Visual Perturbation and Adaptive Hard Negative Contrastive Learning for Compositional Reasoning in Vision-Language Models

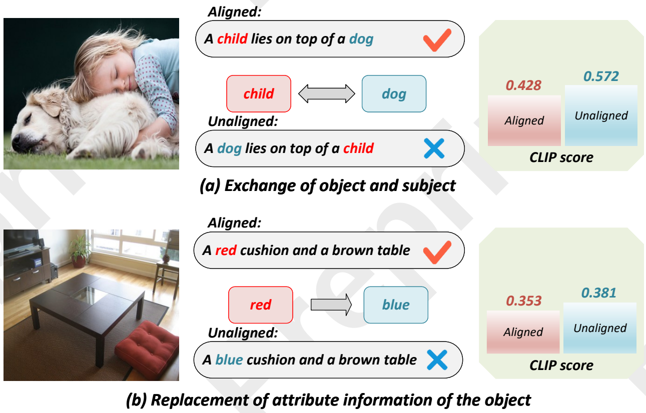
Xin Huang[1,3] , **Ruibin Li**[1,3] , **Tong Jia**[2] , **Wei Zheng**[1,3] , **Ya Wang**[1,3*]
[1]School of Artificial Intelligence and Software Engineering, Nanyang Normal University, Henan, China
[2]Institute for Artificial Intelligence, Peking University, Beijing, China
[3]Collaborative Innovation Center of Intelligent Explosion-proof Equipment, Henan, China
huangxin@nynu.edu.cn, liruibin199810@nynu.edu.cn, jia.tong@pku.edu.cn,
zhengwei821@nynu.edu.cn, wangya@nynu.edu.cn

Abstract

Vision-Language Models (VLMs) are essential for multimodal tasks, especially compositional reasoning (CR) tasks, which require distinguishing fine-grained semantic differences between visual and textual embeddings. However, existing methods primarily fine-tune the model by generating text-based hard negative samples, neglecting the importance of image-based negative samples, which results in insufficient training of the image encoder and ultimately impacts the overall performance of the model. Moreover, negative samples are typically treated uniformly, without considering their difficulty levels, and the alignment of positive samples is insufficient, which leads to challenges in aligning difficult sample pairs. To address these issues, we propose Adaptive Hard Negative Perturbation Learning (AHNPL). AHNPL translates text-based hard negatives into the visual domain to generate semantically disturbed image-based negatives for training the model, thereby enhancing its overall performance. AHNPL also introduces a contrastive learning approach using a multimodal hard negative loss to improve the model's discrimination of hard negatives within each modality and a dynamic margin loss that adjusts the contrastive margin according to sample difficulty to enhance the distinction of challenging sample pairs. Experiments on three public datasets demonstrate that our method effectively boosts VLMs' performance on complex CR tasks. The source code is available at https://github.com/nynu-BDAI/AHNPL.

1 Introduction

In recent years, Vision-Language Models (VLMs) have achieved significant zero-shot performance advancements in multimodal tasks such as retrieval [Huang *et al.*, 2017; Huang *et al.*, 2025], classification [Metzen *et al.*, 2023; Novack *et al.*, 2023] and segmentation [Xu *et al.*, 2022; Lüddecke and Ecker, 2022], establishing themselves as a foundation for multimodal tasks. However, despite their impressive performance, VLMs still face notable challenges in handling compositional reasoning (CR) tasks [Doveh *et al.*, 2023a; Thrush *et al.*, 2022]. CR tasks demand that models accurately understand and reason about fine-grained relationships among objects, attributes, and actions in images or texts. For instance, they must be able to distinguish sentences differing only in word order or accurately identify objects in images that have similar colors but different attributes. Such tasks go beyond the need for coarse-grained semantic understanding, requiring precise reasoning in complex scenarios.



(a) Exchange of object and subject

(b) Replacement of attribute information of the object

Figure 1: CLIP scores of the image and corresponding descriptions. The "Aligned" descriptions correctly reflect the image content, while the "Unaligned" descriptions represent incorrect descriptions of the image.

Yuksekgonul et al. [Yuksekgonul *et al.*, 2022] highlight the limitations of existing VLMs in CR tasks. They found that current state-of-the-art VLMs often perform poorly in distinguishing semantic differences caused by subtle changes in word order or attribute variations. As demonstrated in Fig. 1, the CLIP scores show that, in some cases, the unaligned descriptions receive higher scores, indicating that VLMs still struggle to distinguish fine-grained semantic differences in object relationships and attributes. To address this issue, researchers attempted to enhance models' ability to discern complex semantic relationships by generating hard negative samples. These hard negative samples are semantically similar to the positive samples but have subtle semantic differences. By generating these challenging samples to fine-tune the model, its performance on CR tasks is improved.

However, current methods for incorporating hard negative samples into contrastive learning face several issues. First, some methods focus solely on hard negatives in the text modality, overlooking equally important hard negatives in the visual modality [Huang *et al.*, 2024], resulting in insufficient training of the image encoder and thereby affecting the overall model performance. Second, these methods typically treat all negative samples equally, without considering their varying difficulty levels [Zhang *et al.*, 2024c; Zhang *et al.*, 2024a]. This prevents the model from effectively capturing subtle differences in hard negatives during training, thus hindering its ability to learn fine-grained distinctions. Additionally, these methods often neglect the adjustment of alignment for positive samples, which may cause the model to align only simple positive pairs while struggling to align more challenging ones.

