This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

No Hard Negatives Required: Concept Centric Learning Leads to Compositionality without Degrading Zero-shot Capabilities of Contrastive Models

Hai X. Pham* David T. Hoffmann Ricardo Guerrero Brais Martinez
Samsung AI Center Cambridge, UK

Abstract

Contrastive vision-language (V&L) models remain a popular choice for various applications. However, several limitations have emerged, most notably the limited ability of V&L models to learn compositional representations. Prior methods often addressed this limitation by generating custom training data to obtain hard negative samples. Hard negatives have been shown to improve performance on compositionality tasks, but are often specific to a single benchmark, do not generalize, and can cause substantial degradation of basic V&L capabilities such as zero-shot or retrieval performance, rendering them impractical. In this work we follow a different approach. We identify two root causes that limit compositionality performance of V&Ls: 1) Long training captions do not require a compositional representation; and 2) The final global pooling in the text and image encoders lead to a complete loss of the necessary information to learn binding in the first place. As a remedy, we propose two simple solutions: 1) We obtain short concept centric caption parts using standard NLP software and align those with the image; and 2) We introduce a parameter-free cross-modal attention-pooling to obtain concept centric visual embeddings from the image encoder. With these two changes and simple auxiliary contrastive losses, we obtain SOTA performance on standard compositionality benchmarks, while maintaining or improving strong zero-shot and retrieval capabilities. This is achieved without increasing inference cost. We release the code for this work at https://github.com/SamsungLabs/concept_centric_clip.

1. Introduction

Contrastive vision-language (V&L) models have been a cornerstone of computer vision and machine learning since the publication of CLIP [24]. While contrastive V&L models have achieved remarkable performance on a large variety of tasks and enabled deployment of vision models to open world settings, various limitations have been found. Most notably, they struggle when learning compositional representations, such as binding a noun to its attributes, understanding the relations between objects, or recognizing when one of multiple objects has been replaced by another object [6, 10, 32]. Compositionality is a core capability of vision systems and results in various failure cases if not learned. Therefore, many works have focused on ways to improve compositionality post-hoc [1, 4–6, 8, 10, 15, 22, 27, 28, 32, 34], most of which [4, 5, 22, 27, 34] rely on clever ways of constructing hard negatives for fine-tuning purposes.

Previous works correctly identified one of the reasons why contrastive models do not learn compositional representations: A simple Bag-of-Words (BoW) representation is completely sufficient to retrieve the right image from a batch, as the caption is typically long and detailed. They have proposed solving this issue by training with hard negatives generated by following simple rules. Models trained on these negatives learn a compositional representation in the narrow domain defined by the set of rules used to generate hard negatives. Training with hard negatives has indeed been shown to yield in-distribution improvements, but these gains often fail to generalize to hard negatives that exhibit slightly different structures [10, 15]. At the same time, the improvements in compositionality achieved via hard negative training frequently come at the expense of generalization to other downstream tasks; precisely the property that originally made contrastive V&L models popular.

Here, we follow a different approach and focus on better exploiting existing pre-training data. In particular, we focus on noun-phrases, which are groups of words that function as a concept in a sentence, e.g. “a red couch” or “the tall building”, and define two auxiliary loss terms, a contrastive concept loss and a cross-attentive pooling loss, which combine to rectify the root causes of the contrastive learning behavior. Our contrastive concept loss focuses on contrasting noun-phrases and images. Noun-phrases are short and

