Interpretable Composition Attribution Enhancement for Visio-linguistic Compositional Understanding

Anonymous ACL submission

Abstract

Contrastively trained vision-language models such as CLIP have achieved remarkable progress in vision and language representation learning. Despite the promising progress, their proficiency in compositional reasoning over attributes and relations (*e.g.*, distinguishing between "the car is underneath the person" and "the person is underneath the car") remains notably inadequate. We investigate the cause for this deficient behavior is the *composition attribution issue*, where the attribution scores (*e.g.*, attention scores or GradCAM scores) for relations (*e.g.*, underneath) or attributes (*e.g.*, red) in text are substantially lower than those for object terms. In this work, we show such issue is mitigated via a novel framework called **CAE** (**C**omposition **A**ttribution **E**nhancement). This generic framework incorporates various interpretable attribution methods to encourages the model to pay greater attention on composition words denoting relationships and attributes within the text. Detailed analysis shows that our approach enables the models to adjust and rectify the attribution on the texts. Extensive experiments across seven benchmarks reveal that our framework significantly enhances the ability to discern intricate details and construct more sophisticated interpretations of combined visual and linguistic elements.

1 Introduction

The field of vision-language research has made great advancements in recent years (Radford et al., 2021; Jia et al., 2021b; Rombach et al., 2022; Alayrac et al., 2022). Vision-Language foundation models, such as CLIP, have exhibited remarkable performance across a broad range of well-established evaluation tasks (Deng et al., 2009; Agrawal et al., 2019; Lin et al., 2014; Ramesh et al., 2021), directly or indirectly fostering progress in numerous areas, such as text-to-image generation (Ramesh et al., 2022), video recognition (Ni et al., 2022) and multi-modal large language models (Zhu et al., 2023; Liu et al., 2024).

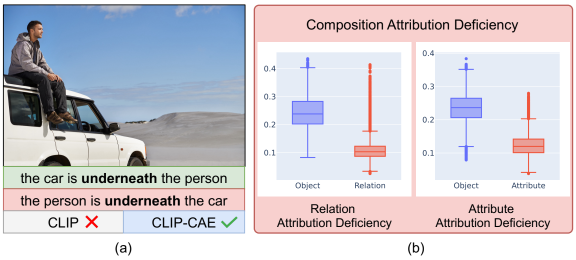


Figure 1: **(Left)** a) An illustrative example from the Winoground benchmark for assessing relation understanding of VLMs. VLMs exhibit difficulty in accurately matching the image with the correct caption (denoted in green). **(Right)** b) The issue of composition attribution deficiency. The distribution of GradCAM-based attribution scores for object tokens is significantly lower than which of the composition tokens (relation and attribute).

Despite these advances, a notable limitation still persists: VLMs such as CLIP exhibit significant challenges in understanding visio-linguistic concepts beyond object nouns, in particular relations and attributes (Thrush et al., 2022; Yuksekgonul et al., 2022; Zhao et al., 2022). Specifically, they struggle with understanding relations between objects, binding correct attributes to the correct objects. For example, as illustrated in Fig. 1 (Left), given an image and two similar textual descriptions (containing the same set of words but composed differently), such as "the car is underneath the person" and "the person is underneath the car", humans can effortlessly discern the contextual differences between the two sentences. However, VLMs tend to struggle, highlighting a significant challenge in compositional reasoning (Thrush et al., 2022; Yuksekgonul et al., 2022).

To further investigate the factors impeding the compositional understanding capabilities of VLMs such as CLIP, we employ various model attribution techniques, such as attention-based and GradCAM-based methods (Chefer et al., 2021), to analyze the attribution scores assigned by the model to object and non-object words when performing image-text matching. As show in Fig. 1 (Right), our investigation reveal a consistent pattern across four different attribution scores: the attribution scores for object words are significantly higher than those for relation and attribute words. For example, the mean attribution score of object tokens is 0.244, which is two times than the relation tokens (0.111). This indicates that the model disproportionately emphasizes object words, neglecting fine-grained details such as relations and attributes in the text. This phenomenon aligns with the recent studies (Yuksekgonul et al., 2022; Kamath et al., 2023) which argued the presence of shortcuts in contrastive learning pretraining. Specifically, the models distinguish the correct image-text pairs from distinctly incorrect ones through simple object recognition, without need to comprehend finer-grained details such as relations and attributes in the texts. In this work, we further identify that the primary issue for compositional understanding is the unfair attribution for relation and attribute words. We refer to this as the issue of *composition attribution deficiency*.

However, the existing methods to improve visio-linguistic compositional understanding are not designed to adjust the attribution for different texts. (Yuksekgonul et al., 2022) introduces captions with perturbed word order and nearest neighboring images into each batch, to force models to distinguish correct and hard negative samples. (Doveh et al., 2024) use LLMs for hard negative mining and (Cascante-Bonilla et al., 2023) explore using synthetic datasets to compose hard negative samples. Regardless of the methods of hard negative mining, existing methods do not endow the models with proportionate attribution across different texts, neglecting the attribution issues.

Inspired by our observation, we propose a novel framework, named **CAE** (**C**omposition **A**ttribution **E**nhancement), to enhance the compositional understanding of VLMs without constructing any hard negative samples explicitly. Specifically, in addition to a task-specific loss, **CAE** adds a new loss that aligns the attribution scores distribution of different types of text tokens during the training process. This encourages the model to pay more attention on fine-grained details (relations or attributes) within the text beyond object nouns. We propose four instances of our framework: attention-based, GradCAM-based, perturbation-based and gradient-based attribution. In each instance, the model's compositional understanding abilities is naturally improved. Furthermore, our approach can be easily integrated with hard negative samples, leading to additional performance gains.

We summarize our contributions as follows:

1. We introduce a simple yet effective novel method to enhance the VLMs' compositional understanding without introducing any hard negative samples explicitly.

2. Extensive experiments across four attribution methods and seven widely-used vision-language compositional benchmarks demonstrate the effectiveness of our method.