To address these issues, we propose Adaptive Hard Negative Perturbation Learning (AHNPL) to enhance VLMs' performance in CR tasks. Our AHNPL method explicitly maps the subtle semantic variations of text-based hard negatives into the visual space, generating corresponding image embeddings for these negative texts, thereby improving the model's ability to discern subtle differences in the visual modality, which ultimately enhances the overall performance of the model. Moreover, AHNPL introduces dynamic hard negative contrastive learning, comprising a multimodal hard negative loss and a dynamic margin contrastive loss. The multimodal hard negative loss reduces the similarity between negative and positive samples across both text and visual modalities, improving the model's ability to distinguish hard negatives. The dynamic margin contrastive loss adjusts the contrastive margin based on sample complexity, enabling the model to focus more on challenging negatives and positives, which improves the alignment quality of challenging sample pairs.

Overall, our contributions are as follows:

- We propose a novel method that generates image-based hard negative samples by mapping subtle semantic shifts from text-based negatives into the visual domain, fine-tuning the model to enhance its overall performance.
- We introduce a dynamic hard negative contrastive learning approach. This includes a multimodal hard negative loss to distinguish hard negatives across text and image, and a dynamic margin loss that adapts to sample difficulty, focusing on challenging examples.

2 Related Work

Contrastive Learning. The objective of contrastive representation learning is to learn representations that are close to each other for similar samples and distant from each other for dissimilar samples. As one of the representative examples, CLIP [Radford *et al.*, 2021] has become a milestone in this field. CLIP is a Transformer-based [Vaswani *et al.*, 2017] model that consists of an image encoder and a text encoder, which are trained simultaneously. The objective is to maximize the cosine similarity of the image and text embeddings from the correct image-text pairs and to minimize the similarity between the incorrect pairs. A batch of N training samples (i.e., matching image-text pairs) results in a similarity matrix for each image-text combination. The main diagonal indicates the correct pair matches; the remaining entries correspond to negative entries. The InfoNCE loss is applied to the N × N similarity score matrix, and through contrastive learning, the image and text encoders are more effectively aligned across the two modalities, thus enhancing the performance of VLMs in multimodal tasks.

Compositional Reasoning. CR in vision and language tasks evaluates a model's ability to understand and manipulate complex ideas by breaking them into simpler components and recombining them in new ways. While VLMs excel at handling multimodal data, research shows they struggle with analyzing complex information, such as object relationships. Thrush et al. [Thrush *et al.*, 2022] first identified this issue, demonstrating that VLMs often fail to distinguish semantic differences caused by changes in word order, with performance sometimes no better than random guessing. This type of task, which tests a model's ability to understand fine-grained semantic structures, is referred to as CR. Many researchers have proposed methods of generating negative samples to fine-tune models to improve their performance in CR tasks. For example, Doveh et al. [Doveh *et al.*, 2023a] introduced DAC, which fine-tunes the model to improve its performance in CR tasks by using a Large Language Model (LLM) to generate texts of different scenes. Zhang et al. [Zhang *et al.*, 2024b] proposed CE-CLIP, which refines contrastive objectives with diverse negative samples to enhance semantic understanding. While effective, we believe current methods haven't fully explored hard negative mining, as they overlook the importance of image-based negative samples and fail to adjust the alignment process of positive sample pairs. To address this issue, we propose AHNPL, which translates text-based hard negatives into the visual domain, generating semantically disturbed image-based negatives, and dynamically adjusts the training process based on the similarity of each positive sample pair.

3 Methodology

In this section, we provide a detailed introduction to our proposed AHNPL method. Fig. 2 shows an overview of the entire pipeline.

3.1 Hard Negative Generation

In contrastive learning, hard negatives are samples that are semantically similar to positive samples but have subtle semantic differences. For example, consider the caption: "A boy wearing a red hat is playing on the beach". A potential hard negative could be: "A boy wearing a blue hat is playing on the beach". While this hard negative accurately describes most elements in the image, it differs from the positive sample in the color of the hat. Including such hard negatives in the training process can help the model recognize subtle differences, thereby improving overall accuracy and performance.

To generate these hard negatives, we use the natural language processing tool Spacy [Honnibal, 2017] to parse the captions and assign part-of-speech tags to each word. Specifically, we generate two types of hard negatives. The first type involves swapping two nouns in a sentence to generate hard