*Correspondence to: <pham.xuan.hai@outlook.com>

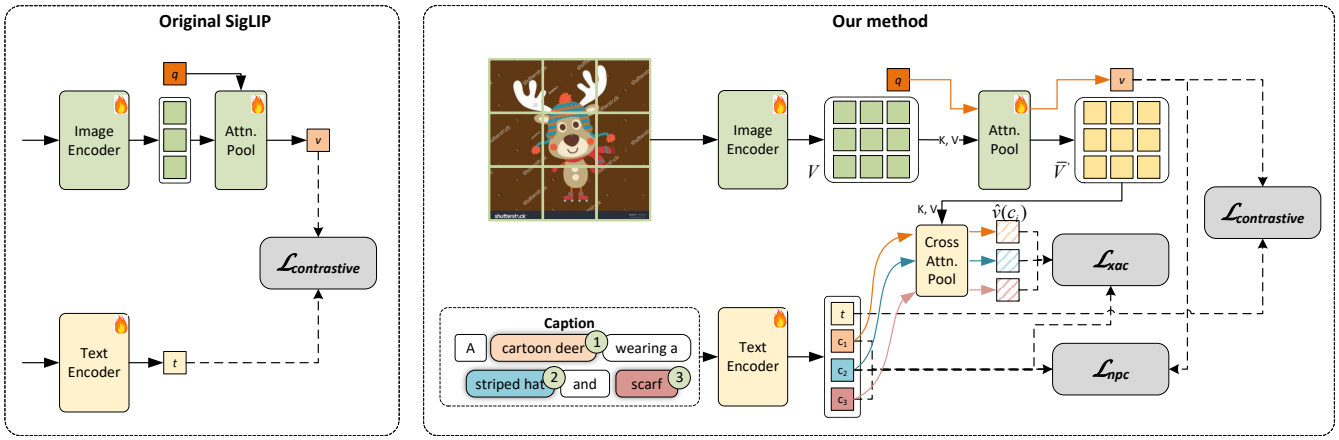


Figure 1. **Method overview**. **(a)** SigLIP uses a learnable query token in combination with an attention layer to pool the visual tokens into a single token. Aligning only global representations hampers the learning of a compositional representation. **(b)** Similar to SigLIP, our method aligns the global representations v and t. To simplify learning of a compositional representation, our method extends SigLIP by first, pooling the text encoder output tokens into concept embeddings $\{c_k\}$, which are used to attention-pool concept-specific information from the visual tokens $\bar{V}$, resulting in $\hat{v}(c)$. $\hat{v}(c_k)$ and the corresponding c_k are aligned using L_{xac}. Furthermore, the global visual representation v is aligned with all c_k via $\mathcal{L}_{\textbf{npc}}$, a multi-positive variant of SigLIP loss. Similarly, the global image and text representations v and t are aligned via $\mathcal{L}_{contrastive}$.

thus are not solvable via a BoW representation. To provide an example: For the noun-phrase "a red couch", a BoW representation is likely insufficient, as the probability of an image containing a couch and anything colored in red is quite high. Thus, the model must learn a latent representation that is more discriminative. Since our method relies on real data positives instead of synthetically generated hard negatives, our method is less likely to learn shortcuts that do not generalize to broader domains.

Despite the shortcuts when training with long captions, the overall architectural design of contrastive V&L models also hinders the learning of a good compositional representation. Visual and textual representations are pooled into single-vector embeddings, which mix nouns and adjectives from different regions of the image or different parts of the caption. This global pooling operation results in a complete loss of associations between attributes and nouns. Fine-tuning with post-pooling hard negative losses can yield improvements by aligning the model to in-domain distributions, however, this will not substantially change the underlying behavior of the encoders.

Instead of focusing on encouraging binding after the global pooling with hard negatives, we focus on learning binding **before** global pooling. Once learned, the model only needs to preserve a representation of that binding through the pooling operation. To encourage binding before the pooling operation, we introduce a cross-modal attention pooling function that extracts noun-phrase-specific information from the full visual representation, thereby producing multi-word, concept-specific embeddings learned with an auxiliary contrastive loss. As the attention-pooling function has no learnable parameters, this directly propagates the learning signal to the penultimate representation, i.e., it encourages a compositional representation before the global pooling operation. A schematic of our complete method can be seen in Fig. 1.

Our method, C^{2}LIP (Concept-centric CLIP), is simple to use, comes with no additional trainable parameters, and requires only fine-tuning for a few epochs without the incurred cost of complicated hard negative generation pipelines. We show quantitatively that this simple method leads to improvements over its base model and reaches SOTA on the SugarCrepe [10] and SugarCrepe++ [6] benchmarks. At the same time, our method leads to improvements on image caption retrieval datasets while incurring only modest drops in classification performance on ImageNet. Note that the drop in performance on ImageNet can be attributed to two factors: First, fine-tuning is performed on a less diverse dataset than the one used for pre-training. Second, the shift toward a more scene centric representation (required for compositional reasoning) conflicts with ImageNet's objective of focusing on the most salient object. Overall, C^{2}LIP demonstrates high performance across a variety of benchmarks and tasks. In contrast, the baselines usually excel at a single task but perform considerably poorer on the others.

