

# Decoupled Global-Local Alignment for Improving Compositional Understanding

Xiaoxing Hu\*

xiaoxinghhh@gmail.com

Beijing Institute of Technology

School of Information and Electronics

Beijing, China

Kaicheng Yang\*

kaichengyang@deepglint.com

DeepGlint

Beijing, China

Jun Wang

junwang@deepglint.com

DeepGlint

Beijing, China

Haoran Xu

xhr964691257@163.com

Zhejiang University

Zhejiang Province, China

Ziyong Feng

ziyongfeng@deepglint.com

DeepGlint

Beijing, China

Yupei Wang†

wangyupei2019@outlook.com

Beijing Institute of Technology

School of Information and Electronics

Beijing, China

## Abstract

Contrastive Language-Image Pre-training (CLIP) has achieved success on multiple downstream tasks by aligning image and text modalities. However, the nature of global contrastive learning limits CLIP's ability to comprehend compositional concepts, such as relations and attributes. Although recent studies employ global hard negative samples to improve compositional understanding, these methods significantly compromise the model's inherent general capabilities by forcibly distancing textual negative samples from images in the embedding space. To overcome this limitation, we introduce a **Decoupled Global-Local Alignment (DeGLA)** framework that improves compositional understanding while substantially mitigating losses in general capabilities. To optimize the retention of the model's inherent capabilities, we incorporate a self-distillation mechanism within the global alignment process, aligning the learnable image-text encoder with a frozen teacher model derived from an exponential moving average. Under the constraint of self-distillation, it effectively mitigates the catastrophic forgetting of pretrained knowledge during fine-tuning. To improve compositional understanding, we first leverage the in-context learning capability of Large Language Models (LLMs) to construct about 2M high-quality negative captions across five types. Subsequently, we propose the Image-Grounded Contrast (IGC) loss and Text-Grounded Contrast (TGC) loss to enhance vision-language compositionally. Experimental results across both general and compositional reasoning tasks validate the effectiveness of the DeGLA framework. Our code is released at <https://github.com/xiaoxing2001/DeGLA>.

\*Equal contribution.

†Corresponding author.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

MM '25, Dublin, Ireland

© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 979-8-4007-2035-2/2025/10

<https://doi.org/10.1145/3746027.3755032>

## CCS Concepts

- Computing methodologies → Computer vision representations.

## Keywords

Vision-Language Model, Multi-Modal, Compositional Understanding

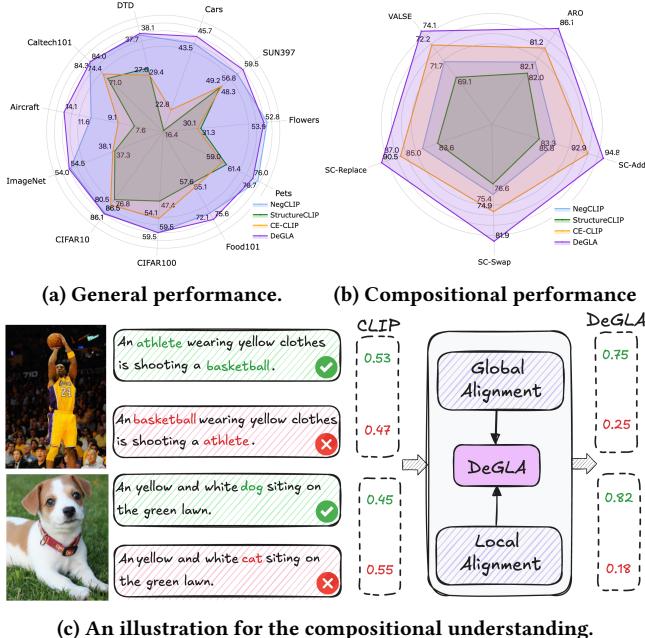
### ACM Reference Format:

Xiaoxing Hu, Kaicheng Yang, Jun Wang, Haoran Xu, Ziyong Feng, and Yupei Wang. 2025. Decoupled Global-Local Alignment for Improving Compositional Understanding. In *Proceedings of the 33rd ACM International Conference on Multimedia (MM '25), October 27–31, 2025, Dublin, Ireland*. ACM, New York, NY, USA, 14 pages. <https://doi.org/10.1145/3746027.3755032>

## 1 Introduction

The rapid expansion of mobile networks and social platforms significantly boosts the large-scale generation of image-text pairs, laying a crucial foundation for vision-language pre-training [15, 16]. In particular, Contrastive Language-Image Pre-training [51] (CLIP) leverages two distinct unimodal encoders for images and texts, and employs contrastive loss [45] for representation learning. After pretraining on extensive image-text pairs sourced from the internet, CLIP exhibits strong transferability and is extensively applied across various tasks, including image captioning [1, 32, 40], object detection [27, 38, 56], and segmentation [4, 17, 19, 31, 41, 57, 62, 64].

As CLIP attracts increasing attention from researchers, several enhanced methods based on CLIP have been proposed [14, 34, 69]. SLIP [43] introduces a multitask learning framework that integrates self-supervised learning with CLIP pretraining. FILIP [61] enhances the fine-grained alignment between image patches and textual words by refining the contrastive loss, while maintaining the capability for offline pre-computation of image and text representations during inference, ensuring efficiency in both large-scale training and inference phases. ALIP [59] proposes the use of synthetic captions and adaptive contrastive loss to mitigate the influence of noisy data and improve the efficacy of pre-training data utilization. SigLIP [67] utilizes a sigmoid loss, which supports scaling up batch sizes and also performs effectively at smaller batch sizes. However, global contrastive learning inadequately leverages compositional



**Figure 1: (a) General performance comparison across 11 classification datasets. (b) Compositional performance comparison on VALSE, ARO, and SugarCrepe benchmarks. (c) CLIP scores for images with aligned and unaligned captions. DeGLA demonstrates significantly enhanced capabilities in compositional understanding.**

structures within image-text pairs, thereby limiting CLIP’s capability to capture nuanced compositional information in multimodal data [63]. As shown in Figure 1c, CLIP struggles to discern the relationship between “athlete” and “basketball”.

Recent studies [7, 8, 22, 63, 68] have sought to improve the compositional understanding of CLIP. NegCLIP [63] introduces the Attribution, Relation, and Order (ARO) benchmark and, for the first time, proposes a fine-tuning framework that integrates hard negative samples to enhance CLIP’s compositional understanding. Structure-CLIP [22] integrates Scene Graph Knowledge (SGK) to augment multimodal structured representations. Hard-positives [25] incorporates hard positive samples during fine-tuning to strengthen the model’s capacity to capture subtle but semantically related variations among similar instances. CE-CLIP [68] introduces a simple yet effective fine-tuning framework with two fine-grained alignment losses to enhance compositional understanding, achieving state-of-the-art performance on multiple vision-language compositional reasoning benchmarks. However, our analysis reveals that existing methods improve CLIP’s compositional understanding at the cost of general capabilities, exhibiting significant catastrophic forgetting of pre-trained knowledge.

To address the limitation of previous works, we innovatively introduce a **Decoupled Global-Local Alignment (DeGLA)** framework, which aims to improve compositional understanding while substantially mitigating losses in general capabilities. To optimize the retention of the inherent capabilities of the model, we incorporate a self-distillation mechanism within the global alignment process,

aligning the learnable image-text encoder with a teacher model derived from an exponential moving average. To improve compositional understanding, we first leverage the in-context learning capability of Large Language Models (LLMs) to construct about 2M high-quality negative captions across five types. Subsequently, we propose the Image-Grounded Contrast (ICC) and Text-Grounded Contrast (TCC) to improve fine-grained understanding. We conduct extensive experiments and demonstrate that our method achieves new state-of-the-art performance in both compositional and general tasks. The main contributions are summarized as follows:

- We observe that previous methods, while enhancing CLIP’s compositional understanding, often **compromise its general understanding capabilities**.
- We propose a **simple yet effective negative caption generation pipeline** that leverages the context learning capability of Large Language Models (LLMs) to generate high-quality negative captions.
- We introduce the **DeGLA framework**, which employs a self-distillation mechanism within the global alignment to maintain the model’s inherent general comprehension capabilities. Additionally, it combines Image-Grounded Contrast (IGC) loss and Text-Grounded Contrast (TGC) loss to improve vision-language compositional understanding.
- We conduct **extensive experiments and demonstrate that DeGLA achieves new state-of-the-art performance** in both compositional and general tasks.

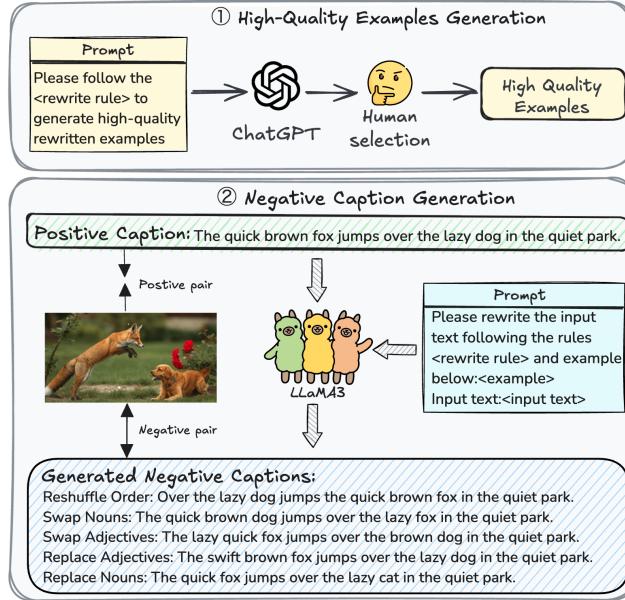
## 2 Related work

### 2.1 Vision-Language Contrastive Learning

Vision-language representation learning aims to develop robust representations by pretraining models on extensive image-text pair datasets. A key development in this field, CLIP [52], pretrained on 400 million web-collected image-text pairs using contrastive learning, demonstrates strong zero-shot generalization across a wide range of vision tasks. Recent advancements include several methodologies that build on CLIP’s framework. UniCLIP [30] improves data efficiency by integrating contrastive losses from various domains into a single universal space. HiCLIP [11] incorporates hierarchy-aware attention mechanisms in both visual and linguistic branches of CLIP, significantly enhancing cross-modal alignment. LaCLIP [9] employs large language models to rewrite text, thereby increasing sentence structure and vocabulary diversity while preserving essential concepts and meanings. Nonetheless, these developments do not rectify the limitations in CLIP’s compositional understanding, constrained by the inherent shortcomings of global contrastive learning [63].

