioms to theoretically support the fairness and trustworthiness of its calculation (Grabisch et al., 2016; Ren et al., 2023). Please see Appendix A for details.

3.2. Quantifying the compositional knowledge of VLMs

Based on the definition of the Harsanyi dividend, we then elaborate on how to quantify the compositional knowledge of VLMs with it. To limit the scope of discussion in this paper, we mainly focus on the aspects of compositionality as follows: relations, objects and attributes. To this end, we believe that **a VLM with a comprehensive and excellent compositional reasoning capability should be sensitive to the changes of objects, relations, attributes, and also, being sensitive to the changes of interactions among them**. To be specific, let us take the samples in Figure 1 for an example.

Given the two input captions in Figure 1 showing changes regarding the textual compositionality, they both contain the *same* words but these words have *different* interactions with each other, describing different relations between objects. To this end, a VLM with an excellent compositional reasoning capability should learn that the object words alone (*i.e.*, {*child's hand, deer's nose*}) or the relation words alone

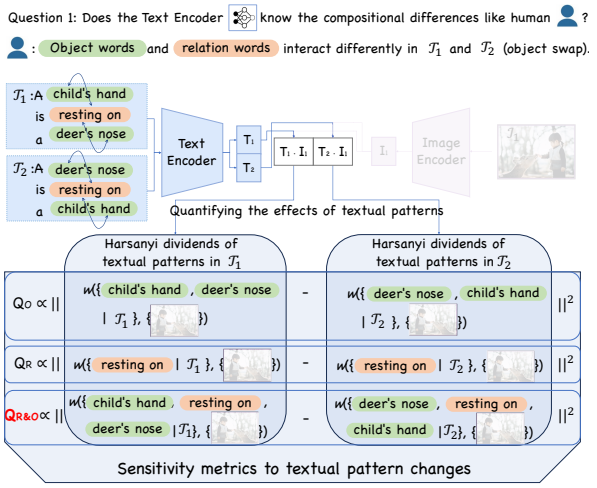

Figure 2. Evaluating the sensitivities of text encoders of VLMs to the changes of textual patterns. Specifically, given captions $\mathcal{T}_1$ and $\mathcal{T}_2$ with object words being swapped, we design Q_O, Q_R and $Q_{R\&O}$ to assess whether text encoders react correctly to the fine-grained changes of compositionality. In this case, $Q_{R\&O}$, which measures the interaction changes between object words and relation words, should be of a greater value than Q_O and Q_R. Here $\mathrm{T}_1 \cdot \mathrm{I}_1$ represents the cosine similarity between the normalized text embedding T_1 and normalized image embedding I_1.

(*i.e.*, {*resting on*}), have almost the same casual effect on the understanding of each caption. By contrast, the interactions between the object words and the relation words in each caption should have different casual effects on the understanding of each caption correspondingly.

Besides, given the two input images in Figure 1 showing changes in terms of the visual compositionality, they both contain objects of the *same* identity (*i.e.*, image regions showing the same child and the same deer in these two images.), but these objects interact *differently* with each other. In this way, between these two images, a VLM with an excellent compositional reasoning capability should learn that image regions of the same object alone (*i.e.*, image regions of the child or the deer alone) should make similar casual effects on the representation of each image. On the contrary, the interactions between these two object regions should have different casual effects.

In order to meticulously examine the above fine-grained understanding of compositionality inside VLMs, we propose to exploit the Harsanyi dividend to first quantitatively measure the effect of each visual and textual pattern on the output of VLMs. In this way, we can further quantitatively evaluate whether VLMs show sharp sensitivities to the changes of these patterns, which relate to different aspects of compositionality. Specifically, we can analogously consider the inference process of VLMs as a game $v(\cdot,\cdot)$ with two sets of players $\mathcal{N}^{\mathcal{I}} = \{1, 2, ..., n^{\mathcal{I}}\}$ (*e.g.*, all the visual concepts on an image) and $\mathcal{N}^{\mathcal{T}} = \{1, 2, ..., n^{\mathcal{T}}\}$ (*e.g.*, all the words in a caption). Here, $v(\cdot,\cdot)$ represents the output of VLMs, which measures the matching similarity between an input image and an input text, *e.g.*, the cosine similarity between an input image embedding and an input text embedding for CLIP (Radford et al., 2021). The casual effect of the pattern $\mathcal{S}^{\mathcal{I}} \subseteq \mathcal{N}^{\mathcal{I}}$ and $\mathcal{S}^{\mathcal{T}} \subseteq \mathcal{N}^{\mathcal{T}}$ defined by the Harsanyi dividend is then calculated as,

$$w(\{\mathcal{S}^{\mathcal{I}}, \mathcal{S}^{\mathcal{T}}\}|\{\mathcal{N}^{\mathcal{I}}, \mathcal{N}^{\mathcal{T}}\}) = \sum_{\substack{\mathcal{S}^{\mathcal{I}'} \subseteq \mathcal{S}^{\mathcal{I}}, \\ \mathcal{S}^{\mathcal{T}'} \subseteq \mathcal{S}^{\mathcal{T}}}} (-1)^{|\mathcal{S}^{\mathcal{I}'}|-|\mathcal{S}^{\mathcal{I}}|+|\mathcal{S}^{\mathcal{T}'}|-|\mathcal{S}^{\mathcal{T}}|} \cdot v(\mathcal{S}^{\mathcal{I}'}, \mathcal{S}^{\mathcal{T}'}). \quad (2)$$

where $\mathcal{S}^{\mathcal{I}'}$ represents the input image with only the visual concepts in the subset $\mathcal{S}^{\mathcal{I}'}$ while masking other visual concepts in $\mathcal{N}^{\mathcal{I}} \backslash \mathcal{S}^{\mathcal{I}'}$; $\mathcal{S}^{\mathcal{T}'}$ represents the input caption with only the words in the subset $\mathcal{S}^{\mathcal{T}'}$ while masking other words in $\mathcal{N}^{\mathcal{T}} \backslash \mathcal{S}^{\mathcal{T}'}$[2]. For simplicity, we denote $w(\{\mathcal{S}^{\mathcal{I}}, \mathcal{S}^{\mathcal{T}}\}|\{\mathcal{N}^{\mathcal{I}}, \mathcal{N}^{\mathcal{T}}\})$ as $w(\{\mathcal{S}^{\mathcal{I}}, \mathcal{S}^{\mathcal{T}}\})$ in the following sections.

With Equation 2 measuring the effects of visual/textual patterns, we then design metrics to evaluate VLMs' sensitivities to the changes of these patterns, so as to fully examine the compositional reasoning capabilities of VLMs in a fine-grained manner. In the following sections, we start from each unimodal representations of VLMs *separately*, and then to the multimodal representations of VLMs *jointly*.

4. Can text encoders of VLMs understand texts compositionally?

To fully examine the compositional reasoning capabilities of VLMs in a fine-grained manner, we first explore whether text encoders of VLMs encode reliable compositional knowledge in the first place. Specifically, we propose several metrics based on the Harsanyi dividend to quantitatively evaluate the sensitivities of text encoders to the changes of textual patterns, which relate to different aspects of compositionality.

As shown in Figure 2, given an image-text pair $\{\mathcal{I}_1, \mathcal{T}_1\}$ and a perturbed text $\mathcal{T}_2$, which is generated from swapping object words in $\mathcal{T}_1$, VLMs like CLIP (Radford et al., 2021) would output two matching scores $v(\mathcal{N}^{\mathcal{I}_1}, \mathcal{N}^{\mathcal{T}_1})) = \mathrm{I}_1 \cdot \mathrm{T}_1$ and $v(\mathcal{N}^{\mathcal{I}_1}, \mathcal{N}^{\mathcal{T}_2})) = \mathrm{I}_1 \cdot \mathrm{T}_2$, where I_1 represents the normalized image embedding of $\mathcal{I}_1$ and $\mathrm{T}_{1/2}$ represents the normalized text embedding of $\mathcal{T}_{1/2}$. In this way, the variations between $\mathrm{I}_1 \cdot \mathrm{T}_1$ and $\mathrm{I}_1 \cdot \mathrm{T}_2$ can be considered as the

[2]In this paper, the baseline value for masking out image regions/text tokens was set as zero, following previous studies (Ancona et al., 2019; Wang et al., 2021; Zhang et al., 2020b; 2021b; Dong et al., 2022).

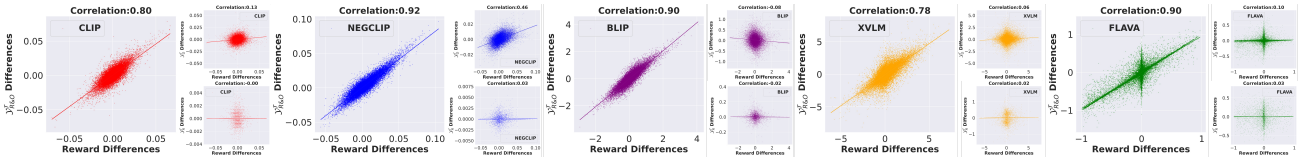

Figure 3. The Pearson correlation coefficients $\rho(\mathcal{X}^{\mathcal{T}}, \mathcal{Y}^{\mathcal{T}})$ between the reward differences $\mathcal{X}^{\mathcal{T}}$ and interaction effect differences on the Visual Genome Relation dataset, *i.e.*, $\mathcal{Y}_R^{\mathcal{T}}$, $\mathcal{Y}_O^{\mathcal{T}}$ and $\mathcal{Y}_{R\&O}^{\mathcal{T}}$. Each point represents a data sample containing two captions and one image. Results show that the reward differences between the two captions were mainly related to the interaction changes of object words and relation words, demonstrating that text encoders of VLMs reacted correctly to the textual compositional differences in this dataset.

coarse measurement on the sensitivities of text encoders to the changes of a mixture of textual patterns. However, it still remains uncertain whether text encoders of VLMs react to each specific texture pattern in a correct manner individually. To this end, we expect to take a step further and comprehensively analyze *fine-grained* sensitivities of text encoders, regarding different aspects of compositionality. Mathematically, we propose the sensitivity metric as follows,

$$Q(\mathcal{T}_1, \mathcal{T}_2, \mathcal{I}_1) = \frac{1}{\mathcal{Z}_{sens}^{\mathcal{T}}} ||w(\{\mathcal{N}^{\mathcal{I}_1}, \mathcal{S}^{\mathcal{T}_1}\}) - w(\{\mathcal{N}^{\mathcal{I}_1}, \mathcal{S}^{\mathcal{T}_2}\})||^2 \quad (3)$$

where $\mathcal{Z}_{sens}^{\mathcal{T}} = E_{\mathcal{T}' \in \{\mathcal{T}_1, \mathcal{T}_2\}} E_{\mathcal{S}^{\mathcal{T}'}} ||w(\{\mathcal{N}^{\mathcal{I}_1}, \mathcal{S}^{\mathcal{T}'}\})||^2$ is used for normalization. Intuitively, this metric measures detailed textual casual effect changes between text $\mathcal{T}_1$ and $\mathcal{T}_2$ when given the image $\mathcal{I}_1$[3]. In this way, given different subsets of words $\mathcal{S}^{\mathcal{T}_1}$ and $\mathcal{S}^{\mathcal{T}_2}$, we mainly compute five types of sensitivity metrics in implementations.

• **The sensitivity metric of relation words** Q_R. This metric measures the casual effect changes of only relation words on the output of text encoders of VLMs, where $\mathcal{S}^{\mathcal{T}_1}$ and $\mathcal{S}^{\mathcal{T}_2}$ contain relation words only.

• **The sensitivity metric of attribute words** Q_A. This metric measures the casual effect changes of only attribute words on the output of text encoders of VLMs, where $\mathcal{S}^{\mathcal{T}_1}$ and $\mathcal{S}^{\mathcal{T}_2}$ contain attribute words only.

• **The sensitivity metric of object words** Q_O. This metric measures the casual effect changes of only object words on the output of text encoders of VLMs, where $\mathcal{S}^{\mathcal{T}_1}$ and $\mathcal{S}^{\mathcal{T}_2}$ contain object words only.

• **The sensitivity metric of interaction between relation words and object words** $Q_{R\&O}$. This metric measures the casual effect changes of interactions between relation words and object words, where $\mathcal{S}^{\mathcal{T}_1}$ and $\mathcal{S}^{\mathcal{T}_2}$ contain both relation words and object words.