3. Our proposed method can be seamlessly integrated with hard-sample mining, thereby further boosting the model's capability of compositional understanding.

2 Related Works

Contrastive Vision-Language Models. Modern VLMs undergo pre-training on large-scale and noisy multimodal datasets (Radford et al., 2021; Jia et al., 2021b; Alayrac et al., 2022; Singh et al., 2022; Li et al., 2022), and then are applied to downstream tasks in a zero-shot manner, achieving remarkable success. Among these models, CLIP (Radford et al., 2021) stands out, which utilizes a contrastive learning method for pretraining. Our focus on CLIP is motivated by two primary factors. Firstly, image-text contrastive learning has become a prevalent and highly successful strategy for VLM pretraining (Jia et al., 2021a; Sun et al., 2023), catalyzing a series of subsequent CLIP-like models. Secondly, CLIP demonstrates extensive applicability across various domains. Therefore, enhancing CLIP can effectively extend its benefits to a wider range of vision-language applications.

Vision-Language Compositionality. Despite the impressive advancements achieved in VLMs, recent studies (Zhao et al., 2022; Yuksekgonul et al., 2022; Thrush et al., 2022) show that existing VL models exhibit limited compositional reasoning abilities. Yuksekgonul et al. (Yuksekgonul et al., 2022) argue that image-text contrastive learning learns shortcuts and does not learn enough compositional information such as relation and attribute. To

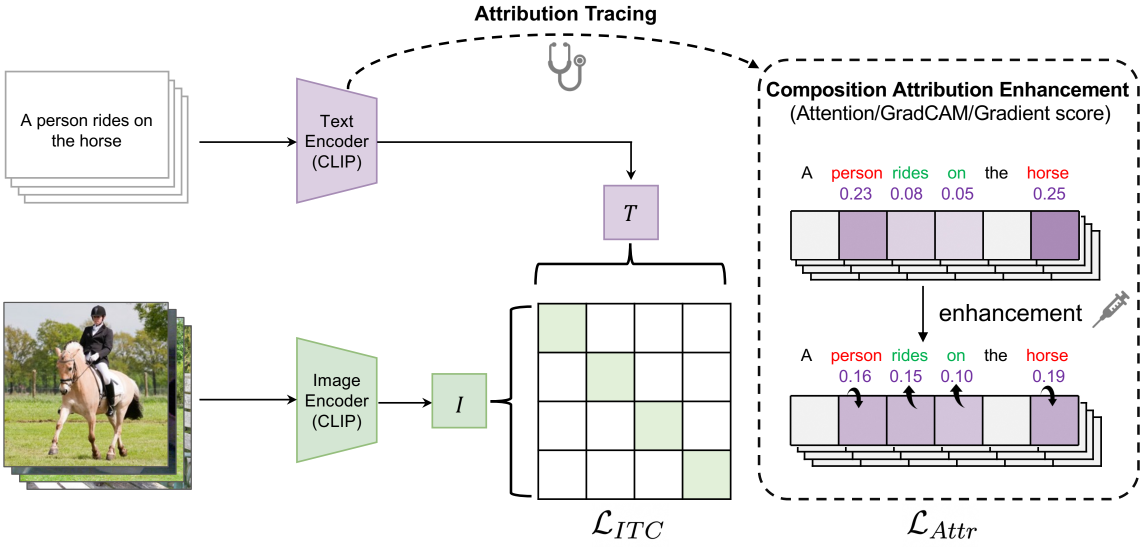


Figure 2: Overview of our method.

address this limitation, exisiting approaches mostly investigate how to augment the text captions or images in contrastive learning to enhance the ability of compositional understanding (Yuksekgonul et al., 2022; Singh et al., 2023; Doveh et al., 2024; Zhang et al., 2024; Doveh et al., 2023; Sahin et al., 2024; Cascante-Bonilla et al., 2023). Yuksekgonul et al. (Yuksekgonul et al., 2022) firstly proposed a simple and straightforward fix: mining hard negatives, which can improves the model's performance. Smith et al. (Basu et al., 2023) enhances CLIP's visio-linguistic reasoning via introducing a distillation objective from text-to image generative models such as Stable-Diffusion.

Enhance Models with Interpretation Methods. In both natural language processing and computer vision community, some previous works have been proposed to use interpretation methods to augment model. For instance, (Ghaeini et al., 2019) introduces an loss function that encourages the gradient of the input to positively influence the ground truth. (Huang et al., 2021) designs a method that constrains the model to focus more on rationales than non-rationales. (Ebrahimi et al., 2021) addresses the issue of catastrophic forgetting in continual learning by encouraging the model to concentrate on its initial decision-making explanations. In the realm of medical imaging, (Simpson et al., 2019) proposes a regularization method that penalizes visual saliency maps derived from classifier gradients when these maps are inconsistent with lesion segmentation, thereby mitigating overfitting issues. Furthermore, (Yang et al., 2023) enhances the model's visual grounding capability by constraining visual gradient-based explanations to be consistent with region-level annotations provided by humans.

3 Method

Our approach employs an attribution method to derive attribution scores on the text, subsequently optimizing these attribution scores to enhance model's capability for compositional reasoning.

