Interpretable Composition Attribution Enhancement for Visio-linguistic Compositional Understanding

Wei Li[1,2*], Zhen Huang[2], Xinmei Tian[1†], Le Lu[2],
Houqiang Li[1], Xu Shen[2†], Jieping Ye[2]
[1]MoE Key Laboratory of Brain-inspired
Intelligent Perception and Cognition, University of Science and Technology of China
[2]Alibaba Cloud
lwzkd@mail.ustc.edu.cn, {xinmei, lihq}@ustc.edu.cn
{muzhao.hz, shenxu.sx, yejieping.ye}@alibaba-inc.com

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 the 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 encourage the model to pay greater attention to composition words denoting relationships and attributes within the text. Detailed analysis shows that our approach enables the models to adjust and rectify the attribution of 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).

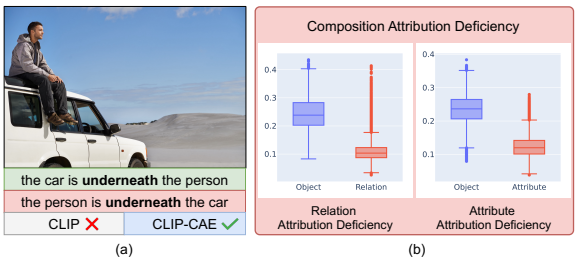

Figure 1: **(Left)** a) An illustrative example from the Winoground benchmark for assessing relation understanding of VLMs. VLMs exhibit difficulty in accurately matching the image with the correct caption (denoted in green). **(Right)** b) The issue of composition attribution deficiency. The distribution of GradCAM-based attribution scores for object tokens is significantly higher than that of the composition tokens (relation and attribute). Similar phenomena are consistently observed across the other three attribution methods, as presented in Appendix A.

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, and 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).

*This work was done when the author was visiting Alibaba Cloud as a research intern.
†Corresponding author.

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 14616–14632
November 12-16, 2024

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 shown in Fig. 1 (Right), our investigation reveals 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 the 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 to 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-negative 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. (2022) argue that

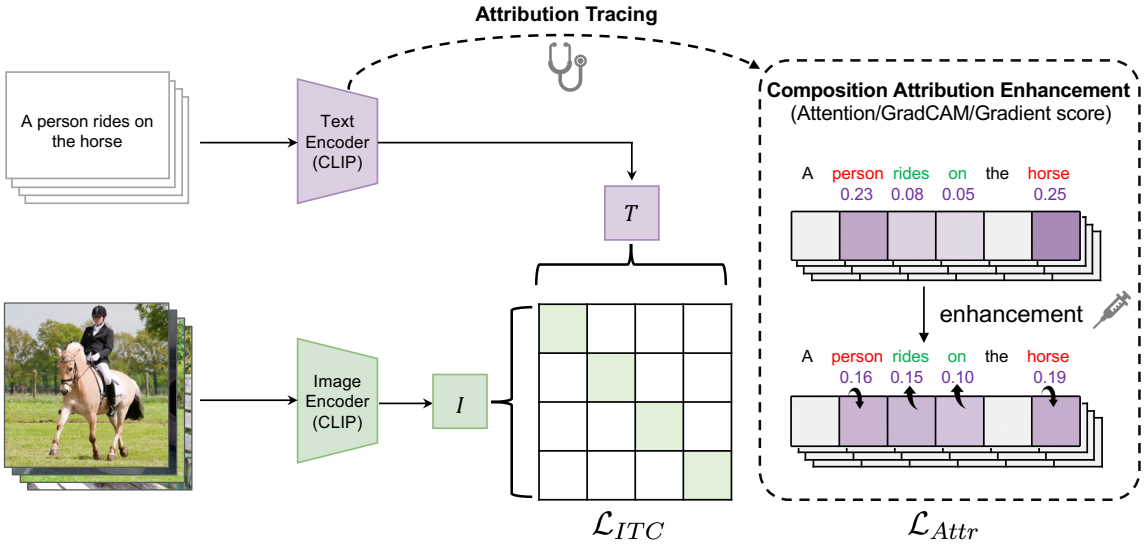

Figure 2: Overview of our method.

image-text contrastive learning learns shortcuts and does not learn enough compositional information such as relation and attribute. To address this limitation, existing 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. (2022) firstly proposed a simple and straightforward fix: mining hard negatives, which can improve the model's performance. 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 communities, some previous works have been proposed to use interpretation methods to augment models. For instance, Ghaeini et al. (2019) proposes a loss function to constrain the gradient values of important words in the input are greater than zero when predicting the groundtruth to encourage the important words in the input to positively influence the right prediction. 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 the 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 similarity score is 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 contrastive 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:

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

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

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. With 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 a 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 optimizes the text attribution score to encourage the model to pay more attention to relation or attribute tokens. An intuitive approach is to make the two items as close as possible, an idea that is also reflected in (Huang et al., 2021). Therefore, we define the attribution loss as follows:

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

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:

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

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 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 an 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 $\bar{\mathbf{E}}_\mathbf{i}$ for each layer i.

$$\bar{\mathbf{E}}_\mathbf{i} = \sum_{j=1}^{h} (\nabla \mathbf{A}^\mathbf{i}_\mathbf{j} \odot \mathbf{A}^\mathbf{i}_\mathbf{j})^+, \quad (7)$$

where $\odot$ is the Hadamard product, $\mathbf{A}^\mathbf{i}_\mathbf{j}$ denote the attention matrix of the head j in layer i, $\nabla \mathbf{A}^\mathbf{i}_\mathbf{j} := \frac{\partial S(I,T)}{\partial \mathbf{A}^\mathbf{i}_\mathbf{j}}$ 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.

$$\mathbf{R} \leftarrow \mathbf{R} + \bar{\mathbf{E}}_\mathbf{i} \cdot \mathbf{R}. \quad (8)$$

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 compute 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 current token with another distinct token. Given the 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. More details can be found in Appendix I. The attribution score for the current token is then defined as the average drop in similarity score $S(I,T)$ resulting from multiple perturbations:

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

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 more detailed description of the relation and attribute content in the images. For evaluation, we use ARO (Yuksekgonul et al., 2022), SugarCrepe (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 OpenCLIP 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 an NVIDIA Tesla V100 GPU. For details on the computational budget, please refer to Appendix C.

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, 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 3.7% on the Winoground image score and an average absolute improvement of 2.5% on the Winoground group score (most difficult average metric). Additional results and analyses about the performance on Winoground are available in Appendix H. We also conduct hyperparameter ablation studies. Please refer to Appendix E for details.

Additionally, our method demonstrates a slight

	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	64.7	66.3	71.1	77.5	60.8	67.4	67.4	84.1	70.7	33.1	8.8	5.5
CLIP-CAE (Attention-Based)	**69.8**	65.3	**72.1**	**79.2**	**65.4**	**68.4**	**69.1**	**84.5**	**72.8**	**33.9**	**12.5**	**8.0**
CLIP-CAE (GradCAM-Based)	**71.0**	65.3	**72.7**	**77.8**	**67.0**	**68.4**	**68.2**	83.6	**73.4**	29.6	**9.8**	**6.8**
CLIP-CAE (Perturbation-Based)	**69.6**	65.2	**74.1**	**79.7**	**67.7**	**69.9**	**69.0**	**84.2**	**73.1**	28.0	8.3	**5.8**
CLIP-CAE (Gradient-Based)	**67.7**	65.8	**73.2**	**79.0**	**61.6**	**67.7**	**68.7**	83.7	**72.6**	29.3	**9.0**	**6.3**

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. We report the average results of each method over five different random seeds. The variances and confidence intervals can be found in the Appendix D.

	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*.

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 captions in this dataset is subtly ambiguous. These alignments necessitate meticulous discernment even for humans, 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 exhibits 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 yields superior performance improvements. For instance, CLIP-Neg obtains 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 enhances 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. The experimental results for the combination with another representative hard-negative mining method (CE-CLIP (Zhang et al., 2024)) are provided in Appendix F, showing similar outcomes.

This phenomenon is plausible, as our method enables the model to pay more attention to 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 relationships in text. These results

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		STS-Benchmark	
	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 STS-Benchmark.** Highlighted in bold denote an improvement over CLIP-FT, while the underlined ones indicate a performance degradation compared to CLIP-FT.

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) suggests that improvement in compositional understanding may negatively affect the model's performance on image-text retrieval, which is also verified in Table 10 in the Appendix. 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 the text encoder in our method. The overall improvements in text-to-image retrieval present the potential of our plug-and-play method in enhancing the general text embedding models. Additional results and analyses in comparison with other methods can be found in Appendix G.

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.

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 a pretrained CLIP model to conduct experiments on the ARO-

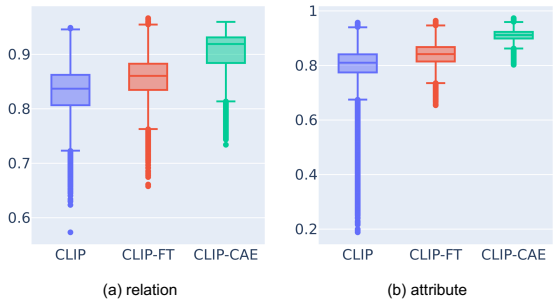

Figure 3: The similarity distribution between the text embeddings obtained by encoding the entire text and those derived from encoding specific relations or attributes within the text.

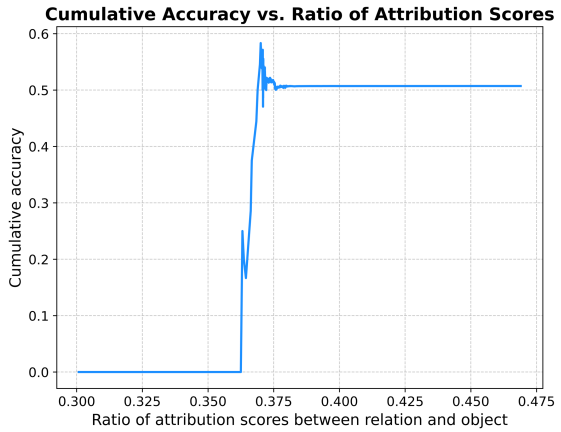

Figure 4: Cumulative accuracy on ARO-Relation of CLIP vs. the attribution score ratio between relation and object tokens.

