Contrasting intra-modal and ranking cross-modal hard negatives to enhance visio-linguistic compositional understanding

Le Zhang[1,2] Rabiul Awal[1] Aishwarya Agrawal[1,2,3]
Mila - Quebec AI Institute[1]
Université de Montréal[2]
Canada CIFAR AI Chair[3]

Abstract

Vision-Language Models (VLMs), such as CLIP, exhibit strong image-text comprehension abilities, facilitating advances in several downstream tasks such as zero-shot image classification, image-text retrieval, and text-to-image generation. However, the compositional reasoning abilities of existing VLMs remains subpar. The root of this limitation lies in the inadequate alignment between the images and captions in the pretraining datasets. Additionally, the current contrastive learning objective fails to focus on fine-grained grounding components like relations, actions, and attributes, resulting in "bag-of-words" representations. We introduce a simple and effective method to improve compositional reasoning in VLMs. Our method better leverages available datasets by refining and expanding the standard image-text contrastive learning framework. Our approach does not require specific annotations and does not incur extra parameters. When integrated with CLIP, our technique yields notable improvement over state-of-the-art baselines across five vision-language compositional benchmarks. [1]

1. Introduction

The field of vision-language research has experienced remarkable progress over recent years, thanks to the introduction of vast datasets [14, 56], the adaptation of attention mechanism, and the pioneering objectives such as contrastive learning. Impressively, these models demonstrate a notable capability in zero-shot generalization, as seen in areas like Visual Question Answering (VQA) [18], captioning [1, 33, 42], and image-text retrieval [52, 62, 71]. Strong Vision-Language Models (VLMs), such as CLIP [52], are even pushing the boundaries in text-to-image generation (CLIP is used to guide image generation given the input prompt) [50, 53, 55]. However, despite these advances, a notable limitation persists: these models often miss the intricate compositional nuances of relationships, attributes, objects, and actions [63, 70]. A clear manifestation of this shortcoming is their difficulty in distinguishing between captions with the same set of words but composed differently like "Horse is eating the grass" and "Grass is eating the horse" [70] when paired with relevant images. Such compositional understanding remains a critical frontier for continued advancement in vision-language integration.

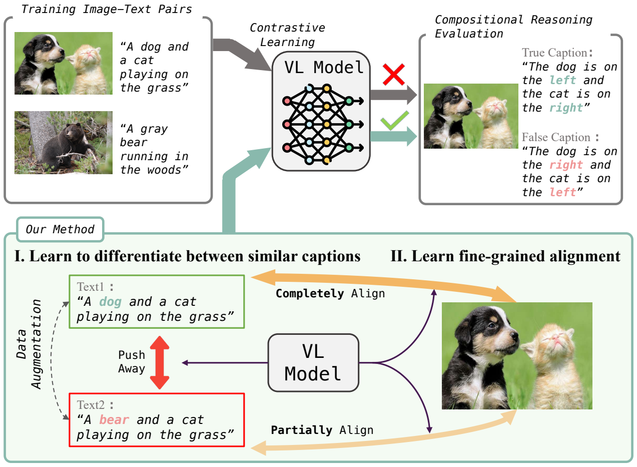

Training Image–Text Pairs
"A dog and a cat playing on the grass"
"A gray bear running in the woods"
Contrastive Learning
VL Model
Compositional Reasoning Evaluation
True Caption: "The dog is on the left and the cat is on the right"
False Caption : "The dog is on the right and the cat is on the left"
Our Method
I. Learn to differentiate between similar captions
II. Learn fine-grained alignment
Data Augmentation
Text1 : "A dog and a cat playing on the grass"
Push Away
Text2 : "A bear and a cat playing on the grass"
VL Model
Completely Align
Partially Align

Figure 1. Models trained with standard image-text contrastive learning lack sufficient compositional reasoning abilities. Our method teaches the model to better differentiate between similar captions and learn fine-grained alignment between images and text to improve compositional reasoning.

A primary factor impeding compositional understanding in current VLMs stems from their learning methodology and training dataset. These models are usually trained on huge image-text pairs crawled from the web using contrastive learning [52]. The caption is short and noisy; the image-text contrastive objective optimizes the model to distinguish between correct image-text pairs and a vast array of incorrect ones. However, because the incorrect pairs are often markedly distinct, the model primarily distinguishes them

[1] We open-source our code at https://github.com/lezhang7/Enhance-FineGrained.

through simple object recognition, without needing to comprehend fine-grained details such as attributes and relations. Fig. 1 depicts a scenario where CLIP struggles with the compositional reasoning of “left” and “right” concepts.

Earlier studies, including NegCLIP [70], have employed phrase swapping to produce additional captions for training. This underscores the importance of incorporating hard negatives in vision-language contrastive learning. However, simply incorporating additional samples into standard image-text contrastive learning does not fully leverage hard negatives. In this work, we refine and expand the contrastive learning objective for hard negative captions (see Fig. 1), which vary in semantics like relations, attributes, actions, and objects. We focus on two dimensions. First, we advocate for a clearer distinction in the representations of positive and hard-negative captions, aiming to boost the model’s ability to recognize nuanced semantic variations. Second, we maintain a minimum similarity gap between authentic image-text pairs and their challenging hard-negative counterparts to encourage the learning of fine-grained image-text alignment. Consequently, we propose two objectives: i) intra-modal contrast, and ii) cross-modal rank, built on the hinge loss [9] approach. The latter incorporates an adaptive threshold during the fine-tuning phase. This means as the model becomes more adept, the threshold increases, reflecting both the growing difficulty of the task and the model’s increasing competency. This approach not only resonates with curriculum learning principles but also ensures a more stable training process.

To validate the effectiveness, we conduct experiments on two models: the versatile CLIP and the strong X-VLM [71]. Our evaluation across various compositional datasets consistently reveals performance enhancements, establishing our method as a new *state-of-the-art* across all assessed benchmarks. Specifically, training CLIP with our method on the COCO dataset leads to an improvement of 23.7% and 13.5% respectively on the Relation and Attribution splits of the ARO benchmark [70], 7.2% on the VALSE benchmark [49], 5.9% on the VL-CheckList benchmark [74], and a significant improvement of 12.1% on the recently developed SugarCrepe benchmark [24]. We also achieve modest improvements of 0.5%, 2.5% respectively on the ARO Relation and Attribution splits, 1.3% on VALSE and 2.1% on VL-CheckList on top of the already strong X-VLM model upon application of our method. Finally we also evaluate our method on the conventional image-text retrieval and image classification benchmarks, resulting in 7.5% improvement in image-text retrieval and a small 1.6% decrease in image classification.