Preliminary: Consider a training example consisting of an image $\boldsymbol{I}$ and its corresponding caption $\boldsymbol{T}$. Contrastive Loss CLIP consists of a text encoder $f_t : T \longrightarrow \mathbb{R}^d$ and an image encoder $f_i : I \longrightarrow \mathbb{R}^d$ to encode image and text into embedding space $\mathbb{R}^d$ separately. The image-text similaity score are computed as:

$$S(I,T) = \frac{f_i(I) \cdot f_t(T)}{||f_i(I)|| \cdot ||f_t(T)||} / \tau, \qquad (1)$$

where temperature τ is a learnable parameter. Consider a batch $\mathcal{B}$ consisting of N pairs of images and texts sampled from the training dataset. The Image-Text Contrastive (ITC) loss $\mathcal{L}_{ITC}$ contains an image-to-text constrastive loss $\mathcal{L}_{i2t}$ and a text-to-image contrastive loss $\mathcal{L}_{i2t}$ that

$$\mathcal{L}_{ITC} = (\mathcal{L}_{i2t} + \mathcal{L}_{i2t})/2. \qquad (2)$$

The image-to-text contrastive loss $\mathcal{L}_{i2t}$ and text-to-image contrastive loss $\mathcal{L}_{t2i}$ are formulated as follows:

$$\mathcal{L}_{i2t} = \sum_{(I,T)\in\mathcal{B}} -\log \frac{\exp^{S(I,T)}}{\sum_{T_i \in \mathcal{B}} \exp^{S(I,T_i)}}, \qquad (3)$$

$$\mathcal{L}_{t2i} = \sum_{(I,T)\in\mathcal{B}} -\log \frac{\exp^{S(I,T)}}{\sum_{I_j\in\mathcal{B}} \exp^{S(I_j,T)}}. \tag{4}$$

Formulation Firstly, we utilize a widely-used text scene graph parser (Wu et al., 2019) to parse the caption $\boldsymbol{T}$, extracting the relations, attributes and objects present within the text. Which little cost, this process effectively categorizes which tokens in $\boldsymbol{T}$ pertain to relations or attributes, and which pertain to objects. Note the parsing process is only applied to the training samples. The trained CLIP is used in the same way as the original one.

For an VLM such as CLIP, the attribution score a_i for each token T_i in the caption indicates the contribution or importance of each token to the output image-text similarity. A higher magnitude of a_i signifies a greater importance of T_i to the final output. Given our knowledge of the positions of object tokens and relation/attribute tokens in the text, we can obtain the attribution scores for these tokens. For each sample, we derive the object attribution score a_{obj} by averaging the attribution scores of all object tokens. Similarly, we obtain the compositional attribution score a_{comp} for each sample by averaging the attribution scores of all relation/attribute tokens. For the entire batch, we define $A_{obj} = [a^0_{obj}, a^1_{obj}, a^2_{obj}, ..., a^n_{obj}]$, $A_{comp} = [a^0_{comp}, a^1_{comp}, a^2_{comp}, ..., a^n_{comp}]$, n is the batch size. The proposed CAE introduces an extra learning objective $\mathcal{L}_{Attr}$ that optimize the text attribution score to encourage the model to pay more attention on relation or attribute tokens. An intuitive approach is to make the two items as close as possible, a idea that is also reflected in (Huang et al., 2021). Therefore, we define the attribution loss as follows:

$$\mathcal{L}_{Attr} = max(A_{obj} - A_{comp} + \epsilon, 0), \tag{5}$$

where ϵ denotes the margin hyper-parameter, and is set to 0 default for all our experiments. The overall objective function is formulated as follows:

$$\mathcal{L}_{total} = \mathcal{L}_{ITC} + \lambda \cdot \mathcal{L}_{Attr}, \tag{6}$$

where λ is a hyper-parameter balancing the two objectives.

In the following subsections, we introduce four instances with different attribution types.

3.1 Attention-Based Attribution

In this method, for a given batch of data, we initially extract the attention matrices from each layer of the text encoder (averaging across all heads). Subsequently, we isolate the attention scores of the CLS token with respect to the other tokens within these matrices, designating them as the attribution scores for the current layer. Then, we average of the attribution scores across all layers to obtain the final attribution a_i for each token in the sentence. Finally, we compute the average attribution score for all object tokens to get a_{obj} and similarly for all relation and attribute tokens to get a_{comp}.

3.2 GradCAM-Based Attribution

In this method, we follow the attribution approach proposed in (Chefer et al., 2021) to obtain a attribution score for each text token, given the calculated image-text similarity score.

Firstly, we initialize the text attribution map R as an identity matrix, the dimensions of which correspond to the size of the attention matrix at each layer of the text encoder. Subsequently, we compute the gradients of the attention weights by leveraging the image-text similarity computed from paired image-text inputs and average them across all attention heads. This procedure yields an explainability map $\mathbf{\bar{E}_i}$ for each layer i.

$$\mathbf{\bar{E}_i} = \sum_{j=1}^{h} (\nabla \mathbf{A^i_j} \odot \mathbf{A^i_j})^{+}, \tag{7}$$

where $\odot$ is the Hadamard product, $\mathbf{A^i_j}$ denote the attention matrix of the head j in layer i, $\nabla \mathbf{A^i_j} := \frac{\partial S(I,T)}{\partial \mathbf{A^i_j}}$ for $S(I,T)$ which is the the similarity score computed for the text T with the image I.

Finally, we aggregate the explainability maps of all layers using the propagation rule as presented in (Chefer et al., 2021) to derive the final text attribution map.

$$\mathbf{R} \leftarrow \mathbf{R} + \mathbf{\bar{E}_i} \cdot \mathbf{R}. \tag{8}$$

Then, we use the row of R that corresponds to the CLS token to get the object attribution score A_{obj} and compositional attribution score A_{comp} of each sample similar to Attention-Based method.

3.3 Perturbation-Based Attribution

Consider a paired image $\boldsymbol{T}$ and text $\boldsymbol{I}$, CLIP can computes their similarity score $S(I,T)$. To obtain attribution scores for each token in $\boldsymbol{T}$, inspired by the "Input Marginalization" methodology (Kim et al., 2020), we perturb the input text while keeping the image fixed. Specifically, we replace a