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 the 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 differentiates the compositional nuances in the 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 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

To investigate whether our model more accurately associates the regions in images corresponding to compositional relationships with the relevant words in the text compared to CLIP, we employ the Grad-CAM (Chefer et al., 2021) tool to visualize the region-level associations between the image and the text. We randomly selected 200 action relation samples from ARO-Relation (with actions that can be grounded in images, *i.e.*, "cutting", "eating", "feeding", "holding", "leaning on", "lying on", "riding", "sitting on", "touching"). We invite 5 volunteers to take a blind setting of which heatmap gets the relation more visibly correct. Specifically, we labeled the heatmaps from CLIP-CAE and CLIP as 0 and 1, respectively. The volunteers are unaware of which model each heatmap came from and only know the labels. We asked the volunteers to choose which heatmap gets the relation more visibly correct for each sample (we also provide a "ambiguous" choice, meaning that these two heatmaps are pretty similar). The average evaluation results from the five volunteers are shown in Fig 5. The result shows that, out of the evaluation results from five volunteers, there is an average of 68 samples for which they think the heatmap of CLIP-CAE more visibly correctly captures the relation compared to CLIP. This interesting phenomenon is reasonable because CLIP-CAE enables the model to focus more on the relations in the text. As a result, there is a higher component of relations in the text features, leading to a better alignment between the text features and the relation regions in the images, which in turn makes the relation regions more visually grounded in the heatmap.

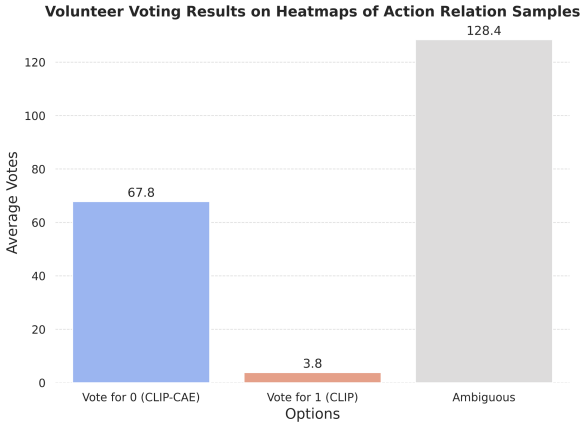

Figure 5: Human voting results.

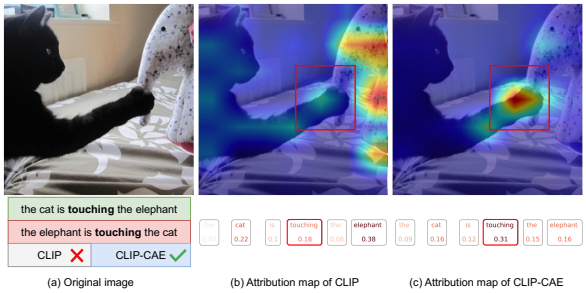

Figure 6: A qualitative visualization case. Both image and text attribution maps are displayed.

In Fig. 6, we show a qualitative example. This case demonstrates 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 relation. 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 image and text that represent compositional concepts, such as the interaction between two objects.

5 Discussion

Orthogonal to hard-negative mining As we know, the three main components of machine learning are data, model, and loss function (Mitchell and Mitchell, 1997). From the perspective of machine learning, existing approaches can primarily be classified into two categories: data-driven and designing new loss functions. A categorization of related work is shown in the Appendix J. Our method falls under the category of loss function, while hard-negative mining methods fall under the category of data-driven. Therefore, as different lines of research, these two approaches improve the model from different dimensions. The hard-negative mining methods focus on constructing hard negative samples in a similar way as the vision-language compositionality benchmarks, which have significantly advanced the field of visual-language compositional understanding. Also, from a machine learning perspective, better data and better loss functions often enable models to learn more effectively (Bishop and Nasrabadi, 2006). This could be one reason why our approach not only enhances the performance of the baseline model, but also combines with state-of-the-art hard-negative mining methods to further boost their performance.

Relationship with ViLT (Kim et al., 2021) ViLT proposes a unified, efficient, and simplest model architecture for vision-and-language pre-training, sharing a similar spirit with CLIP. Both of them tried to achieve cross-modal object-level alignment through general pre-training, *i.e.*, CLIP applied the image-text contrastive learning, while ViLT applied the image-text matching (ITM) and word patch alignment (WPA). Our method introduces a generic framework to encourage the model to pay greater attention to relation and attribute words, which is orthogonal to the model architectures and learning objectives. This means our method possesses the potential to be combined with ViLT. It is also worth noting that ViLT designs a word patch alignment algorithm, using the inexact proximal point method for optimal transports. This technique gives us new ideas to augment our approach further to design a new instantiation of our method. The transportation-based heatmap for a relation or attribute token can be used to represent composition attribution.