[3]Note that to control the input variables for better clarity, in Eq. 3, we do not mask out any image regions when only analysing text pattern casual effect changes.

Table 1. Evaluating the compositional sensitivities of text encoders of VLMs. In the Visual Genome Relation dataset, the object words are swapped to obtain perturbed texts. Results show that $Q_{R\&O}$ (bold) have larger values than Q_R and Q_O in this dataset across different VLMs, demonstrating that text encoders of various VLMs exhibit accurate sensitivities to the changes of textual patterns.

Dataset	Models	Q_O	Q_R	$Q_{R\&O}$
Visual Genome Relation	CLIP	4.5e-3	9.8e-6	**1.3e-2**
	NEGCLIP	3.7e-3	1.1e-5	**2.0e-2**
	BLIP	3.3e-2	1.3e-4	**1.7e-1**
	XVLM	5.0e-2	1.1e-3	**1.4e-1**
	FLAVA	7.4e-2	3.0e-2	**4.3e-1**

• **The sensitivity metric of interaction between attribute words and object words** $Q_{A\&O}$. This metric measures the casual effect changes of interactions between attribute words and object words, where $\mathcal{S}^{\mathcal{T}_1}$ and $\mathcal{S}^{\mathcal{T}_2}$ contain both attribute words and object words.

Experiment protocols. Based on the metrics, we systematically analyzed the compositional knowledge of text encoders of various VLMs. Specifically, we evaluated text encoders of five state-of-the-art VLMs: CLIP (Radford et al., 2021), BLIP (Li et al., 2022b), NEGCLIP (Yuksekgonul et al., 2022), XVLM (Zeng et al., 2022) and FLAVA (Singh et al., 2022). The evaluations were conducted on three popular benchmarks: ARO (Yuksekgonul et al., 2022), SUGARCREPE (Hsieh et al., 2023) and VL-CheckList (Zhao et al., 2022). Due to the page limitation, please see Appendix C for results on SUGARCREPE and VL-CheckList.

Experimental results on ARO benchmark. In this experiment, we mainly focused on the attribute and relation aspects of compositionality, using two of the sub-datasets in the ARO benchmark: Visual Genome Attribution and Visual Genome Relation (Krishna et al., 2017; Hudson & Manning, 2019). Each sample in the dataset includes one image and two captions with minimal differences. We here present results on the Visual Genome Relation dataset in the main paper. Please see Appendix C for results on the Visual Genome Attribution dataset.

In the Visual Genome Relation dataset, the input correct and

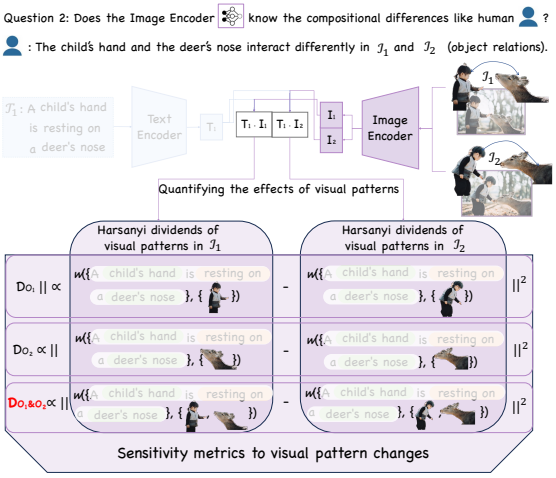

Figure 4. Evaluating the sensitivities of image encoders of VLMs to the changes of visual patterns. Specifically, given images $\mathcal{I}_1$ and $\mathcal{I}_2$ with object relations being altered, we design D_{O_1}, D_{O_2} and $D_{O_1\&O_2}$ to assess whether image encoders react correctly to the fine-grained changes of visual compositionality. In this case, $D_{O_1\&O_2}$, which measures the relation changes between objects, should be of a greater value than D_{O_1} and D_{O_2}.

wrong captions are in the templates of "*[object 1] [relation] [object 2]*" and "*[object 2] [relation] [object 1]*" with the *SWAP* manipulation. To this end, we mainly evaluated the sensitivity metrics of relation words, object words and the interactions between relation words and object words, *i.e.*, Q_R, Q_O and $Q_{R\&O}$. Masks denoting the relation words and object words are obtained directly based on the templates. Based on human understanding, the compositional differences between the correct and wrong captions in this dataset should be largely reflected by the interaction changes between relation words and object words (*i.e.*, $Q_{R\&O}$). By contrast, the casual effect changes of relation words alone and object words alone should be less significant. Results are reported in Table 1, where $Q_{R\&O}$ has the highest values among metrics for various VLMs. **It is to our surprise that despite the poor performance of VLMs on this benchmark, text encoders of VLMs did recognize the dominant compositional differences between captions in the relation-object aspect, similar to human understanding**.

To further examine the compositional knowledge of text encoders, we propose to calculate the Pearson correlation coefficients between the reward differences and interaction effect differences. Specifically, we calculated the reward differences $\mathcal{X}^{\mathcal{T}}$ and interaction effect differences $\mathcal{Y}^{\mathcal{T}}$ as follows: $\mathcal{X}^{\mathcal{T}} = v(\mathcal{N}^{\mathcal{I}_1}, \mathcal{N}^{\mathcal{T}_1}) - v(\mathcal{N}^{\mathcal{I}_1}, \mathcal{N}^{\mathcal{T}_2})$; $\mathcal{Y}^{\mathcal{T}} = w(\mathcal{N}^{\mathcal{I}_1}, \mathcal{S}^{\mathcal{T}_1}) - w(\mathcal{N}^{\mathcal{I}_1}, \mathcal{S}^{\mathcal{T}_2})$. We then calculated the Pearson correlation coefficients $\rho(\mathcal{X}^{\mathcal{T}}, \mathcal{Y}^{\mathcal{T}})$ for different subsets of words $\mathcal{S}^{\mathcal{T}_1}$ and $\mathcal{S}^{\mathcal{T}_2}$, *i.e.*, $\rho(\mathcal{X}^{\mathcal{T}}, \mathcal{Y}^{\mathcal{T}}_O)$, $\rho(\mathcal{X}^{\mathcal{T}}, \mathcal{Y}^{\mathcal{T}}_R)$, $\rho(\mathcal{X}^{\mathcal{T}}, \mathcal{Y}^{\mathcal{T}}_{R\&O})$. Here $\mathcal{Y}^{\mathcal{T}}_O$ represents the casual effect changes when $\mathcal{S}^{\mathcal{T}_1}$ and $\mathcal{S}^{\mathcal{T}_2}$ only contain object words. $\mathcal{Y}^{\mathcal{T}}_R$ represents the casual effect changes when $\mathcal{S}^{\mathcal{T}_1}$ and $\mathcal{S}^{\mathcal{T}_2}$ only contain relation words. $\mathcal{Y}^{\mathcal{T}}_{R\&O}$ represents the casual effect changes when $\mathcal{S}^{\mathcal{T}_1}$ and $\mathcal{S}^{\mathcal{T}_2}$ contain both object and relation words. As shown in Figure 3, $\mathcal{Y}^{\mathcal{T}}_{R\&O}$ has a significant positive correlation with the final reward differences $\mathcal{X}^{\mathcal{T}}$ among various VLMs. Such results demonstrate that the reward differences between correct captions and negative captions were mainly caused by the interaction effect changes between relation words and object words, which was surprisingly consistent with human understanding, despite the unsatisfying performance on this benchmark (Yuksekgonul et al., 2022).

Table 2. Evaluating the compositional sensitivities of image encoders of VLMs with the EQBEN dataset, where object relations alter but each object maintains the same identity within image pairs. The maximum metrics are shown in bold.

Dataset	Models	D_{O_1}	D_{O_2}	$D_{O_1\&O_2}$	ρ_{O_1}	ρ_{O_2}	$\rho_{O_1\&O_2}$
EQBEN	CLIP	1.7e-2	**3.1e-2**	2.0e-2	0.34	**0.56**	0.12
	NEGCLIP	3.7e-2	**7.5e-2**	3.4e-2	0.37	**0.67**	0.05
	BLIP	5.6e-1	**7.4e-1**	4.4e-1	0.36	**0.46**	0.23
	XVLM	8.2e-1	**1.2e0**	8.1e-1	0.39	**0.51**	0.13
	FLAVA	6.9e-1	1.0e0	**1.2e0**	0.20	**0.33**	0.20

5. Can image encoders of VLMs understand images compositionally?

In previous analyses, we surprisingly find text encoders of VLMs demonstrate sharp sensitivities to different aspects of compositionality in texts. Such experimental results significantly reduce the accountability of text encoders for the poor compositionality performance of VLMs. To rule out other potential causes, we then turn our concentration on exploring whether image encoders should be held accountable for this weakness.

Similar to the analyses on text encoders of VLMs, we hope to answer the following key question in this section: do image encoders correctly know the fine-grained compositional differences between two images? As shown in Figure 4, given an image-text pair $\{\mathcal{I}_1, \mathcal{T}_1\}$ and a perturbed image $\mathcal{I}_2$, where the relation between the child and the deer varies. It is expected that the relation changes between the child and the deer should be considered the essential compositional differences between images $\mathcal{I}_1$ and $\mathcal{I}_2$. In the meantime, casual effect changes of the child alone or the deer alone should be considered less significant. To diagnose such detailed visual compositional understanding of VLMs, we propose to calculate the following sensitivity metric,

$$D(\mathcal{I}_1, \mathcal{I}_2, \mathcal{T}_1) = \frac{1}{\mathcal{Z}^{\mathcal{I}}_{sens}} ||w(\{\mathcal{S}^{\mathcal{I}_1}, \mathcal{N}^{\mathcal{T}_1}\}) - w(\{\mathcal{S}^{\mathcal{I}_2}, \mathcal{N}^{\mathcal{T}_1}\})||^2 \quad (4)$$

where $\mathcal{Z}^{\mathcal{I}}_{sens} = E_{\mathcal{I}' \in \{\mathcal{I}_1, \mathcal{I}_2\}} E_{\mathcal{S}^{\mathcal{I}'}} ||w(\{\mathcal{S}^{\mathcal{I}'}, \mathcal{N}^{\mathcal{T}_1}\})||^2$ is calculated for normalization. Intuitively, this metric measures detailed visual casual effect changes between $\mathcal{I}_1$ and $\mathcal{I}_2$ when given the text $\mathcal{T}_1$[4]. To limit the discussion in this paper, we only focus on image pairs $\mathcal{I}_1$ and $\mathcal{I}_2$, which contain the same object pair but show different object relations as in Figure 4. We leave analyses on more complex visual cases for future studies. To this end, we mainly calculated two types of sensitivity metrics given different sets of visual concepts $\mathcal{S}^{\mathcal{I}_1}$ and $\mathcal{S}^{\mathcal{I}_2}$ in implementations.

• **The sensitivity metric of each object** D_{O_1}/D_{O_2}. This metric measures the casual effect changes of each individual object region on the output of image encoders of VLMs, where $\mathcal{S}^{\mathcal{I}_1}$ and $\mathcal{S}^{\mathcal{I}_2}$ contain the object regions of the same identity, *i.e.*, object O_1/O_2.

• **The sensitivity metric of relation within a pair of objects** $D_{O_1\&O_2}$. This metric measures the casual effect change of the relation within a pair of objects O_1 and O_2, where $\mathcal{S}^{\mathcal{I}_1}$ and $\mathcal{S}^{\mathcal{I}_2}$ contain both object regions.

Experiment protocols. Following previous analyses, we continued to evaluate the compositional knowledge of image encoders for CLIP, NEGCLIP, BLIP, XVLM and FLAVA. As for benchmarks, we exploited the EQBEN dataset (Wang et al., 2023) collected from natural videos (Zhou et al., 2018; Ji et al., 2020; Wang et al., 2022) or synthetic engines (Rombach et al., 2022; Hertz et al., 2022; Greff et al., 2022). In this dataset, each sample contains two image-text matching pairs sharing minimal differences. However, to conduct quantitative evaluations with Eq. 4, we need to obtain the pixel-wise mask for each described object on images, which is not provided in the original dataset. To this end, we carefully selected a subset of data samples (290 image-text pairs) from the original dataset and utilized SAM (Kirillov et al., 2023) to help obtain the final mask for each described object in images. Please see Appendix B for annotations[5].