	ARO		Sugar-Crepe		VL-Checklist		VALSE	SVO-Probes	ComVG	Winoground		
Model	Relation	Attribute	Relation	Attribute	Relation	Attribute	Relation	Relation	Relation	Text	Image	Group
Random Chance	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	25.0	25.0	16.7
CLIP (Radford et al., 2021)	58.7	62.7	68.8	70.8	63.6	67.7	66.1	79.5	66.7	31.6	11.1	9.4
SDS-CLIP (Basu et al., 2023)	53.0	62.0	-	-	-	-	-	-	-	-	-	
CLIP-FT	64.9	66.3	70.8	77.5	60.8	67.5	67.2	84.1	70.8	33.9	8.2	5.3
CLIP-CAE (Attention-Based)	**69.7**	65.3	**72.0**	**79.2**	**65.4**	**68.4**	**69.1**	**84.5**	**72.5**	**33.9**	**13.5**	**8.2**
CLIP-CAE (GradCAM-Based)	**70.9**	65.2	**73.0**	**77.9**	**66.8**	**68.3**	**67.6**	83.7	**72.7**	29.8	**9.9**	**7.0**
CLIP-CAE (Perturbation-Based)	**69.8**	65.3	**74.3**	**79.7**	**67.8**	**69.8**	**68.9**	84.0	**73.2**	28.7	**8.2**	**5.3**
CLIP-CAE (Gradient-Based)	**68.1**	65.8	**73.7**	**79.0**	**61.7**	**67.9**	**69.2**	83.6	**72.4**	29.8	**8.8**	**5.9**

Table 1: **Results on ARO, Sugar-Crepe, VL-Checklist, VALSE, SVO-Probes, ComVG and Winoground**. Highlighted in **bold** denote an improvement over CLIP-FT, while the underlined ones indicate a performance degradation compared to CLIP-FT. Empty scores mean that the model's code has not been released.

current token with another distinct token. Given characteristic of our task, we further constrain the perturbation range. For tokens representing objects, relations, or attributes, we randomly select an alternative concept from a corresponding candidate set as the replacement token. The attribution score for the current token is then defined as the average drop in similarity score $S(I, T)$ resulting from multiple perturbation:

$$a_i = \mathbb{E}_p[S(T, (\texttt{stopgrad}(I)) - S(T_p, (\texttt{stopgrad}(I))] \quad (9)$$

where $\mathbb{E}_p$ is the mean across multiple perturbation.

This approach allows us to calculate attribution scores for each object, relation, or attribute token within the sentence. Then, we compute the average attribution score for all object tokens to get a_{obj} and for all relation and attribute tokens to get a_{comp}.

3.4 Gradient-Based Attribution

The attribution score a_i is defined as a function of the gradient of the input text token $\mathbf{x}_i$. Specifically, we sum the absolute values of the gradients across the input embedding dimensions to obtain the gradient for each input text token:

$$a_i = \sum_{j=1}^{d} \| \frac{\partial S(I, T)}{\partial \mathbf{x}_{ij}} \|_1 \quad (10)$$

where $\mathbf{x}_{ij}$ represents the j-th dimension of token $\mathbf{x}_i$, $S(I, T)$ denote the image-text similarity computed by the model for a paired text $\boldsymbol{T}$ and image $\boldsymbol{I}$.

Subsequently, a softmax function is applied to normalize all token gradient values. The attribution score for each token is thus defined as the normalized gradient value.

4 Experiments

Datasets. For training, we use the approximately 110k image-text pairs from MSCOCO (Lin et al., 2014) given that its captions are less noisy and provide a richer description of the relation and attribute content in the images. For evaluation, we use ARO (Yuksekgonul et al., 2022), Sugar-Crepe (Hsieh et al., 2024), VL-Checklist (Zhao et al., 2022), Winoground (Thrush et al., 2022), VALSE (Parcalabescu et al., 2021), SVO-Probes (Hendricks and Nematzadeh, 2021) and ComVG (Jiang et al., 2022). The details of these seven datasets are in the Appendix B.

Implementation Detail. We used the popular ViT-B/32 OpenAI CLIP (Radford et al., 2021) as our model in all the experiments using the Open-CLIP repository (Ilharco et al., 2021). We finetune it for 5 epochs with a batch size of 256. We use a cosine schedule with an initial learning rate of 5e-7 and use 50 steps for warm up. AdamW (Kingma and Ba, 2014) optimizer is used with a weight decay of 0.2. All the experiments are conducted on a NVIDIA Tesla V100 GPU.

Baseline. Our approach is mainly compared against two distinct baselines: (i) a pre-trained CLIP model; (ii) a CLIP model fine-tuned on MSCOCO utilizing only the contrastive loss, devoid of our proposed attribution optimization loss. It is imperative to emphasize that the second baseline, (ii), plays a critical role in mitigating the influence of image-text pairs derived from MSCOCO during the finetuning process.

4.1 Main Results

Table 1 presents the comparative performance of our proposed method against the baseline across seven evaluation benchmarks comprehensively designed for compositional understanding. All our CLIP-CAE models are trained on the same dataset and with the same training hyperparameters as CLIP-FT. Without bells and whistles, our method,

	ARO		Sugar-Crepe		VL-Checklist		VALSE	SVO-Probes	ComVG	Winoground		
Model	Relation	Attribute	Relation	Attribute	Relation	Attribute	Relation	Relation	Relation	Text	Image	Group
Random Chance	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	25.0	25.0	16.7
CLIP	58.7	62.7	68.8	70.8	63.6	67.7	66.1	79.5	66.7	31.6	11.1	9.4
CLIP-FT *with HN*	80.5	71.4	73.3	79.8	71.9	70.0	75.5	83.7	70.2	29.2	8.8	5.3
CLIP-CAE (Attention-Based)	77.9	69.7	**74.4**	**81.6**	**73.0**	69.9	74.5	**84.4**	**71.4**	**33.9**	**12.9**	**8.2**
CLIP-CAE (GradCAM-Based)	80.0	68.9	**74.8**	**80.8**	**74.0**	**70.1**	74.3	**84.1**	**73.0**	26.3	**9.9**	**5.9**
CLIP-CAE (Perturbation-Based)	79.8	69.9	**74.2**	**83.4**	**75.1**	**71.1**	72.9	**84.7**	**73.9**	26.9	**9.9**	**7.0**
CLIP-CAE (Gradient-Based)	**80.8**	**71.6**	**74.1**	**80.7**	**73.0**	69.8	75.4	**84.1**	**71.0**	27.5	8.8	**5.9**