6 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 methods 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 improve the semantic representations of texts.

7 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.

8 Acknowledgements

This work was supported in part by NSFC No. 62222117. We sincerely thank the meta-reviewer and the anonymous reviewers for their constructive and insightful feedback.

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*.

Christopher M Bishop and Nasser M Nasrabadi. 2006. *Pattern recognition and machine learning*, volume 4. Springer.

Jirayu Burapacheep, Ishan Gaur, Agam Bhatia, and Tristan Thrush. 2024. Colorswap: A color and word order dataset for multimodal evaluation. *arXiv preprint arXiv:2402.04492*.

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 bi-modal 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*.

Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. 2017. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2901–2910.

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*.

Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *International conference on machine learning*, pages 5583–5594. PMLR.

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.

Martha Lewis, Nihal Nayak, Peilin Yu, Jack Merullo, Qinan Yu, Stephen Bach, and Ellie Pavlick. 2024. Does clip bind concepts? probing compositionality in large image models. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1487–1500.

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.

Tom M Mitchell and Tom M Mitchell. 1997. *Machine learning*, volume 1. McGraw-hill New York.

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. 7 illustrates the distribution of attribution scores for object tokens and composition tokens derived from other three distinct attribution methods (attention-based, perturbation-based and gradient-based). 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 Evaluation datasets

To comprehensively evaluate the effectiveness of our method, we conduct experiments on a total of 7 most commonly used benchmark datasets. The following are the specific details for each dataset.

(1) ARO (Yuksekgonul et al., 2022) is a comprehensive benchmark designed to assess the compositional reasoning capabilities of vision-language (VL) models. It consists of two distinct subsets aimed at evaluating relation and attribute understanding, namely the Visual Genome Relation (VG-Relation) and Visual Genome Attribution (VG-Attribution) datasets. VG-Relation contains 48 distinct relation categories, encompassing 23, 937 test

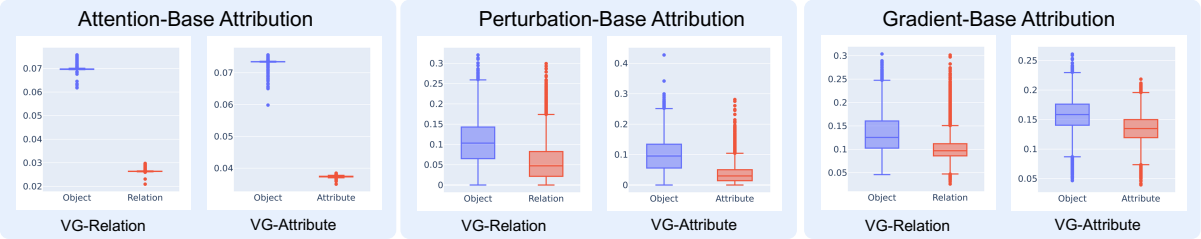

Figure 7: The attribution score distribution of object and composition token using attention-based, perturbation-based, and gradient-based attribution method.

cases, while VG-Attribution includes 117 unique attribute pairs, with a total of 28,748 test cases. Each test case in these datasets is accompanied by an image, a corresponding correct caption, and a swapped mismatched caption.

(2) VL-Checklist (Zhao et al., 2022) a large-scale benchmark, comprising approximately $410k$ images, constructed by integrating four existing datasets: Visual Genome (Krishna et al., 2017), SWiG (Pratt et al., 2020), VAW (Pham et al., 2021), and HAKE (Li et al., 2019). For each image in these datasets, two captions are provided: a positive caption, which accurately describes the image and is sourced from the original dataset, and a negative caption, created by altering one word in the positive caption. 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 exhibit non-compositional characteristics, ambiguity, reliance on invisible details, or association with rare images or texts, necessitating complex reasoning that extends beyond simple compositionality. Consequently, we report our results on this "clean" subset following (Cascante-Bonilla et al., 2023).

(5) VALSE (Parcalabescu et al., 2021) 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 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.

C Computational budget

In Table 6, we present the model sizes, as well as the training and evaluation budgets for all the models in our paper. All experiments were conducted on an NVIDIA Tesla V100 GPU with 32 GB of memory.

D Error bars and confidence intervals

We have run the experiments for each method with five different random seeds, and we report the means, variances, and confidence intervals in Table 7. From the results above, it can be observed that the performance of our CLIP-CAE is consistently higher than that of CLIP-FT by a non-trivial

model	CLIP	CLIP-FT	SDS-CLIP	CLIP-CAE (Attention-Based)	CLIP-CAE (GradCAM-Based)	CLIP-CAE (Perturbation-Based)	CLIP-CAE (Grad-Based)
#Params	151M	151M	151M	151M	151M	151M	151M
Training budge	-	1.1h	-	1.1h	1.1h	1.1h	1.1h
Evaluation budge	0.4h	0.4h	0.4h	0.4h	0.4h	0.4h	0.4h

Table 6: Model size and computational budge.

margin, and the variation in experimental results across different seeds is relatively small (86.7% of the experiments have a standard deviation of less than 0.5).