To summarize, we present three key contributions: (1) We propose a simple yet effective solution to better leverage available image-text datasets to improve VLMs’ compositional understanding without introducing any additional parameters. This is achieved by extending the contrastive learning framework: introducing intra-modal contrast and cross-modal rank objectives. (2) Our adaptive threshold strategy induces curriculum learning during fine-tuning, leading to improved results and stable training without the need for labour-some and time-consuming parameter tuning. (3) We demonstrate the effectiveness of our approach through its state-of-the-art performance on five benchmarks. Furthermore, we conduct a thorough analysis of each component of our model, providing insights for future research and a deeper understanding of our methodology through extensive experiments.

2. Related Work

Contrastive Vision-Language Models Vision-language models have garnered remarkable success in both the vision and multimodal domains. Modern VLMs are pretrained on large and noisy multi-modal datasets [56, 57] and then applied to downstream tasks in a zero-shot manner. Among them, CLIP [52] stands out, employing a contrastive learning method for pretraining. Our reasons to focus on CLIP are twofold: firstly, image-text contrastive learning has become a prevalent strategy for VLM pretraining [25, 59, 61, 69, 71, 72]; secondly, CLIP boasts extensive applicability, spanning various domains. This includes zero-shot image classification [15, 44, 47, 77], object detection [45], semantic segmentation [64, 68, 76, 78], text-image retrieval, evaluation of text-image alignment [8, 22], and text-to-image generation [50, 53, 55]. Furthermore, the vision encoder from CLIP can serve as a strong backbone for generative vision-language models [2, 33, 37, 80]. Therefore, enhancements on CLIP can effectively radiate to a broader range of vision-language applications.

Vision-Language Compositionality While Vision-Language Models exhibit remarkable strength in handling multimodal data, recent investigations suggest that these models tend to learn a “bag of words” representation, which hampers their compositional understanding [12, 70]. A number of benchmarks have emerged to evaluate the performance of VLMs, focusing on various dimensions like relations, attributes, objects, among others. For instance, ARO [70] emphasizes the understanding of attributes and relations, while VL-checklist [74] drills down into finer subcategories such as size, color, action, and spatial relations. VALSE [49] targets linguistic phenomena like existence, counting, plurality, and coreference, whereas Winoground [63] delves into complex reasoning, encompassing commonsense and external knowledge. SugarCrepe [24] aims to address the hackability issue where pure-text models without image information can outshine robust VLMs on several compositional benchmarks, attributing to a significant distribution gap between positive and hard

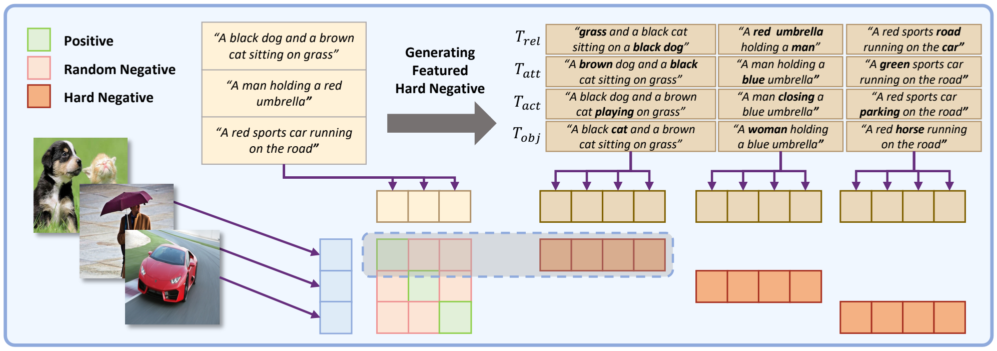

Positive
Random Negative
Hard Negative
"A black dog and a brown cat sitting on grass"
"A man holding a red umbrella"
"A red sports car running on the road"
Generating Featured Hard Negative
T_{rel}
T_{att}
T_{act}
T_{obj}
"grass and a black cat sitting on a black dog"
"A red umbrella holding a man"
"A red sports road running on the car"
"A brown dog and a black cat sitting on grass"
"A man holding a blue umbrella"
"A green sports car running on the road"
"A black dog and a brown cat playing on grass"
"A man closing a blue umbrella"
"A red sports car parking on the road"
"A black cat and a brown cat sitting on grass"
"A woman holding a blue umbrella"
"A red horse running on the road"

Figure 2. (Top) An overview of our method's pipeline and hard negative generation examples. Losses are applied on the shaded boxes.

negative captions. All these benchmarks are structured as cross-modal retrieval tasks – discern between correct and incorrect captions given an image, and evaluations are based on accuracy metrics.

The quest to augment VLMs' compositional understanding has ignited substantial interest within the community. The DAC approach [13] proposes to enhance caption density and quality by utilizing an off-the-shelf caption model [33] and a segmentation model [28]. Conversely, SGVL [21] and MosaiCLIP [60] employ additional scene graph annotations to guide model learning on compositional relations. Although these methods demonstrate effectiveness, they necessitate either a specific model (like a segmentation model) or additional annotations (such as a scene graph). A distinct line of research explores hard negative mining methodology [27], where SLVC [12], Paiss et al. [48] and NegCLIP [70] enrich samples with negative text via random word-swapping. We perceive negative augmentation as a refined method since it does not hinge on extra resources (model or data) and postulate that the current methodologies do not entirely harness the potential of hard negative mining, and thus, we introduce two additional losses atop our featured hard negatives to further bolster the compositional understanding capability.

3. Method

In the proposed method, we expand upon image-text contrastive learning and introduce two loss functions specifically applied to the automatically generated hard negatives. In this section, we first discuss the process of hard negative generation, followed by a detailed description of our loss functions. Fig. 2 illustrates the overview of pipeline and Fig. 3 illustrates proposed losses.

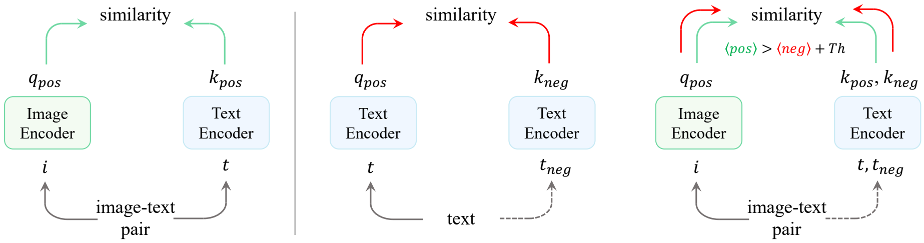
3.1. Featured Hard Negative Generation

In contrastive learning, *hard negatives* refer to instances that exhibit high similarity to positive samples, yet do not qualify as positive themselves. Consider the following caption as an example: "A gray cat sits on top of a **wooden** chair near a plant." A potential hard negative could be: "A gray cat sits on top of a **plastic** chair near a plant." [12] While the hard negative correctly identifies the majority of elements in the image, it diverges from the positive sample with regards to the chair's material. Incorporating hard negatives into the training process can enable models to discern subtle distinctions, thereby enhancing their overall accuracy and performance [16, 20, 26, 41, 51, 54].