Experimental results on EQBEN benchmark. Besides the proposed metrics, we also calculated the Pearson correlation coefficients ρ_{O_1}, ρ_{O_2} and $\rho_{O_1\&O_2}$ for further evaluations, following Sec. 4. Here $\rho_{O_1/O_2} = \rho(\mathcal{X}^{\mathcal{I}}, \mathcal{Y}^{\mathcal{I}}_{O_1/O_2})$ and $\rho_{O_1\&O_2} = \rho(\mathcal{X}^{\mathcal{I}}, \mathcal{Y}^{\mathcal{I}}_{O_1\&O_2})$, where $\mathcal{X}^{\mathcal{I}} = v(\mathcal{N}^{\mathcal{I}_1}, \mathcal{N}^{\mathcal{T}_1}) - v(\mathcal{N}^{\mathcal{I}_2}, \mathcal{N}^{\mathcal{T}_1})$; $\mathcal{Y}^{\mathcal{I}} = w(\mathcal{N}^{\mathcal{I}_1}, \mathcal{S}^{\mathcal{T}_1}) - w(\mathcal{N}^{\mathcal{I}_2}, \mathcal{S}^{\mathcal{T}_1})$. $\mathcal{Y}^{\mathcal{I}}_{O_1/O_2}$ represents the casual effect changes when $\mathcal{S}^{\mathcal{I}_1}$ and $\mathcal{S}^{\mathcal{I}_2}$ only contain the same object O_1/O_2. $\mathcal{Y}^{\mathcal{I}}_{O_1\&O_2}$ represents the casual effect changes when $\mathcal{S}^{\mathcal{I}_1}$ and $\mathcal{S}^{\mathcal{I}_2}$ contain both objects O_1 and O_2.

Results are summarized in Table 2. On the one hand, different from the trend in Table 1 where text encoders show sharp sensitivities to the relation changes between objects (*i.e.*, $Q_{R\&O}$ being larger than Q_O and Q_R.), $D_{O_1\&O_2}$ had a smaller value than D_{O_1} and D_{O_2} for most VLMs, which shows that image encoders demonstrated less accurate compositional sensitivities regarding the changes of object relations. They were more sensitive to the mild and less significant changes of each object alone, instead of the major relation changes between objects. Besides, the coefficients results show that the reward differences within the image pairs were not mostly related to the relation changes between objects, *i.e.*, $\rho_{O_1\&O_2}$ did not have the largest value than ρ_{O_1} and ρ_{O_2}, which was less consistent with human understanding compared to text encoders of VLMs. In summary, the above results show that **image encoders demonstrated weaker compositional reasoning capabilities, which may partially result in the overall poor compositional performance of VLMs.**

6. Do text encoders and image encoders have matching compositional knowledge?

Based on previous analyses, we find that text encoders and image encoders of VLMs both demonstrate certain sensitivities to the compositional changes of input, though image encoders are less sensitive. To further comprehensively examine the compositional knowledge of VLMs, we then evaluate whether text encoders and image encoders have mutually matching compositional knowledge. In other words, do image encoders correctly consider the interaction between object words and relation words in texts as the relations between objects in images, or mistakenly consider as each object alone? Similarly, do text encoders correctly consider the relations between objects in images as the interactions between object words and relation words in texts, or mistakenly consider as the object/relation words alone?

To evaluate the correspondence between the compositional knowledge encoded inside text encoders and image encoders, we propose to compute the following modified metrics for image-text pairs describing relations between two objects,

$$Q_{\mathcal{T}:R\&O\rightarrow\mathcal{I}:(\cdot)} = \frac{1}{\hat{\mathcal{Z}}^{\mathcal{T}}_{sens}} ||w(\{\mathcal{S}^{\mathcal{I}_1}, \mathcal{S}^{\mathcal{T}_1}_{R\&O}\}) - w(\{\mathcal{S}^{\mathcal{I}_1}, \mathcal{S}^{\mathcal{T}_2}_{R\&O}\})||^2 \quad (5)$$

$$D_{\mathcal{I}:O_1\&O_2\rightarrow\mathcal{T}:(\cdot)} = \frac{1}{\hat{\mathcal{Z}}^{\mathcal{I}}_{sens}} ||w(\{\mathcal{S}^{\mathcal{I}_1}_{O_1\&O_2}, \mathcal{S}^{\mathcal{T}_1}\}) - w(\{\mathcal{S}^{\mathcal{I}_2}_{O_1\&O_2}, \mathcal{S}^{\mathcal{T}_1}\})||^2 \quad (6)$$

where $\hat{\mathcal{Z}}^{\mathcal{T}}_{sens} = E_{\mathcal{T}' \in \{\mathcal{T}_1, \mathcal{T}_2\}} E_{\mathcal{S}^{\mathcal{T}'}} E_{\mathcal{S}^{\mathcal{I}_1}} ||w(\{\mathcal{S}^{\mathcal{I}_1}, \mathcal{S}^{\mathcal{T}'}_{R\&O}\})||^2$ and $\hat{\mathcal{Z}}^{\mathcal{I}}_{sens} = E_{\mathcal{I}' \in \{\mathcal{I}_1, \mathcal{I}_2\}} E_{\mathcal{S}^{\mathcal{I}'}} E_{\mathcal{S}^{\mathcal{T}_1}} ||w(\{\mathcal{S}^{\mathcal{I}'}_{O_1\&O_2}, \mathcal{S}^{\mathcal{T}_1}\})||^2$ are used for normalization. Here, for $\mathcal{T}' \in \{\mathcal{T}_1, \mathcal{T}_2\}$, $\mathcal{S}^{\mathcal{T}'}_{R\&O}$ denotes the object words and relation words in texts; for

[4] Similar to Eq. 3, we do not mask out any text tokens when only analysing visual pattern casual effect changes for clarity.

[5] New annotations are available here.

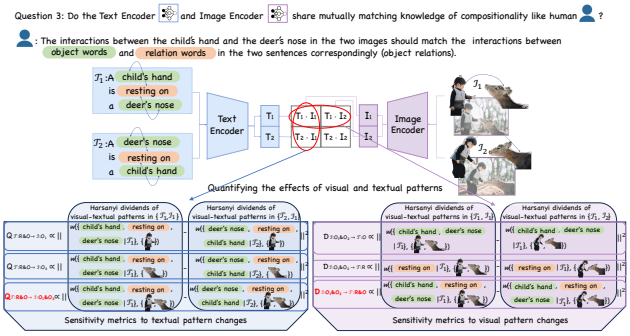

Figure 5. Evaluating whether image encoders and text encoders of VLMs possess mutually matching compositional knowledge with modified sensitivity metrics. Specifically, given image-text pairs $\{\mathcal{I}_1, \mathcal{T}_1\}$ and $\{\mathcal{I}_2, \mathcal{T}_2\}$ sharing minimal differences of object relations, we design $Q_{\mathcal{T}:R\&O\rightarrow\mathcal{I}:O_1}$, $Q_{\mathcal{T}:R\&O\rightarrow\mathcal{I}:O_2}$ and $Q_{\mathcal{T}:R\&O\rightarrow\mathcal{I}:O_1\&O_2}$ to assess whether image encoders obtain the corresponding compositional knowledge for text encoders. Besides, we also design $D_{\mathcal{I}:O_1\&O_2\rightarrow\mathcal{T}:O}$, $D_{\mathcal{I}:O_1\&O_2\rightarrow\mathcal{T}:R}$ and $D_{\mathcal{I}:O_1\&O_2\rightarrow\mathcal{T}:R\&O}$ to assess whether text encoders obtain the corresponding compositional knowledge for image encoders. Please zoom in for better visualization.

$\mathcal{I}' \in \{\mathcal{I}_1, \mathcal{I}_2\}$, $\mathcal{S}^{\mathcal{I}'}_{O_1\&O_2}$ denotes the image regions of object O_1 and object O_2.

Intuitively, Eq. 5 aims to examine VLMs by measuring what components of images are more related to the interactions between object words and relation words. As shown in Figure 5, we mainly calculated $Q_{\mathcal{T}:R\&O\rightarrow\mathcal{I}:O_1/O_2}$ and $Q_{\mathcal{T}:R\&O\rightarrow\mathcal{I}:O_1\&O_2}$. Taking $Q_{\mathcal{T}:R\&O\rightarrow\mathcal{I}:O_1}$ as an example, it measures how the interaction changes between object words and relation words within $\mathcal{T}_1$ and $\mathcal{T}_2$ would affect the output of VLMs when only given the image region of the child in $\mathcal{I}_1$. In this case, $Q_{\mathcal{T}:R\&O\rightarrow\mathcal{I}:O_1\&O_2}$ should be of the greatest value, showing that only when the visual pattern demonstrates the interactions between objects, swapping object words to change their relations in the textual pattern would cause significant influence on the output of VLMs.

Similarly, Eq. 6 aims to examine VLMs by measuring what components of texts are more related to the interaction between objects on images. As shown in Figure 5, we mainly calculated $D_{\mathcal{I}:O_1\&O_2\rightarrow\mathcal{T}:O/R}$ and $D_{\mathcal{I}:O_1\&O_2\rightarrow\mathcal{T}:R\&O}$. Taking $D_{\mathcal{I}:O_1\&O_2\rightarrow\mathcal{T}:O}$ as an example, it measures how the relation changes between objects within $\mathcal{I}_1$ and $\mathcal{I}_2$ would affect the output of VLMs, when only given the object words in $\mathcal{T}_1$. In this case, $D_{\mathcal{I}:O_1\&O_2\rightarrow\mathcal{T}:R\&O}$ should be of the greatest value, showing that only when the textual pattern demonstrates the relations between objects, altering the relations of objects in images would put significant effects on the output of VLMs.

Experiment results. Following previous analyses, we continued analyzing CLIP, NEGCLIP, BLIP, XVLM, FLAVA and harnessed our annotated EQBEN sub-dataset for evaluations, considering each sample contains two images and two texts both with minimal differences. Results in Tab. 3 show that $Q_{\mathcal{T}:R\&O\rightarrow\mathcal{I}:O_1\&O_2}$ did not maintain the largest among all three metrics, showing that in terms of object relations, image encoders of VLMs did not learn corresponding visual patterns to match the textual object relation patterns encoded inside text encoders. Instead, image encoders tended to associate the representations of mere objects with the interactions between object words and relation words learned by text encoders (*e.g.*, $Q_{\mathcal{T}:R\&O\rightarrow\mathcal{I}:O_2}$ showed the largest value for all VLMs.). Meanwhile, $D_{\mathcal{I}:O_1\&O_2\rightarrow\mathcal{T}:R\&O}$ failed to maintain the largest among all three metrics across all VLMs as well, showing that regarding the object relations, text encoders also did not associate the corresponding textual patterns to the visual object relation patterns learned by image encoders. They sometimes considered the representations of object words alone to be more related to the object relations depicted in images (*e.g.*, $D_{\mathcal{I}:O_1\&O_2\rightarrow\mathcal{T}:O}$ had the largest value for CLIP/NEGCLIP.). In summary, **these models did not exhibit mutually matching compositional knowledge from the text and visual sides, which may also partially account for the poor compositional capabilities of VLMs.**

Table 3. Evaluating whether image encoders and text encoders of VLMs possess mutually matching compositional knowledge with the EQBEN dataset, where each sample contains two images and two texts both with minimal differences in the relation aspect. The maximum metrics are shown in bold.