Table 2: **Results on ARO, Sugar-Crepe, VL-Checklist, VALSE, SVO-Probes, ComVG and Winoground when combined with hard negative samples**. Highlighted in **bold** denote an improvement over CLIP-FT *with HN*, while the underlined ones indicate a performance degradation compared to CLIP-FT *with HN*.

incorporating four distinct attribution variants, consistently demonstrates significant improvements over CLIP-FT across nearly all seven benchmarks. Notably, on the highly challenging visio-linguistic reasoning benchmark, Winoground, our method exhibits superior performance. For instance, the CLIP-CAE (Attention-Based) model achieves an average absolute improvement of 5.3% on the Winoground image score and an average absolute improvement of 2.9% on the Winoground group score (most difficult average metric). Additionally, our method demonstrates a slight performance decline on ARO-Attribute compared to CLIP-FT (though still better than pretrained CLIP). Upon meticulous examination of certain failure cases within the dataset, we observed that the alignment between images and corresponding true caption in this dataset is subtly ambiguous.

These alignments necessitate meticulous discernment even for human, thereby indicating a higher level of difficulty and the presence of noise within the dataset. This observation is consistent with related works (Cascante-Bonilla et al., 2023) that utilize hard-negative samples, also exhibiting negligible performance fluctuations on ARO-Attribute.

4.2 Combined with Hard-Negative samples

Given that our method is orthogonal to hard-sample mining, we sought to further verify the generality and efficacy of our approach by integrating it with hard negative samples. The experiment results are presented in Table 2. Compared to the pretrained model, utilizing hard negative samples substantially improves model performance across most datasets. However, performance also exhibit considerable decline on the out-of-domain and challenging Winoground. When compared to using hard negative samples alone, the combination of our method and hard negative samples yield superior performance improvements. For instance, CLIP-Neg obtain a remarkable 71.9% accuracy on VL-Checklist-Relation, which is further elevated to 75.1% with our combined approach, surpassing CLIP-Neg by 3.2%. Notably, on the Winoground, the integration of our method significantly enhance performance over CLIP-Neg, with absolute improvements up to 4.7% in text score, 4.1% in image score, and 2.9% in group score.

This phenomenon is plausible, as our method enables the model to pay more attention on concepts beyond object words. Consequently, when combined with hard negative samples, the model can more effectively discern nuance semantic differences in positive and negative text samples, especially words related to different relations and attributes, thereby enhancing its understanding of compositional relationship in text. These results further validate the effectiveness and plug-and-play nature of our method.

4.3 Results on Downstream Retrieval Tasks

In practical applications, CLIP is often utilized for image-text retrieval. Previous study (Cascante-Bonilla et al., 2023; Yuksekgonul et al., 2022) suggest that improvement in compositional understanding may negatively affect the model's performance on image-text retrieval. To investigate this, we evaluate our model on the downstream image-text retrieval task. As shown in Table 3, our approach shows overall improvements in text-to-image retrieval, albeit it exhibits a minor underperformance in image-to-text retrieval on the Flickr30K. This discrepancy could be due to the exclusive regularization imposed on text encoder in our method. The overall improvements in test-to-image retrieval present the potential of our plug-and-play method in enhancing the general text embeddings models.

Model	MSCOCO				Flickr30K			
	T2I		I2T		T2I		I2T	
	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5
CLIP	30.2	55.7	50.1	74.9	59.0	83.5	78.4	95.1
CLIP-CAE (AB)	**38.3**	**65.2**	**54.9**	**78.7**	**64.4**	**87.1**	**80.2**	**95.6**
CLIP-CAE (GCB)	**34.8**	**61.9**	48.5	74.8	**59.8**	**84.2**	73.5	92.5
CLIP-CAE (PB)	**36.5**	**63.5**	**52.7**	**77.0**	**63.3**	**87.1**	**78.9**	94.7
CLIP-CAE (GB)	**38.8**	**65.6**	**54.4**	**78.2**	**65.1**	**87.6**	**80.1**	94.9
Avg.	**37.1**	**64.1**	**52.6**	**77.2**	**63.2**	**86.5**	78.2	94.4

Table 3: **Downstream results on MSCOCO and Flickr30K**. Highlighted in **bold** denote an improvement over baseline, while the underlined ones indicate a performance degradation compared to baseline.

Model	SICK-R		STSBenchmark	
	spearman	pearson	spearman	pearson
CLIP	67.9	68.6	61.5	59.1
CLIP-FT	68.0	73.5	66.3	64.0
CLIP-CAE	**69.3**	71.6	**66.5**	**65.2**

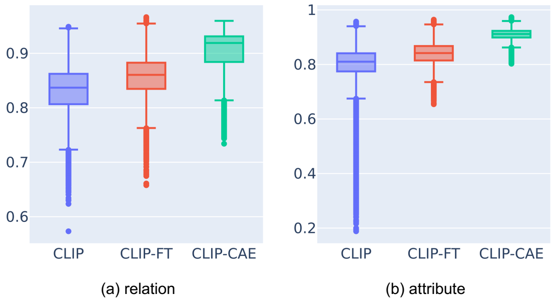
Table 4: **Semantic Textual Similarity results on SICK-R and STSBenchmark.** Highlighted in bold denote an improvement over CLIP-FT, while the underlined ones indicate a performance degradation compared to CLIP-FT.

4.4 Analysis

4.4.1 Analysis of text embedding

Semantic Textual Similarity We evaluate our text encoder and text encoder of CLIP and CLIP-FT on the task of Semantic Textual Similarity (STS), using two widely-used benchmarks: the STS-Benchmark (Cer et al., 2017) and SICK-R (Marelli et al., 2014). As indicated in Table 4, our text encoder consistently outperforms CLIP-FT across both benchmarks, especially on SICK-R. Our CLIP-CAE significantly surpasses both CLIP-FT and CLIP , with CLIP-FT exhibiting only a nominal 0.1 improvement over CLIP. A slight decrease compared to CLIP-FT in pearson correlation on SICK-R may be due to the non-linear nature existing in high-dimensional embedding space. These results demonstrate that our text encoder excels in capturing nuanced semantic differences and complex semantic relationships within texts, resulting in embeddings with superior semantic representational properties. This indicates that our model not only achieves superior cross-modal image-text alignment but also enhances text representation. Consequently, our method not only boosts multi-modal capabilities but also shows promise for application in uni-modal language tasks, which will be explored in our future work.