### 2.2 Vision-Language Compositionality

Recently, there have been some works [20, 24, 47, 49, 53, 63] aim to enhance the compositional understanding of CLIP. NegCLIP [63] introduces the Attribution, Relation, and Order (ARO) benchmark and utilizes hard negatives, consisting of nearest neighboring images within each batch, to force models to discern fine-grained differences in highly similar scenes. Structure-CLIP [22] incorporates Scene Graph Knowledge (SGK) to enhance multimodal structured representations. CE-CLIP [68] proposes a simple yet



**Figure 2: The overview of our proposed LLM-driven negative caption generation pipeline. We leverage the robust in-context learning capabilities of LLM to generate five types of hard negative captions.**

effective strategy to optimize the utilization of existing image-text datasets, achieving state-of-the-art performance across multiple vision-language compositional reasoning benchmarks. However, although these methods significantly enhance CLIP’s compositional understanding, they frequently compromise its original generalization capabilities. Developing an approach that simultaneously improves both general and compositional understanding continues to be a substantial challenge.

### 2.3 Knowledge Distillation

Knowledge distillation [18] is extensively utilized in various domains, including vision [21, 35, 36] and natural language processing [23]. Recently, a series of knowledge distillation methods tailored specifically for CLIP have been proposed. TinyCLIP [55] introduces affinity mimicking, which explores the interaction between modalities during distillation, enabling student models to replicate the teacher’s behavior in learning cross-modal feature alignment within a visual-linguistic affinity space. CLIP-KD [60] proposes several distillation strategies, including relation, feature, gradient, and contrastive paradigms, aimed at maximizing feature similarity between the teacher and student models. CLIP-CID [60] employs cluster-instance discrimination to facilitate knowledge transfer from the teacher model to the student model, thereby enabling the student model to develop a comprehensive semantic understanding of the pre-training data. Different from the above method, this paper introduces a self-distillation mechanism in global alignment to mitigate catastrophic forgetting during training and preserve the generalization capabilities of the model.

**Table 1: Detailed descriptions of the generated high-quality negative samples.**

Negative types	Rewrite rule
Intra-sentence reshuffle	Subtype1
	Subtype2
	Subtype3
Minimal semantic substitution	Subtype4
	Subtype5

## 3 Method

In subsequent sections, we first present the preliminary knowledge of CLIP in Section 3.1, followed by a detailed exposition of our proposed LLM-driven negative caption generation pipeline in Section 3.2. A comprehensive description of the DeGLA training framework is provided in Section 3.3.

### 3.1 Preliminaries of CLIP

CLIP consists of a text encoder  $\mathcal{E}_T$ , and an image encoder  $\mathcal{E}_I$ . It encodes a batch of image-text pairs  $\{(I_i, T_i)\}_{i=1}^B$  into the feature space  $\{(v_i, t_i)\}_{i=1}^B$ . The model employs the InfoNCE loss [46] to optimize the embeddings by minimizing the distance between corresponding image and text features while maximizing the distance between non-corresponding pairs. The image-to-text contrastive loss function is defined as follows:

$$\mathcal{L}_{I \rightarrow T} = -\log \frac{\exp(v_i \cdot t_i^\top / \tau)}{\sum_{j=1}^B \exp(v_i \cdot t_j^\top / \tau)}, \quad (1)$$

where  $v_i$  serves as the anchor. Given  $t_i$  as the anchor, The symmetric text-to-image contrastive loss is formulated as follows:

$$\mathcal{L}_{T \rightarrow I} = -\log \frac{\exp(t_i \cdot v_i^\top / \tau)}{\sum_{j=1}^B \exp(t_i \cdot v_j^\top / \tau)}, \quad (2)$$

where  $\cdot$  denotes the dot product used to calculate the similarity between normalized image and text embeddings, and  $\tau$  is a learnable temperature parameter. The overall CLIP loss is:

$$\mathcal{L}_{CLIP} = \frac{1}{2} (\mathcal{L}_{I \rightarrow T} + \mathcal{L}_{T \rightarrow I}). \quad (3)$$

CLIP encodes all semantic content from images and texts into a global representation during its training process, which constrains its capability for compositional understanding. While recent works [22, 63, 68] have sought to enhance this ability by introducing hard negative samples to promote the learning of local semantics, our analysis reveals experimental results demonstrate that such enhancements often degrade the model’s inherent general comprehension capabilities. Consequently, this paper focuses on improving compositional understanding without compromising the original general comprehension abilities of the model.

### 3.2 LLM-Driven Negative Caption Generation

To improve compositional understanding of CLIP, we first leverage the in-context learning capability of Large Language Models (LLMs) to construct about 2M high-quality negative captions across multiple types. Previous methods for generating compositional negative captions can be classified into two categories: rule-based generation [22, 63] and unmasking-based generation [7, 8, 68]. However,

rule-based approaches are limited in their ability to produce complex compositional negatives, such as those involving minimal semantic substitutions. In contrast, unmasking-based methods can generate diverse negatives but often fail to avoid producing hard positive samples such as “the cat is eating the food” → “the cat is eating the hot food”.

To address these limitations, we propose an LLM-driven generation method to enhance and standardize the generation of compositional negative captions. Specifically, we categorize negative captions into two primary types: (1) intra-sentence reshuffling, which improves the VLM’s sensitivity to sentence order, and (2) minimal semantic substitution, which enhances the VLM’s ability to discern subtle semantic variations. For intra-sentence reshuffling, we examine three subtypes: (a) full sentence reordering, (b) noun swapping, and (c) adjective swapping. For minimal semantic substitution, we investigate two subtypes: (a) adjective replacement and (b) noun replacement. Detailed descriptions of each negative sample type are presented in Table 1. As shown in Figure 2, we initially generate high-quality negative samples to fully leverage the in-context learning capabilities of LLMs. For each type, we first use ChatGPT4-Turbo to generate 200 rewritten examples according to rewritten rules, then manually select the 50 highest-quality examples through rigorous evaluation. Building upon these examples, we employ the Llama-3.1-8B-Instruct model<sup>1</sup> to generate large-scale compositional negatives. To ensure semantic divergence from original sentences and avoid hard positives, we incorporate explicit filtering constraints in the prompts (e.g., “generated sentences must exhibit distinct semantics”). Complete prompt templates and representative samples are provided in the appendix.

### 3.3 DeGLA Framework

**Global Alignment.** Building upon the base setting [63], we integrate negative captions into the Equation 1 for global alignment, yielding the augmented image-to-text contrastive:

$$\mathcal{L}_{I \rightarrow T} = -\log \frac{\exp(v_i \cdot t_i^\top / \tau)}{\sum_{j=1}^B \left( \exp(v_i \cdot t_j^\top / \tau) + \sum_{k=1}^K \exp(v_i \cdot \bar{t}_{i,k}^\top / \tau) \right)}, \quad (4)$$

where  $\bar{t}_{i,k}$  denotes the  $k$ -th negative text related to the  $j$ -th text. In this work, we set  $K = 4$  to denote the number of negative samples paired with each positive sample during training, due to the merging of two subtypes, as detailed in the appendix. Then we utilize the hard negative-aware image-text contrastive loss as our base loss which can be formulated as:

$$\mathcal{L}_{base} = \frac{1}{2} (\mathcal{L}_{I \rightarrow T} + \mathcal{L}_{T \rightarrow I}), \quad (5)$$

However, directly incorporating hard negative text into global contrastive learning significantly compromises the model’s general capabilities, as indicated in Section 4. In this paper, we introduce a self-distillation mechanism within the global alignment framework to mitigate the catastrophic forgetting of pre-trained knowledge. To enhance model robustness and prevent rapid forgetting of pre-trained knowledge during training, we employ an Exponential Moving Average (EMA) strategy to update the weights of the frozen

<sup>1</sup><https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct>

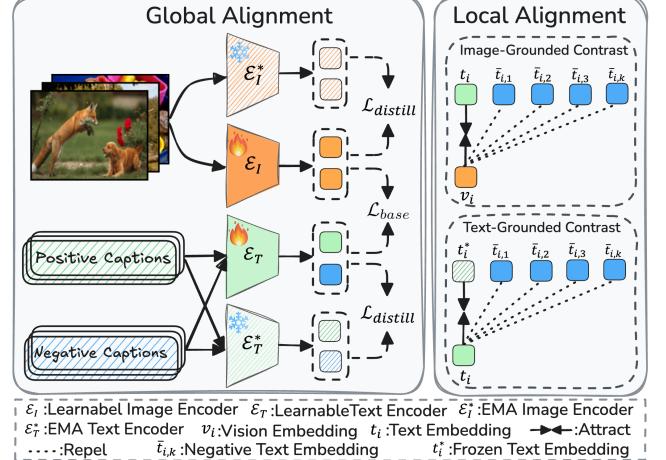


Figure 3: The proposed training framework of DeGLA.