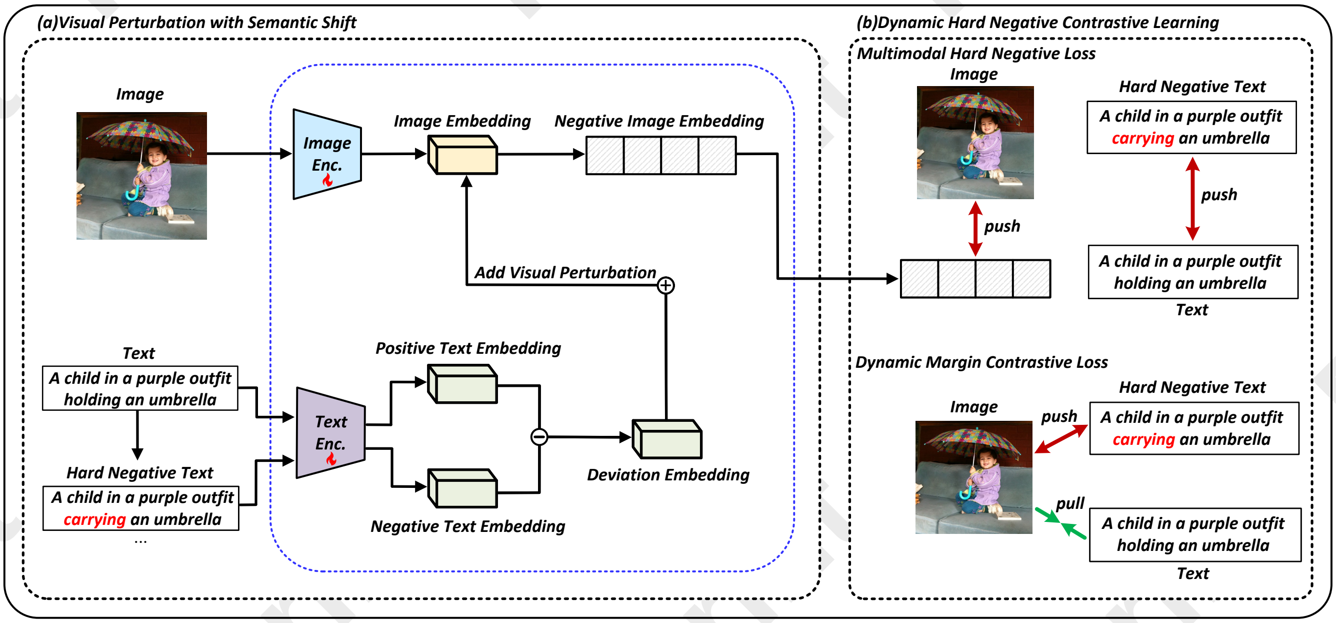

(a)Visual Perturbation with Semantic Shift
Image
Image Enc.
Image Embedding
Negative Image Embedding
Add Visual Perturbation
Text
A child in a purple outfit holding an umbrella
Hard Negative Text
A child in a purple outfit carrying an umbrella
...
Text Enc.
Positive Text Embedding
Negative Text Embedding
Deviation Embedding
(b)Dynamic Hard Negative Contrastive Learning
Multimodal Hard Negative Loss
Image
Hard Negative Text
A child in a purple outfit carrying an umbrella
push
push
A child in a purple outfit holding an umbrella
Text
Dynamic Margin Contrastive Loss
Image
Hard Negative Text
push
A child in a purple outfit carrying an umbrella
pull
A child in a purple outfit holding an umbrella
Text

Figure 2: The proposed Adaptive Hard Negative Perturbation Learning (AHNPL). (a) Performing visual perturbation by computing the semantic shift in text to generate a deviation embedding. (b) The dynamic hard negative contrastive learning method makes targeted adjustments to the strategies for alignment of positive samples and the distinction of hard negative samples.

negatives. This aims to alter the relationships between entities, and thus train the model to learn how to distinguish between different entity relationships. The second type involves randomly masking words in the sentence based on their part-of-speech tags (such as nouns, verbs, or adjectives), and then using the RoBERTa [Liu *et al.*, 2019] model to predict and fill in the masked parts. This approach aims to generate hard negatives that are similar in context and sentence structure, which helps the model enhance its sensitivity to semantic changes and lexical variations.

3.2 Visual Perturbation with Semantic Shift

In the previous section, we identify keywords in the captions through part-of-speech parsing and construct textual negative samples. Furthermore, we design a method for generating visual negative samples to train the model's image encoder, thereby enhancing the model's overall performance.

Existing negative sample generation methods often overly focus on textual negative samples, neglecting the role of visual negative samples, which leads to insufficient training of the image encoder and subsequently affects the overall model performance. To solve this problem, we propose a visual perturbation with semantic shift method for generating image negative samples. Specifically, we first capture semantic changes by computing the deviation embedding between the original text and the negative text. Given the original text T_{orig} and the negative text T_{neg}, we use a pretrained CLIP model to extract their high-dimensional embedding vectors $e_{T_{orig}}$ and $e_{T_{neg}}$, respectively. These vectors represent the semantic information of the texts within the multimodal embedding space of the CLIP model. The deviation embedding Δ_e is computed as follows:

$$\Delta_e = e_{T_{neg}} - e_{T_{orig}} \tag{1}$$

where Δ_e reflects the semantic shift from the positive caption to the negative caption in the text. This shift not only represents changes in specific words or phrases within the text, but also implies adjustments in the contextual semantic structure. By utilizing this deviation embedding, we can capture the subtle semantic changes in the text.

Next, we directly incorporate the generated deviation embedding Δ_e into the image embedding space by adding it to the original image's embedding vector $e_{I_{orig}}$ to create the embedding representation $e_{I_{neg}}$ for the image negative sample:

$$e_{I_{neg}} = e_{I_{orig}} + \Delta_e \tag{2}$$

This approach ensures that the generated negative samples maintain appropriate semantic relevance to the original image in the embedding space, while deviating from the original image in terms of semantics. The image negative samples generated by this method maintain semantic consistency with the textual negative samples, providing more challenging training examples for the model.

3.3 Dynamic Hard Negative Contrastive Learning

Existing contrastive learning methods have some obvious limitations. First, these methods fail to adapt their alignment strategy according to the varying difficulties of positive samples, specifically by failing to adaptively enhance the learning signal for hard positive pairs that exhibit lower similarity, which causes the model to overly focus on aligning simple positive samples while lacking the capability to handle more

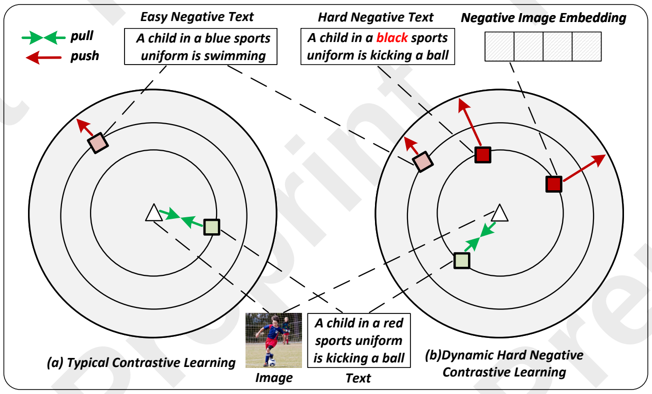

pull
push
Easy Negative Text
A child in a blue sports uniform is swimming
Hard Negative Text
A child in a black sports uniform is kicking a ball
Negative Image Embedding
A child in a red sports uniform is kicking a ball
Image
Text
(a) Typical Contrastive Learning
(b)Dynamic Hard Negative Contrastive Learning

Figure 3: (a) The typical contrastive learning framework. (b) The proposed Dynamic Hard Negative Contrastive Learning, in which adjustments are specifically applied to the strategies for the distinction of hard negative samples and the alignment of positive samples.

challenging positive pairs. Additionally, these methods typically not only treat all negative samples as equivalent, disregarding the difficulty levels of negative samples during training, but also fail to effectively leverage multimodal information to distinguish the fine-grained semantic differences between hard negative samples and their corresponding positive samples. This approach makes it difficult for the model to establish clear decision boundaries and leads to misjudgments when processing hard negative samples that are semantically highly similar to positive samples, thereby affecting the effectiveness of contrastive learning.