Dataset	Models	$Q_{\mathcal{T}:R\&O\rightarrow\mathcal{I}:O_1}$	$Q_{\mathcal{T}:R\&O\rightarrow\mathcal{I}:O_2}$	$Q_{\mathcal{T}:R\&O\rightarrow\mathcal{I}:O_1\&O_2}$
EQBEN	CLIP	2.8e-1	**6.9e-1**	2.4e-1
	NEGCLIP	5.1e-1	**1.3e0**	3.8e-1
	BLIP	5.1e-1	**6.7e-1**	4.3e-1
	XVLM	1.2e-1	**1.9e-1**	8.5e-2
	FLAVA	5.4e-1	**9.2e-1**	6.5e-1
Dataset	Models	$D_{\mathcal{I}:O_1\&O_2\rightarrow\mathcal{T}:R}$	$D_{\mathcal{I}:O_1\&O_2\rightarrow\mathcal{T}:O}$	$D_{\mathcal{I}:O_1\&O_2\rightarrow\mathcal{T}:R\&O}$
EQBEN	CLIP	8.0e-1	**2.4e0**	1.3e0
	NEGCLIP	1.0e0	**3.0e0**	1.3e0
	BLIP	6.7e-1	1.5e0	**3.6e0**
	XVLM	1.0e0	1.6e0	**3.2e0**
	FLAVA	1.0e0	2.1e0	**2.6e0**

7. Conclusion

In this paper, we have conducted systematical analyses on the compositionality reasoning capabilities of Vision Language Models (VLMs), which are widely considered as important characteristics of human intelligence. To this end, we have progressively diagnosed the compositional knowledge of each unimodal encoder *separately* and then the multimodal compositional knowledge *jointly*. A number of new metrics from a novel game-theoretic view have been proposed to conduct fine-grained compositionality knowl-

edge diagnoses. In this way, we have obtained and validated several insights regarding the causes for the poor compositional performance of VLMs, which may help provide useful guidance on future explorations.

Acknowledgements

This paper is partially supported by the National Key R&D Program of China No.2022ZD0161000.

Impact Statement

This paper presents work whose goal is to diagnose the compositional knowledge of vision language models in a holistic manner. One positive impact of our work is that our analyses may help instruct future studies to effectively improve the compositional performance of vision language models, so as to advance the development of many downstream unimodal and multimodal applications. However, it is also crucial to recognize the potential negative impacts, which may result from the malicious misuse of downstream algorithms.

References

Ancona, M., Oztireli, C., and Gross, M. Explaining deep neural networks with a polynomial time algorithm for shapley value approximation. In *International Conference on Machine Learning*, pp. 272–281. PMLR, 2019.

Cascante-Bonilla, P., Shehada, K., Smith, J. S., Doveh, S., Kim, D., Panda, R., Varol, G., Oliva, A., Ordonez, V., Feris, R., et al. Going beyond nouns with vision & language models using synthetic data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20155–20165, 2023.

Chen, L., Lou, S., Zhang, K., Huang, J., and Zhang, Q. Harsanyinet: Computing accurate shapley values in a single forward propagation. *arXiv preprint arXiv:2304.01811*, 2023.

Chen, Y.-C., Li, L., Yu, L., El Kholy, A., Ahmed, F., Gan, Z., Cheng, Y., and Liu, J. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pp. 104–120. Springer, 2020.

Cresswell, M. J. Logics and languages. 1973.

Deng, H., Ren, Q., Zhang, H., and Zhang, Q. Discovering and explaining the representation bottleneck of dnns. *arXiv preprint arXiv:2111.06236*, 2021.

Dong, S., Wang, J., Liang, J., Fan, H., and Ji, R. Explaining deepfake detection by analysing image matching. In *European Conference on Computer Vision*, pp. 18–35. Springer, 2022.

Doveh, S., Arbelle, A., Harary, S., Schwartz, E., Herzig, R., Giryes, R., Feris, R., Panda, R., Ullman, S., and Karlinsky, L. Teaching structured vision & language concepts to vision & language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2657–2668, 2023.

Du, Y., Wei, F., Zhang, Z., Shi, M., Gao, Y., and Li, G. Learning to prompt for open-vocabulary object detection with vision-language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14084–14093, 2022.

Fan, L., Krishnan, D., Isola, P., Katabi, D., and Tian, Y. Improving clip training with language rewrites. *Advances in Neural Information Processing Systems*, 36, 2023.

Gao, Y., Liu, J., Xu, Z., Zhang, J., Li, K., Ji, R., and Shen, C. Pyramidclip: Hierarchical feature alignment for vision-language model pretraining. *Advances in neural information processing systems*, 35:35959–35970, 2022.

Goel, S., Bansal, H., Bhatia, S., Rossi, R., Vinay, V., and Grover, A. Cyclip: Cyclic contrastive language-image pretraining. *Advances in Neural Information Processing Systems*, 35:6704–6719, 2022.

Grabisch, M. and Roubens, M. An axiomatic approach to the concept of interaction among players in cooperative games. *International Journal of game theory*, 28(4):547–565, 1999.

Grabisch, M. et al. *Set functions, games and capacities in decision making*, volume 46. Springer, 2016.

Greff, K., Belletti, F., Beyer, L., Doersch, C., Du, Y., Duckworth, D., Fleet, D. J., Gnanapragasam, D., Golemo, F., Herrmann, C., et al. Kubric: A scalable dataset generator. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3749–3761, 2022.

Gu, X., Lin, T.-Y., Kuo, W., and Cui, Y. Open-vocabulary object detection via vision and language knowledge distillation. In *International Conference on Learning Representations*, 2021.

Harsanyi, J. C. A simplified bargaining model for the n-person cooperative game. *International Economic Review*, 4(2):194–220, 1963.

Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., and Cohen-or, D. Prompt-to-prompt image editing with cross-attention control. In *The Eleventh International Conference on Learning Representations*, 2022.

Herzig, R., Mendelson, A., Karlinsky, L., Arbelle, A., Feris, R., Darrell, T., and Globerson, A. Incorporating structured representations into pretrained vision & language models using scene graphs. *arXiv preprint arXiv:2305.06343*, 2023.

Honnibal, M. and Montani, I. spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. *To appear*, 7 (1):411–420, 2017.

Hsieh, C.-Y., Zhang, J., Ma, Z., Kembhavi, A., and Krishna, R. Sugarcrepe: Fixing hackable benchmarks for vision-language compositionality. In *Thirty-Seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023.

Huang, K., Sun, K., Xie, E., Li, Z., and Liu, X. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023a.

Huang, Y., Tang, J., Chen, Z., Zhang, R., Zhang, X., Chen, W., Zhao, Z., Lv, T., Hu, Z., and Zhang, W. Structure-clip: Enhance multi-modal language representations with structure knowledge. *arXiv preprint arXiv:2305.06152*, 2023b.

Hudson, D. A. and Manning, C. D. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6700–6709, 2019.

Ji, J., Krishna, R., Fei-Fei, L., and Niebles, J. C. Action genome: Actions as compositions of spatio-temporal scene graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10236–10247, 2020.

Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., Le, Q., Sung, Y.-H., Li, Z., and Duerig, T. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pp. 4904–4916. PMLR, 2021.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., Dollár, P., and Girshick, R. Segment anything. *arXiv:2304.02643*, 2023.

Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.-J., Shamma, D. A., et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73, 2017.

Li, B., Weinberger, K. Q., Belongie, S., Koltun, V., and Ranftl, R. Language-driven semantic segmentation. In *International Conference on Learning Representations*, 2022a. URL https://openreview.net/forum?id=RriDjddCLN.

Li, J., Selvaraju, R., Gotmare, A., Joty, S., Xiong, C., and Hoi, S. C. H. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34: 9694–9705, 2021a.

Li, J., Li, D., Xiong, C., and Hoi, S. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pp. 12888–12900. PMLR, 2022b.

Li, L. H., Yatskar, M., Yin, D., Hsieh, C.-J., and Chang, K.-W. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019a.

Li, M. and Zhang, Q. Does a neural network really encode symbolic concept? In *ICML*, 2023.

Li, Y., Liang, F., Zhao, L., Cui, Y., Ouyang, W., Shao, J., Yu, F., and Yan, J. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. In *International Conference on Learning Representations*, 2021b.

Li, Y., Fan, H., Hu, R., Feichtenhofer, C., and He, K. Scaling language-image pre-training via masking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23390–23400, 2023.

Li, Y.-L., Xu, L., Liu, X., Huang, X., Xu, Y., Chen, M., Ma, Z., Wang, S., Fang, H.-S., and Lu, C. Hake: Human activity knowledge engine. *arXiv preprint arXiv:1904.06539*, 2019b.

Liu, H., Li, C., Wu, Q., and Lee, Y. J. Visual instruction tuning, 2023.

Ma, Z., Hong, J., Gul, M. O., Gandhi, M., Gao, I., and Krishna, R. Crepe: Can vision-language foundation models reason compositionally? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10910–10921, 2023.

Momeni, L., Caron, M., Nagrani, A., Zisserman, A., and Schmid, C. Verbs in action: Improving verb understanding in video-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15579–15591, 2023.

OpenAI. Chatgpt. 2022.

Peng, W., Xie, S., You, Z., Lan, S., and Wu, Z. Synthesize, diagnose, and optimize: Towards fine-grained vision-language understanding. In *CVPR*, 2024.

Pham, K., Kafle, K., Lin, Z., Ding, Z., Cohen, S., Tran, Q., and Shrivastava, A. Learning to predict visual attributes in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 13018–13028, 2021.

Pratt, S., Yatskar, M., Weihs, L., Farhadi, A., and Kembhavi, A. Grounded situation recognition. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pp. 314–332. Springer, 2020.

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

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

Ray, A., Radenovic, F., Dubey, A., Plummer, B. A., Krishna, R., and Saenko, K. Cola: How to adapt vision-language models to compose objects localized with attributes? *arXiv preprint arXiv:2305.03689*, 2023.

Ren, J., Zhang, D., Wang, Y., Chen, L., Zhou, Z., Chen, Y., Cheng, X., Wang, X., Zhou, M., Shi, J., et al. A unified game-theoretic interpretation of adversarial robustness. *arXiv preprint arXiv:2111.03536*, 4, 2021.

Ren, J., Zhou, Z., Chen, Q., and Zhang, Q. Can we faithfully represent absence states to compute shapley values on a dnn? In *The Eleventh International Conference on Learning Representations*, 2022.

Ren, J., Li, M., Chen, Q., Deng, H., and Zhang, Q. Defining and quantifying the emergence of sparse concepts in dnns. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20280–20289, 2023.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Sahin, U., Li, H., Khan, Q., Cremers, D., and Tresp, V. Enhancing multimodal compositional reasoning of visual language models with generative negative mining. *arXiv preprint arXiv:2311.03964*, 2023.

Shapley, L. S. A value for n-person games, contributions to the theory of games, 2, 307–317, 1953.

Singh, A., Hu, R., Goswami, V., Couairon, G., Galuba, W., Rohrbach, M., and Kiela, D. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15638–15650, 2022.

Sundararajan, M., Dhamdhere, K., and Agarwal, A. The shapley taylor interaction index. In *International conference on machine learning*, pp. 9259–9268. PMLR, 2020.

Tan, H. and Bansal, M. Lxmert: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 5100–5111, 2019.

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

Wang, T., Lin, K., Li, L., Lin, C.-C., Yang, Z., Zhang, H., Liu, Z., and Wang, L. Equivariant similarity for vision-language foundation models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11998–12008, October 2023.

Wang, X., Ren, J., Lin, S., Zhu, X., Wang, Y., and Zhang, Q. A unified approach to interpreting and boosting adversarial transferability. In *International Conference on Learning Representations*, 2020.

Wang, X., Lin, S., Zhang, H., Zhu, Y., and Zhang, Q. Interpreting attributions and interactions of adversarial attacks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1095–1104, 2021.

Wang, Y., Gao, D., Yu, L., Lei, W., Feiszli, M., and Shou, M. Z. Geb+: A benchmark for generic event boundary captioning, grounding and retrieval. In *European Conference on Computer Vision*, pp. 709–725. Springer, 2022.

Xu, J., De Mello, S., Liu, S., Byeon, W., Breuel, T., Kautz, J., and Wang, X. Groupvit: Semantic segmentation emerges from text supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18134–18144, 2022.

Yao, K., Wang, J., Diao, B., and Li, C. Towards understanding the generalization of deepfake detectors from a game-theoretical view. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2031–2041, 2023.

Yao, L., Huang, R., Hou, L., Lu, G., Niu, M., Xu, H., Liang, X., Li, Z., Jiang, X., and Xu, C. Filip: Fine-grained interactive language-image pre-training. *arXiv preprint arXiv:2111.07783*, 2021.