image encoder  $\mathcal{E}_I^*$  and text encoder  $\mathcal{E}_T^*$ :

$$\begin{aligned} \mathcal{E}_I^* &= \alpha \mathcal{E}_I^* + (1 - \alpha) \mathcal{E}_I, \\ \mathcal{E}_T^* &= \alpha \mathcal{E}_T^* + (1 - \alpha) \mathcal{E}_T, \end{aligned} \quad (6)$$

where the hyperparameter  $\alpha$  controls the update speed of the frozen model’s parameters. Given a batch of image-text pairs  $\{I_i, T_i, \bar{T}_i\}$ , we obtain the image and text embeddings  $\{v_i, t_i, \bar{t}_i\}$  from the learnable encoders ( $\mathcal{E}_I, \mathcal{E}_T$ ) and  $\{v_i^*, t_i^*, \bar{t}_i^*\}$  from the frozen EMA encoders ( $\mathcal{E}_I^*, \mathcal{E}_T^*$ ), respectively. Following L2 normalization, we compute the squared distance between the image-text embeddings from the learnable and frozen EMA encoders to quantify the knowledge discrepancy between the fine-tuned and pre-trained models:

$$\mathcal{L}_{Distill} = \sum_{i=1}^B \left( \|v_i - v_i^*\|_2^2 + \|t_i - t_i^*\|_2^2 + \sum_{k=1}^K \|\bar{t}_{i,k} - \bar{t}_{i,k}^*\|_2^2 \right). \quad (7)$$

Under the constraint of Equation 7, the image-text embeddings undergo subtle refinements within the pre-trained representation space, thereby preserving the model’s general capabilities from excessive degradation.

**Local Alignment.** In addition to the global alignment of  $\mathcal{L}_{base}$ , to further enhance compositional understanding, we propose a Local Alignment, which includes Image-Grounded Contrast Loss(IGC) and Text-Grounded Contrast Loss(TGC). We initially introduce the Image-Grounded Contrast loss to attract image embeddings towards positive text embeddings and repel them from negative text embeddings in the feature space (as shown in Figure 3). Specifically, given an image-text pair  $(I_i, T_i)$  and corresponding hard negative texts  $\bar{T}_i$ , the IGC loss  $\mathcal{L}_{IGC}$  is defined as follows:

$$\mathcal{L}_{IGC} = -\log \frac{\exp(v_i \cdot t_i^\top / \tau)}{\exp(v_i \cdot t_i^\top / \tau) + \sum_{k=1}^K \exp(v_i \cdot \bar{t}_{i,k}^\top / \tau)}, \quad (8)$$

where  $\tau$  is a temperature parameter that controls the sharpness of the distribution. Switching to the textual perspective, hard negative texts, which are minor modifications of the originals, remain proximate to the positive texts within the feature space of the CLIP text encoder. This closeness may lead to mismatches during image embedding alignment with positive text embeddings, potentially

**Table 2: Results (%) on VALSE.** The best results are marked in bold, and the second-best results are underlined. The number represents the improvement of our method compared to the CE-CLIP.

Model	#Params	Existence quantifiers	Plurality number	Counting	Sp.rel. relations	Actions repl.	Actions antact swap	Coreference standard	Coreference clean	Foil-it!	Avg.
BLIP[33]	583M	86.3	73.2	68.1	71.5	77.2	61.1	53.8	48.2	93.8	70.0
BEIT3[54]	1.9B	77.4	74.6	68.8	74.0	86.7	65.2	50.0	44.2	96.0	70.4
BLIP2[32]	3.4B	55.5	71.5	66.0	62.4	83.6	51.6	48.6	51.9	95.9	65.4
MiniGPT-4[70]	>9B	65.5	72.5	67.4	68.4	83.2	58.8	52.6	51.0	95.8	68.4
<i>Hard Negative based method</i>											
XVLM-coco[65]	216M	83.0	75.6	67.5	70.2	73.8	68.6	46.4	49.6	94.8	69.5
CE-XVLM[68]	216M	83.5	72.8	72.1	68.7	71.8	69.1	51.0	46.8	93.8	70.8
CLIP[52]	151M	68.7	57.1	61.0	65.4	77.8	71.8	54.1	51.0	89.8	65.3
CyCLIP[12]	151M	69.3	58.3	61.0	66.4	78.1	72.0	53.2	51.6	88.8	65.5
NegCLIP[63]	151M	76.8	72.0	65.2	72.7	81.6	84.8	58.9	54.8	91.8	71.7
Structure-CLIP [22]	151M	75.6	67.1	62.0	68.2	80.4	88.3	44.5	58.7	91.2	69.1
CE-CLIP[68]	151M	78.6	77.6	64.3	74.0	81.2	88.5	54.9	52.9	93.7	72.2
DeGLA (ours)	151M	<b>82.4</b>	<u>73.8</u>	<b>68.3</b>	<u>75.3</u>	<b>82.6</b>	<b>88.8</b>	<u>58.5</u>	<u>54.8</u>	<b>93.8</b>	<b>74.1(+1.9)</b>

**Table 3: Results(%) on SugarCrepe.** Vera and Grammar are text-only models. The best results are marked in bold, and the second-best results are underlined. The number represents the improvement of our method compared to the CE-CLIP.

Model	REPLACE				SWAP				ADD		
	Object	Attribute	Relation	Avg.	Object	Attribute	Avg.	Object	Attribute	Avg.	
Human	100.0	99.0	97.0	98.7	99.0	100.0	99.5	99.0	99.0	99.0	
Vera [37]	49.4	49.6	49.1	49.4	49.4	49.2	49.3	49.4	49.6	49.5	
Grammar [42]	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	
BLIP2 [32]	-	-	-	86.7	-	-	69.8	-	-	86.5	
<i>Hard Negative based method</i>											
CLIP [52]	90.9	80.0	69.2	80.2	62.7	61.4	64.0	77.2	68.2	72.7	
NegCLIP [63]	92.7	85.9	76.5	85.0	75.3	75.2	75.4	88.8	82.8	85.8	
Structure-CLIP [22]	91.4	85.0	74.4	83.6	72.7	<u>80.5</u>	<u>76.6</u>	85.5	81.1	83.3	
CE-CLIP [68]	<u>93.1</u>	<u>88.8</u>	<u>79.0</u>	<u>87.0</u>	<u>72.8</u>	77.0	74.9	92.4	<u>93.4</u>	<u>92.9</u>	
DeGLA (ours)	<b>94.5</b>	<b>92.6</b>	<b>84.2</b>	<b>90.5(+3.5)</b>	<b>81.6</b>	<b>82.1</b>	<b>81.9(+6.9)</b>	<b>93.8</b>	<b>95.7</b>	<b>94.8(+1.9)</b>	

**Table 4: Results (%) on ARO.** The best results are marked in bold, and the second-best results are underlined. The number represents the improvement of our method compared to the CE-CLIP.

Model	Relation	Attribute	COCO-order	Flickr-order	Avg.
BLIP[33]	59.0	88.0	-	-	-
BEIT3[54]	60.6	74.6	-	-	-
BLIP2[32]	41.2	71.3	-	-	-
MiniGPT-4[70]	46.9	55.7	-	-	-
<i>Hard Negative based method</i>					
CLIP [52]	59.2	62.9	48.4	59.1	57.4
CyCLIP[12]	59.1	65.4	-	-	-
NegCLIP[63]	80.4	70.5	<u>86.9</u>	<u>90.5</u>	<u>82.1</u>
Structure-CLIP [22]	<u>81.8</u>	<b>80.5</b>	81.7	83.9	82.0
CE-CLIP[68]	<b>83.9</b>	<u>76.4</u>	80.9	83.7	81.2
DeGLA (ours)	81.6	74.3	<b>93.8</b>	<b>94.7</b>	<b>86.1(+4.9)</b>

impairing compositional understanding. Different from the Image-Grounded Contrast loss, the Text-Grounded Contrast (TGC) loss  $\mathcal{L}_{TGC}$  operates solely within the text modality. Its function is to enhance the text encoder's ability to more effectively discriminate between positive and negative texts, thus improving the compositional understanding of vision language models. The  $\mathcal{L}_{TGC}$  is formulated as:

$$\mathcal{L}_{TGC} = -\log \frac{\exp(t_i \cdot t_i^{*\top} / \tau)}{\exp(t_i \cdot t_i^{*\top} / \tau) + \sum_{k=1}^K \exp(t_i \cdot \bar{t}_{i,k}^\top / \tau)}, \quad (9)$$

where  $t_i^*$  represents the progressively frozen text embedding obtained from the frozen EMA text encoder  $\mathcal{E}_T^*$ . This embedding is used as an anchor in local alignment for two main reasons: (1) It acts as the positive sample in contrastive learning, enabling the model to discriminate between positive and compositional negative samples; and (2) It mitigates potential overfitting of the text encoder to the feature space of the fine-tuning dataset. Finally, the overall loss function is defined as:

$$\mathcal{L}_{all} = \mathcal{L}_{Base} + \lambda_1 \mathcal{L}_{IGC} + \lambda_2 \mathcal{L}_{TGC} + \lambda_3 \mathcal{L}_{Distill}. \quad (10)$$

where  $\lambda_1, \lambda_2, \lambda_3$  are loss weights to balance the influence of different loss functions.