To address these limitations, we propose dynamic hard negative contrastive learning (illustrated in Fig. 3), which dynamically adjusts its learning strategy based on sample difficulty, specifically enhancing the model's ability to distinguish difficult samples, and also establishes crucial semantic differences between hard negative samples and their corresponding positive samples in both visual and textual modalities, further improving the model's capacity to discern fine-grained semantic distinctions between them.

Contrastive Loss Our method uses a contrastive loss, which is applied to a text and image pair (T, I) as input and consists of two components: (i) an image encoder $e_I = f_v(I)$; (ii) a text encoder $e_T = f_t(T)$. In this setting, the text-to-image similarity score is computed as:

$$S(T, I) = \frac{e_T^T e_I}{\|e_T\| \|e_I\|} \tag{3}$$

As with most contemporary VLMs, we employ the contrastive loss as one of our losses for each batch:

$$\begin{aligned} L_{cont} = -\sum_i \Bigg(\log \left(\frac{\exp(S(T_i, I_i)/\tau)}{\sum_j \exp(S(T_i, I_j)/\tau)} \right) \\ + \log \left(\frac{\exp(S(T_i, I_i)/\tau)}{\sum_k \exp(S(T_k, I_i)/\tau)} \right) \Bigg) \end{aligned} \tag{4}$$

where τ is a temperature parameter.

Multimodal Hard Negative Loss. To enhance the model's ability to distinguish negative samples, we specially introduce a negative sample loss term. By simultaneously handling negative samples from both visual and textual perspectives, we help the model more accurately distinguish these challenging negative samples, thereby improving the effectiveness of contrastive learning.

The visual negative loss is designed to enhance the similarity differences between the visual negative embeddings I_n and the original image I. Specifically, it is computed as:

$$L_{neg}^{visual} = \sum_{(I,T) \in B} -\log \left(\frac{1}{\sum_{I_n \in I_{hs}} \exp(S(I, I_n))} \right) \tag{5}$$

where I_{hs} is the set of all such generated visual hard negative embeddings I_n for an original image I, and $S(I, I_n)$ represents the similarity between the original image I and each visual hard negative I_n in the set I_{hs}. By addressing the subtle semantic differences between hard visual negatives, this loss encourages the model to better distinguish visually similar negative samples.

Next, the textual negative loss is defined. This loss enhances the similarity differences between the original text T and the hard negative texts T_n. The textual negative loss is given as follows:

$$L_{neg}^{textual} = \sum_{(I,T) \in B} -\log \left(\frac{1}{\sum_{T_n \in T_{hs}} \exp(S(T, T_n))} \right) \tag{6}$$

where T_{hs} is the set of all types of such textual hard negative samples T_n generated for an original text T, and $S(T, T_n)$ denotes the similarity between the original text T and each textual hard negative T_n. This loss encourages the model to distinguish the original text from semantically similar hard negative texts by minimizing their similarity.

Finally, the total negative loss is obtained by combining the visual and textual negative losses:

$$L_{neg} = L_{neg}^{visual} + L_{neg}^{textual} \tag{7}$$

By integrating the losses for both textual and visual negative samples, this approach aims to comprehensively reduce the similarity between negative and positive samples across both modalities, thereby enhancing the model's ability to distinguish negative samples. This not only improves the model's performance in cross-modal alignment tasks but also boosts its robustness by focusing on difficult-to-distinguish negative samples in both image and text domains.

Dynamic Margin Contrastive Loss. In contrastive learning, the alignment of positive samples and the distinction of negative samples are key to improving model performance. Existing methods typically use a fixed margin (Margin Loss) threshold to adjust the similarity between positive and negative samples, but they still perform poorly when handling samples of varying difficulty.

When it comes to positive samples, such fixed margin approaches actually hold the assumption that all samples have the same level of difficulty. As a result, when handling challenging samples, the model tends to treat them the same as