Yuksekgonul, M., Bianchi, F., Kalluri, P., Jurafsky, D., and Zou, J. When and why vision-language models behave like bags-of-words, and what to do about it? In *The Eleventh International Conference on Learning Representations*, 2022.

Zang, Y., Li, W., Zhou, K., Huang, C., and Loy, C. C. Open-vocabulary detr with conditional matching. In *European Conference on Computer Vision*, pp. 106–122. Springer, 2022.

Zeng, Y., Zhang, X., and Li, H. Multi-grained vision language pre-training: Aligning texts with visual concepts. In *International Conference on Machine Learning*, pp. 25994–26009. PMLR, 2022.

Zhang, D., Zhang, H., Zhou, H., Bao, X., Huo, D., Chen, R., Cheng, X., Wu, M., and Zhang, Q. Building interpretable interaction trees for deep nlp models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 14328–14337, 2021a.

Zhang, H., Cheng, X., Chen, Y., and Zhang, Q. Game-theoretic interactions of different orders. *arXiv preprint arXiv:2010.14978*, 2020a.

Zhang, H., Li, S., Ma, Y., Li, M., Xie, Y., and Zhang, Q. Interpreting and boosting dropout from a game-theoretic view. In *International Conference on Learning Representations*, 2020b.

Zhang, H., Xie, Y., Zheng, L., Zhang, D., and Zhang, Q. Interpreting multivariate shapley interactions in dnns. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 10877–10886, 2021b.

Zhao, T., Zhang, T., Zhu, M., Shen, H., Lee, K., Lu, X., and Yin, J. Vl-checklist: Evaluating pre-trained vision-language models with objects, attributes and relations, 2022. URL https://arxiv.org/abs/2207.00221.

Zhou, C., Loy, C. C., and Dai, B. Extract free dense labels from clip. In *European Conference on Computer Vision*, pp. 696–712. Springer, 2022.

Zhou, L., Xu, C., and Corso, J. Towards automatic learning of procedures from web instructional videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Zhu, D., Chen, J., Shen, X., Li, X., and Elhoseiny, M. Minigpt-4: Enhancing vision-language understanding with advanced large language models, 2023.

A. Axioms of the Harsanyi dividend

The Harsanyi dividend (Harsanyi, 1963) satisfies many axioms, which provide solid theoretical foundations for our explanations in the main paper. The axioms are as follows.

- *Linearity axiom.* Given a game t combined by a game u and a game v, *i.e.*, $t(\cdot) = u(\cdot) + v(\cdot)$, the Harsanyi dividend of any subset of players $\mathcal{S}$ in the game t is equal to the sum of the Harsanyi dividends in the game u and v, *i.e.*, $w_t(\mathcal{S}|\mathcal{N}) = w_u(\mathcal{S}|\mathcal{N}) + w_v(\mathcal{S}|\mathcal{N})$.
- *Dummy axiom.* If $\forall \mathcal{S} \subseteq \mathcal{N} \backslash \{i\}, v(\mathcal{S} \cup \{i\}) = v(\mathcal{S}) + v(\{i\})$, then player i is a dummy player, having no interactions with other players,*i.e.*, $w(\mathcal{S} \cup \{i\}|\mathcal{N}) = 0$.
- *Symmetry axiom.* Given two players i and j, if $\forall \mathcal{S} \subseteq \mathcal{N} \backslash \{i, j\}, v(\mathcal{S} \cup \{i\}) = v(\mathcal{N} \cup \{j\})$, then they have the same interaction effects with other players, *i.e.*, $\forall \mathcal{S} \subseteq \mathcal{N} \backslash \{i, j\}, w(\mathcal{S} \cup \{i\}|\mathcal{N}) = w(\mathcal{S} \cup \{j\}|\mathcal{N})$.
- *Efficiency axiom.* The overall output of the game $v(\mathcal{N})$ can be disentangled into the interaction effects of different subsets of players $\mathcal{S}$, *i.e.*, $v(\mathcal{N}) = \sum_{\mathcal{S} \subseteq \mathcal{N}} w(\mathcal{S}|\mathcal{N})$.

Besides, the Harsanyi dividend is also related to the Shapley value (Shapley, 1953), as described in the following theorem.

Theorem 1 (proven in (Harsanyi, 1963; Ren et al., 2022)). *Let $\phi(i)$ represents the Shapley value of the player $i \in \mathcal{N}$. In this way, we have $\phi(i) = \sum_{\mathcal{S} \subseteq \mathcal{N} \backslash \{i\}} \frac{1}{|S|+1} w(S \cup \{i\}|\mathcal{N})$, showing that the Shapley value can be considered as the uniform allocations from the numerical values of Harsanyi dividends.*

B. Visualizing the new annotations of EQBEN

In this section, we provide samples from our new object annotations in the subset of the EQBEN dataset. The visualization results are presented in Figure 6.

C. More results on diagnosing the compositional knowledge of text encoders

Experimental results on Visual Genome Attribution dataset. In this dataset, the correct captions and wrong captions are in the templates of *"the [attribute 1] [object 1] and the [attribute 2] [object 2]"* and *"the [attribute 2] [object 1] and the [attribute 1] [object 2]"* with the *SWAP* manipulation. Therefore, we mainly measured the sensitivity metrics of attribute words, object words and the interactions between attribute words and object words, *i.e.*, Q_A, Q_O and $Q_{A\&O}$. Masks denoting the attribute words and object words were obtained based on the templates. Intuitively, this type of caption pair is majorly different in the aspect of the interaction between attribute words and object words, rather than attribute/object words alone. Results in Table 4 are consistent with this human intuition, showing $Q_{A\&O}$ has the largest value among other metrics. **Such results indicate that text encoders also recognize the prominent compositional differences between captions in the attribute-object aspect.**

Besides, we also calculated the Pearson correlation coefficients $\rho(\mathcal{X}^{\mathcal{T}}, \mathcal{Y}^{\mathcal{T}})$ between reward differences and interaction effect differences in the Visual Genome Attribution dataset, *i.e.*, $\rho(\mathcal{X}^{\mathcal{T}}, \mathcal{Y}_O^{\mathcal{T}})$, $\rho(\mathcal{X}^{\mathcal{T}}, \mathcal{Y}_A^{\mathcal{T}})$, $\rho(\mathcal{X}^{\mathcal{T}}, \mathcal{Y}_{A\&O}^{\mathcal{T}})$. $\mathcal{Y}_A^{\mathcal{T}}$ represents the casual effect changes when $\mathcal{S}^{\mathcal{T}_1}$ and $\mathcal{S}^{\mathcal{T}_2}$ only contain attribute words. $\mathcal{Y}_{A\&O}^{\mathcal{T}}$ represents the casual effect changes when $\mathcal{S}^{\mathcal{T}_1}$ and $\mathcal{S}^{\mathcal{T}_2}$ contain both object and attribute words. Results in Figure 7 show that the interaction effect changes between attribute words and object words played a major part in the final reward differences, showing that text encoders correctly reflected the prominent compositional differences within each caption pair.

Experimental results on VL-CheckList benchmark. The VL-CheckList benchmark (Zhao et al., 2022) contains a large scale of images and captions (over 100K) combined from 4 datasets: Visual Genome (Krishna et al., 2017), SWiG (Pratt et al., 2020), VAW (Pham et al., 2021), and HAKE (Li et al., 2019b) datasets. In each sample, there exist one image, one correct caption and one wrong caption. The wrong caption is generated by *REPLACING* one compositionality aspect of the correct caption, including objects, relations and attributes. Furthermore, these samples were then divided into 9 categories: 1) Object: the location and size of it, 2) Relation: action or spatial relation between objects, 3) Attribute: color, material, size, state and action. In experiments, we exploited Spacy (Honnibal & Montani, 2017) for part-of-speech tagging to roughly divide each sentence into object words, relation words and attribute words. Other irrelevant words were then treated as the constant text background.

Table 4. Evaluating the compositional sensitivities of text encoders of VLMs. In the Visual Genome Attribution dataset, the attribute words are swapped to obtain perturbed texts. Results show that $Q_{A\&O}$ (bold) have larger values than Q_A and Q_O in this dataset across different VLMs, showing that text encoders of various VLMs exhibit accurate sensibilities to the changes of textual patterns.

Dataset	Models	Q_O	Q_A	$Q_{A\&O}$
Visual Genome Attribution	CLIP	1.2e-4	2.9e-3	**1.6e-2**
	NEGCLIP	2.6e-4	4.5e-3	**2.9e-2**
	BLIP	3.0e-3	2.5e-2	**1.9e0**
	XVLM	3.3e-3	2.0e-2	**4.5e-1**
	FLAVA	2.0e-2	9.6e-2	**2.9e-1**

We report the results in Table 5 and Table 6. Intuitively, by replacing the correct caption with new object/relation/attribute words, not only does the effect of object/relation/attribute words change, but also the interactions among them should vary significantly. As shown in Table 5, on the object aspect, $Q_{R\&O}$ and Q_O (as well as $\rho_{R\&O}$ and ρ_O) had the highest values across different text encoders of VLMs, demonstrating their capabilities to recognize the replacing effect of object words. Besides, Q_O usually had a slightly higher value than $Q_{R\&O}$, showing that text encoders considered that the replacement of object words affected less on the relation changes between objects.

On the relation aspect in Table 5, results showed that Q_R and $Q_{R\&O}$ (as well as ρ_R and $\rho_{R\&O}$) had the largest values, showing that text encoders of VLMs recognized the changes of relation words and the interaction changes between relation words and object words. Besides, for action relations, results show that text encoders of CLIP and NEGCLIP considered that the replacement of relation words caused less impact on the interaction changes between object words and relation words (*i.e.*, $Q_{R\&O}/\rho_{R\&O}$ having a smaller value than Q_R/ρ_R). Meanwhile, text encoders of BLIP, XVLM and FLAVA had opposite understandings on the replacement of relation words (*i.e.*, $Q_{R\&O}/\rho_{R\&O}$ having a larger value than Q_R/ρ_R). As for spatial relations, all text encoders of these VLMs considered the replacement of relation words affected more on interaction changes between object words and relation words.

In Table 6, we present the results on the attribute aspect in the VL-CheckList dataset. Similar to the results in Table 5, Q_A and $Q_{A\&O}$ (as well as ρ_A and $\rho_{A\&O}$) had the largest values in general, indicating that text encoders of VLMs successfully recognized the changes of attribute words and the interaction changes between attribute words and object words. Besides, it is noteworthy that for replacing attribute words in captions, text encoders of CLIP and NEGCLIP usually paid more attention to the changes of attribute words alone (*i.e.*, $Q_{A\&O}/\rho_{A\&O}$ having a smaller value than Q_A/ρ_A), which may expose certain deficiency of CLIP on understanding the close binding relationship between attribute words and object words. Such results may provide certain explanations for the phenomenon that Stable Diffusion models (Rombach et al., 2022) struggled to generate correct images given the type of attribute binding text prompts (Huang et al., 2023a), since they learned their text encoders from CLIP.

In summary, results in Table 5 and Table 6 indicate that text encoders of VLMs recognized the effects of replacing words in texts similar to human understanding. Besides, experimental results on such a large-scale dataset further strengthen the trustworthiness of our explanations on diagnosing the compositional reasoning capabilities of text encoders.

Experimental results on SUGARCREPE benchmark. The SUGARCREPE benchmark (Hsieh et al., 2023) contains a significant amount of image-text samples, each of which contains one image and two captions. Compared to previous benchmarks, each wrong caption in SUGARCREPE is manipulated with one of the following manipulations: *SWAP*, *REPLACE* and *ADD*, and debiased to obtain a plausible and fluent text (OpenAI., 2022). Such a comprehensive benchmark presents a more significant compositional challenge to VLMs. In experiments, we harnessed Spacy (Honnibal & Montani, 2017) to roughly obtain the masks denoting object words, relation words and attribute words. Results are summarized from Table 7 to 8.

Generally speaking, given such complex caption pairs, text encoders of VLMs still demonstrated similar behavior on the *REPLACE* manipulation and the *SWAP* manipulation as in previously evaluated benchmarks. As for the new *ADD* manipulation, it is expected that adding new object/attribute words should not only change the effect of object/attribute words alone, but also vary the interactions they contributed to. Results in Table 7 and 8 are consistent with the above intuitions, further demonstrating that text encoders of VLMs were capable of recognizing the compositional differences between captions in a fine-grained manner.