## 4 Experiments

### 4.1 Implement details

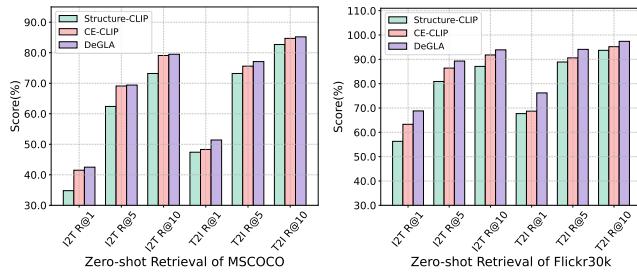
**Training Setup.** To ensure a fair comparison with previous studies [63, 68], we employ LLaMA3.1-instruct-8B [13] to generate compositional negative samples from the MSCOCO dataset, facilitating direct comparisons with NegCLIP [63] and CE-CLIP [68]. We utilize the CLIP-ViT/B-32 model as the foundation vision-language model, initializing it with pretrained weights from CLIP [52]. The model is fine-tuned on 8 NVIDIA V100 (32G) GPUs for 5 epochs using a batch size of 256, consistent with the protocols of previous

**Table 5: Zero-shot classification performance on 11 datasets. The best results are marked in bold, and the second-best results are underlined. The number represents the improvement of our method compared to the CE-CLIP.**

Model	CIFAR10	CIFAR100	Food101	Pets	Flowers	SUN397	Cars	DTD	Caltech101	Aircraft	ImageNet	Avg.
<i>Pretrained model</i>												
CLIP [52]	86.5	61.0	78.5	79.6	58.4	59.9	48.8	38.7	86.3	15.3	57.9	61.0
<i>Hard negative based method</i>												
NegCLIP [63]	<u>86.1</u>	<b>59.9</b>	<u>72.1</u>	<b>78.7</b>	<b>53.9</b>	<u>56.8</u>	<u>43.5</u>	<u>37.7</u>	<b>84.3</b>	<u>11.6</u>	<u>54.0</u>	<u>58.1</u>
Structure-CLIP [22]	76.8	47.4	55.1	61.4	31.3	48.3	16.4	29.4	71.0	7.6	37.3	43.8
CE-CLIP [68]	80.5	54.1	57.6	59.0	30.1	49.2	22.8	27.6	74.4	9.1	38.1	45.7
DeGLA (ours)	<b>86.5</b>	<u>59.5</u>	<b>75.6</b>	<u>76.0</u>	52.8	<u>59.5</u>	<b>45.7</b>	<b>38.1</b>	<u>84.0</u>	<b>14.1</b>	<u>54.5</u>	<b>58.7(+13.0)</b>

**Table 6: Liner probe performance on 11 datasets. The best results are marked in bold, and the second-best results are underlined. The number represents the improvement of our method compared to the CE-CLIP.**

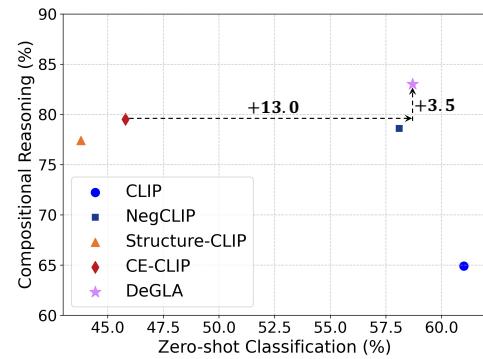
Model	CIFAR10	CIFAR100	Food101	Pets	Flowers	SUN397	Cars	DTD	Caltech101	Aircraft	ImageNet	Avg.
<i>Pretrained model</i>												
CLIP [52]	95.0	80.1	88.5	89.3	94.6	74.1	80.8	73.6	90.5	44.8	74.3	80.5
<i>Hard negative based method</i>												
NegCLIP [63]	<u>94.6</u>	<u>80.0</u>	<u>86.1</u>	<b>89.6</b>	<u>93.9</u>	<u>72.9</u>	<b>78.8</b>	<u>72.9</u>	<b>90.0</b>	<u>43.2</u>	<u>72.9</u>	<u>79.5</u>
Structure-CLIP [22]	91.9	75.5	81.2	86.2	89.6	69.0	67.4	67.7	65.2	37.7	67.7	72.7
CE-CLIP [68]	94.3	78.5	84.3	88.1	92.6	71.0	<u>74.1</u>	71.8	88.3	39.6	70.7	77.6
DeGLA (ours)	<b>95.1</b>	<u>80.5</u>	<b>86.7</b>	<u>89.5</u>	<b>94.6</b>	<u>74.0</u>	<b>78.8</b>	<u>73.0</u>	<u>89.6</u>	<b>43.5</b>	<u>73.4</u>	<b>79.9(+2.3)</b>

**Figure 4: Zero-shot image-text retrieval performance comparison on MSCOCO and Flickr30k.**

works [8, 63, 68]. We employ AdamW as the optimizer, initialized with a learning rate of  $1 \times 10^{-6}$  and a weight decay of 0.1. The parameters  $\alpha$ ,  $\beta_1$ , and  $\beta_2$  are set to 0.9996, 0.9, and 0.98, respectively. We perform a hyperparameter search for  $\lambda_1, \lambda_2, \lambda_3$ , with optimal values of  $\lambda_1 = 0.1$ ,  $\lambda_2 = 0.1$ , and  $\lambda_3 = 0.005$ .

**Evaluation Setup.** To comprehensively evaluate DeGLA, we conduct comparative experiments on three compositional reasoning benchmarks: ARO [63], VALSE [47], and SugarCrepe [20]. To verify that DeGLA retains general capabilities, we further assess its performance on zero-shot classification, liner probe, and retrieval tasks. For zero-shot classification, we evaluate on eleven datasets: CIFAR-10, CIFAR-100 [29], Food101 [2], Oxford Pets [48], Flowers102 [44], SUN397 [58], Stanford Cars [28], DTD [5], Caltech101 [10], FGVC-Aircraft [39], and ImageNet [6]. The datasets used for evaluating the linear probe is the same as that used for zero-shot classification. For zero-shot image-text retrieval, we report results on MSCOCO [3] and Flickr30k [50].

We compare DeGLA with three kinds of models: (1)Text-only models, including vera [37] and Grammar [42]. (2) State-of-the-art generative vision-language models, including BLIP [33], BLIP-2 [32], MiniGPT-4 [70]; (3) High-performance vision-language understanding models, such as BEIT-3 [54], XVLM [66]; (4) Specialized

**Figure 5: Performance trade off between compositional reasoning (average performance on VALSE, SugarCrepe, and ARO benchmarks) and zero-shot classification.**

compositional improvement methods NegCLIP [63], CyCLIP [12], Structure-CLIP [22], and CE-CLIP [68].

## 4.2 Compositional Reasoning

**Performance on VALSE.** We first evaluate DeGLA on VALSE, a benchmark specifically designed to assess pre-trained vision-language models’ sensitivity to foiled instances. This benchmark systematically tests fundamental linguistic phenomena across visual and linguistic modalities, including existence, plurality, counting, spatial relations, actions, and entity coreference. As indicated in Table 2, DeGLA outperforms other models in existence, counting, and spatial relations, and ranks second in plurality, actions replacement, coreference standard, and foil-it. Compared to CLIP, DeGLA exhibits an 8.8% average performance improvement, demonstrating that our approach more effectively leverages hard negatives through local and global alignment strategies. Thanks to the diverse negative samples generated in this study and the employment of both image-grounded and text-grounded contrastive learning, DeGLA

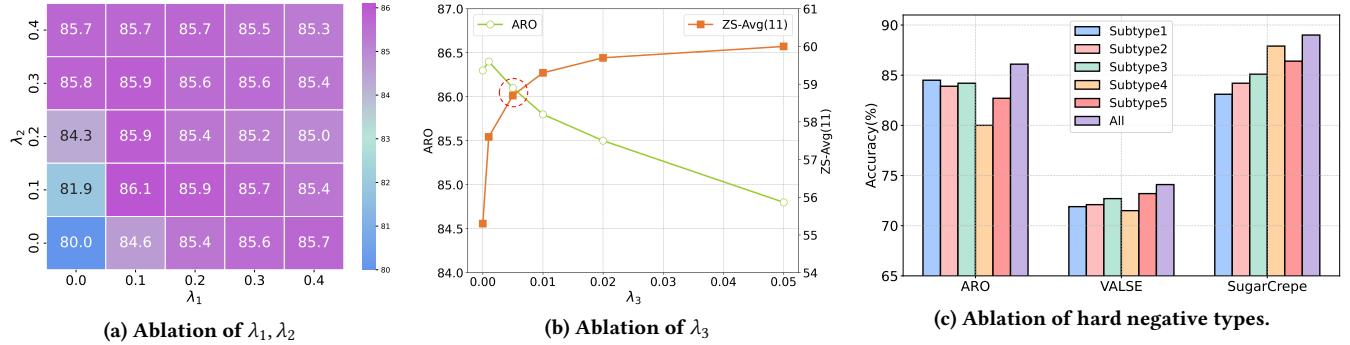


Figure 6: Ablation of the different loss weights and hard negative types.

significantly surpasses CE-CLIP, achieving an overall performance increase of 1.9%.

**Performance on SugarCrepe.** SugarCrepe is a benchmark designed to reduce language bias in existing benchmarks and provide a more accurate assessment of a model’s compositional understanding. As indicated in Table 3, DeGLA achieves new state-of-the-art results across all metrics. DeGLA shows significant improvements over the CLIP model, with increases of 10.3% on Replace, 17.9% on Swap, and 22.1% on Add. These results suggest that DeGLA effectively distinguishes between positive and negative samples through local alignment. Compared to CE-CLIP, our method achieves performance enhancements of 3.5%, 6.9%, and 1.9% on Replace, Swap, and Add, respectively. This substantial improvement is attributed to two primary factors: First, the negative sample generation method we introduced leverages Large Language Models (LLMs) to produce high-quality negative text descriptions, reducing the generation of noisy and challenging positive samples. Second, we implement image-grounded and text-grounded contrast mechanisms that enhance the model’s discriminative capabilities by drawing positive sample pairs closer and pushing negative sample pairs further apart in the feature space, thus improving the model’s compositional understanding.