In summary, our contributions are: 1) We introduce a hard negative-free pipeline to fine-tune existing contrastive V&L models to improve compositionality. 2) We show that training on short noun-phrases instead of long de-

tailed captions leads to improvements on compositionality while maintaining zero-shot capabilities. 3) We introduce a parameter-free cross-modal attention pooling mechanism that is used during training to encourage concept binding before the global pooling operation. 4) The proposed method is simple to train, adds minimal training overhead, and retains unchanged inference. 5) We show quantitatively that our method C^2LIP reaches SOTA on standard compositionality tasks, while at the same time maintaining high performance on standard retrieval and zero-shot benchmarks. It is particularly appealing that C^2LIP is among the top performers on all benchmarks and metrics tested, instead of excelling on some at the cost of others, as frequently observed for other methods.

2. Related work

Despite the great success of contrastive models for zero-shot classification and retrieval, several issues emerge when they are applied to more sophisticated tasks. For example, Yüksekgönül et al. [32] showed that CLIP tends to learn BoW representations. Even more concerning, Tang et al. [28] found that CLIP matches an image of an eggplant and a lemon to the caption "a purple lemon" rather than the correct "a yellow lemon".

Various strategies have been tested to prevent contrastive models from learning BoW representations [4, 5, 22, 23, 27, 34]. Many of these follow Yüksekgönül et al. [32] and use spaCy [9] to swap words in the captions, thereby creating hard negatives [4, 5, 23, 34] for the contrastive loss. Others achieve the same goal by employing a BERT model to replace words [4, 34]. More sophisticated methods generate hard negatives with an LLM, for instance via in-context learning [22], however, this approach incurs the risk of failures during hard negative creation as well as a significant computational cost. However, Hsieh et al. [10], Lewis et al. [15] showed that designing a caption augmentation pipeline that generalizes to unseen cases is difficult, and many approaches merely exploit biases in the test datasets. This questions whether the reliance on hand-crafted hard negatives is a promising direction in the first place. In parallel with hard negative captions, several works explore hard negative images generated by text-to-image models [16, 22]. These approaches typically generate hard negative captions and use these as input to a text-to-image model to obtain a hard negative image. Producing a large number of hard-negative images is very costly, as it requires both the creation of hard-negative texts and the execution of the text-to-image model, and thus inflates the dataset size, leading to substantially higher computational demands.

Many of the methods described above rely on custom contrastive losses, such as an extension of the MIL-NCE loss [5], an adaptive margin loss [16], or a dynamic threshold loss [34]. These losses are necessary to mitigate the negative impact of noisy or potentially erroneous synthetic samples, but each introduces its own set of challenges. Other works [5, 8, 20, 36] argue that the problem lies in caption quality and therefore obtain dense captions [5, 20, 36] for the images. Unfortunately, generating such dense captions is expensive, and additional tricks [20, 36] are required to make use of long captions in contrastive learning.

Another line of work attributes the problem to the model architecture, in particular the late interaction between modalities [1, 30]. Inspired by object-centric learning, Assouel et al. [1] use cross-attention to obtain caption-conditioned visual representations during both training and inference. However, their method requires an LLM to decompose a caption into a scene graph, effectively outsourcing most of the binding problem to the LLM, and requires an independent forward pass for each sub-caption both during training and inference, increasing the cost of training and inference drastically. Our approach also uses attention to produce text-conditioned visual representations, but it is considerably simpler and does neither rely on an LLM, nor a complicated binding module, nor multiple forward passes. Instead, we only need to detect noun-phrases in regular captions (e.g., using spaCy [9]) to use in our two auxiliary training objectives. The inference pipeline is identical to the vanilla CLIP/SigLIP pipeline.

3. Concept-centric contrastive learning

3.1. Preliminaries: SigLIP model

In this work, we utilize SigLIP [33] as base model, as it already provides higher initial performance than CLIP. In this section we provide the necessary background on SigLIP by briefly describing the model architecture, especially its attention pooling layer and the loss function, which we exploit in our concept-aware objectives. However, our method does not require any component specific to SigLIP. Thus, it is in principle applicable to any CLIP-like model. We leave such explorations for future work.