Table 5. Evaluating the compositional knowledge encoded in text encoders of VLMs with the VL-CheckList dataset (object/relation aspect). Here ρ_O, ρ_R and $\rho_{R\&O}$ represents $\rho(\mathcal{X}^{\mathcal{T}}, \mathcal{Y}_O^{\mathcal{T}})$, $\rho(\mathcal{X}^{\mathcal{T}}, \mathcal{Y}_R^{\mathcal{T}})$ and $\rho(\mathcal{X}^{\mathcal{T}}, \mathcal{Y}_{R\&O}^{\mathcal{T}})$ for short. The top two largest metrics are in bold. On the object aspect, results show that Q_O and $Q_{R\&O}$ (as well as ρ_O and $\rho_{R\&O}$) had the largest values, showing that text encoders of VLMs recognized both the changes of object words and the interaction changes between object words and relation words, when replacing the object words in captions. On the relation aspect, results show that Q_R and $Q_{R\&O}$ (as well as ρ_R and $\rho_{R\&O}$) had the largest values, showing that when replacing the relation words in captions, text encoders of VLMs successfully recognized both the changes of relation words and the interaction changes between relation words and object words.

Dataset	Category	Models	Q_O	Q_R	$Q_{R\&O}$	ρ_O	ρ_R	$\rho_{R\&O}$
VL-CheckList (Object)	Location	CLIP	**3.7e-1**	1.1e-2	**3.1e-2**	**0.86**	-0.01	**0.22**
		NEGCLIP	**3.6e-1**	6.5e-3	**2.1e-2**	**0.87**	0.00	**0.18**
		BLIP	**3.8e0**	4.3e-2	**9.3e-1**	**0.65**	-0.05	**0.44**
		XVLM	**1.4e0**	9.7e-2	**2.9e-1**	**0.55**	0.03	**0.28**
		FLAVA	**2.5e0**	3.3e-1	**1.0e0**	**0.43**	0.05	**0.29**
	Size	CLIP	**3.2e-1**	1.5e-2	**3.6e-2**	**0.81**	0.00	**0.22**
		NEGCLIP	**3.2e-1**	6.0e-3	**2.0e-2**	**0.88**	0.01	**0.18**
		BLIP	**3.7e0**	4.4e-2	**9.0e-1**	**0.66**	-0.04	**0.44**
		XVLM	**1.5e0**	9.6e-2	**2.9e-1**	**0.55**	0.03	**0.27**
		FLAVA	**2.3e0**	3.4e-1	**1.0e0**	**0.44**	0.05	**0.28**
VL-CheckList (Relation)	Action	CLIP	4.5e-2	**1.0e-1**	**5.0e-2**	0.02	**0.46**	**0.21**
		NEGCLIP	1.8e-2	**9.2e-2**	**3.5e-2**	-0.01	**0.56**	**0.32**
		BLIP	3.0e-1	**7.3e-1**	**1.5e0**	-0.02	**0.36**	**0.61**
		XVLM	2.6e-1	**3.0e-1**	**4.1e-1**	-0.05	**0.17**	**0.47**
		FLAVA	3.9e-1	**8.6e-1**	**1.5e0**	0.16	**0.19**	**0.51**
	Spatial	CLIP	1.5e-2	**4.3e-2**	**5.4e-2**	0.02	**0.11**	**0.53**
		NEGCLIP	9.2e-3	**1.3e-2**	**2.1e-2**	0.03	**0.23**	**0.64**
		BLIP	1.7e-1	**2.3e-1**	**9.7e-1**	0.01	**0.12**	**0.79**
		XVLM	9.3e-2	**2.2e-1**	**2.9e-1**	-0.01	**0.15**	**0.43**
		FLAVA	2.6e-1	**6.7e-1**	**7.2e-1**	-0.05	**0.29**	**0.51**

D. More results on evaluating the mutual compositional knowledge between text encoders and image encoders

In this section, we provide more results to further evaluate the mutual compositional knowledge between text encoders and image encoders. To this end, we exploited the Visual Genome Relation dataset to conduct further analyses on a larger scale. Since each sample in the dataset contains one image and two captions individually, we mainly used $Q_{\mathcal{T}:R\&O\to\mathcal{I}:(\cdot)}$ to evaluate how image encoders considered the compositional knowledge of text encoders. For the Visual Genome Relation dataset, we obtained the segmentation mask for each object with the help of SAM (Kirillov et al., 2023), similar to our annotated EQBEN dataset. The newly annotated dataset contained 2000 samples in total. The newly added segmentation masks are visualized in Figure 8. In this way, as shown in Table 9, $Q_{\mathcal{T}:R\&O\to\mathcal{I}:O_1\&O_2}$ failed to have the largest value than $Q_{\mathcal{T}:R\&O\to\mathcal{I}:O_1}$ and $Q_{\mathcal{T}:R\&O\to\mathcal{I}:O_2}$, showing that in terms of object relations, image encoders did not have the corresponding compositional knowledge of text encoders. Instead, image encoders tended to consider the interaction between object words and relation words learned by text encoders as the object representations on images (*i.e.*, $Q_{\mathcal{T}:R\&O\to\mathcal{I}:O_1}$ being the largest among metrics). The above results were consistent with the results in Table 3 in terms of $Q_{\mathcal{T}:R\&O\to\mathcal{I}:(\cdot)}$.

E. More results on recent VLMs

In this section, we further conducted experiments on two more recent VLMs: FLIP (Li et al., 2023) and LaCLIP (Fan et al., 2023). Specifically, we followed our proposed methods from Section 4 to 6 with the same evaluation benchmarks, comprehensively evaluating the compositional knowledge of text encoders, the compositional knowledge of image encoders and also, whether text encoders and image encoders have mutually matching compositional knowledge. We summarized the results in Table 10, Table 11 and Table 12 respectively. Experimental results are consistent with our findings in the main paper. Specifically, (1) text encoders of FLIP and LaCLIP showed excellent compositional reasoning capabilities, able to recognize the dominant compositional differences between input texts like human understanding; (2) image encoders of FLIP and LaCLIP demonstrated weaker compositional reasoning capabilities; (3) image encoders and text encoders of FLIP and LaCLIP did not exhibit mutually matching compositional knowledge. The above experiments further strengthen the reliability of the findings in the main paper.

Table 6. Evaluating the compositional knowledge encoded in text encoders of VLMs with the VL-CheckList dataset (attribute aspect). Here ρ_O, ρ_A and $\rho_{A\&O}$ represents $\rho(\mathcal{X}^{\mathcal{T}}, \mathcal{Y}_O^{\mathcal{T}})$, $\rho(\mathcal{X}^{\mathcal{T}}, \mathcal{Y}_A^{\mathcal{T}})$ and $\rho(\mathcal{X}^{\mathcal{T}}, \mathcal{Y}_{A\&O}^{\mathcal{T}})$ for short. The top two largest metrics are in bold. Results show that Q_A and $Q_{A\&O}$ (as well as ρ_A and $\rho_{A\&O}$) had the largest values in general, indicating that when replacing the attribute words in captions, text encoders successfully recognized both the changes of attribute words and the interaction changes between object words and attribute words.

Dataset	Category	Models	Q_O	Q_A	$Q_{A\&O}$	ρ_O	ρ_A	$\rho_{A\&O}$
VL-CheckList (Attribute)	Color	CLIP	4.4e-3	**7.6e-2**	**4.0e-2**	0.02	**0.65**	**0.30**
		NEGCLIP	3.1e-3	**8.5e-2**	**2.0e-2**	0.00	**0.81**	**0.05**
		BLIP	4.9e-2	**7.7e-1**	**2.3e0**	0.04	**0.45**	**0.84**
		XVLM	7.8e-2	**3.4e-1**	**6.2e-1**	0.02	**0.18**	**0.59**
		FLAVA	1.4e-1	**8.1e-1**	**1.1e0**	0.14	**0.16**	**0.46**
	Material	CLIP	1.3e-2	**8.6e-2**	**4.8e-2**	-0.03	**0.58**	**0.35**
		NEGCLIP	3.6e-3	**9.2e-2**	**3.0e-2**	0.00	**0.70**	**0.13**
		BLIP	1.0e-1	**1.3e0**	**1.9e0**	0.11	**0.42**	**0.71**
		XVLM	1.5e-1	**5.6e-1**	**4.5e-1**	0.05	**0.17**	**0.39**
		FLAVA	3.0e-1	**1.0e0**	**1.2e0**	0.21	**0.35**	**0.41**
	Size	CLIP	1.1e-2	**3.6e-2**	**3.6e-2**	0.06	**0.32**	**0.49**
		NEGCLIP	6.1e-3	**4.0e-2**	**2.5e-2**	0.03	**0.53**	**0.23**
		BLIP	4.8e-2	**6.5e-1**	**1.2e0**	0.03	**0.33**	**0.77**
		XVLM	4.5e-2	**2.9e-1**	**3.0e-1**	0.04	**0.23**	**0.51**
		FLAVA	9.0e-2	**4.3e-1**	**8.7e-1**	-0.04	**0.19**	**0.46**
	State	CLIP	9.1e-3	**4.4e-2**	**3.5e-2**	0.04	**0.50**	**0.48**
		NEGCLIP	4.7e-3	**6.8e-2**	**2.3e-2**	0.04	**0.71**	**0.16**
		BLIP	4.0e-2	**9.9e-1**	**1.0e0**	-0.03	**0.55**	**0.73**
		XVLM	8.1e-2	**4.6e-1**	**3.6e-1**	-0.01	**0.22**	**0.44**
		FLAVA	1.7e-1	**8.1e-1**	**1.2e0**	0.06	**0.18**	**0.42**
	Action	CLIP	1.8e-2	**1.4e-1**	**4.8e-2**	0.01	**0.68**	**0.10**
		NEGCLIP	1.4e-2	**1.7e-1**	**5.5e-2**	**0.07**	**0.72**	-0.05
		BLIP	1.1e-1	**1.7e0**	**9.6e-1**	-0.02	**0.64**	**0.57**
		XVLM	1.2e-1	**7.2e-1**	**4.0e-1**	-0.01	**0.27**	**0.33**
		FLAVA	2.1e-1	**1.1e0**	**1.2e0**	-0.01	**0.15**	**0.38**

F. Experimental details

In the main paper, we evaluated the following state-of-the-art Vision Language Models (VLMs): CLIP[6] (Radford et al., 2021), NEGCLIP[7] (Yuksekgonul et al., 2022), BLIP[8] (Li et al., 2022b), XVLM[9] (Zeng et al., 2022), FLAVA[10] (Singh et al., 2022), all of which were obtained from their officially released checkpoints. For reference, we also provide the compositional performance of these VLMs with the benchmarks exploited in the main paper. Specifically, we used $ACC_{\mathcal{T}}$ to measure the accuracy of VLMs picking up the correct caption when given one image with two captions. We also used $ACC_{\mathcal{I}}$ to measure the accuracy of VLMs picking up the correct image when given one caption with two images. Results in Tab. 13-17 show that these state-of-the-art VLMs performed poorly on these benchmarks, exposing their weakness in understanding the compositional information of input variables.