**Performance on ARO.** We present the performance of our proposed DeGLA on the ARO benchmark in Table 4. DeGLA achieves substantial improvements, registering an average performance increase of 28.7% over CLIP and 4.9% over CE-CLIP. It is noteworthy that while DeGLA exhibits significant overall performance gains, it still trails CE-CLIP and Structure-CLIP in the domains of relations and attributes. The primary reason for this discrepancy is that our negative sample generation method prioritizes the production of a diverse array of negative samples, rather than focusing specifically on enhancing the understanding of relations and attributes.

### 4.3 General Understanding

**Zero-shot classification.** In Table 5, we present the zero-shot classification performance across 11 datasets. Notably, while Structure-CLIP and CE-CLIP exhibit robust compositional understanding, they significantly reduce the model’s original general capabilities, with average accuracies decreasing by 17.2% and 15.3%, respectively, compared to CLIP. In contrast, our proposed DeGLA employs a self-distillation constraint module for global alignment, effectively minimizing the loss of general capability. Our method shows a

Table 7: Ablation of different components. CN: compositional negatives. SD: self-distillation.

Model	CN	IGC	TGC	SD	ARO	ZS-Avg.(11)
CLIP [52]					57.4	61.0
CE-CLIP [68]					81.2	45.7
DeGLA(ours)	✓				80.8	55.4
	✓	✓			84.8	55.1
	✓		✓		82.2	57.2
	✓	✓	✓		<b>86.2</b>	56.9
	✓	✓	✓	✓	86.1	<b>58.7</b>

13.0% improvement in average accuracy over CE-CLIP, while substantially preserving the model’s inherent general capabilities. As illustrated in Figure 5, DeGLA achieves a more effective balance between compositional reasoning and general comprehension capabilities compared to other methods.

**Linear probe.** In Table 5, we detail the linear probe performance across 11 datasets. Consistent with the zero-shot classification results, Structure-CLIP and CE-CLIP significantly reduce the model’s original general capabilities, with average accuracies decreasing by 7.8% and 2.9%, respectively, compared to CLIP. Conversely, our proposed DeGLA model not only demonstrates superior compositional understanding relative to CE-CLIP but also achieves a 2.3% average performance improvement in linear probe tasks across these datasets.

**Zero-shot image text retrieval.** We compare the zero-shot retrieval performance of DeGLA, Structure CLIP, and CE-CLIP on the MSCOCO and Flickr30k datasets. As illustrated in Figure 4, DeGLA surpasses previous models on both datasets, demonstrating superior performance in standard retrieval tasks. This finding is consistent with our earlier experimental results, further substantiating DeGLA’s excellence in both general capabilities and compositional reasoning.

### 4.4 Analysis

**Trade-off Analysis.** As illustrated in Figure 5, we analyze the trade-off between general capabilities and compositional reasoning. For the evaluation of compositional reasoning, we utilize the average scores from three benchmarks. Typically, enhancements in compositional reasoning are accompanied by declines in general

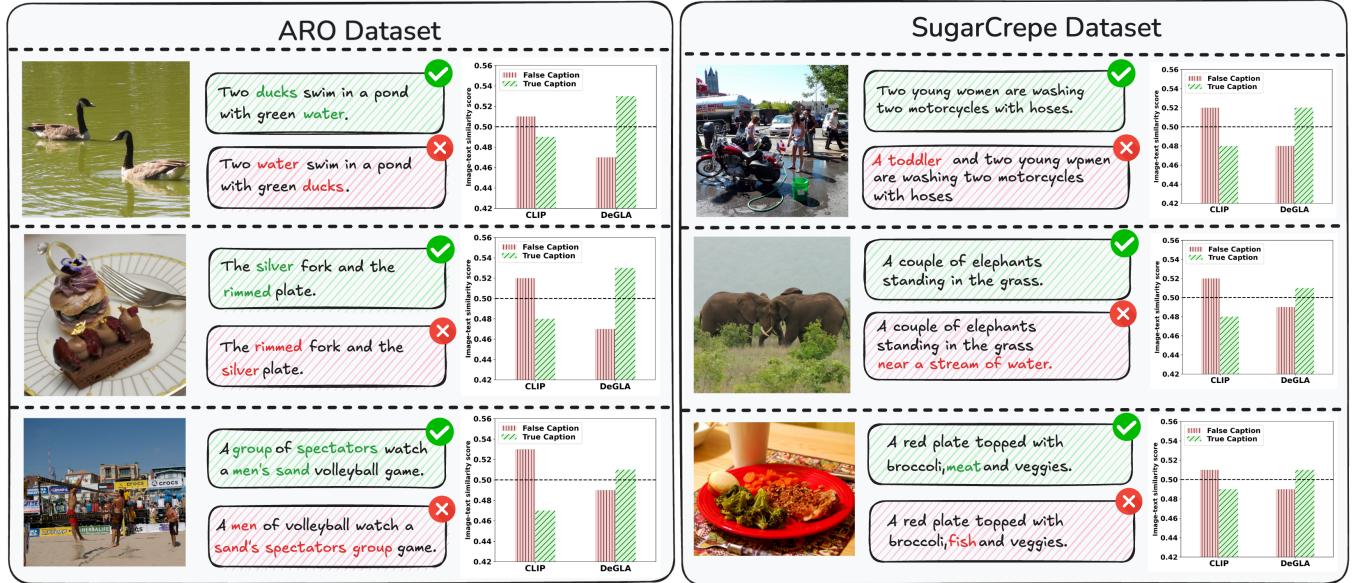


Figure 7: The case study between CLIP and DeGLA. Green ✓ indicates true caption, while red ✗ indicates false caption. The bar chart represents the model’s prediction results.

capabilities, as demonstrated by CE-CLIP, which, despite significant gains in compositional reasoning compared to CLIP, suffers a substantial reduction in general capabilities. In contrast, through its meticulously designed data generation pipeline and training framework, DeGLA achieves an optimal balance. Compared to the baseline CE-CLIP, DeGLA records a 13% improvement in zero-shot classification and an average enhancement of 3.5% in compositional reasoning tasks, further underscoring its superiority.

**Ablation of different components.** To validate the efficacy of the proposed method, we perform comprehensive ablation studies on key components using the ARO benchmark in Table 7. By combining all modules, DeGLA achieves the most balanced performance, attaining an 86.1% average score on ARO with 4.9% higher than the previous SOTA (CE-CLIP). For zero-shot classification, it outperforms CE-CLIP by 13.0%, demonstrating superior generalization. Analysis of component contributions reveals that integrating high-quality negative samples increases ARO performance by 23.4% but decreases zero-shot classification performance by 5.6%. The local alignment losses, IGC and TGC, improve ARO performance by 4.0% and 1.4%, respectively. After Combining further boosting ARO performance by 5.4%, while zero-shot classification still decreases by 4.1%. Notably, TGC mitigates declines in general capability by using frozen text embeddings from the EMA text encoder, ensuring minimal deviation from the pretrained CLIP model during training. The introduction of a self-distillation mechanism significantly enhances general capability, improving zero-shot classification by 1.8% with a negligible 0.1% loss in ARO performance.

**Ablation of compositional negative texts.** To verify the effectiveness of our proposed instruction-based negative generation pipeline, we conduct an ablation analysis on five subtypes of negative texts, as depicted in Figure 6c. The results show that combining these subtypes achieves optimal performance, thereby validating

the pipeline’s effectiveness. This success is attributed to the representativeness of our samples, which effectively enhance the compositional understanding of CLIP.

## 4.5 Case Study

Figure 7 compares CLIP and DeGLA on ARO and SugarCrepe. Consistent with NegCLIP [63], CLIP shows a “bag-of-words” issue, struggling with structural text changes like word order swaps or substitutions—for example, failing to distinguish between “Two ducks swim in a pond with green water” and “Two swans swim in a pond with green ducks.” In contrast, DeGLA correctly differentiates these based on the image. After training on our compositionality-focused hard negatives (Section 3.2), DeGLA demonstrates improved compositional reasoning over pretrained CLIP, validating the effectiveness of our data generation pipeline.

## 5 Conclusion

In this paper, we observe that while previous methods enhance CLIP’s compositional understanding, they often compromise its general capabilities. To overcome this limitation, we introduce a **Decoupled Global-Local Alignment (DeGLA)** framework, which not only improves compositional understanding but also significantly reduces losses in general capabilities. To optimize the retention of the model’s inherent capabilities, we incorporate a self-distillation mechanism within the global alignment process, aligning the learnable image-text encoder with a frozen teacher model derived from an exponential moving average. For enhancing compositional understanding, we leverage the in-context learning capabilities of Large Language Models (LLMs) to generate approximately 2M high-quality negative captions across five types. We further propose the Image-Grounded Contrast (IGC) loss and Text-Grounded Contrast (TGC) loss to improve vision-language compositionality.