Model	**#Params**	**ARO**			**VALSE**									
		Relation	**Attribute**	**Avg**	**Existence**	**Plurality**	**Counting**	**Sp.rel.**	**Actions**		**Coreference**		**Foil-it!**	**Avg**
					quantifiers	number		relations	repl.	actant swap	standard	clean		
BLIP [Li *et al.*, 2022]	583M	59.0	88.0	73.5	86.3	73.2	68.1	71.5	77.2	61.1	53.8	48.2	93.8	70.0
BEIT3 [Wang *et al.*, 2023]	1.9B	60.6	74.6	67.6	77.4	74.6	68.8	74.0	86.7	65.2	50.0	44.2	96.0	70.4
BLIP2 [Li *et al.*, 2023]	3.4B	41.2	71.3	56.3	55.5	71.5	66.0	62.4	83.6	51.6	48.6	51.9	95.9	65.4
MiniGPT-4 [Zhu *et al.*, 2023]	>9B	46.9	55.7	51.3	65.5	72.5	67.4	68.4	83.2	58.8	52.6	51.0	95.8	68.4
Scene Graph relied method														
syn-CLIP [Cascante-Bonilla *et al.*, 2023]	151M	71.4	66.9	69.2	-	-	-	-	-	-	-	-	-	-
Segmentation & LLM relied method														
DAC-LLM [Doveh *et al.*, 2023a]	151M	81.3	73.9	77.6	-	-	-	-	-	-	-	-	-	-
DAC-SAM [Doveh *et al.*, 2023a]	151M	77.2	70.5	73.9	-	-	-	-	-	-	-	-	-	-
Hard Negative based method														
CLIP [Radford *et al.*, 2021]	151M	59.3	62.9	61.1	68.7	57.1	61.0	65.4	77.8	71.8	54.1	51.0	89.8	65.3
CyCLIP [Goel *et al.*, 2022]	151M	59.1	65.4	62.3	69.3	58.3	61.0	66.4	78.1	72.0	53.2	51.6	88.8	65.5
SDS-CLIP [Basu *et al.*, 2024]	151M	53.0	62.0	57.5	-	-	-	-	-	-	-	-	-	-
NegCLIP [Yuksekgonul *et al.*, 2022]	151M	80.2	70.5	75.4	76.8	71.7	65.0	72.9	81.6	84.7	58.6	53.8	91.9	71.6
CLIP-SVLC [Doveh *et al.*, 2023b]	151M	80.6	73.0	76.8	-	-	-	-	-	-	-	-	-	-
CE-CLIP [Zhang *et al.*, 2024b]	151M	83.0	76.4	79.7	78.6	77.7	64.4	74.4	81.2	88.6	54.7	54.8	93.7	72.5
Ours	151M	**83.8**	**77.0**	**80.4**	**84.0**	**78.7**	**64.9**	**76.3**	**81.9**	**88.1**	**56.0**	**58.6**	**94.4**	**75.9**

Table 1: Results (%) on **ARO** and **VALSE**. Our proposed AHNPL achieves new state-of-the-art (SOTA) results on the VALSE benchmark.

simple samples, making it difficult to capture the subtle semantic relationships, ultimately reducing the model's ability to align positive samples effectively. For negative samples, these methods generally introduce them into the contrastive learning framework without considering the variations in their difficulty. Although hard negatives are semantically similar to positive samples, they are characterized by subtle semantic differences. This subtle semantic difference makes it challenging for the model to effectively distinguish between positive samples and hard negatives, thereby limiting the model's performance.

To effectively enhance the model's capability in aligning hard positive samples, we introduce a learnable parameter a, which is initialized as a random value from a standard normal distribution. We set a lower bound of 0.2 for a to ensure it maintains a positive value throughout the entire training process. When the similarity between positive samples is low or there are significant semantic differences, the model automatically increases its focus on these challenging positive pairs, thus improving alignment accuracy.

Specifically, we compare the similarity $S(I, T)$ of the positive sample pairs with the learnable parameter a to adjust the alignment strategy for positive samples:

$$L^{+}_{mar} = \sum_{(I,T)\in B} \max(0, a - S(I, T)) \tag{8}$$

This loss function ensures that positive samples with greater difficulty receive more attention during training, enabling the model to learn more effectively from challenging positive samples, thereby improving its overall performance in aligning positive samples.

To enable the model to better distinguish difficult negative samples, we introduce an adaptive threshold updating mechanism in margin loss based on the difficulty of the samples. The specific loss function is defined as follows:

$$L^{-}_{mar} = \sum_{(I,T)\in B} \sum_{T_n \in T_{hs}} \max(0, S(I, T_n) - S(I, T) + M^{t}_{n}) \tag{9}$$

where M^t_n is the adaptive threshold computed for each textual hard negative T_n in the training step t. This threshold is computed as follows:

$$M^{t}_{n} = \frac{1}{|B|} \sum_{(I,T)\in B} \left(S^{t-1}(I, T) - S^{t-1}(I, T_n)\right) \tag{10}$$

where B represents the set of samples in the current training batch, $S^{t-1}(I, T)$ and $S^{t-1}(I, T_n)$ represent the similarity between image I and the positive text T, as well as the similarity between the image I and a textual hard negative T_n, respectively, in the previous training step.

The final margin loss combining both the negative margin loss and the positive margin loss is defined as:

$$L_{mar} = L^{+}_{mar} + L^{-}_{mar} \tag{11}$$

This adaptive threshold updating strategy allows the model to dynamically adjust its learning strategy during the training process, ensuring effective alignment of positive samples and proper distinction of negative samples.

The final loss function can be expressed as follows:

$$L_{total} = L_{cont} + L_{neg} + L_{mar} \tag{12}$$

4 Experiments

4.1 Experimental Settings

All experiments are conducted using the PyTorch framework on a single NVIDIA A40 GPU and an Intel® Xeon® Gold 6330 CPU, running on Ubuntu 22.10. During the training phase, we initialize a pretrained CLIP model and fine-tune it on the MSCOCO [Lin *et al.*, 2014] dataset for 10 epochs with a batch size of 128. The learning rate is set to 2×10^{-5}, and weight decay is set to 0.1.

4.2 Dataset

Training Datasets. The CLIP model is originally pretrained on 400 million image-text pairs sourced from the web, and we continue to fine-tune the model based on this pretraining. In our experiments, we use the widely adopted

Model	**REPLACE**				**SWAP**			**ADD**		
	Obj.	Att.	Rel.	Avg.	Obj.	Att.	Avg.	Obj.	Att.	Avg.
Human	100	99.0	97.0	98.7	99.0	100	99.5	99.0	99.0	99.0
Vera [Liu *et al.*, 2023]	49.4	49.6	49.1	49.4	49.4	49.2	49.3	49.4	49.6	49.5
Grammar [Morris *et al.*, 2020]	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0
BLIP2	-	-	-	86.7	-	-	69.8	-	-	86.5
CLIP	90.9	80.0	69.2	80.2	61.4	64.0	62.7	77.2	68.2	72.7
NegCLIP	92.7	85.9	76.5	85.0	75.2	75.4	75.3	88.8	82.8	85.8
CE-CLIP	93.1	88.8	79.0	87.0	72.8	77.0	74.9	92.4	93.4	92.9
Ours	**93.2**	**88.9**	**80.1**	**87.4**	**76.3**	**75.8**	**76.1**	**97.2**	**94.5**	**95.9**

Table 2: Results (%) on **SugarCrepe**. Vera and Grammar are text-only models.

cross-modal text-image retrieval dataset MSCOCO [Lin *et al.*, 2014]. MSCOCO is chosen for its rich annotations of diverse objects, attributes, and relationships, which enhance the model's semantic understanding in multimodal tasks.