To bolster the compositional understanding of our model, we deliberately create hard negatives that embody various alterations to the original captions. These adjustments encompass changes in the relationship, attributes, and action of the image's objects. Furthermore, we produce hard negatives where we replace an object name with another, encouraging the model to distinguish between different objects. To generate these hard negatives, we employ Part-Of-Speech (POS) parsing and Language Models. Utilizing Spacy [23], we parse the captions and assign POS tags to each word. For relational hard negative, we interchange the positions of two noun words. For attribution, action, and object name alterations, we randomly mask an adjective, verb, or noun word, and subsequently fill in the masked area using the RoBERTa [39], examples are shown in Fig 2. For each caption, we generate all four types of hard negatives, replacing any examples in which the requisite words or two objects are absent from the caption with a placeholder string. This approach ensures a comprehensive and robust training dataset for enhancing our model's performance.

3.2. Expanded Losses

Preliminaries Contrastive VLMs consist of a image encoder $f_i : X_{image} \longrightarrow \mathbb{R}^d$ and a text encoder $f_t : X_{text} \longrightarrow \mathbb{R}^d$. The cosine similarity between two inputs I, T using their encoders f_i, f_t are computed as: $S(I,T) = \frac{f_i(I)\cdot f_t(T)}{||f_i(I)||\cdot||f_t(T)||}/\tau$ where $\cdot$ represents inner product and τ is a trainable temperature parameter. The image-text contrastive loss is applied on the computed similarity. Considering image-text pairs (I, T) within a batch $\mathcal{B}$, the computation of the Image-Text Contrastive (ITC) loss is formulated as



(a) Image-text Contrast (b) Intra-Modal Contrast (c) Cross-Modal Rank

Figure 3. **Conceptual loss comparison**. Red arrows denote minimizing similarity, while green arrows denote maximize it; Dotted arrow represents data augmentation. (a) Standard image-text contrastive learning applied in [52]. (b) Proposed intra-modal contrast applied on generated hard negative texts and (c) cross-modal rank applied on positive and hard negative pairs with adaptive threshold.

follows:

$$\mathcal{L}_{itc} = \sum_{(I,T)\in\mathcal{B}} -\left(\log\frac{\exp^{S(I,T)}}{\sum\limits_{T_i\in\mathcal{B}}\exp^{S(I,T_i)}} + \log\frac{\exp^{S(I,T)}}{\sum\limits_{I_j\in\mathcal{B}}\exp^{S(I_j,T)}}\right) \tag{1}$$

For each image-text pair (I, T), prior research methodologies generate a single hard negative caption T_{hn} through the random swapping of a word. This generated caption is subsequently treated as an additional random negative [12, 48, 70]. Thus, the formulation of the Image-Text Contrastive loss with the inclusion of a hard negative can be described as follows:

$$\begin{aligned}\mathcal{L}_{itc(hn)} &= \sum_{(I,T)\in\mathcal{B}} -\log\frac{\exp^{S(I,T)}}{\sum\limits_{I_j\in\mathcal{B}}\exp^{S(I_j,T)}} \\ &+ \sum_{(I,T)\in\mathcal{B}} -\log\frac{\exp^{S(I,T)}}{\sum\limits_{T_i\in\mathcal{B}}\exp^{S(I,T_i)} + \sum\limits_{T_k\in\mathcal{T}_{hn}}\exp^{S(I,T_k)}}\end{aligned} \tag{2}$$

Intra-Modal Contrastive Adhering to the aforementioned notations and given an image-text pair (I, T) within batch $\mathcal{B}$, our method, as outlined in Section 3.1, generates four distinct hard negatives $\mathcal{T}_{hn} = \{T_{rel}, T_{att}, T_{act}, T_{obj}\}$ corresponding to changes in relation, attribute, action and object entity. The primary motivation behind employing intra-modal contrastive (IMC) loss is to promote the model's ability to differentiate between hard negative captions to the maximum extent and contrastive loss is well-suited for this purpose. Consequently, the formulation is:

$$\mathcal{L}_{imc} = \sum_{(I,T)\in\mathcal{B}} -\log\frac{1}{\sum\limits_{T_k\in\mathcal{T}_{hn}}\exp^{S(T,T_k)}} \tag{3}$$

Cross-Modal Rank with Adaptive Threshold Hard negative captions retain some elements of truth about the image, indicating a partial correctness in the image-text alignment. The model is designed to discern the similarity between a true image-text pair and a hard negative pair to a certain extent; i.e. it stops further optimization using hard negative pairs once the similarity difference exceeds a predefined threshold. To achieve this, we employ a ranking loss with a threshold. This threshold ensures that the similarity score for an image-text pair, $S(I, T)$, is greater than the similarity score for that image and any hard negative caption, $S(I, T_k)$, by at least a threshold value Th_k corresponding to the type of hard negative. This concept is formally represented as follows:

$$S(I,T) > \{S(I,T_k) + Th_k | T_k \in \mathcal{T}_{hn}\}$$

Inspired by the hinge loss concept [9], we employ this threshold in the loss function, which we call Cross-modal Rank (CMR) loss, defined as follows:

$$\mathcal{L}_{cmr} = \sum_{(I,T)\in\mathcal{B}}\sum_{T_k\in\mathcal{T}_{hn}} max(0, S(I,T_k) - S(I,T) + Th_k) \tag{4}$$

Determining an appropriate threshold for hinge loss is challenging [65]. Inspired by existing research on adaptive thresholds [6, 36, 73, 75] that posit that an effective threshold should evolve in accordance with the training progress, we adapt this principle to the multi-modal learning domain. Our approach models the threshold using the difference in the model's similarity scores between the true and *hard negatives* pairs, serving as a indicator of the model's compositional understanding capability. Especially during the initial training phase, when differentiating between the *hard negatives* and true pairs is tough, a lower threshold is appropriate. As training advances, and the model refines its understanding, this score disparity grows. This progressive threshold adaptation, aligning with curriculum learning principles, aims for smoother optimization, avoidance of local minima, and improved generalization [67]. Consequently, the threshold encapsulates both task intricacy and model proficiency. Thus, at training step t, the threshold for each type $\{k | k \in (rel, att, act, obj)\}$ is computed as:

$$Th_k^t = \frac{1}{|\mathcal{B}|}\sum_{(I,T)\in\mathcal{B}}(S^{t-1}(I,T) - S^{t-1}(I,T_k)) \tag{5}$$

Another unique aspect of our approach is that we implement distinct thresholds for different types of hard negatives, each tailored to a specific "curriculum", while most existing approaches utilizing adaptive thresholds in non-multimodal domains [6, 36, 73, 75] employ just one threshold. The adaptive Cross-modal Rank loss at step t is defined as:

$$\mathcal{L}_{cmr} = \sum_{(I,T)\in\mathcal{B}} \sum_{T_k\in\mathcal{T}_{hn}} max(0, S(I,T_k) - S(I,T) + Th_k^t) \quad (6)$$