E Hyperparameter ablation

We conduct comprehensive and separate ablation studies about the hyperparameters, including the learning rate and loss weight. The experiment results in Table 8 make two observations: From 1e−7 to 5e−6, the overall optimal learning rates for different finetuning objectives are the same, which is also the best learning rate of vanilla CLIP-FT. This implies that our finetuning objective shares the same learning dynamics and loss landscape as the normal finetuning. The optimal loss weights for different finetuning objectives varies from 0.2 to 50, *e.g.*, 0.2 for perturbation-based and 50 for GradCAM-based methods. This is related to the differences in the gradient scales of different types of loss. For example, the average gradient of the parameters with respect to the contrastive loss is approximately $1/3$ of that with respect to the perturbation-based attribution loss, while the gradient ratio is about 60 times when change to GradCAM-based attribution loss, which aligns with the corresponding optimal loss weights.

F More results when combined with hard-negative mining methods

As our method is orthogonal to hard-negative mining methods. We conduct experiments to combine our approach with two state-of-the-art methods (NegCLIP (Yuksekgonul et al., 2022) and CE-CLIP (Zhang et al., 2024)). The experimental results are shown in Table 9. From these results, it can be observed that integrating our method with other approaches consistently enhances overall performance. This indicates that our method can effectively combine with other methods to collectively improve the model's compositional understanding capabilities, particularly these mainstream methods based on hard-negative mining.

G More results on downstream retrieval task

We compared our approach on the MSCOCO retrieval task with three representative state-of-the-art hard-negative mining methods, *i.e.*, CE-CLIP (Zhang et al., 2024), DAC (Doveh et al., 2024), TSVLC (Doveh et al., 2023). The experimental results are presented in Table 10, the numbers in the bracket are the gain compared with CLIP. From the results, there are two observations: (1) Considering the text-to-image and image-to-text retrieval, all methods introduce more improvement for the text-to-image retrieval. And the improvements brought by DAC and TSVLC are relatively minor. (2) The hard-negative mining methods considerably degrade performance on image-to-text retrieval task, with DAC and TSVLC exhibiting the most pronounced reductions, ranging from 11% to 28%. In contrast, our method incurs minimal negative impact and, compared to CLIP, delivers substantial improvements in both text-to-image and image-to-text retrieval tasks. These results underscore the effectiveness of our attribution-based objective relative to the hard-negative mining approaches.

It is worth noting that further hyperparameter tuning can enhance the performance of our method on downstream retrieval tasks. However, this improvement comes at the cost of a decline in compositional understanding performance. This suggests the need for a trade-off between compositional understanding capabilities and downstream retrieval performance.

H Analysis about the improvement on winoground

As Winoground is an extremely challenging and relatively small dataset, to better see the confidence of the improvement on it, we follow the Winoground paper to compute the confidence interval. Specifically, we divide the dataset into 4 groups, then obtained 4 scores for each model and score type, and used Student's t-distribution to compute the 95% confidence intervals. The results are shown in Table 11 (average scores and confidence intervals).

		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
SDS-CLIP		53.0	62.0	-	-	-	-	-	-	-	-	-	
CLIP-FT	mean±std	64.74±0.33	66.28±0.10	71.06±0.33	77.54±0.16	60.84±0.15	67.44±0.08	67.4±0.42	84.08±0.04	70.72±0.12	33.1±0.75	8.78±0.82	5.54±0.29
	confience interval	[64.33, 65.15]	[66.16, 66.40]	[70.65, 71.47]	[77.34, 77.74]	[60.65, 61.03]	[67.34, 67.54]	[66.88, 67.92]	[84.03, 84.13]	[70.57, 70.87]	[32.17, 34.03]	[7.76, 9.80]	[5.18, 5.90]
CLIP-CAE (Attention-Based)	mean±std	69.84±0.10	65.30±0.13	72.08±0.23	79.16±0.10	65.42±0.16	68.44±0.05	69.14±0.34	84.52±0.16	72.80±0.19	33.90±0.54	12.54±0.72	7.96±0.48
	confience interval	[69.72, 69.96]	[65.14, 65.46]	[71.79, 72.37]	[79.04, 79.28]	[65.22, 65.62]	[68.38, 68.50]	[68.72, 69.56]	[84.32, 84.72]	[72.56, 73.04]	[33.23, 34.57]	[11.65, 13.43]	[7.36, 8.56]
CLIP-CAE (GradCAM-Based)	mean±std	71.02±0.19	65.32±0.07	72.74±0.16	77.80±0.18	67.00±0.14	68.44±0.12	68.16±0.30	83.58±0.19	73.36±0.36	29.58±0.78	9.82±0.41	6.78±0.44
	confience interval	[70.78, 71.26]	[65.23, 65.41]	[72.54, 72.94]	[77.58, 78.02]	[66.83, 67.17]	[68.29, 68.59]	[67.79, 68.53]	[83.34, 83.82]	[72.91, 73.81]	[28.61, 30.55]	[9.31, 10.33]	[6.23, 7.33]
CLIP-CAE (Perturbation-Based)	mean±std	69.64±0.15	65.22±0.12	74.08±0.17	79.70±0.15	67.72±0.12	69.92±0.07	68.98±0.15	84.16±0.10	73.06±0.31	27.98±0.88	8.32±0.70	5.76±0.42
	confience interval	[69.45, 69.83]	[65.07, 65.37]	[73.87, 74.29]	[79.51, 79.89]	[67.57, 67.87]	[69.83, 70.01]	[68.79, 69.17]	[84.04, 84.28]	[72.68, 73.44]	[26.89, 29.07]	[7.45, 9.19]	[5.24, 6.28]
CLIP-CAE (Grad-Based)	mean±std	67.72±0.29	65.80±0.13	73.20±0.35	79.00±0.22	61.60±0.24	67.74±0.10	68.68±0.32	83.66±0.05	72.60±0.19	29.32±0.24	9.02±0.58	6.32±0.41
	confience interval	[67.36, 68.08]	[65.64, 65.96]	[72.77, 73.63]	[78.73, 79.27]	[61.48, 61.72]	[67.62, 67.86]	[68.28, 69.08]	[83.60, 83.72]	[72.36, 72.84]	[29.02, 29.62]	[8.30, 9.74]	[5.81, 6.83]