Evaluation Dataset. We evaluate our method on several vision-language compositional benchmarks: ARO, VALSE [Parcalabescu *et al.*, 2022], and SugarCrepe [Hsieh *et al.*, 2023] (a bias-mitigated version of CREPE [Yu *et al.*, 2023]). Each test example in these datasets includes an image along with a corresponding correct description and a modified incorrect description. The model's task is to determine which description is correct.

- **ARO** [Yuksekgonul *et al.*, 2022]. This is a CR benchmark that includes positive or negative captions and evaluates sensitivity to word order. Word-order negative sentences are created by reordering words, which changes sentence semantics in attributes, relationships, and word order meaning.
- **VALSE** [Parcalabescu *et al.*, 2022]. This benchmark evaluates VLMs' understanding of linguistic phenomena. It includes six tests on structures: morphological syntax, verb-argument structure, and word order. These tests require models to accurately link visual elements to linguistic descriptions, assessing visual understanding and alignment of these phenomena.
- **SugarCrepe** [Hsieh *et al.*, 2023]. This benchmark uses LLMs to generate fluent and meaningful negative samples, avoiding biases introduced by traditionally used rule-based templates. It employs an adversarial improvement mechanism to minimize evaluation biases and ensures the reliability of results.

4.3 Overall Results

We present the evaluation results for the ARO and VALSE benchmarks in Tab. 1, and Tab. 2 shows the evaluation results on the SugarCrepe benchmark. As a widely used evaluation benchmark in this field, VALSE provides a solid basis for comparing our model against various mainstream models. Our AHNPL demonstrates effective improvements over all methods that utilize hard negatives, achieving excellent results across all benchmarks.

As shown in Tab. 1, our method achieves new state-of-the-art (SOTA) results on the VALSE dataset, with an average matching score of 75.9%, outperforming the current SOTA model CE-CLIP by 3.4%. Other models, such as NegCLIP, also perform well with a score of 71.6%, but our model still leads by 4.3%, particularly excelling in the quantifiers and clean subtasks, where it outperforms CE-CLIP by 5.4% and 3.8%, respectively. On the ARO benchmark, AHNPL outperforms CE-CLIP in the ARO-Relation task, indicating an improvement in relationship understanding. In the ARO-Attribute task, AHNPL slightly surpasses CE-CLIP, suggesting that our model is comparable to the current SOTA model in attribute understanding.

Model	*negatives*	*MHNL*	*DMCL*	**ARO-R**	**ARO-A**	**VALSE**	**SugarCrepe**	**Avg**
CLIP				59.3	62.9	67.0	73.0	65.6
	✓			81.6	72.0	74.2	80.3	77.0
	✓	✓		82.1	72.9	73.0	85.1	78.3
	✓		✓	80.1	71.9	72.7	87.5	78.1
Ours	✓	✓	✓	**83.8**	**77.0**	**75.9**	**93.6**	**82.3**

Table 3: Ablation of losses, where *negatives* represent image-text contrastive with additional hard negatives.

SugarCrepe is a debiased dataset that effectively avoids the biases introduced by traditional rule-based negative sample generation, making it a benchmark of relatively high difficulty. We choose only the best-performing and most representative models for evaluation on this benchmark. On the SugarCrepe benchmark, AHNPL achieves higher overall average scores than CE-CLIP, particularly in the SWAP and ADD tasks for the Object category, where it outperforms CE-CLIP by 3.5% and 4.8%, respectively. These results further demonstrate that our model has effective advantages in handling complex semantic relationships and reducing language bias.

4.4 Ablation Study

We conduct ablation experiments on multiple enhanced versions of the CLIP base model to evaluate the contribution and effectiveness of each component in our method. The specific impact of different loss functions on model performance is shown in Tab. 3. First, the results indicate that the introduction of hard negative samples effectively improves the model's performance, with an improvement of up to 11.4% compared to the zero-shot CLIP, highlighting their key role in contrastive learning. Furthermore, compared to merely introducing negative samples into contrastive learning, each individual loss function we introduce shows varying degrees of performance improvement across all benchmarks. When all the loss functions are combined, the model achieves optimal performance. This improvement is fully reflected in our final model, demonstrating the superiority of the method in extending and enhancing contrastive learning objectives.

4.5 Case Study

We present several case studies illustrating the performance of CLIP and AHNPL across four subtasks of the SugarCrepe benchmark, as shown in Tab. 4. SugarCrepe uses LLMs to generate fluent captions with common sense, thereby challenging VLMs to distinguish negative captions effectively. In the "Swap Object" task, where models must understand object relationships (e.g., "A painting of a vase with a sunflower on a table" vs. "A painting of a sunflower with a vase on a