Empirically, we find that adding the term $-S(T, T_{rel})$ to CMR offers benefits and without threshold constraints, the value of relation hard negatives escalates rapidly, hindering training. This is because these negatives, unlike others, are not formed by substituting words with feasible alternatives, leading to easily distinguishable, implausible sentences. Consequently, there is a marked difference in similarity scores. For stable training, an upper bound u on the threshold is crucial:

$$Th_k^t = min\left(u, \frac{1}{N}\sum_{(I,T)\in\mathcal{B}} (S^{t-1}(I,T) - S^{t-1}(I,T_k))\right) \quad (7)$$

Subsequently, incorporating the loss weight hyper-parameters α and β, the final loss function can be expressed as follows:

$$\mathcal{L} = \mathcal{L}_{itc(hn)} + \alpha \cdot \mathcal{L}_{imc} + \beta \cdot \mathcal{L}_{cmr} \quad (8)$$

4. Experiments

We assess the performance of our method using two models. Firstly, we employ CLIP [52], a foundational model in the vision-language domain. Additionally, we experiment with X-VLM [71], a resilient model trained on multi-grained objectives, known for its notable performance in compositional understanding [4].

4.1. Setup

Training We refer to the CLIP finetuned with our proposed losses as the **C**ompositional **E**nhanced CLIP (CE-CLIP). We train in two configurations: (1) CE-CLIP, using only the COCO dataset [34], for direct comparison with NegCLIP [70], and (2) CE-CLIP+, which leverages a combined dataset of COCO, CC3M [58], and Visual Genome [29] aiming for heightened performance.

We employ the *CLIP-VIT/32-B* from the Open-CLIP implementation and the *X-VLM-16M* from its primary code repository for evaluation purposes.[2] Both models undergo fine-tuning over 5 epochs following previous works [12, 70] using 2 A100 GPUs. We allocate batch sizes of 256 for CLIP and 64 for X-VLM fine-tuning. All training parameters, like learning rate, decay rate, etc., remain at default values. We conducted a hyper-parameter search for α, β with optimal values of $\alpha = 0.2$ and $\beta = 0.4$.

Evaluation We evaluate our method on several vision-language(vl)-compositional benchmarks: ARO[70], VL-CheckList[74], VALSE[49], and SugarCrepe[24] (bias-mitigated version of CREPE[40]). Although Winoground was designed to test compositional reasoning, Diwan et al. [11] highlights other challenges posed by this dataset, like commonsense reasoning and unique image/text understanding. As these are not focus of our work, we excluded Winoground from our evaluations. We evaluate our methods in zero-shot settings. Each evaluation involves classifying positive and negative captions for a given image, with a random success probability of 50%.

For a comprehensive evaluation, we selected robust baselines: (1) Cutting-edge generative vision-language models such as BLIP [32], BLIP2 [33], and MiniGPT-4 [80]; (2) High-performing vision-language understanding models like BEIT3 [66], ALBEF [31], UNITER [7], CyCLIP [17], and X-VLM [71]; (3) Compositional improvement methods such as syn-CLIP [5] and CLIP-SGVL [21] (both leveraging scene graph annotations), DAC [13] (utilizing segmentation models and LLMs), and NegCLIP [70] and CLIP-SVLC [12] that employ hard negative.

4.2. Compositional reasoning enhancement

We present results for ARO and VALSE in Tab. 1, VL-CheckList in Tab. 2, and SugarCrepe in Tab. 3. Our CE-CLIP model, which is trained on the same dataset as NegCLIP, surpasses all methods utilizing hard negatives across all benchmarks. It demonstrates significant improvements over the baseline CLIP model: 23.7% on ARO-Relation, 13.5% on ARO-Attribute, 7.2% on VALSE, 5.2% on VL-CheckList, and 12.1% on SugarCrepe. This indicates that our approach more effectively utilizes hard negatives through intra-modal contrasting and cross-modal ranking. Notably, the smallest absolute improvement was observed on VL-CheckList, likely because this benchmark presents an out-of-distribution challenge for our CE-CLIP, given that it is only fine-tuned on COCO, while VL-CheckList integrates several diverse datasets. Conversely, we note a substantial improvement on the ARO benchmark, which could be attributed to the hard negative types in our model that are specifically tailored to enhance the understanding of objects and attributes. Additionally, the significant gains observed on SugarCrepe, a benchmark designed to mitigate language bias in other benchmarks and provide a more accurate reflection of a model's compositional understanding, are particularly noteworthy.

The CE-CLIP+, trained on a more comprehensive dataset, achieves superior performance, with an average improvement of 24.3% and 14.2% on ARO Relation and Attribution splits, 11.4% on VALSE, 9.2% on VL-CheckList, and 14.4% on SugarCrepe which translates to an impressive average accuracy of 87.5%. Similar to CE-CLIP, the greatest

[2] https://github.com/mlfoundations/open_clip

Model	#Params	ARO		VALSE									
		Relation	Attribute	Existence quantifiers	Plurality number	Counting	Sp.rel. relations	Actions repl.	 actant swap	Coreference standard	 clean	Foil-it!	Avg.
Random Chance		50											
BLIP[32]	583M	59.0	88.0	86.3	73.2	68.1	71.5	77.2	61.1	53.8	48.2	93.8	70.0
BEIT3[66]	1.9B	60.6	74.6	77.4	74.6	68.8	74.0	86.7	65.2	50.0	44.2	96.0	70.4
BLIP2[33]	3.4B	41.2†	71.3†	55.5	71.5	66.0	62.4	83.6	51.6	48.6	51.9	95.9	65.4
MiniGPT-4[33]	>9B	46.9†	55.7†	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[5]†	151M	71.4	66.9	-	-	-	–	-	-	-	-	-	
Segmentation & LLM relied method													
DAC-LLM[13]†	151M	81.3	73.9	-	-	-	-	-	-	-	-	-	-
DAC-SAM[13]†	151M	77.2	70.5	-	-	-	-	-	-	-	-	-	-
Hard Negative based method													
XVLM-coco[71]	216M	73.4	86.8	83.0	75.6	67.5	70.2	73.8	68.6	46.4	49.6	94.8	69.5
CE-XVLM	216M	**73.9**$_{+0.5}$	**89.3**$_{+2.5}$	83.5	72.8	72.1	68.7	71.8	69.1	51.0	46.8	93.8	**70.8**$_{+1.3}$
CLIP[52]	151M	59.3	62.9	68.7	57.1	61.0	65.4	77.8	71.8	54.1	51.0	89.8	65.3
CyCLIP[17]†	151M	59.1	65.4	69.3	58.3	61.0	66.4	78.1	72.0	53.2	51.6	88.8	65.5
SDS-CLIP[3] †	151M	53.0	62.0	-	-	-	-	-	-	-	-	-	-
NegCLIP[70]	151M	80.2	70.5	76.8	71.7	65.0	72.9	81.6	84.7	58.6	53.8	91.9	71.6
CLIP-SVLC[12] †	151M	80.61	73.03	-	-	-	-	-	-	-	-	-	-
CE-CLIP	151M	83.0$_{+23.7}$	76.4$_{+13.5}$	78.6	77.7	64.4	74.4	81.2	88.6	54.7	54.8	93.7	72.5$_{+7.2}$
CE-CLIP+	151M	**83.6**$_{+24.3}$	**77.1**$_{+14.2}$	84.5	79.2	67.8	76.4	83.4	89.4	56.7	57.8	94.7	**76.7**$_{+11.4}$

Table 1. **Results (%) on ARO and VALSE**. The best scores for each section are highlighted in bold. † represents scores are extracted from papers. Empty scores suggest that the model's codebase has not been released.