Table 7: Means, variances, and confidence intervals over five different random seeds.

		ARO		Sugar-Crepe		VL-Checklist		VALSE	SVO-Probes	ComVG	Winoground		
Model	Hyperparameter	Relation	Attribute	Relation	Attribute	Relation	Attribute	Relation	Relation	Relation	Text	Image	Group
CLIP-FT	$lr = 1 \times 10^{-7}$	67.9	65.9	69.5	76.9	63.1	66.8	66.3	81.7	70.5	33.9	8.8	7.0
	$\boldsymbol{lr = 5 \times 10^{-7}}$	**65.1**	**66.1**	**71.0**	**77.7**	**60.9**	**67.5**	**67.5**	**84.1**	**70.4**	**32.8**	**7.6**	**5.3**
	$lr = 1 \times 10^{-6}$	63.2	65.3	72.1	78.0	60.4	67.9	68.0	84.8	71.1	33.9	8.8	5.9
	$lr = 2 \times 10^{-6}$	61.6	64.9	72.7	78.2	59.8	68.3	67.0	85.2	71.4	31.0	9.9	6.4
	$lr = 5 \times 10^{-6}$	58.2	63.9	73.0	78.6	59.9	69.3	66.3	85.3	73	31.0	12.3	8.2
CLIP-CAE (Attention-Based) $lr = 5 \times 10^{-7}$	$\lambda = 0.5$	66.2	66.3	71.7	78.4	62.0	67.5	67.4	84.5	71.3	32.8	9.9	7.0
	$\lambda = 1.0$	67.4	66.3	71.8	78.5	62.3	67.6	67.9	84.9	71.6	33.3	10.5	7.6
	$\lambda = 10.0$	69.3	65.0	72.3	78.6	63.9	68.1	68.6	84.9	72.4	31.6	13.5	8.8
	$\boldsymbol{\lambda = 30.0}$	**70.0**	**65.4**	**72.3**	**79.0**	**65.0**	**68.3**	**69.7**	**84.7**	**72.7**	**35.7**	**13.5**	**8.2**
	$\lambda = 50.0$	69.7	65.3	72.0	79.2	65.4	68.4	69.1	84.5	72.5	33.9	13.5	9.2
CLIP-CAE (GradCAM-Based) $lr = 5 \times 10^{-7}$	$\lambda = 1.0$	67.4	65.8	71.9	77.7	61.6	67.4	68.0	84.7	71.9	32.8	8.8	7.0
	$\lambda = 10.0$	67.8	64.4	72.9	77.6	63.0	68.2	68.1	85.0	73.5	30.4	8.2	5.3
	$\lambda = 20.0$	68.9	64.6	71.6	77.7	64.1	68.5	69.4	84.6	73.5	31.6	8.8	4.7
	$\boldsymbol{\lambda = 50.0}$	**70.1**	**64.5**	**72.2**	**78.0**	**66.0**	**68.6**	**69.9**	**84.2**	**73.1**	**32.8**	**10.5**	**7.0**
	$\lambda = 100.0$	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) $lr = 5 \times 10^{-7}$	$\lambda = 0.05$	65.7	66.5	72.2	78.1	62.9	68.4	67.6	84.7	72.1	31.6	5.9	4.1
	$\lambda = 0.1$	66.7	66.3	72.8	78.2	65.5	68.9	68.8	84.6	72.2	31.0	8.2	4.7
	$\boldsymbol{\lambda = 0.2}$	**69.8**	**65.3**	**74.3**	**79.7**	**67.8**	**69.8**	**68.9**	**84.0**	**73.2**	**28.7**	**8.2**	**5.3**
	$\lambda = 0.5$	70.0	63.1	73.2	81.9	68.4	70.0	68.9	81.7	72.3	30.4	9.4	6.4
	$\lambda = 1.0$	71.4	62.1	72.9	82.2	70.8	69.9	66.1	80.0	71.0	25.2	7.6	6.4
CLIP-CAE (Gradient-Based) $lr = 5 \times 10^{-7}$	$\lambda = 1.0$	67.0	66.6	72.1	78.6	61.0	67.5	68.3	84.3	71.4	30.4	6.4	4.1
	$\lambda = 5.0$	67.6	66.3	71.4	78.7	60.8	67.5	68.7	83.8	72.5	29.8	7.0	5.3
	$\boldsymbol{\lambda = 10.0}$	**68.1**	**65.8**	**73.7**	**79.0**	**61.7**	**67.9**	**69.2**	**83.6**	**72.4**	**29.8**	**8.8**	**5.9**
	$\lambda = 20.0$	67.3	65.3	74.0	79.7	62.7	67.9	69.2	83.6	72.4	28.7	8.8	5.3
	$\lambda = 50.0$	66.9	65.0	75.8	80.0	64.2	67.7	68.1	83.7	71.6	28.1	9.9	5.9