G. Future studies

In this paper, we have obtained and validated several insights to explain the poor compositional reasoning capabilities of VLMs, which we believe could provide beneficial guidance for future studies. However, considering the large gap between theory and practice, significant research efforts are still required to achieve the breakthrough in practice based on our analyses, which we plan to leave to future endeavors. To this end, we introduce two promising solutions as follows: (1) improving image encoders' sensitivities to compositionality changes, instead of text encoders. To this end, one possible solution is to design modules to specifically approximate the Harsanyi dividends of different visual patterns in image encoders, drawing inspiration from (Chen et al., 2023). This approach can allow us to explicitly extract the casual effects of different visual patterns during the training phase, making the improvements of visual compositionality sensitivities more accessible; (2) enhancing the alignment of compositional knowledge between text encoders and image encoders of VLMs. To this end, one possible solution is to leverage our metrics in Eq. 3- Eq. 6 as an auxiliary training loss for VLMs. However, implementing this approach requires a large scale of training data featuring subtle changes of compositionality with detailed

[6]https://github.com/openai/CLIP/

[7]https://github.com/mertyg/vision-language-models-are-bows

[8]https://github.com/salesforce/BLIP

[9]https://github.com/zengyan-97/X-VLM/

[10]https://github.com/apsdehal/flava-tutorials/blob/main/winoground-flava-example.ipynb

Table 7. Evaluating the compositional knowledge encoded in text encoders of VLMs with the SUGARCREPE dataset (object/relation aspect). Here ρ_O, ρ_R and $\rho_{R\&O}$ represents $\rho(\mathcal{X}^{\mathcal{T}}, \mathcal{Y}_O^{\mathcal{T}})$, $\rho(\mathcal{X}^{\mathcal{T}}, \mathcal{Y}_R^{\mathcal{T}})$ and $\rho(\mathcal{X}^{\mathcal{T}}, \mathcal{Y}_{R\&O}^{\mathcal{T}})$ for short. On the object aspect, for the *ADD* manipulation, results show that Q_O and $Q_{R\&O}$ (as well as ρ_O and $\rho_{R\&O}$) had the largest values (shown in bold), showing that text encoders of VLMs recognized both the changes of object words and the interaction changes between object words and relation words, when adding new object words in captions. As for the *REPLACE* manipulation on the object aspect, Q_O and $Q_{R\&O}$ (as well as ρ_O and $\rho_{R\&O}$) still had the largest values (shown in bold), showing that text encoders of VLMs recognized both the changes of object words and the interaction changes between relation words and object words, when replacing object words in captions. As for the *SWAP* manipulation on the object aspect, $Q_{R\&O}$ (as well as $\rho_{R\&O}$) had the largest value (shown in bold), showing that text encoders of VLMs recognized the interaction changes between relation words and object words, when swapping object words in captions. On the relation aspect, for the *REPLACE* manipulation, results show that Q_R and $Q_{R\&O}$ (as well as ρ_R and $\rho_{R\&O}$) had the largest values, showing that when replacing the relation words in captions, text encoders of VLMs successfully recognized both the changes of relation words and the interaction changes between relation words and object words. The above results show that text encoders of VLMs recognized the compositional differences between complex captions in the SUGARCREPE dataset, similar to our intuitions across different types of manipulations.

Dataset	Manipulation	Models	Q_O	Q_R	$Q_{R\&O}$	ρ_O	ρ_R	$\rho_{R\&O}$
SUGARCREPE (Object)	*ADD*	CLIP	**3.3e-2**	1.1e-2	**2.1e-2**	**0.55**	0.07	**0.37**
		NEGCLIP	**4.2e-2**	7.6e-3	**1.3e-2**	**0.66**	0.12	**0.17**
		BLIP	**2.1e-1**	3.0e-2	**3.1e-1**	**0.32**	-0.03	**0.58**
		XVLM	**2.0e-1**	9.9e-2	**1.8e-1**	**0.32**	0.07	**0.23**
		FLAVA	**5.5e-1**	3.0e-1	**7.5e-1**	**0.06**	0.02	**0.10**
	REPLACE	CLIP	**1.6e-1**	9.8e-3	**2.9e-2**	**0.80**	0.09	**0.37**
		NEGCLIP	**1.5e-1**	7.4e-3	**1.7e-2**	**0.84**	0.11	**0.32**
		BLIP	**8.1e-1**	2.3e-2	**5.7e-1**	**0.59**	0.07	**0.52**
		XVLM	**5.8e-1**	4.7e-2	**2.3e-1**	**0.51**	0.03	**0.24**
		FLAVA	**6.8e-1**	1.3e-1	**5.7e-1**	**0.16**	-0.02	**0.30**
	SWAP	CLIP	9.5e-3	9.5e-3	**1.3e-2**	0.29	0.01	**0.58**
		NEGCLIP	7.1e-3	3.5e-3	**1.7e-2**	0.14	0.01	**0.51**
		BLIP	4.0e-2	1.4e-2	**1.3e-1**	0.23	-0.05	**0.65**
		XVLM	6.6e-2	3.3e-2	**1.3e-1**	0.10	0.01	**0.47**
		FLAVA	1.3e-1	1.1e-1	**3.6e-1**	0.13	0.02	**0.17**
SUGARCREPE (Relation)	*REPLACE*	CLIP	3.9e-3	**3.9e-2**	**3.0e-2**	0.03	**0.47**	**0.36**
		NEGCLIP	1.4e-3	**4.3e-2**	**2.4e-2**	-0.05	**0.53**	**0.13**
		BLIP	1.3e-2	**1.3e-1**	**3.1e-1**	0.05	**0.45**	**0.76**
		XVLM	3.5e-2	**1.8e-1**	**2.2e-1**	0.00	**0.26**	**0.44**
		FLAVA	7.6e-2	**4.0e-1**	**4.7e-1**	-0.01	**0.12**	**0.13**

textual and visual annotations, which underscores the need for additional research efforts in this direction.

Table 8. Evaluating the compositional knowledge encoded in text encoders of VLMs with the SUGARCERPE dataset (attribute aspect). Here ρ_O, ρ_A and $\rho_{A\&O}$ represents $\rho(\mathcal{X}^{\mathcal{T}}, \mathcal{Y}_O^{\mathcal{T}})$, $\rho(\mathcal{X}^{\mathcal{T}}, \mathcal{Y}_A^{\mathcal{T}})$ and $\rho(\mathcal{X}^{\mathcal{T}}, \mathcal{Y}_{A\&O}^{\mathcal{T}})$ for short. On the attribute aspect, for the *ADD* manipulation, results show that Q_A and $Q_{A\&O}$ (as well as ρ_A and $\rho_{A\&O}$) had the largest values (shown in bold), showing that text encoders of VLMs recognized both the changes of attribute words and the interaction changes between object words and attribute words, when adding new attribute words in captions. As for the *REPLACE* manipulation on the attribute aspect, Q_A and $Q_{A\&O}$ (as well as ρ_A and $\rho_{A\&O}$) still had the largest values (shown in bold), showing that text encoders of VLMs recognized both the changes of attribute words and the interaction changes between attribute words and object words, when replacing attribute words in captions. As for the *SWAP* manipulation on the attribute aspect, $Q_{A\&O}$ (as well as $\rho_{A\&O}$) had the largest value (shown in bold), showing that text encoders of VLMs recognized the interaction changes between attribute words and object words, when swapping attribute words in captions. The above results show that text encoders of VLMs also correctly recognized the attribute-wise compositional difference within complex caption pairs in the SUGARCERPE dataset.

Dataset)	Manipulation	Models	Q_O	Q_A	$Q_{A\&O}$	ρ_O	ρ_A	$\rho_{A\&O}$
SUGARCREPE (Attribute)	*ADD*	CLIP	2.8e-3	**2.4e-2**	**2.5e-2**	0.05	**0.38**	**0.34**
		NEGCLIP	1.8e-3	**3.5e-2**	**1.8e-2**	0.00	**0.46**	**0.14**
		BLIP	2.7e-2	**1.1e-1**	**2.8e-1**	-0.03	**0.49**	**0.66**
		XVLM	5.1e-2	**1.3e-1**	**2.1e-1**	0.00	**0.22**	**0.35**
		FLAVA	2.2e-1	**3.3e-1**	**5.3e-1**	-0.03	**0.13**	**0.07**
	REPLACE	CLIP	3.3e-3	**6.1e-2**	**2.4e-2**	0.10	**0.67**	**0.36**
		NEGCLIP	2.9e-3	**8.0e-2**	**1.8e-2**	0.09	**0.72**	**0.16**
		BLIP	5.6e-3	**3.8e-1**	**5.0e-1**	-0.03	**0.50**	**0.71**
		XVLM	1.3e-2	**2.8e-1**	**2.6e-1**	-0.06	**0.31**	**0.47**
		FLAVA	4.8e-2	**4.7e-1**	**4.5e-1**	-0.02	**0.29**	**0.20**
	SWAP	CLIP	1.7e-2	1.7e-2	**2.2e-2**	0.13	0.17	**0.47**
		NEGCLIP	9.8e-3	1.3e-2	**2.0e-2**	0.20	0.09	**0.43**
		BLIP	4.5e-2	7.5e-2	**7.7e-1**	-0.01	0.23	**0.82**
		XVLM	8.8e-2	1.3e-1	**4.3e-1**	-0.01	0.16	**0.56**
		FLAVA	9.8e-2	2.0e-1	**4.3e-1**	-0.04	0.06	**0.18**

Table 9. Evaluating the mutual compositional knowledge encoded in image encoders and text encoders of VLMs with the Visual Genome Relation dataset. The maximum metric values are shown in bold.

Dataset	Models	$Q_{\mathcal{T}:R\&O\to\mathcal{I}:O_1}$	$Q_{\mathcal{T}:R\&O\to\mathcal{I}:O_2}$	$Q_{\mathcal{T}:R\&O\to\mathcal{I}:O_1\&O_2}$
Visual Genome Relation	CLIP	**3.7e-1**	3.3e-1	1.6e-1
	NEGCLIP	**7.3e-1**	5.9e-1	2.2e-1
	BLIP	**1.8e-1**	1.7e-1	8.5e-2
	XVLM	**3.4e-1**	2.9e-1	1.5e-1
	FLAVA	**3.2e-1**	3.1e-1	2.7e-1

Table 10. Evaluating the compositional knowledge of text encoders of recent VLMs.

Dataset	Models	Q_O	Q_R	$Q_{R\&O}$
Visual Genome Relation	FLIP	1.6e-2	9.9e-5	**4.2e-2**
	LaCLIP	3.7e-3	2.5e-6	**8.3e-3**

Table 11. Evaluating the compositional knowledge of image encoders of recent VLMs.

Dataset	Models	D_{O_1}	D_{O_2}	$D_{O_1\&O_2}$
EQBEN	FLIP	3.4e-2	**5.2e-2**	3.3e-2
	LaCLIP	1.5e-2	**2.2e-2**	2.1e-2

Table 12. Evaluating whether image encoders and text encoders of recent VLMs possess mutually matching compositional knowledge.

Dataset	Models	$Q_{\mathcal{T}:R\&O\to\mathcal{I}:O_1}$	$Q_{\mathcal{T}:R\&O\to\mathcal{I}:O_2}$	$Q_{\mathcal{T}:R\&O\to\mathcal{I}:O_1\&O_2}$
EQBEN	FLIP	2.2e-1	**5.5e-1**	4.0e-1
	LaCLIP	5.8e-1	**2.0e0**	7.4e-1
Dataset	Models	$D_{\mathcal{I}:O_1\&O_2\to\mathcal{T}:R}$	$D_{\mathcal{I}:O_1\&O_2\to\mathcal{T}:O}$	$D_{\mathcal{I}:O_1\&O_2\to\mathcal{T}:R\&O}$
EQBEN	FLIP	8.1e-1	**2.1e0**	1.0e0
	LaCLIP	7.8e-1	**4.0e0**	8.8e-1