Model	#Params	Attribute						Object			Relation			Avg
		Action	Color	Material	Size	State	Avg	Location	Size	Avg	Action	Spatial	Avg	
Random Chance		50												
ALBEF[31] †	210M	81.7	84.2	87.3	69.5	72.08	79.3	81.7	80.5	81.1	70.5	64.6	66.5	75.6
UNITER[7]†	300M	72.6	76.2	75.8	63.5	68.1	71.3	82.4	81.5	81.9	69.2	61.5	64.7	72.6
BLIP[32]†	583M	79.5	83.2	84.7	59.8	68.8	75.2	83.0	81.3	82.2	59.5	75.7	70.5	75.7
BEIT3[66]	1.9B	79.6	78.5	80.1	63	68.4	73.9	85.2	83.8	84.5	76.6	62.3	69.4	75.3
BLIP2[33]†	3.4B	81.0	86.2	90.3	61.7	70.1	77.8	85.4	84.3	84.9	84.9	56.2	70.6	77.8
MiniGPT-4[79] †	>9B	-	-	-	-	-	71.3	-	-	84.2	84.1	-	-	-
Scene Graph relied method														
CLIP-SGVL[21]†	>151M	76.6	78.0	80.6	59.7	61.2	71.2	83.0	81.3	82.6	79.0	-	-	-
syn-CLIP[5] †	151M	-	-	-	-	-	70.4	-	-	-	-	-	69.4	-
Segmentation & LLM relied method														
DAC-LLM[13]†	151M	-	-	-	-	-	77.3	-	-	87.3	86.4	-	-	-
DAC-SAM[13]†	151M	-	-	-	-	-	75.8	-	-	88.5	89.8	-	-	-
Hard Negative based method														
XVLM-coco[71]	216M	80.4	81.1	83.1	60.3	70.8	**75.1**	86.3	85.3	85.8	79.0	61.8	70.4	76.5
CE-XVLM	216M	80.5	76.0	80.6	67.2	69.8	74.8$_{-0.3}$	87.3	86.6	**86.9**$_{+1.1}$	80.8	78.6	**79.7**$_{+9.3}$	**78.6**$_{+2.1}$
CLIP[52]	151M	70.5	69.4	69.5	60.7	67	67.4	80.2	79.7	80.0	72.2	53.8	63.0	69.2
CLIP-SVLC[12]†	151M	69.4	77.5	77.4	73.4	62.3	72.0	-	-	85.0	74.7	63.2	68.95	74.2
NegCLIP[70]	151M	72.1	75.7	78.1	61.3	67.3	70.9	84.4	83.8	84.1	80.7	57.1	68.9	73.4
CE-CLIP	151M	75.6	72.7	79.7	65.3	69.8	72.6$_{+5.2}$	84.8	84.5	84.6$_{+4.6}$	78.5	65.0	71.8$_{+8.8}$	75.1$_{+5.9}$
CE-CLIP+	151M	78.5	83.5	85.2	65.8	70.8	**76.7**$_{+9.3}$	86.7	85.9	**86.3**$_{+6.3}$	81.0	68.4	**74.7**$_{+11.7}$	**78.4**$_{+9.2}$

Table 2. **Results (%) on VL-CheckList.** The best scores for each section are highlighted in bold. † represents scores are extracted from papers. Empty scores suggest that the model's codebase has not been released.

and smallest improvements were observed in ARO and VL-CheckList, respectively, reinforcing our initial hypothesis. The out-of-distribution challenge observed in CE-CLIP has been substantially mitigated in CE-CLIP+ through training on a varied range of data distributions. For example, in the Attribute evaluation split, CE-CLIP showed a modest 5.2% improvement on VL-CheckList and a significant 13.5% increase on ARO. Impressively, CE-CLIP+ outperforms CE-CLIP by 0.7% (76.4→77.1) on ARO-Attribute and an exceptional 4.1% (72.6→76.7) on the VL-CheckList Attribute split. This underscores the challenges of out-of-distribution evaluation encountered by CE-CLIP and illustrates the effectiveness of augmenting dataset size as a remedy. Overall, CE-CLIP+ demonstrates robust performance, surpassing models with significantly larger parameters or those trained with extra resources and annotations across the majority of benchmarks. This strengthens the potential scalability of our method within extensive pre-training frameworks, although we acknowledge the necessity for further investigation.

X-VLM shows a modest improvement compared with

Model	REPLACE				SWAP			ADD		
	Object	Attribute	Relation	Avg	Object	Attribute	Avg	Object	Attribute	Avg
Human	100	99	97	98.7	99	100	99.5	99	99	99
Vera[38]	49.4	49.6	49.1	49.4	49.4	49.2	49.3	49.4	49.6	49.5
Grammar[46]	50	50	50	50	50	50	50	50	50	50
BLIP2[33] †	-	-	-	86.7	-	-	69.8	-	-	86.5
CLIP	90.9	80	69.2	80.2	61.4	64	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	**87.0**$_{+6.8}$	72.8	77	74.9$_{+12.2}$	92.4	93.4	**92.9**$_{+20.2}$
CE-CLIP+	93.8	90.8	83.2	**89.3**$_{+9.1}$	76.8	79.3	**78.0**$_{+15.3}$	93.8	94.9	**94.4**$_{+21.7}$

Table 3. **Results(%) on SugarCrepe**. Vera and Grammar are text-only models.

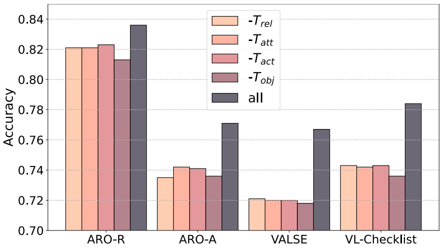


Figure 4. Ablations on hard-negative types

large improvement gained on CE-CLIP, primarily due to differences in pretraining approaches. X-VLM is pretrained on multiple fine-grained tasks that necessitate specific object bounding box annotations, whereas CLIP is trained directly on automatically crawled, noisy image-text pairs. Our simple annotation-free method can bolster the already strong X-VLM, further emphasizing its distinctive characteristics in learning compositionality. However, our method is most beneficial for CLIP like models that do not already benefit from object annotations during pre-training.