## Acknowledgments

This work is supported by National Natural Science Foundation of China (Grant No. 62301046)

## References

- [1] 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. *NIPS* 35 (2022), 23716–23736.
- [2] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. 2014. Food-101—mining discriminative components with random forests. In *Computer vision—ECCV 2014: 13th European conference, zurich, Switzerland, September 6–12, 2014, proceedings, part VI*. Springer, 446–461.
- [3] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325* (2015).
- [4] Seokju Cho, Heeseong Shin, Sungwhan Hong, Anurag Arnab, Paul Hongsuck Seo, and Seungryong Kim. 2024. Cat-seg: Cost aggregation for open-vocabulary semantic segmentation. In *CVPR*. 4113–4123.
- [5] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. 2014. Describing textures in the wild. In *CVPR*. 3606–3613.
- [6] 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*. Ieee, 248–255.
- [7] Sivan Doveh, Assaf Arbelle, Sivan Harary, Roei Herzig, Donghyun Kim, Paula Cascante-Bonilla, Amit Alfassy, Rameswar Panda, Raja Giryes, Rogerio Feris, et al. 2023. Dense and aligned captions (dac) promote compositional reasoning in vl models. *NIPS* 36 (2023), 76137–76150.
- [8] 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 *CVPR*. 2657–2668.
- [9] Lijie Fan, Dilip Krishnan, Phillip Isola, Dina Katabi, and Yonglong Tian. 2023. Improving clip training with language rewrites. *NIPS* 36 (2023), 35544–35575.
- [10] Li Fei-Fei, Rob Fergus, and Pietro Perona. 2004. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *CVPR*. IEEE, 178–178.
- [11] Shijie Geng, Jianbo Yuan, Yu Tian, Yuxiao Chen, and Yongfeng Zhang. 2023. HiCLIP: Contrastive language-image pretraining with hierarchy-aware attention. *ICLR* (2023).
- [12] Shashank Goel, Hritik Bansal, Sumit Bhatia, Ryan Rossi, Vishwa Vinay, and Aditya Grover. 2022. Cyclip: Cyclic contrastive language-image pretraining. *NIPS* 35 (2022), 6704–6719.
- [13] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783* (2024).
- [14] Tiancheng Gu, Kaicheng Yang, Xiang An, Ziyong Feng, Dongnan Liu, Weidong Cai, and Jiankang Deng. 2024. RWKV-CLIP: a robust vision-language representation learner. *arXiv preprint arXiv:2406.06973* (2024).
- [15] Tiancheng Gu, Kaicheng Yang, Ziyong Feng, Xingjun Wang, Yanzhao Zhang, Dingkun Long, Yingda Chen, Weidong Cai, and Jiankang Deng. 2025. Breaking the Modality Barrier: Universal Embedding Learning with Multimodal LLMs. *arXiv preprint arXiv:2504.17432* (2025).
- [16] Tiancheng Gu, Kaicheng Yang, Chaoyi Zhang, Yin Xie, Xiang An, Ziyong Feng, Dongnan Liu, Weidong Cai, and Jiankang Deng. 2025. RealSyn: An Effective and Scalable Multimodal Interleaved Document Transformation Paradigm. *arXiv preprint arXiv:2502.12513* (2025).
- [17] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. 2021. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921* (2021).
- [18] Geoffrey Hinton. 2015. Distilling the Knowledge in a Neural Network. *arXiv preprint arXiv:1503.02531* (2015).
- [19] Lukas Hoyer, David Joseph Tan, Muhammad Ferjad Naeem, Luc Van Gool, and Federico Tombari. 2025. Semivil: Semi-supervised semantic segmentation with vision-language guidance. In *ECCV*. Springer, 257–275.
- [20] Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. 2024. Sugarcrape: Fixing hackable benchmarks for vision-language compositionality. *NIPS* 36 (2024).
- [21] Libo Huang, Yan Zeng, Chuanguang Yang, Zhulin An, Boyu Diao, and Yongjun Xu. 2024. eTag: Class-Incremental Learning via Embedding Distillation and Task-Oriented Generation. In *AAAI*, Vol. 38. 12591–12599.
- [22] Yufeng Huang, Jiji Tang, Zhuo Chen, Rongsheng Zhang, Xinfeng Zhang, Weijie Chen, Zeng Zhao, Zhou Zhao, Tangjie Lv, Zhipeng Hu, et al. 2024. Structure-CLIP: Towards Scene Graph Knowledge to Enhance Multi-Modal Structured Representations. In *AAAI*, Vol. 38. 2417–2425.
- [23] Xiaoli Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2019. Tinybert: Distilling bert for natural language understanding. *arXiv preprint arXiv:1909.10351* (2019).
- [24] 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* (2023).
- [25] Amita Kamath, Cheng-Yu Hsieh, Kai-Wei Chang, and Ranjay Krishna. 2024. The hard positive truth about vision-language compositionality. In *ECCV*. Springer, 37–54.
- [26] Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *CVPR*. 3128–3137.
- [27] Prannay Kaul, Weidi Xie, and Andrew Zisserman. 2023. Multi-modal classifiers for open-vocabulary object detection. In *ICML*. PMLR, 15946–15969.
- [28] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 2013. 3d object representations for fine-grained categorization. In *ICCV*. 554–561.
- [29] Alex Krizhevsky, Geoffrey Hinton, et al. 2009. Learning multiple layers of features from tiny images. (2009).
- [30] Janghyeon Lee, Jongxu Kim, Hyounguk Shon, Bumsoo Kim, Seung Hwan Kim, Honglak Lee, and Junmo Kim. 2022. Uniclip: Unified framework for contrastive language-image pre-training. *NIPS* 35 (2022), 1008–1019.
- [31] Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and René Ranftl. 2022. Language-driven semantic segmentation. *arXiv preprint arXiv:2201.03546* (2022).
- [32] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*. PMLR, 19730–19742.
- [33] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*. PMLR, 12888–12900.
- [34] Yanghao Li, Haoqi Fan, Ronghang Hu, Christoph Feichtenhofer, and Kaiming He. 2023. Scaling language-image pre-training via masking. In *CVPR*. 23390–23400.
- [35] Zheng Li, Xiang Li, Lingfeng Yang, Borui Zhao, Renjie Song, Lei Luo, Jun Li, and Jian Yang. 2023. Curriculum temperature for knowledge distillation. In *AAAI*, Vol. 37. 1504–1512.
- [36] Zheng Li, Jingwen Ye, Mingli Song, Ying Huang, and Zhigeng Pan. 2021. Online knowledge distillation for efficient pose estimation. In *ICCV*. 11740–11750.
- [37] Jiacheng Liu, Wenya Wang, Dianzhuo Wang, Noah A Smith, Yejin Choi, and Hannaneh Hajishirzi. 2023. Vera: A general-purpose plausibility estimation model for commonsense statements. *arXiv preprint arXiv:2305.03695* (2023).
- [38] Chufan Ma, Yi Jiang, Xin Wen, Zehuan Yuan, and Xiaojuan Qi. 2023. Codet: Co-occurrence guided region-word alignment for open-vocabulary object detection. *NIPS* 36 (2023), 71078–71094.
- [39] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. 2013. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151* (2013).
- [40] Oscar Mañas, Pau Rodriguez, Saba Ahmadi, Aida Nematzadeh, Yash Goyal, and Aishwarya Agrawal. 2022. Mapl: Parameter-efficient adaptation of unimodal pre-trained models for vision-language few-shot prompting. *arXiv preprint arXiv:2210.07179* (2022).
- [41] Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, et al. 2022. Simple open-vocabulary object detection. In *ECCV*. Springer, 728–755.
- [42] John X Morris, Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin, and Yanjun Qi. 2020. Textattack: A framework for adversarial attacks, data augmentation, and adversarial training in nlp. *arXiv preprint arXiv:2005.05909* (2020).
- [43] Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. 2022. Slip: Self-supervision meets language-image pre-training. In *ECCV*. Springer, 529–544.
- [44] Maria-Elena Nilsback and Andrew Zisserman. 2008. Automated flower classification over a large number of classes. In *2008 Sixth Indian conference on computer vision, graphics & image processing*. IEEE, 722–729.
- [45] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748* (2018).
- [46] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748* (2018).
- [47] Letitia Parcalabescu, Michele Cafagna, Lilita Muradjan, Anette Frank, Lacer Calixto, and Albert Gatt. 2021. VALSE: A task-independent benchmark for vision and language models centered on linguistic phenomena. *arXiv preprint arXiv:2112.07566* (2021).
- [48] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. 2012. Cats and dogs. In *CVPR*. IEEE, 3498–3505.
- [49] Wujian Peng, Sicheng Xie, Zuyao You, Shiyi Lan, and Zuxuan Wu. 2024. Synthesize Diagnose and Optimize: Towards Fine-Grained Vision-Language Understanding. In *CVPR*. 13279–13288.
- [50] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *ICCV*. 2641–2649.
- [51] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark,