Text Embedding Ingredients We conduct an analysis on ARO-Relation and ARO-Attribute datasets to validate that our text encoder can capture relations and attributes within captions more effectively. Specifically, for each sample, we separately encode the correct caption and the relation or attribute phrase annotated within these captions to obtain their respective text embeddings. Subsequently, we calculate the cosine similarity between the embeddings derived from the full caption and the relations or attributes phrases. As shown in Fig. 3, it can be observed that the embeddings generated by the text encoder of CLIP-CAE exhibit a significantly higher overall similarity compared to those produced by CLIP and CLIP-FT. This finding indicates that the text encoder of CLIP-CAE places greater emphasis on relations and attributes when encoding text, resulting in embeddings that encapsulate more information about these semantic elements.



(a) relation

(b) attribute

Figure 3: The similarity distribution between the text embeddings obtained by encoding the entire text and those derived from encoding specific relations or attributes within the text.

4.4.2 Relationship Between Attribution Score and Performance

We investigate the variations in the model's performance as a function of attribution scores for relations or attributes. We utilize pretrained CLIP model to conduct experiments on the ARO-Relation. The focus level of the model on relations is quantified by the ratio of the attribution score for the relation tokens to that of the object tokens. Concurrently, we assess the model's accuracy across all samples with ratios below current value. As illustrated in Fig. 4, the ratios for all samples predominantly fall within the range of 0.36 to 0.38. Within this interval, a higher attribution score ratio corresponds to a cumulative increase in accuracy. This trend indicates that the more attention the model allocates to the relation tokens, the better it dif-

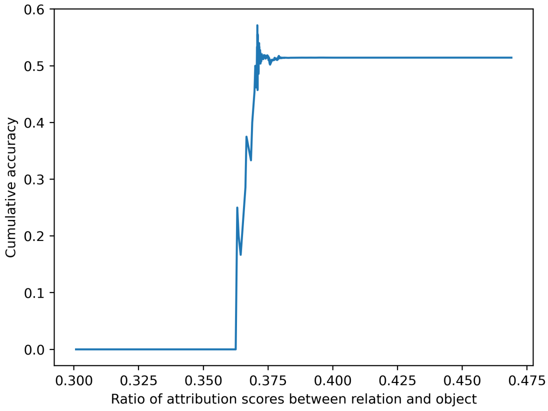


Figure 4: Cumulative accuracy on ARO-Reletion of CLIP vs. the attribution score ratio between relation and object tokens.

ferentiates the compositional nuances in correct caption and false caption, thereby exhibiting better compositional understanding capabilities. This phenomenon further substantiates the reasonability of our proposed method.

4.4.3 Ablation

We conducted ablation studies under various training configurations, with the results presented in Table 5. It can be observed that that utilizing only $\mathcal{L}_{Attr}$ results in a performance decline across multiple benchmarks. This performance degradation can be due to the absence of $\mathcal{L}_{ITC}$, which serves as a constraint to align image and text features. Without this constraint, features may undergo excessive deviation, thereby compromising the original alignment performance. When the $\mathcal{L}_{ITC}$ is combined with our attribution loss, the model exhibits superior performance across all benchmarks, thus demonstrating the effectiveness of our approach.

Model	$\mathcal{L}_{ITC}$	$\mathcal{L}_{Attr}$	**ARO**	**Sugar-Crepe**	**VL-CheckList**	**VALSE**	**ComVG**	**Avg.**
CLIP			60.7	69.8	65.7	66.1	66.7	65.8
CLIP-FT		✓	59.5	71.5	**69.6**	64.3	63.0	65.6
CLIP-FT	✓		65.6	74.2	64.2	67.2	70.8	68.4
CLIP-CAE	✓	✓	**67.5**	**75.6**	66.9	**69.1**	**72.5**	**70.3**

Table 5: **Ablation of losses.** $\mathcal{L}_{ITC}$ represents image-text contrastive loss, $\mathcal{L}_{Attr}$ denote our proposed attribution loss.

4.4.4 Case Study

In Fig. 5, we employ the GradCAM tool (Chefer et al., 2021) to visualize a qualitative example from the ARO-Relation, generating attribution maps for both image and text. It is evident that the original CLIP model excessively attends to object-specific regions in both modalities. In contrast, our proposed CLIP-CAE directs the model's attention beyond objects to areas representing relationship. For instance, in this example, CLIP-CAE more effectively focuses on the regions depicting the interaction between the cat and the elephant, specifically the area where they touch, as well as the word "touch" in the text. This demonstrates that our model is better at capturing regions in images and text that represent compositional concepts, such as the interaction between two objects.

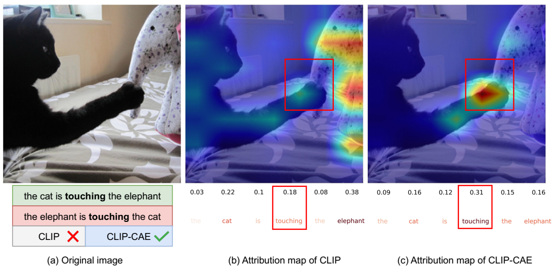


Figure 5: A qualitative visualization case. Both image and text attribution maps are displayed.

5 Conclusion

In this work, we present an intuitive and novel method to enhance the composition attribution and the compositional reasoning ability of contrastive vision-language models such as CLIP. Extensive experiments across variant attribution method and seven benchmarks show the effectiveness of our method. Our method can be easily integrated with existing hard negative mining techniques to further boost the performance. We hope our methods can provide useful insights to solve the compositional understanding dilemma of VLMs and improves the semantic representations of texts.