4.3. Emergence of curriculum learning

In this section, we illustrate how the adaptive threshold in the cross-modal loss facilitates curriculum learning during fine-tuning. We analyze the evolution of the threshold values and losses over time, with the curve in Fig. 5d showing a sharp increase in the *Threshold Relation* value. This rise is mainly due to the semantic and grammatical errors in relation-swap hard negatives (e.g., sentences in Fig. 2), simplifying the model's task of differentiating authentic captions from hard negatives. Consequently, the elevated threshold counters this by increasing the task difficulty, providing a stronger supervisory signal and compelling the model to discern greater differences between these captions.

The threshold, calculated as the average gap between true and hard negative similarity scores, mirrors the task's complexity and the model's discernment capability. CE-CLIP+'s training loss curve (Fig. 5b) indicates that CMR loss stabilizes after initial fluctuations, striking a balance between escalating task difficulty and the model's adaptive capacity, thereby highlighting the inherent curriculum learning.

The emergence of curriculum learning achieves satisfactory outcomes without needing extensive hyper-parameter tuning. In contrast, a fixed threshold strategy would require impractical n^4 trials for exploring n different values across four thresholds. Fig. 5a compares CE-CLIP+ results across 5 benchmarks using various thresholds, showing adaptive approach outperforms the fixed ones and converges faster. Initially, the adaptive strategy provides a smaller supervision signal compared to the fixed approach but as the training progresses, it adjusts the threshold according to the task complexity and model capacity. This adjustment enhances learning efficiency and generalization.

Model	*itc(hn)*	*IMC*	*CMR*	ARO-R	ARO-A	VALSE	VLCheckList	Avg
CLIP				59.3	62.9	67.0	69.2	64.6
	✓			81.6	72.0	74.2	73.6	75.4
	✓	✓		82.6	75.8	75.9	76.6	77.7
	✓		✓	82.3	72.6	75.5	77.8	77.1
CE-CLIP+	✓	✓	✓	**83.6**	**77.1**	**76.7**	**78.4**	**79.0**

Table 4. **Ablation of losses.** *itc(hn)* represents image-text contrastive with additional hard negatives.

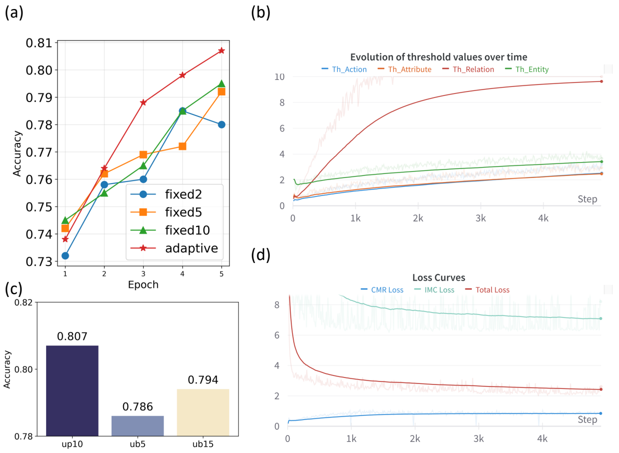


Figure 5. **Ablation studies.** (a) Adaptive vs Fixed threshold with values 2, 5, 10; (b) Evolution of threshold over time ; (c) Performance with different upper bounds on threshold. (d) Loss curves showing stabilization of the CMR loss after initial training steps.

4.4. Ablation studies

We present ablation studies to understand the effectiveness of different components of our method. We conduct these ablations using our best model CE-CLIP+.

Losses. The impact of each proposed loss is detailed in Tab 4. Notably, the introduction of hard negatives led to tremendous performance gains, highlighting their pivotal role in contrastive learning. Each individual loss we introduced showed significant improvements as well across all benchmarks. The best performance is achieved when all losses are combined, thus demonstrating the effectiveness of our approach.

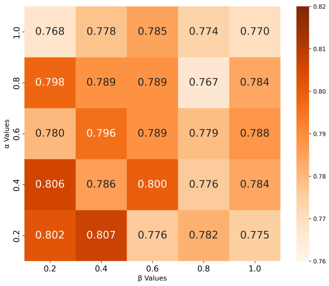


Figure 6. **Ablations of α, β.**

Task	CLIP	CLIP-FT	CE-CLIP	CE-CLIP+
compositional tasks				
ARO-R[70]	59.3	61.7	83	83.6(+24.3)
ARO-A[70]	62.9	66.1	76.4	77.1(+14.2)
VALSE[49]	65.3	71.8	72.5	76.7(+11.4)
VLChecklist[74]	69.2	68.6	75.1	78.4(+9.2)
SugarCrepe[24]	73.1	77.2	85.2	87.5(+14.4)
standard tasks				
T2I R@5[35]	56.0	66.2	69.4	72.3(+13.4)
I2T R@5[35]	75.0	78.3	74.3	76.1(+1.1)
ImageNet1K	93.2	92.8	92.6	92.7(-0.5)
CIFAR10	94.2	94.2	93.8	93.8(-0.4)
CIFAR100	79.0	79.1	78.0	78.1(-0.9)

Table 5. Performance on standard image-text retrieval and image classification. Improvements in green are calculated w.r.t CLIP.

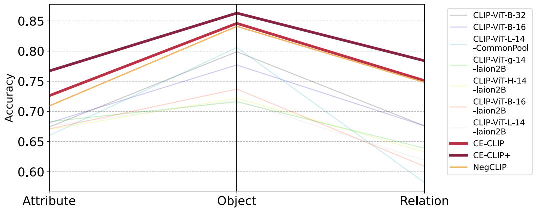


Figure 7. Impact of scaling-up the model on VL-CheckList performance.

Hard Negative Types As shown in Fig. 4, each type of hard-negative uniquely benefits the model, the object hard negatives benefitting the most. Combining all types yields the best results. The success of our flexible approach indicates that incorporating additional types, such as numerical negatives [48], may further boost performance.

Upper Bound on Threshold. Setting a threshold upper bound prevents training collapse. Our ablation study, as detailed in Fig. 5c, demonstrates that an upper bound of 10 yields optimal performance by effectively constraining the maximum value of the *Threshold Relation* (Fig. 5b), thereby ensuring stability during the training process.

Loss Weight. Fig. 5d shows the divergence of CMR loss scale from IMC loss, highlighting the importance of proper loss weight selection for training. Fig. 6 reveals that our method is robust across 5 benchmarks with varying α and β values, though larger α and β decrease performance. Optimal outcomes occur at $\alpha = 0.2$ and $\beta = 0.4$.