Table 8: Hyper-parameter search.

	ARO		Sugar-Crepe		VL-Checklist		VALSE	SVO-Probes	ComVG
Model	Relation	Attribute	Relation	Attribute	Relation	Attribute	Relation	Relation	Relation
NegCLIP	80.5	71.4	73.3	79.8	71.9	70.0	75.5	83.7	70.2
+ CAE	<u>80.0</u>	<u>68.9</u>	**74.8**	**80.8**	**74.0**	**70.1**	**74.3**	**84.1**	**73.0**
CE-CLIP	82.4	72.7	72.3	79.8	75.2	69.4	74.8	82.7	69.1
+ CAE	<u>82.0</u>	<u>72.5</u>	**74.3**	**82.9**	**79.6**	**71.5**	**75.1**	**84.3**	**73.7**

Table 9: Additional experiment results when combined with two representative methods using hard-negative mining (CE-CLIP and NegCLIP). Note that for CE-CLIP, the hard-negative data for training is the same as NegCLIP.

Model	IR-R@1	IR-R@5	TR-R@1	TR-R@5
CLIP	30.2	55.7	50.1	74.9
CE-CLIP	42.2 (+12.0%)	69.4 (+13.7%)	47.1 (-3%)	74.6 (-0.3%)
DAC	31.4 (+1.2%)	57.0 (+1.3%)	23.9 (-26.2%)	47.0 (-27.9%)
TSVLC	33.3 (+3.1%)	58.2 (+2.5%)	37.2 (-12.9%)	64.3 (-10.6%)
CLIP-CAE	**38.3 (+8.1%)**	**65.2 (+9.5%)**	**54.9 (+4.8%)**	**78.7 (+3.8%)**

Table 10: Comparision with three state-of-the-art hard-negative mining methods on MSCOCO retrieval tasks.

Model	Winoground		
	Text Score	Image Score	Group Score
NegCLIP	29.2 [21.88,36.67]	8.8 [0.35,17.35]	5.3 [0.00,12.53]
CE-CLIP	22.8 [10.18,26.12]	14.6 [0.00,23.54]	8.2 [0.00,7.59]
DAC	20.5 [12.00,29.05]	6.4 [2.99,9.91]	1.8 [0.00,5.32]
TSVLC	28.1 [15.41,40.79]	8.8 [0.35,17.35]	5.3 [0.49,10.11]
CLIP-CAE	**33.9** [17.15,50.85]	**13.5** [7.71,19.24]	**8.2** [0.00,17.09]

Table 11: Means and confidence intervals on Winoground are reported, we follow the Winoground paper to compute the confidence interval.

From the results, we can observe that our method indeed shows a non-trivial improvement over other methods. We suspect that part of the reason for Winoground's performance being lower than random chance stems from its extremely strict metric calculation method. This method requires that for a given sample (which includes two image-text pairs, *i.e.*, image_0, text_0, and image_1, text_1), the model must simultaneously satisfy Similarity(image_0, text_0) > Similarity(image_0, text_1) and Similarity(image_1, text_1) > Similarity(image_1, text_0) for the text score of the sample to be 1. The same applies to the image score. For the group score to be 1, it requires that both the text score and image score be 1. Then we simplify the original metric calculation method to a more common one, where for every image-text pair, if Similarity(image_0, text_0) > Similarity(image_0, text_1) is satisfied, the text score is 1, and similarly for the image score. The calculation method for the group score remains unchanged. We then re-evaluate different models using this new metric method, and the results are shown in Table 12. In the context of this common computation method, all models perform better than random chance (e.g., 50% for image score*), and our method remains the best, outperforming others by $4.7\% \sim 10.0\%$ (group score).