- et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*. PMLR, 8748–8763.
- [52] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*. PMLR, 8748–8763.
- [53] 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 *CVPR*. 5238–5248.
- [54] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhrojit Som, et al. 2022. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. *arXiv preprint arXiv:2208.10442* (2022).
- [55] Kan Wu, Houwen Peng, Zhenghong Zhou, Bin Xiao, Mengchen Liu, Lu Yuan, Hong Xuan, Michael Valenzuela, Xi Stephen Chen, Xinggang Wang, et al. 2023. Tinyclip: Clip distillation via affinity mimicking and weight inheritance. In *ICCV*. 21970–21980.
- [56] Sizhe Wu, Wenwei Zhang, Lumin Xu, Sheng Jin, Xiangtai Li, Wentao Liu, and Chen Change Loy. 2024. CLIPself: Vision Transformer Distills Itself for Open-Vocabulary Dense Prediction. In *The Twelfth International Conference on Learning Representations*. <https://openreview.net/forum?id=DjzvJCRsVf>
- [57] Yao Wu, Mingwei Xing, Yachao Zhang, Yuan Xie, and Yanyun Qu. 2024. Clip2uda: Making frozen clip reward unsupervised domain adaptation in 3d semantic segmentation. In *ACM MM*. 8662–8671.
- [58] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. 2010. Sun database: Large-scale scene recognition from abbey to zoo. In *CVPR*. IEEE, 3485–3492.
- [59] Kaicheng Yang, Jiankang Deng, Xiang An, Jiawei Li, Ziyong Feng, Jia Guo, Jing Yang, and Tongliang Liu. 2023. Alip: Adaptive language-image pre-training with synthetic caption. In *ICCV*. 2922–2931.
- [60] Kaicheng Yang, Tiancheng Gu, Xiang An, Haiqiang Jiang, Xiangzi Dai, Ziyong Feng, Weidong Cai, and Jiankang Deng. 2024. Clip-cid: Efficient clip distillation via cluster-instance discrimination. *AAAI* (2024).
- [61] Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. 2021. Filip: Fine-grained interactive language-image pre-training. *arXiv preprint arXiv:2111.07783* (2021).
- [62] Qihang Yu, Ju He, Xueqing Deng, Xiaohui Shen, and Liang-Chieh Chen. 2023. Convolutions die hard: Open-vocabulary segmentation with single frozen convolutional clip. *NIPS* 36 (2023), 32215–32234.
- [63] Mert Yuksekoglu, 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? *arXiv preprint arXiv:2210.01936* (2022).
- [64] Yuhang Zang, Wei Li, Kaiyang Zhou, Chen Huang, and Chen Change Loy. 2022. Open-vocabulary detr with conditional matching. In *ECCV*. Springer, 106–122.
- [65] Yan Zeng, Xinsong Zhang, and Hang Li. 2021. Multi-Grained Vision Language Pre-Training: Aligning Texts with Visual Concepts. *arXiv preprint arXiv:2111.08276* (2021).
- [66] Yan Zeng, Xinsong Zhang, and Hang Li. 2021. Multi-grained vision language pre-training: Aligning texts with visual concepts. *arXiv preprint arXiv:2111.08276* (2021).
- [67] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *ICCV*. 11975–11986.
- [68] Le Zhang, Rabiu Awal, and Aishwarya Agrawal. 2024. Contrasting Intra-Modal and Ranking Cross-Modal Hard Negatives to Enhance Visio-Linguistic Compositional Understanding. In *CVPR*. 13774–13784.
- [69] Lu Zhang, Ke Yan, and Shouhong Ding. 2024. AlignCLIP: Align Multi Domains of Texts Input for CLIP models with Object-IoU Loss. In *ACM MM*. 1092–1100.
- [70] 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* (2023).

Dataset	Classes	Train size	Test size	Evaluation metric
Food101	102	75,750	25,250	accuracy
CIFAR10	10	50,000	10,000	accuracy
CIFAR100	100	50,000	10,000	accuracy
SUN397	397	19,850	19,850	accuracy
Cars	196	8,144	8,041	accuracy
Aircraft	100	6,667	3,333	mean per class
DTD	47	3,760	1,880	accuracy
Pets	37	3,680	3,669	mean per class
Caltech101	101	3,000	5,677	mean-per-class
Flowers	102	2,040	6,149	mean per class
ImageNet	1000	1,281,167	50,000	accuracy

**Table 9: List of linear probe datasets with the data distribution and evaluation metrics.**

Dataset	Test Images	Captions per Image	Evaluation Protocol
MSCOCO	5,000	5	Image-to-Text & Text-to-Image
Flickr30k	1,000	5	Image-to-Text & Text-to-Image

**Table 10: Zero-shot image-text retrieval evaluation settings.**

## A Supplementary Material

### A.1 Training Details

Details of the hyperparameter configurations are presented in Table 8. An asterisk (\*) indicates that the batch size for our positive image-text pairs is 256, and each positive text is paired with four negative texts during training.

Hyperparameters	
Batch size	256*
Optimizer	AdamW
Weight decay	0.1
Adam $\beta$	(0.9, 0.98)
Adam $\epsilon$	1e-6
Learning rate	1e-6
Learning rate schedule	cosine decay
Ema $\alpha$	0.9996
$(\lambda_1, \lambda_2, \lambda_3)$	(0.1, 0.1, 0.005)
Epochs	5
Training GPUs	8×V100

**Table 8: Detailed hyper-parameters for training DeGLA.**

### A.2 Details of General Understanding Downstream Datasets

**Zero-shot Classification.** Following the previous work RWKV-CLIP [14], we evaluate the zero-shot classification performance of the models on 11 datasets. The prompt used in zero-shot classification is presented in Table 13.

**Linear Probe.** The datasets used for the linear probe evaluation are the same as those used for zero-shot classification. Details on each dataset and the corresponding evaluation metrics are provided in Table 9.

**Zero-shot Image-Text Retrieval.** The detail of zero-shot image-text retrieval benchmark is presented in Table 10. We evaluate retrieval performance on MS-COCO and Flickr30k following standard protocols. For MS-COCO, we use the 1K test set from [26],

while Flickr30k employs 1,000 images with 5 captions each. In both datasets, we assess bidirectional retrieval (image-to-text and text-to-image) using pre-trained model embeddings without fine-tuning. Performance is measured by Recall@K (R@1, R@5, R@10), representing the percentage of correct matches in the top-K results.

### A.3 Details of Compositional Reasoning Benchmarks

In Table 11, we summarize an overview of the three compositional reasoning benchmarks employed in this work.

**ARO.** ARO [63] benchmark is used to probe the VLM’s understanding of relations, attributes, and order in image-text data. Specifically, the Relation task involves swapping objects in the text, the Attribute task involves switching attributes in the text, and the Order task involves disrupting the sequence of the entire sentence.

**VALSE.** The VASLE [47] benchmark assesses the compositional understanding of Visual Language Models (VLMs) across six dimensions: existence, plurality, counting, spatial relations, actions, and entity coreference. The existence subset employs a single cue to evaluate whether models can discern the presence or absence of specific entities in images. The plurality subset similarly uses a single cue to determine if models can differentiate between singular and plural noun phrases depicted in images, such as “exactly one flower” vs. “some flowers”. The counting subset, which includes balanced, adversarial, and small-number scenarios, tests the models’ ability to confirm the accuracy of a stated number of entities against those shown in the image. The spatial relations subset uses a single cue to assess model performance in identifying different spatial relationships, creating foils by modifying the spatial preposition in the original caption. The actions subset, encompassing action replacement and actant swap settings, evaluates the models’ capability to (i) match the described action with the depicted action, and (ii) accurately identify the agents performing and receiving the action. Finally, the coreference subset examines the models’ proficiency in resolving pronominal references within multimodal contexts, covering both pronouns linked to noun phrases visually grounded in the image and deictic or image-level references.

**SugarCrepe.** The SugarCrepe [47] benchmark encompasses three modalities: Replace, Swap, and Add, enhancing the assessment of visual-linguistic compositional understanding. In the Replace modality, a hard negative is generated by substituting a single atomic concept in a positive caption, which then mismatches with the corresponding image. We classify the replacements according to their conceptual types: REPLACE-OBJ, REPLACE-ATT, and REPLACE-REL. The Swap modality produces hard negatives by interchanging two atomic concepts of the same type within a positive caption, maintaining the original content. This includes SWAP-OBJ and SWAP-ATT categories, excluding relation swaps to avoid generating incoherent text. The Add modality creates hard negatives by introducing a new atomic concept into a positive caption, leading to a discrepancy with the image. It includes ADD-OBJ and ADD-ATT, while relation insertions are excluded due to their implausibility.

### A.4 Details of Negative Caption Generation

Following CE-CLIP [68], we adopt 2014 train split of COCO [3] as our base dataset and employ the LLM-driven negative caption

Negative type	Prompt for ChatGPT	Prompt for LLaMA3.1-8B-instruct
Intra-sentence reshuffle	Please generate 200 high-quality rewritten examples for everyday life scenarios based on the following rule: <Reshuffle the overall order of the sentence>, ensuring that the rewritten sentences have significantly different or opposite meanings from the original sentences.	Please rewrite the following input caption based on the rewrite rule and examples: <Reshuffle the overall order of the sentence>. Here are the examples: <example1>,<example2>,<example3>,<example4>,<example5>. ensuring that the rewritten sentences have significantly different or opposite meanings from the original sentences. Input text: <input text>
	Please generate 200 high-quality rewritten examples for everyday life scenarios based on the following rule: <Swap the nouns in the sentence>, ensuring that the rewritten sentences have significantly different or opposite meanings from the original sentences.	Please rewrite the following input caption based on the rewrite rule and examples: <Swap the nouns in the sentence>. Here are the examples: <example1>,<example2>,<example3>,<example4>,<example5>. ensuring that the rewritten sentences have significantly different or opposite meanings from the original sentences. Input text: <input text>
	Please generate 200 high-quality rewritten examples for everyday life scenarios based on the following rule: <Swap the adjectives in the sentence>, ensuring that the rewritten sentences have significantly different or opposite meanings from the original sentences.	Please rewrite the following input caption based on the rewrite rule and examples: <Swap the adjectives in the sentence>. Here are the examples: <example1>,<example2>,<example3>,<example4>,<example5>. ensuring that the rewritten sentences have significantly different or opposite meanings from the original sentences. Input text: <input text>
Minimal semantic substitution	Please generate 200 high-quality rewritten examples for everyday life scenarios based on the following rule: <Replace the adjectives in the sentence>, ensuring that the rewritten sentences have significantly different or opposite meanings from the original sentences.	Please rewrite the following input caption based on the rewrite rule and examples: <Replace the adjectives in the sentence>. Here are the examples: <example1>,<example2>,<example3>,<example4>,<example5>. ensuring that the rewritten sentences have significantly different or opposite meanings from the original sentences. Input text: <input text>
	Please generate 200 high-quality rewritten examples for everyday life scenarios based on the following rule: <Replace the nouns in the sentence>, ensuring that the rewritten sentences have significantly different or opposite meanings from the original sentences.	Please rewrite the following input caption based on the rewrite rule and examples: <Replace the nouns in the sentence>. Here are the examples: <example1>,<example2>,<example3>,<example4>,<example5>. ensuring that the rewritten sentences have significantly different or opposite meanings from the original sentences. Input text: <input text>

Figure 8: Details of the prompt utilized for generating high-quality examples and negative captions

Benchmark	#Examples	Compositionalablity
ARO [63]	82695	Relation,Attribution,Order
VALSE [47]	8309	Existence, Plurality, Counting, Spatial Relations, Actions,Entity Coreference
SugarCrepe [20]	5948	Relation,Attribution,Order,Semantic Substitution

Table 11: Details of compositional reasoning benchmarks.