4.5. Performance on standard benchmarks

Previous studies [12, 70], suggest that advancements in compositional understanding might negatively affect performance on standard image-text retrieval and image classification tasks. To investigate this, we evaluate our method on zero-shot image-text retrieval on COCO and linear probing on ImageNet-1k [10] and CIFAR [30]. As shown in Tab. 5, our results demonstrate improvements in text-to-image retrieval with minimal impact on image classification accuracy. By prioritizing compositional understanding, our CE-CLIP and CE-CLIP+ enhance performance across all evaluated benchmarks. Furthermore, to demonstrate that our enhancements in COCO image-text retrieval are not merely a result of fine-tuning on COCO, we include comparative results from CLIP-FT, a COCO fine-tuned variant of CLIP. Our findings indicate that both CE-CLIP and CE-CLIP+ outperform CLIP-FT in text-to-image retrieval, albeit with a slight underperformance in image-to-text retrieval. We hypothesize this could be due to our method's exclusive reliance on textual hard negatives.

We investigated the impact of integrating CE-CLIP into MAPL [43] on the Visual Question Answering (VQA) task by training on the COCO dataset and conducting a zero-shot evaluation on VQAv2 [19]. The findings indicate that CE-CLIP, achieving an accuracy of 39.82, closely matches the original CLIP's performance of 39.78. This demonstrates that CE-CLIP preserves the visual strengths of CLIP.

4.6. Can scaling-up alone solve compositionality

To substantiate our assertion in Fig. 1 that standard contrastive learning as implemented in CLIP fails to grasp compositionality, we tested several scaled-up versions of CLIP models including the largest ViT-G/14 trained on LAION-2B from Open-CLIP, on the VLChecklist benchmark. As Fig. 7 shows, none of these scaled-up models surpass our base-sized CE-CLIP model. This shows that scaling-up the model alone is not enough for comprehending compositionality, underscoring the significance of our work and the need for more research in this field.

5. Conclusion

Our study addresses the challenge of compositional understanding in VLMs , we expand image-text contrastive loss and introduce two losses that infuse compositional supervision into pretrained VLMs using a featured hard negative generation strategy. Our intra-modal contrastive loss mitigates high intra-modal similarity while our cross-modal rank loss ensures a minimum semantic distance between true and hard negative image-text pairs, with the adaptive threshold functioning as curriculum learning to enhance performance. Empirically, our method achieves superior performance in 5 compositional benchmarks, surpassing previous methods without requiring additional annotations or resources. Scaling the dataset size further boosts performance, highlighting our method's potential for VLMs and its promise for broader applications and capabilities.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning, 2022. 1

[2] Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, Jenia Jitsev, Simon Kornblith, Pang Wei Koh, Gabriel Ilharco, Mitchell Wortsman, and Ludwig Schmidt. Openflamingo: An open-source framework for training large autoregressive vision-language models, 2023. 2

[3] Samyadeep Basu, Maziar Sanjabi, Daniela Massiceti, Shell Xu Hu, and Soheil Feizi. Augmenting clip with improved visio-linguistic reasoning. *ArXiv*, abs/2307.09233, 2023. 6

[4] Emanuele Bugliarello, Laurent Sartran, Aishwarya Agrawal, Lisa Anne Hendricks, and Aida Nematzadeh. Measuring progress in fine-grained vision-and-language understanding, 2023. 5

[5] Paola Cascante-Bonilla, Khaled Shehada, James Smith, Sivan Doveh, Donghyun Kim, Rameswar Panda, Gül Varol, Aude Oliva, Vicente Ordonez, Rogério Schmidt Feris, and Leonid Karlinsky. Going beyond nouns with vision & language models using synthetic data. *ArXiv*, abs/2303.17590, 2023. 5, 6

[6] Weihua Chen, Xiaotang Chen, Jianguo Zhang, and Kaiqi Huang. Beyond triplet loss: a deep quadruplet network for person re-identification, 2017. 4, 5

[7] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning, 2020. 5, 6

[8] Jaemin Cho, Seunghyun Yoon, Ajinkya Kale, Franck Dernoncourt, Trung Bui, and Mohit Bansal. Fine-grained image captioning with clip reward. *ArXiv*, abs/2205.13115, 2022. 2

[9] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20:273–297, 1995. 2, 4

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 8

[11] Anuj Diwan, Layne Berry, Eunsol Choi, David Harwath, and Kyle Mahowald. Why is winoground hard? investigating failures in visuolinguistic compositionality, 2022. 5

[12] Sivan Doveh, Assaf Arbelle, Sivan Harary, Rameswar Panda, Roei Herzig, Eli Schwartz, Donghyun Kim, Raja Giryes, Rogerio Feris, Shimon Ullman, and Leonid Karlinsky. Teaching structured vision language concepts to vision language models, 2022. 2, 3, 4, 5, 6, 8

[13] Sivan Doveh, Assaf Arbelle, Sivan Harary, Roei Herzig, Donghyun Kim, Paola Cascante-bonilla, Amit Alfassy, Rameswar Panda, Raja Giryes, Rogerio Feris, Shimon Ullman, and Leonid Karlinsky. Dense and aligned captions (dac) promote compositional reasoning in vl models, 2023. 3, 5, 6

[14] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, Eyal Orgad, Rahim Entezari, Giannis Daras, Sarah Pratt, Vivek Ramanujan, Yonatan Bitton, Kalyani Marathe, Stephen Mussmann, Richard Vencu, Mehdi Cherti, Ranjay Krishna, Pang Wei Koh, Olga Saukh, Alexander Ratner, Shuran Song, Hannaneh Hajishirzi, Ali Farhadi, Romain Beaumont, Sewoong Oh, Alex Dimakis, Jenia Jitsev, Yair Carmon, Vaishaal Shankar, and Ludwig Schmidt. Datacomp: In search of the next generation of multimodal datasets, 2023. 1

[15] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision*, pages 1–15, 2023. 2

[16] Weifeng Ge, Weilin Huang, Dengke Dong, and Matthew R. Scott. Deep metric learning with hierarchical triplet loss. In *European Conference on Computer Vision*, 2018. 3

[17] Shashank Goel, Hritik Bansal, Sumit Bhatia, Ryan A. Rossi, Vishwa Vinay, and Aditya Grover. Cyclip: Cyclic contrastive language-image pretraining, 2022. 5, 6

[18] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering, 2017. 1

[19] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering, 2017. 8

[20] Ben Harwood, B. V. Kumar, G. Carneiro, Ian D. Reid, and Tom Drummond. Smart mining for deep metric learning. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2840–2848, 2017. 3

[21] Roei Herzig, Alon Mendelson, Leonid Karlinsky, Assaf Arbelle, Rogerio Feris, Trevor Darrell, and Amir Globerson. Incorporating structured representations into pretrained vision - language models using scene graphs, 2023. 3, 5, 6

[22] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning, 2022. 2

[23] Matthew Honnibal and Ines Montani. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear, 2017. 3

[24] Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. Sugarcrepe: Fixing hackable benchmarks for vision-language compositionality, 2023. 2, 5, 8

[25] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision, 2021. 2

[26] Yannis Kalantidis, Mert Bulent Sariyildiz, No'e Pion, Philippe Weinzaepfel, and Diane Larlus. Hard negative mixing for contrastive learning. *ArXiv*, abs/2010.01028, 2020. 3