6 Limitation

Despite our approach effectively enhances the model's compositional understanding ability across various attribution methods without employing hard negative samples, our method does not impose explicit constraints or enhancements on the visual component of VLMs. Analyzing and explicitly enhancing the visual model through diverse attribution and interpretation methods will be a focus of our future work. Furthermore, we also intend to employ our approach to further interpret and analyze existing model deficiencies, thereby enabling precise optimization and enhancement.

References

Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. 2019. Nocaps: Novel object captioning at scale. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8948–8957.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736.

Samyadeep Basu, Maziar Sanjabi, Daniela Massiceti, Shell Xu Hu, and Soheil Feizi. 2023. Augmenting clip with improved visio-linguistic reasoning. *arXiv preprint arXiv:2307.09233*.

Paola Cascante-Bonilla, Khaled Shehada, James Seale Smith, Sivan Doveh, Donghyun Kim, Rameswar Panda, Gul Varol, Aude Oliva, Vicente Ordonez, Rogerio Feris, et al. 2023. Going beyond nouns with vision & language models using synthetic data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20155–20165.

Daniel Cer, Mona Diab, Eneko Agirre, Inigo Lopez-Gazpio, and Lucia Specia. 2017. Semeval-2017 task 1: Semantic textual similarity-multilingual and cross-lingual focused evaluation. *arXiv preprint arXiv:1708.00055*.

Hila Chefer, Shir Gur, and Lior Wolf. 2021. Generic attention-model explainability for interpreting bimodal and encoder-decoder transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 397–406.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255.

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

Sivan Doveh, Assaf Arbelle, Sivan Harary, Roei Herzig, Donghyun Kim, Paola Cascante-Bonilla, Amit Alfassy, Rameswar Panda, Raja Giryes, Rogerio Feris, et al. 2024. Dense and aligned captions (dac) promote compositional reasoning in vl models. *Advances in Neural Information Processing Systems*, 36.

Sivan Doveh, Assaf Arbelle, Sivan Harary, Eli Schwartz, Roei Herzig, Raja Giryes, Rogerio Feris, Rameswar Panda, Shimon Ullman, and Leonid Karlinsky. 2023. Teaching structured vision & language concepts to vision & language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2657–2668.

Sayna Ebrahimi, Suzanne Petryk, Akash Gokul, William Gan, Joseph E Gonzalez, Marcus Rohrbach, and Trevor Darrell. 2021. Remembering for the right reasons: Explanations reduce catastrophic forgetting. *Applied AI letters*, 2(4):e44.

Reza Ghaeini, Xiaoli Z Fern, Hamed Shahbazi, and Prasad Tadepalli. 2019. Saliency learning: Teaching the model where to pay attention. *arXiv preprint arXiv:1902.08649*.

Lisa Anne Hendricks and Aida Nematzadeh. 2021. Probing image-language transformers for verb understanding. *arXiv preprint arXiv:2106.09141*.

Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. 2024. Sugarcrepe: Fixing hackable benchmarks for vision-language compositionality. *Advances in Neural Information Processing Systems*, 36.

Quzhe Huang, Shengqi Zhu, Yansong Feng, and Dongyan Zhao. 2021. Exploring distantly-labeled rationales in neural network models. *arXiv preprint arXiv:2106.01809*.

Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, et al. 2021. Openclip. *If you use this software, please cite it as below*, page 1.

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

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021b. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, volume 139, pages 4904–4916.

Kenan Jiang, Xuehai He, Ruize Xu, and Xin Eric Wang. 2022. Comclip: Training-free compositional image and text matching. *arXiv preprint arXiv:2211.13854*.

Amita Kamath, Jack Hessel, and Kai-Wei Chang. 2023. What's" up" with vision-language models? investigating their struggle with spatial reasoning. *arXiv preprint arXiv:2310.19785*.

Siwon Kim, Jihun Yi, Eunji Kim, and Sungroh Yoon. 2020. Interpretation of nlp models through input marginalization. *arXiv preprint arXiv:2010.13984*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al.

2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR.

Yong-Lu Li, Liang Xu, Xinpeng Liu, Xijie Huang, Yue Xu, Mingyang Chen, Ze Ma, Shiyi Wang, Hao-Shu Fang, and Cewu Lu. 2019. Hake: Human activity knowledge engine. *arXiv preprint arXiv:1904.06539*.

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

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Marco Marelli, Luisa Bentivogli, Marco Baroni, Raffaella Bernardi, Stefano Menini, and Roberto Zamparelli. 2014. Semeval-2014 task 1: Evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment. In *Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014)*, pages 1–8.

Bolin Ni, Houwen Peng, Minghao Chen, Songyang Zhang, Gaofeng Meng, Jianlong Fu, Shiming Xiang, and Haibin Ling. 2022. Expanding language-image pretrained models for general video recognition. In *European Conference on Computer Vision*, pages 1–18. Springer.

Letitia Parcalabescu, Michele Cafagna, Lilitta Muradjan, Anette Frank, Iacer Calixto, and Albert Gatt. 2021. Valse: A task-independent benchmark for vision and language models centered on linguistic phenomena. *arXiv preprint arXiv:2112.07566*.

Khoi Pham, Kushal Kafle, Zhe Lin, Zhihong Ding, Scott Cohen, Quan Tran, and Abhinav Shrivastava. 2021. Learning to predict visual attributes in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13018–13028.

Sarah Pratt, Mark Yatskar, Luca Weihs, Ali Farhadi, and Aniruddha Kembhavi. 2020. Grounded situation recognition. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 314–332. Springer.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *ICML*, volume 139, pages 8748–8763.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *International conference on machine learning*, pages 8821–8831. Pmlr.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695.

Ugur Sahin, Hang Li, Qadeer Khan, Daniel Cremers, and Volker Tresp. 2024. Enhancing multimodal compositional reasoning of visual language models with generative negative mining. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5563–5573.