I More details about the perturbation method

In the perturbation-based method, we use the sets of objects, relations, and attributes from the Visual Genome dataset (Krishna et al., 2017) as the candidate sets for perturbing the concepts in the captions. For the object and attribute candidate sets, we directly use their original sets, which respectively contain 150 objects and 200 attributes. For the relation candidate set, we eliminate some insignificant prepositions, such as "of", "for", and "with", and remove redundant relation terms with similar meanings, like "on" and "above". Finally, we retain 36 relations that represent distinct meanings.

For each concept to be perturbed, we replace it with another concept from the corresponding candidate set (for instance, "standing" –> "riding"). This ensures that the perturbed sentence continues to preserve grammaticality and remains plausible. Despite it is possible that perturbations may replace words with their synonyms. However, due to the pre-processing steps, the number of words with identical meanings in our candidate set is minimal, and since substitutions are made randomly, the probability of replacing a word with one of the same meaning is pretty low.

Model	Winoground		
	Text Score*	Image Score*	Group Score*
NegCLIP	59.6	53.5	34.5
CE-CLIP	56.1	53.8	33.0
DAC	56.1	52.9	29.2
TSVLC	58.5	53.8	32.2
CLIP-CAE	**62.6**	**55.6**	**39.2**

Table 12: The performance of different models when using the simplified evaluation metric.

J Category of existing methods

To better clarify the distinctions and connections between existing methods and our work, we have categorized and summarized the relevant representative approaches in Table 13. Further discussion details are provided in the Section 5 of the main text.

K More results on Winoground and ColorSwap

Winoground is an extremely valuable work that allows researchers to gain a deeper understanding of the shortcomings and limitations of VLMs. We further investigate how other representative hard-negative mining methods perform on Winoground. We select four representative state-of-the-art methods for evaluation on Winoground. The results are

Data-driven		Design new loss function	
NegCLIP (Yuksekgonul et al., 2022)	pioneer work of hard-negative mining method	**SDS-CLIP** (Basu et al., 2023)	designing a loss to distill knowledge from text-to-image diffusion model
TSVLC (Doveh et al., 2023)	construct text hard-negative samples	**CLIP-CAE (Ours)**	a framework to enhance the attribution of compositional concepts
syn-CLIP (Cascante-Bonilla et al., 2023)	using game engine to generate synthetic hard-negative samples	-	
DAC (Doveh et al., 2024)	enhance the caption quality and diversity, also including hard-negative samples	-	
CE-CLIP (Zhang et al., 2024)	generate text hard-negative samples more accurately	-	

Table 13: Comparison of different methods in data-driven and loss function design approaches.

Model	Winoground			ColorSwap
	Text score	Image score	Group score	
CLIP	31.6	11.1	9.4	11.7
CLIP-FT	33.9	8.2	5.3	11.0
CE-CLIP	22.8	14.6	8.2	11.0
NegCLIP	29.2	8.8	5.3	10.3
DAC	20.5	6.4	1.8	9.7
TSVLC	28.1	8.8	5.3	7.7
CLIP-CAE	**33.9**	**13.5**	**8.2**	**14.7**

Table 14: Comparison with four state-of-the-art hard-negative mining methods on Winoground and ColorSwap.

shown in the Table 14. Surprisingly, these methods generally performed poorly, with our CLIP-CAE achieving the best overall performance. The reason why our method shows relatively less improvement on Winoground are twofold: on the one hand, Winoground is indeed difficult (Diwan et al., 2022), and on the other hand, evaluation on Winoground may also present an out-of-distribution challenge. We also conduct additional experiments on the recently proposed ColorSwap (Burapacheep et al., 2024), and we report the most difficult group score in Table 14. From the results, CLIP-CAE consistently surpass the CLIP-FT and four state-of-the-art hard-negative mining methods with superior performance on the overall performance of image and text understanding.

L Performance on CLIP-Binding

We conduct experiments on the most difficult dataset (probing complex color binding) in CLIP-Binding (Lewis et al., 2024). The experimental results are shown in Table 15. The results demonstrate that in a zero-shot setting, CLIP-CAE is better than CLIP-FT. This subtle improvement may be because the images used in CLIP-Binding are all

Model	Two-object (color binding)
CLIP	27.58
CLIP-FT	27.18
CLIP-CAE (Attention-Based)	**27.32**
CLIP-CAE (GradCAM-Based))	**28.41**
CLIP-CAE (Perturbation-Based)	**28.65**
CLIP-CAE (Gradient-Based))	**28.08**

Table 15: Experiment results on CLIP-Binding.

from the CLEVER (Johnson et al., 2017) dataset, which has a significant domain gap compared to natural images and their captions. We also observed that the performance of all models on the clip-binding test set is below 30%, indicating that more effort is needed in the future to improve VLMs' abilities of performing complex compositional concept binding. Integrating syntactic structure information from text or relative positional information of objects in images into the model explicitly might be a future direction to consider.