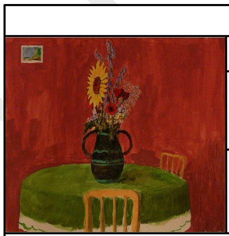
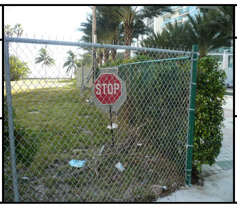
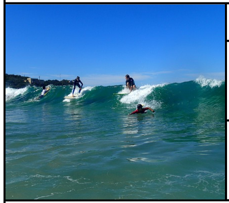
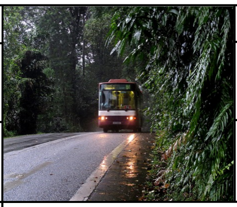
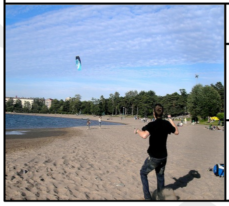
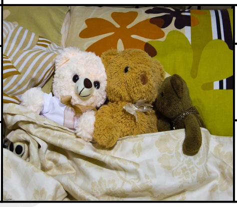
Swap Object							
		CLIP	ours			CLIP	ours
	Caption：A painting of a vase with a sunflower on a table.	0.262	0.305		*Caption：The stop sign is behind the fence instead of on the street.*	0.149	0.231
	Negative caption：A painting of a sunflower with a vase on a table.	0.255	0.292		*Negative caption：The fence is behind the stop sign instead of on the street.*	0.175	0.155
Swap Attribute							
		CLIP	ours			CLIP	ours
	Caption：Four surfers are trying to catch a wave as they stand.	0.172	0.232		*Caption：A city bus that is traveling down a wet country road.*	0.265	0.185
	Negative caption：Trying to catch four waves, surfers stand.	0.187	0.221		*Negative caption：A country bus that is traveling down a wet city road.*	0.280	0.157
Replace Relationship				***Replace Attribute***			
		CLIP	ours			CLIP	ours
	Caption：A man is flying a kite at the beach.	0.134	0.293		*Caption：Three teddy bears laying in bed under the covers.*	0.264	0.216
	Negative caption：A man is holding a kite at the beach.	0.153	0.149		*Negative caption：A single teddy bear laying in bed under the covers.*	0.275	0.142

Table 4: Predictions of CLIP and AHNPL on SugarCrepe tasks: Swap Object, Swap Attribute, Replace Relationship and Replace Attribute. The score represents the similarity score between the caption and the corresponding image as assessed by CLIP/AHNPL. The model selects the caption with the higher similarity score as the correct one.

table”), AHNPL outperforms CLIP. In the “Swap Attribute” task requiring accurate identification of object attributes (e.g., “Four surfers are trying to catch a wave as they stand” vs. “Trying to catch four waves, surfers stand”), CLIP struggles, whereas AHNPL consistently selects the correct caption. For the “Replace Relationship” and “Replace Attribute” tasks, which involve subtle distinctions (e.g., “A man is flying a kite at the beach” vs. “A man is holding a kite at the beach” or “Three teddy bears laying in bed under the covers” vs. “A single teddy bear laying in bed under the covers”), AHNPL effectively handles these nuances through negative caption contrastive learning, outperforming CLIP.

4.6 Visualization

To demonstrate the clear effectiveness of our proposed visual perturbation strategy, a compelling visualization study is provided. Tab. 5 shows two illustrative examples, where each example generates textual negative samples using the methods in Section 3.1. Specifically, for each example, we extract $e_{I_{orig}}$, $e_{T_{orig}}$ and $e_{T_{neg}}$ using zero-shot CLIP, AHNPL at epoch 5 and 10 (training end), respectively. The visual negative embedding $e_{I_{neg}}$ is then computed from these extracted embeddings via our proposed visual perturbation. We illustrate the cosine distances between key pairs of these four embeddings. The changes in cosine distances demonstrate that the visual perturbation strategy effectively pushes the original image features towards negative texts, improving feature distinguishability and thereby helping the model capture subtle positive-negative semantic differences.

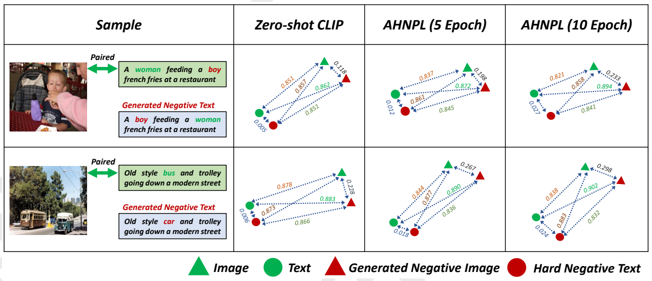


Table 5: Impact of AHNPL training on the embedding space distances of two illustrative examples. Numerical values represent cosine distances between key embedding pairs.

5 Conclusion

This paper proposes AHNPL, enhancing VLMs’ performance in CR tasks via Visual Perturbation with Semantic Shift and Dynamic Hard Negative Contrastive Learning. AHNPL transforms text-based hard negatives into semantically perturbed image negatives, effectively improving the overall performance. Additionally, Dynamic Hard Negative Contrastive Learning strengthens the model’s ability to align positive samples and distinguish hard negatives. Experimental results show that AHNPL outperforms existing methods in downstream tasks, showing robust complex semantic understanding in multimodal scenarios. Future work will explore integrating large-scale knowledge graphs and implicit knowledge to enhance complex semantic relationship acquisition.

Acknowledgments

This work was supported by Humanities and Social Sciences Youth Foundation, Ministry of Education of the People's Republic of China (No. 24YJCZH135), and Science and Technology Projects of Henan Province (Nos. 242102211019 & 252102210027).

References

[Basu *et al.*, 2024] Samyadeep Basu, Shell Xu Hu, Maziar Sanjabi, Daniela Massiceti, and Soheil Feizi. Distilling knowledge from text-to-image generative models improves visio-linguistic reasoning in clip. In *EMNLP*, pages 6105–6113, 2024.

[Cascante-Bonilla *et al.*, 2023] Paola Cascante-Bonilla, Khaled Shehada, James Seale Smith, Sivan Doveh, Donghyun Kim, Rameswar Panda, Gül Varol, Aude Oliva, Vicente Ordonez, Rogério Feris, and Leonid Karlinsky. Going beyond nouns with vision & language models using synthetic data. In *ICCV*, pages 20155–20165, 2023.