Becks Simpson, Francis Dutil, Yoshua Bengio, and Joseph Paul Cohen. 2019. Gradmask: Reduce overfitting by regularizing saliency. *arXiv preprint arXiv:1904.07478*.

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

Harman Singh, Pengchuan Zhang, Qifan Wang, Mengjiao Wang, Wenhan Xiong, Jingfei Du, and Yu Chen. 2023. Coarse-to-fine contrastive learning in image-text-graph space for improved vision-language compositionality. *arXiv preprint arXiv:2305.13812*.

Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. 2023. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. 2022. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5238–5248.

Hao Wu, Jiayuan Mao, Yufeng Zhang, Yuning Jiang, Lei Li, Weiwei Sun, and Wei-Ying Ma. 2019. Unified visual-semantic embeddings: Bridging vision and language with structured meaning representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6609–6618.

Ziyan Yang, Kushal Kafle, Franck Dernoncourt, and Vicente Ordonez. 2023. Improving visual grounding by encouraging consistent gradient-based explanations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19165–19174.

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. 2022. 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*.

Le Zhang, Rabiul Awal, and Aishwarya Agrawal. 2024. Contrasting intra-modal and ranking cross-modal hard negatives to enhance visio-linguistic compositional understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13774–13784.

Tiancheng Zhao, Tianqi Zhang, Mingwei Zhu, Haozhan Shen, Kyusong Lee, Xiaopeng Lu, and Jianwei Yin. 2022. Vl-checklist: Evaluating pre-trained vision-language models with objects, attributes and relations. *arXiv preprint arXiv:2207.00221*.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

A Composition Attribution Deficiency

Fig. 6 illustrates the distribution of attribution scores for object tokens and composition tokens derived from three distinct attribution methods. A consistent pattern emerges across these distinct attribution methods: the attribution scores for object words are markedly higher compared to those for relation and attribute words.

B Appendix: Datasets

(1) ARO (Yuksekgonul et al., 2022) is a large dataset designed to evaluate the compositional understanding ability of VL models. It encompasses two distinct datasets to evaluate relation and attribute understanding respectively: Visual Genome Relation (VG-Relation) and Visual Genome Attribution (VG-Attribution). VG-Relation incorporates 48 distinct relation categories, featuring a total of $23,937$ test cases, whereas VG-Attribution comprises 117 unique attribute pairs, with a total of $28,748$ test cases. Each case within these datasets is accompanied by an image, paired with a matched caption and a swapped mismatched caption.

(2) VL-Checklist (Zhao et al., 2022) is a large-scale dataset containing around 410k images that combines the following four datasets: Visual Genome (Krishna et al., 2017), SWiG (Pratt et al., 2020), VAW (Pham et al., 2021), and HAKE (Li et al., 2019). Each image of these datasets is associated with two captions, a positive and a negative. The positive caption corresponds to the image and is taken from the source dataset. The negative caption is made from the positive caption by changing one word. We report average results for each of the main (Relation and Attribute) groups on VL-Checklist.

(3) Sugar-Crepe (Hsieh et al., 2024) is a recent benchmark designed to avoids ungrammatical and nonsensical negative captions, and generates hard negative captions by swapping, replacing, or adding linguistic elements. In this work, we calculate the accuracy for subsets belonging to the categories of relation and attribute within Sugar-Crepe respectively.

(4) Winoground (Thrush et al., 2022) is a modestly-sized dataset containing 400 samples designed to assess the compositional reasoning capabilities of VL models. Each sample within the dataset consists of two image-text pairs, characterized by overlapping lexical content but distinguished by the alteration of an object, a relation, or both. For every sample, two text-retrieval tasks (text score) and two image-retrieval tasks (image score) are defined, with a combined group score representing overall performance. Recent study (Diwan et al., 2022) has analyzed that successful performance on Winoground necessitates competencies beyond simple compositionality. The study identified a subset of 171 out of the total 400 samples that reliably probe compositional reasoning. In contrast, other samples within the dataset were found to be non-compositional, ambiguous, predicated on invisible details, or associated with highly uncommon images or text, thus requiring more complex reasoning beyond compositionality. Consequently, we report our results on this "clean" subset following (Cascante-Bonilla et al., 2023).

(5) VALSE is a benchmark specifically designed to evaluate the capabilities of VL models across six distinct linguistic phenomena. Each sample within this benchmark comprises an image paired with both a correct caption and a false caption. The

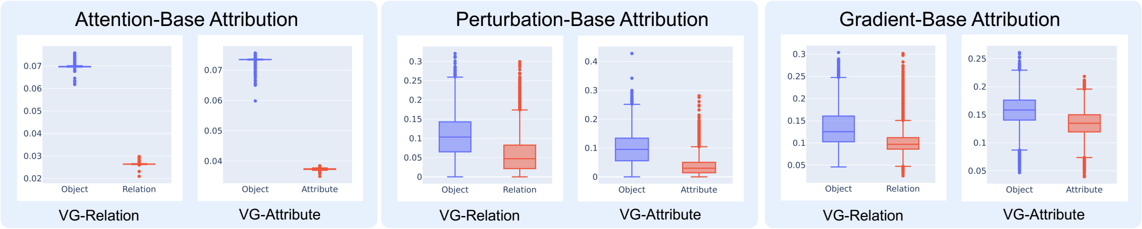


Figure 6: The attribution score distribution of object and composition token using attention-based, perturbation-based and gradient-Based attribution method.

false caption is generated by modifying a word or phrase within the original caption, targeting a particular linguistic phenomenon—such as verb argument structure, spatial relation, or coreference. Three subsets within the benchmark focus on action and spatial relations, aligning closely with our task of compositional understanding. In this study, we report the average accuracy across these three pertinent subsets.

(6) SVO-Probes (Hendricks and Nematzadeh, 2021) and **ComVG** (Jiang et al., 2022) assess VLMs on verb (relation) understanding.