[27] Yannis Kalantidis, Mert Bulent Sariyildiz, Noe Pion, Philippe Weinzaepfel, and Diane Larlus. Hard negative mixing for contrastive learning. *Advances in Neural Information Processing Systems*, 33:21798–21809, 2020. 3

[28] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 3

[29] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Fei-Fei Li. Visual genome: Connecting language and vision using crowdsourced dense image annotations, 2016. 5

[30] Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. Cifar-10 (canadian institute for advanced research). 8

[31] Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Deepak Gotmare, Shafiq Joty, Caiming Xiong, and Steven Hoi. Align before fuse: Vision and language representation learning with momentum distillation, 2021. 5, 6

[32] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, 2022. 5, 6

[33] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023. 1, 2, 3, 5, 6, 7

[34] 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 *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 5

[35] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015. 8

[36] Hao Liu, Xiangyu Zhu, Zhen Lei, and Stan Z. Li. Adaptiveface: Adaptive margin and sampling for face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 4, 5

[37] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023. 2

[38] Jiacheng Liu, Wenya Wang, Dianzhuo Wang, Noah A. Smith, Yejin Choi, and Hannaneh Hajishirzi. Vera: A general-purpose plausibility estimation model for commonsense statements, 2023. 7

[39] 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, 2019. 3

[40] Zixian Ma, Jerry Hong, Mustafa Omer Gul, Mona Gandhi, Irena Gao, and Ranjay Krishna. Crepe: Can vision-language foundation models reason compositionally? *arXiv preprint arXiv:2212.07796*, 2022. 5

[41] R. Manmatha, Chaoxia Wu, Alex Smola, and Philipp Krähenbühl. Sampling matters in deep embedding learning. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2859–2867, 2017. 3

[42] Oscar Mañas, Pau Rodriguez, Saba Ahmadi, Aida Nematzadeh, Yash Goyal, and Aishwarya Agrawal. Mapl: Parameter-efficient adaptation of unimodal pre-trained models for vision-language few-shot prompting, 2023. 1

[43] Oscar Mañas, Pau Rodriguez, Saba Ahmadi, Aida Nematzadeh, Yash Goyal, and Aishwarya Agrawal. Mapl: Parameter-efficient adaptation of unimodal pre-trained models for vision-language few-shot prompting, 2023. 8

[44] Jan Hendrik Metzen, Piyapat Saranrittichai, and Chaithanya Kumar Mummadi. Autoclip: Auto-tuning zero-shot classifiers for vision-language models. *ArXiv*, abs/2309.16414, 2023. 2

[45] Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, Xiao Wang, Xiaohua Zhai, Thomas Kipf, and Neil Houlsby. Simple open-vocabulary object detection with vision transformers, 2022. 2

[46] John X. 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, 2020. 7

[47] Zachary Novack, S. Garg, Julian McAuley, and Zachary Chase Lipton. Chils: Zero-shot image classification with hierarchical label sets. *ArXiv*, abs/2302.02551, 2023. 2

[48] Roni Paiss, Ariel Ephrat, Omer Tov, Shiran Zada, Inbar Mosseri, Michal Irani, and Tali Dekel. Teaching clip to count to ten, 2023. 3, 4, 8

[49] 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 *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8253–8280, Dublin, Ireland, 2022. Association for Computational Linguistics. 2, 5, 8

[50] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 1, 2

[51] Yao Qin, Chiyuan Zhang, Ting Chen, Balaji Lakshminarayanan, Alex Beutel, and Xuezhi Wang. Understanding and improving robustness of vision transformers through patch-based negative augmentation. *ArXiv*, abs/2110.07858, 2021. 3

[52] 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 *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1, 2, 4, 5, 6

[53] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022. 1, 2

[54] Joshua Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. Contrastive learning with hard negative samples. *ArXiv*, abs/2010.04592, 2020. 3

[55] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1, 2

[56] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs, 2021. 1, 2

[57] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. Laion-5b: An open large-scale dataset for training next generation image-text models, 2022. 2

[58] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Annual Meeting of the Association for Computational Linguistics*, 2018. 5

[59] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. Flava: A foundational language and vision alignment model, 2022. 2

[60] Harman Singh, Pengchuan Zhang, Qifan Wang, Mengjiao Wang, Wenhan Xiong, Jingfei Du, and Yu Chen. Coarse-to-fine contrastive learning in image-text-graph space for improved vision-language compositionality. 2023. 3

[61] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale, 2023. 2

[62] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers, 2019. 1

[63] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5238–5248, 2022. 1, 2

[64] Haoxiang Wang, Pavan Kumar Anasosalu Vasu, Fartash Faghri, Raviteja Vemulapalli, Mehrdad Farajtabar, Sachin Mehta, Mohammad Rastegari, Oncel Tuzel, and Hadi Pouransari. Sam-clip: Merging vision foundation models towards semantic and spatial understanding, 2023. 2

[65] Jiang Wang, Yang song, Thomas Leung, Chuck Rosenberg, Jinbin Wang, James Philbin, Bo Chen, and Ying Wu. Learning fine-grained image similarity with deep ranking, 2014. 4

[66] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, and Furu Wei. Image as a foreign language: Beit pretraining for all vision and vision-language tasks, 2022. 5, 6

[67] Xin Wang, Yudong Chen, and Wenwu Zhu. A survey on curriculum learning, 2021. 4

[68] Mengde Xu, Zheng Zhang, Fangyun Wei, Yutong Lin, Yue Cao, Han Hu, and Xiang Bai. A simple baseline for zero-shot semantic segmentation with pre-trained vision-language model. *ArXiv*, abs/2112.14757, 2021. 2

[69] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models, 2022. 2

[70] 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 e-prints*, pages arXiv–2210, 2022. 1, 2, 3, 4, 5, 6, 8

[71] Yan Zeng, Xinsong Zhang, and Hang Li. Multi-grained vision language pre-training: Aligning texts with visual concepts, 2022. 1, 2, 5, 6

[72] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training, 2023. 2

[73] Weichen Zhang, Wanli Ouyang, Wen Li, and Dong Xu. Collaborative and adversarial network for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3801–3809, 2018. 4, 5

[74] Tiancheng Zhao, Tianqi Zhang, Mingwei Zhu, Haozhan Shen, Kyusong Lee, Xiaopeng Lu, and Jianwei Yin. Vl-checklist: Evaluating pre-trained vision-language models with objects, attributes and relations, 2022. 2, 5, 8

[75] Xiaonan Zhao, Huan Qi, Rui Luo, and Larry Davis. A weakly supervised adaptive triplet loss for deep metric learning, 2019. 4, 5

[76] Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from clip, 2022. 2

[77] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. 2

[78] Ziqin Zhou, Bowen Zhang, Yinjie Lei, Lingqiao Liu, and Yifan Liu. Zegclip: Towards adapting clip for zero-shot semantic segmentation, 2023. 2

[79] 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. 6

[80] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models, 2023. 2, 5