Method	Pretrain	Finetune	Hard negative data	Finegrained loss	Hard negative type			Self distillation
					Rule-based	Un-masking based	LLM-based	
CLIP [51]	✓							
NegCLIP [63]		✓	✓		✓			
Structure-CLIP [22]		✓	✓		✓			
CE-CLIP [68]		✓	✓	✓	✓	✓		
DeGLA (ours)	✓	✓	✓	✓			✓	✓

Table 12: Comparison of different methods.

generation pipeline to generate compositionally-enhanced data for fine-tuning. The base dataset has 83k images and 414K captions. Through our augmentation process, we generate five hard negative captions per original caption, resulting in a total of 2.07 million hard negative captions.

**Prompt Details.** In Figure 8, we detail the prompts employed in this study. Notably, the prompts we designed are concise and effective, especially in ensuring that the generated sentences are “hard negative” rather than “hard positive”. To achieve this, we append the instruction “ensuring that the rewritten sentences have significantly different or opposite meanings from the original sentences”

at the end of the prompt, which greatly enhanced the quality of both the examples and the final rewritten captions.

**Examples Generated by ChatGPT.** In Figure 9, we present several representative examples generated by ChatGPT. These examples closely adhere to our rules and serve as high-quality exemplars to guide the large-scale generation of negative samples.

**Rewritten Captions.** In Figure 10, we present examples of negative samples generated by LLaMA3.1-instruct-8B. These examples demonstrate the high quality of captions produced by the LLM-driven negative caption generation pipeline. The captions adhere

Original Text	Rewritten examples	Rewritten Text
The sun rises over the mountains, casting a golden glow.	→	Over a sun casting glow. rises The the golden mountains.
A squirrel darted across the park, searching for acorns.	→	Searching for across A acorns. darted squirrel the park.
The fluffy white cat sat on the soft blue cushion.	→	The matted gray cat sat on the hard red cushion.
A reliable electric car passed the broken gasoline truck.	→	A faulty diesel car trailed the reliable hybrid truck.
The fluffy white cat chased the sleek black dog.	→	The sleek white cat chased the fluffy black dog.
Her smooth silk dress touched the rough wool sweater.	→	Her rough silk dress touched the smooth wool sweater.
The cat chased the dog around the garden.	→	The dog chased the cat around the garden.
A student borrowed a book from the library.	→	A book borrowed a student from the library.
The dog uncovered the bone in the yard.	→	The cat uncovered the bone in the yard.
The rain canceled the outdoor concert.	→	The wind canceled the outdoor concert.

Figure 9: Rewritten examples generated by ChatGPT.

Positive caption	Hard negative examples	Negative caption
 <p>A gray cat is sitting on a wooden bench. On a wooden bench sitting is cat gray a. A gray cat is sitting on a wooden bench. A white cat is sitting on a metal bench. A wooden bench is sitting on a gray cat. A goldfish is sitting on a concrete road.</p>	 <p>A white kite flying over a large field. Field large a over kite flying white a. A large kite flying over a white field. A dark kite flying over a small pond. A field flying over a large kite. A black anchor sinking into a small pond.</p>	 <p>A clean, mediocre motel bathroom with a nice sink. Sink a nice, with mediocre bathroom motel clean, a. A nice, clean motel bathroom with a mediocre sink. A dirty, exceptional motel bathroom with a broken sink. A nice sink with a clean, mediocre motel bathroom. A dirty, luxurious mansion with a broken sink.</p>
	 <p>Two young children playing around a fire hydrant. Fire hydrant around playing a young two children. Two old children playing around a young fire hydrant. Two elderly adults lounging around a broken water fountain. Fire hydrant playing around two young children. Two elderly bureaucrats playing around a freshwater spring.</p>	 <p>A snowy landscape surrounded by lifeless trees and a bench. Trees lifeless by surrounded and bench a snowy landscape. A lifeless landscape surrounded by snowy trees and a bench. A barren landscape surrounded by lush greenery and a hammock. A bench surrounded by lifeless trees and a snowy landscape. A barren wasteland surrounded by thriving gardens and a trash can.</p>

Figure 10: Hard negative examples generated by LLaMA3.1-8B-instruct.

accurately to our rewriting rules, ensuring that the generated sentences possess meanings that are distinct from or opposite to the

original ones. This efficacy can be attributed to the high-quality examples showcased in Figure 9 and the meticulously designed prompts outlined in Figure 8.

**CIFAR 10 & CIFAR 100**

a photo of a {label}.	a blurry photo of a {label}.	a black and white photo of a {label}.	a low contrast photo of a {label}.
a high contrast photo of a {label}.	a bad photo of a {label}.	a good photo of a {label}.	a photo of a small {label}.
a photo of a big {label}.	a photo of the {label}.	a blurry photo of the {label}.	a black and white photo of the {label}.
a low contrast photo of the {label}.	a high contrast photo of the {label}.	a bad photo of the {label}.	a good photo of the {label}.
a photo of the small {label}.	a photo of the big {label}.		

**Food101**

a photo of {label}, a type of food.

**Caltech101**

a photo of a {label}.	a painting of a {label}.	a plastic {label}.	a sculpture of a {label}.
a sketch of a {label}.	a tattoo of a {label}.	a toy {label}.	a rendition of a {label}.
a embroidered {label}.	a cartoon {label}.	a {label} in a video game.	a plushie {label}.
an origami {label}.	art of a {label}.	graffiti of a {label}.	a drawing of a {label}.
a doodle of a {label}.	a photo of the {label}.	a painting of the {label}.	the plastic {label}.
a sculpture of the {label}.	a sketch of the {label}.	a tattoo of the {label}.	the toy {label}.
a rendition of the {label}.	the embroidered {label}.	the cartoon {label}.	the {label} in a video game.
the plushie {label}.	the origami {label}.	art of the {label}.	graffiti of the {label}.
a drawing of the {label}.	a doodle of the {label}.		

**Stanford Cars**

a photo of a {label}.	a photo of the {label}.	a photo of my {label}.	i love my {label}!
a photo of my dirty {label}.	a photo of my clean {label}.	a photo of my new {label}.	a photo of my old {label}.

**DTD**

a photo of a {label} texture.	a photo of a {label} pattern.	a photo of a {label} thing.	a photo of a {label} object.
a photo of the {label} texture.	a photo of the {label} pattern.	a photo of the {label} thing.	a photo of the {label} object.

**FGVC Aircraft**

a photo of a {label}, a type of aircraft.	a photo of the {label}, a type of aircraft.
---	---

**Flowers102**

a photo of a {label}, a type of flower.
---

**Pets**

a photo of a {label}, a type of pet.
--------------------------------------

**SUN39**

a photo of a {label}.	a photo of the {label}.
-----------------------	-------------------------

**ImageNet**

a bad photo of a {label}.	a rendering of a {label}.	graffiti of a {label}.	a bad photo of the {label}.
a low resolution photo of the {label}.	a tattoo of a {label}.	the embroidered {label}.	a photo of a hard to see {label}.
a cropped photo of the {label}.	a photo of a clean {label}.	a photo of a dirty {label}.	a dark photo of the {label}.
a bright photo of a {label}.	a photo of my {label}.	the plastic {label}.	a photo of the cool {label}.
a drawing of a {label}.	a black and white photo of the {label}.	a painting of the {label}.	a painting of a {label}.
a close-up photo of a {label}.	a sculpture of the {label}.	a bright photo of the {label}.	a cropped photo of a {label}.
a pixelated photo of the {label}.	a photo of the dirty {label}.	a jpeg corrupted photo of a {label}.	a blurry photo of the {label}.
a plastic {label}.	a good photo of the {label}.	a rendering of the {label}.	a {label} in a video game.
a photo of the {label}.	a doodle of a {label}.	a close-up photo of the {label}.	a photo of a {label}.
a photo of one {label}.	the {label} in video game.	a sketch of a {label}.	a doodle of the {label}.
the origami {label}.	a low resolution photo of a {label}.	the toy {label}.	a rendition of the {label}.
an origami {label}.	a photo of a large {label}.	a rendition of a {label}.	a photo of a nice {label}.
a photo of the clean {label}.	a blurry photo of a {label}.	a cartoon {label}.	art of a {label}.
a photo of a weird {label}.	a embroidered {label}.	a pixelated photo of a {label}.	itap of the {label}.
a sketch of the {label}.	a good photo of a {label}.	a plushie {label}.	a photo of the nice {label}.
a jpeg corrupted photo of the {label}.	a photo of the weird {label}.	the cartoon {label}.	art of the {label}.
a photo of the small {label}.	a photo of the large {label}.	a black and white photo of a {label}.	the plushie {label}.
a drawing of the {label}.	itap of a {label}.	graffiti of the {label}.	a toy {label}.
a dark photo of a {label}.	a photo of a cool {label}.	a photo of a small {label}.	a tattoo of the {label}.
itap of my {label}.			

**Table 13: Full list of prompts to evaluate the performance of zero-shot classification on 11 visual recognition datasets.**