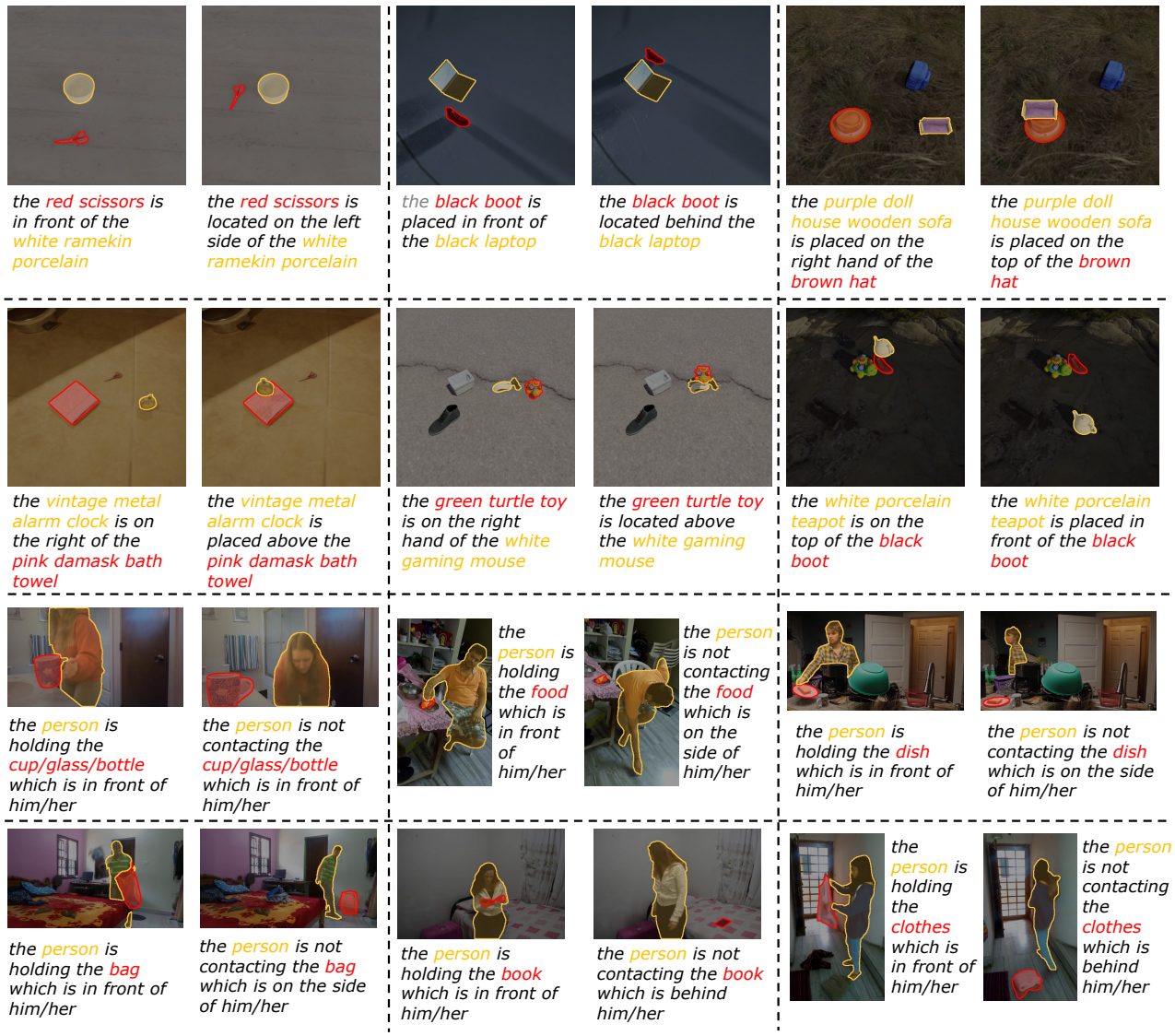

Figure 6. **Visualizing the new annotations on the subset of the EQBEN dataset.** In this paper, we leveraged SAM (Kirillov et al., 2023) to help obtain segmentation results for objects on images. Specifically, we first manually annotated the bounding box for each object described in the corresponding caption. We then harnessed SAM to obtain the segmentation mask for each object. Unsatisfying segmentation results were manually annotated afterward for corrections.

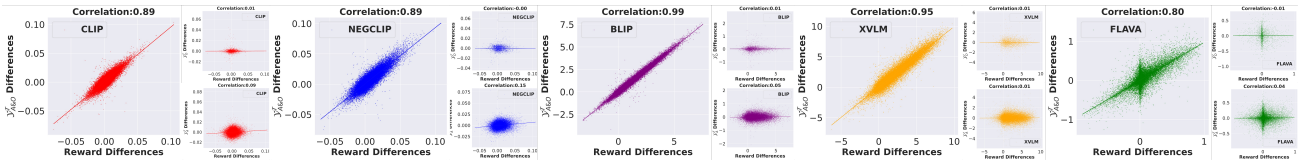

Figure 7. The Pearson correlation coefficients $\rho(\mathcal{X}^{\mathcal{T}}, \mathcal{Y}^{\mathcal{T}})$ between the reward differences $\mathcal{X}^{\mathcal{T}}$ and interaction effect differences on the Visual Genome Attribution dataset, *i.e.*, $\mathcal{Y}_A^{\mathcal{T}}$, $\mathcal{Y}_O^{\mathcal{T}}$ and $\mathcal{Y}_{A\&O}^{\mathcal{T}}$. Each point represents a data sample containing two captions and one image. Results show that the reward differences between the two captions were mainly related to the interaction changes of object words and attribute words.

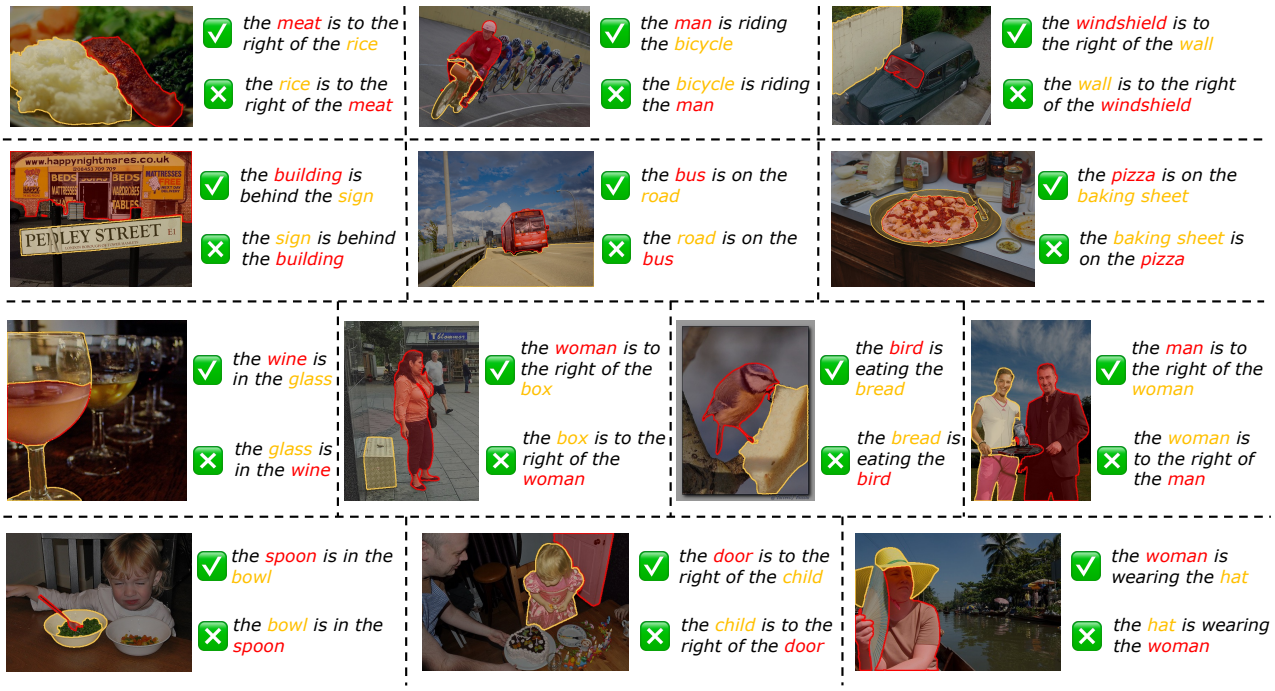

Figure 8. **Visualizing the new annotations on the subset of the Visual Genome Relation dataset.** Specifically, we first exploited the original bounding box annotations from the GQA dataset (Hudson & Manning, 2019) and then used SAM (Kirillov et al., 2023) to obtain the segmentation mask for each object. We then manually annotated unsatisfying segmentation results afterward for corrections.

Table 13. Evaluating the compositional performance of VLMs with the Visual Genome Relation dataset and Visual Genome Attribution dataset.

Dataset	Models	$ACC_{\mathcal{T}}$	Dataset	Models	$ACC_{\mathcal{T}}$
Visual Genome Relation	CLIP	50.65%	Visual Genome Attribution	CLIP	60.98%
	NEGCLIP	58.33%		NEGCLIP	70.87%
	BLIP	55.30%		BLIP	87.12%
	XVLM	58.65%		XVLM	89.51%
	FLAVA	43.79%		FLAVA	68.84%

Table 14. Evaluating the compositional performance of VLMs with the VL-CheckList dataset (attribute aspect).

Dataset	Category	Models	$ACC_{\mathcal{T}}$	Category	Models	$ACC_{\mathcal{T}}$	Category	Models	$ACC_{\mathcal{T}}$	Category	Models	$ACC_{\mathcal{T}}$	Category	Models	$ACC_{\mathcal{T}}$
VL-CheckList (Attribute)	Color	CLIP	71.20%	Material	CLIP	66.32%	Size	CLIP	60.85%	State	CLIP	61.68%	Action	CLIP	73.46%
		NEGCLIP	73.59%		NEGCLIP	75.87%		NEGCLIP	66.06%		NEGCLIP	70.60%		NEGCLIP	75.34%
		BLIP	86.49%		BLIP	89.85%		BLIP	72.13%		BLIP	78.20%		BLIP	80.98%
		XVLM	86.41%		XVLM	87.54%		XVLM	76.67%		XVLM	77.77%		XVLM	82.44%
		FLAVA	69.76%		FLAVA	56.05%		FLAVA	37.77%		FLAVA	50.05%		FLAVA	57.37%

Table 15. Evaluating the compositional performance of VLMs with the SUGARCREPE dataset (object/relation aspect).

Dataset	Manipulation	Models	$ACC_{\mathcal{T}}$	Manipulation	Models	$ACC_{\mathcal{T}}$	Manipulation	Models	$ACC_{\mathcal{T}}$	Dataset	Manipulation	Models	$ACC_{\mathcal{T}}$
SUGARCREPE (Object)	*SWAP*	CLIP	62.70%	*ADD*	CLIP	76.88%	*REPLACE*	CLIP	91.85%	SUGARCREPE (Relation)	*REPLACE*	CLIP	69.18%
		NEGCLIP	77.05%		NEGCLIP	88.12%		NEGCLIP	92.70%			NEGCLIP	74.71%
		BLIP	82.04%		BLIP	95.98%		BLIP	98.16%			BLIP	86.32%
		XVLM	86.12%		XVLM	95.49%		XVLM	97.73%			XVLM	85.89%
		FLAVA	70.20%		FLAVA	90.64%		FLAVA	90.74%			FLAVA	64.07%

Table 16. Evaluating the compositional performance of VLMs with the SUGARCREPE dataset (attribute aspect) and EQBEN dataset (relation aspect). The performance on the EQBEN dataset was evaluated on our newly annotated subset (several samples are shown in Figure 6).

Dataset	Manipulation	Models	$ACC_{\mathcal{T}}$	Manipulation	Models	$ACC_{\mathcal{T}}$	Manipulation	Models	$ACC_{\mathcal{T}}$	Dataset	Manipulation	$ACC_{\mathcal{I}}$	$ACC_{\mathcal{T}}$
SUGARCREPE (Attribute)	*SWAP*	CLIP	64.94%	*ADD*	CLIP	67.60%	*REPLACE*	CLIP	81.94%	EQBEN (Relation)	CLIP	53.85%	55.90%
		NEGCLIP	76.07%		NEGCLIP	81.44%		NEGCLIP	85.21%		NEGCLIP	53.85%	62.15%
		BLIP	94.33%		BLIP	90.18%		BLIP	93.05%		BLIP	59.72%	60.76%
		XVLM	93.10%		XVLM	86.76%		XVLM	91.35%		XVLM	61.46%	66.67%
		FLAVA	81.90%		FLAVA	59.08%		FLAVA	76.15%		FLAVA	54.86%	61.46%

Table 17. Evaluating the compositional performance of VLMs with the VL-CheckList dataset (object/relation aspect).

Dataset	Category	Models	$ACC_{\mathcal{T}}$	Category	Models	$ACC_{\mathcal{T}}$
VL-CheckList (Object)	Location	CLIP	88.80%	Size	CLIP	89.02%
		NEGCLIP	90.72%		NEGCLIP	89.50%
		BLIP	92.52%		BLIP	92.25%
		XVLM	92.85%		XVLM	92.08%
		FLAVA	73.21%		FLAVA	71.34%
VL-CheckList (Relation)	Action	CLIP	77.05%	Spatial	CLIP	55.77%
		NEGCLIP	81.57%		NEGCLIP	60.80%
		BLIP	81.19%		BLIP	61.68%
		XVLM	77.56%		XVLM	74.90%
		FLAVA	33.55%		FLAVA	58.45%