[Doveh *et al.*, 2023a] Sivan Doveh, Assaf Arbelle, Sivan Harary, Roei Herzig, Donghyun Kim, Paola Cascante-Bonilla, Amit Alfassy, Rameswar Panda, Raja Giryes, Rogerio Feris, et al. Dense and aligned captions (DAC) promote compositional reasoning in VL models. In *NeurIPS*, pages 76137–76150, 2023.

[Doveh *et al.*, 2023b] Sivan Doveh, Assaf Arbelle, Sivan Harary, Eli Schwartz, Roei Herzig, Raja Giryes, Rogério Feris, Rameswar Panda, Shimon Ullman, and Leonid Karlinsky. Teaching structured vision & language concepts to vision & language models. In *CVPR*, pages 2657–2668, 2023.

[Goel *et al.*, 2022] Shashank Goel, Hritik Bansal, Sumit Bhatia, Ryan Rossi, Vishwa Vinay, and Aditya Grover. CyCLIP: Cyclic contrastive language-image pretraining. In *NeurIPS*, volume 35, pages 6704–6719, 2022.

[Honnibal, 2017] Matthew Honnibal. spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. *To appear*, 2017.

[Hsieh *et al.*, 2023] Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. Sugarcrepe: fixing hackable benchmarks for vision-language compositionality. In *NeurIPS*, pages 31096–31116, 2023.

[Huang *et al.*, 2017] Xin Huang, Yuxin Peng, and Mingkuan Yuan. Cross-modal common representation learning by hybrid transfer network. In *IJCAI*, pages 1893–1900, 2017.

[Huang *et al.*, 2024] Yufeng Huang, Jiji Tang, Zhuo Chen, Rongsheng Zhang, Xinfeng Zhang, Weijie Chen, Zeng Zhao, Zhou Zhao, Tangjie Lv, Zhipeng Hu, et al. Structure-clip: Towards scene graph knowledge to enhance multi-modal structured representations. In *AAAI*, volume 38, pages 2417–2425, 2024.

[Huang *et al.*, 2025] Xin Huang, Shilong Wang, Tong Jia, Zhihang Gou, and Jingjing Li. Adaptive prompt-based semantic embedding with inspire potential of implicit knowledge for cross-modal retrieval. In *AAAI*, volume 39, pages 17485–17493, 2025.

[Li *et al.*, 2022] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pretraining for unified vision-language understanding and generation. In *ICML*, pages 12888–12900, 2022.

[Li *et al.*, 2023] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pretraining with frozen image encoders and large language models. In *ICML*, pages 19730–19742, 2023.

[Lin *et al.*, 2014] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755, 2014.

[Liu *et al.*, 2019] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[Liu *et al.*, 2023] Jiacheng Liu, Wenya Wang, Dianzhuo Wang, Noah A. Smith, Yejin Choi, and Hannaneh Hajishirzi. Vera: A general-purpose plausibility estimation model for commonsense statements. In *EMNLP*, pages 1264–1287, 2023.

[Lüddecke and Ecker, 2022] Timo Lüddecke and Alexander Ecker. Image segmentation using text and image prompts. In *CVPR*, pages 7086–7096, 2022.

[Metzen *et al.*, 2023] Jan Hendrik Metzen, Piyapat Saranrittichai, and Chaithanya Kumar Mummadi. Autoclip: Autotuning zero-shot classifiers for vision-language models. *arXiv preprint arXiv:2309.16414*, 2023.

[Morris *et al.*, 2020] John Morris, Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin, and Yanjun Qi. Textattack: A framework for adversarial attacks, data augmentation, and adversarial training in nlp. In *EMNLP*, pages 119–126, 2020.

[Novack *et al.*, 2023] Zachary Novack, Julian McAuley, Zachary Chase Lipton, and Saurabh Garg. Chils: Zero-shot image classification with hierarchical label sets. In *ICML*, pages 26342–26362, 2023.

[Parcalabescu *et al.*, 2022] Letitia Parcalabescu, Michele Cafagna, Lilitta Muradjan, Anette Frank, Iacer Calixto, and Albert Gatt. Valse: A task-independent benchmark for vision and language models centered on linguistic phenomena. In *ACL*, pages 8253–8280, 2022.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763, 2021.

[Thrush *et al.*, 2022] Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *CVPR*, pages 5238–5248, 2022.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, pages 5998–6008, 2017.

[Wang *et al.*, 2023] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. Image as a foreign language: Beit pretraining for vision and vision-language tasks. In *CVPR*, pages 19175–19186, 2023.

[Xu *et al.*, 2022] Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas Breuel, Jan Kautz, and Xiaolong Wang. Groupvit: Semantic segmentation emerges from text supervision. In *CVPR*, pages 18134–18144, 2022.

[Yu *et al.*, 2023] Xinyan Yu, Sewon Min, Luke Zettlemoyer, and Hannaneh Hajishirzi. CREPE: open-domain question answering with false presuppositions. In *ACL*, pages 10457–10480, 2023.

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

[Zhang *et al.*, 2024a] Jianrui Zhang, Mu Cai, Tengyang Xie, and Yong Jae Lee. Countercurate: Enhancing physical and semantic visio-linguistic compositional reasoning via counterfactual examples. In *ACL*, pages 15481–15495, 2024.

[Zhang *et al.*, 2024b] Le Zhang, Rabiul Awal, and Aishwarya Agrawal. Contrasting intra-modal and ranking cross-modal hard negatives to enhance visio-linguistic compositional understanding. In *CVPR*, pages 13774–13784, 2024.

[Zhang *et al.*, 2024c] Shuili Zhang, Hongzhang Mu, Tingwen Liu, Qianqian Tong, and Jiawei Sheng. Mskr: Advancing multi-modal structured knowledge representation with synergistic hard negative samples. In *CIKM*, pages 3207–3216, 2024.

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