The SigLIP model [33] follows the conventional contrastive learning framework with two neural networks: the vision encoder f_{img} and text encoder f_{txt} that project the input image-caption (I, T) pair separately into two embedding vectors, (v, t), in the joint subspace $\in \mathbb{R}^D$, where D is the dimensionality of this subspace. The encoders are jointly trained by maximizing the cosine similarity of the positive pairs $s_{pos} = v_{pos} \cdot t_{pos}$, while at the same time minimizing the similarity scores s_{neg} of negative pairs within a training batch. Particularly, the SigLIP model deviates from the preceding CLIP model [24] at two important points: visual attention pooling and sigmoid loss.

Visual attention pooling. The CLIP vision encoder [24] adopted the standard vision transformer (ViT) [13], where

a learnable *[CLS]* token is appended to the input tokens and used to pool features via self-attention through all transformer layers. Whereas in SigLIP, a separate learnable token $q \in \mathbb{R}^D$ is employed to pool features *after* the last transformer layer using an additional attention layer, as demonstrated in Fig. 1a. Let $V \in \mathbb{R}^{M \times D}$ be the outputs of the ViT, in which M is the number of image patches. The visual embedding v is calculated as follows:

$$\bar{q} = f_{query}(q), \bar{K} = f_{key}(V), \bar{V} = f_{value}(V), \quad (1)$$

$$\tilde{q} = \bar{V}^T \cdot \text{attn}(\bar{q}, \bar{K}), \quad (2)$$

$$v = f_{MLP}(\tilde{q}), \quad (3)$$

where f_{query}, f_{key}, f_{value} and f_{MLP} are neural functions.

Contrastive sigmoid loss. We consider a minibatch B comprising paired embeddings $\{(v_i, t_i)|_{i=1}^{|B|}\}$, which we can organize into $|B|$ positive and $|B| \times (|B| - 1)$ negative pairs in total. We define an indicator function z_{ij} with value 1 if $i = j$ and -1 otherwise, reflecting the assumption of pairwise matches between images and captions. The contrastive sigmoid loss is given by

$$\mathcal{L}_{\text{contrastive}} = -\frac{1}{|B|} \sum_{i=1}^{|B|} \sum_{j=1}^{|B|} \log \sigma \left(z_{ij} (\tau v_i \cdot t_j + b) \right), \quad (4)$$

where σ denotes the sigmoid function, τ and b are learnable scalars corresponding to scale and bias terms. The sigmoid loss improves computational efficiency when compared to the more costly batch-wise softmax calculation required by CLIP. However, training only with this objective yields a BoW representation. In the next section we introduce additional concept-aware contrastive losses and explain how these losses are combined with our noun-phrase captions and cross-attention pooling function to promote learning a compositional representation. Note that these modifications do not require any changes to the original model architecture and do not alter the inference procedure.

3.2. Concept-aware contrastive training

We emphasize the binding of the elements in *noun-phrase concepts*, and aligning such concepts to the corresponding images. We start by identifying a set of noun-phrases from each input caption, which can be done offline using standard NLP tools, e.g. spaCy [9].

Concept-aware contrastive loss. The objective of this loss is to explicitly encourage the encoders to bind elements of a noun-phrase together. More specifically, each caption T_i contains K_i concepts (noun-phrases), with latent representations $\{\mathbf{c}_i^k|_{k=1}^{K_i}\}$. These representations are computed by pooling the representations of the text tokens corresponding to the respective noun-phrase, as depicted in Fig. 1b. We then devise an image-text contrastive loss similar to Eq. (4). However, different from Eq. (4), we match each image to all of its corresponding sub-captions (i.e. noun-phrase concepts) simultaneously. This enforces that the model represents all concepts in the global image representation. Our loss extends the standard SigLIP loss to a version that accepts multiple positives per sample, thus the indicator function z'_{ij} must be prepared accordingly. The **n**oun-**p**hrase **c**oncept (npc) loss is given by

$$\mathcal{L}_{npc} = -\frac{1}{\mathcal{K}} \sum_{i=1}^{|B|} \sum_{j=1}^{\mathcal{K}} \log \sigma \left(z'_{ij} (\tau v_i \cdot c_j + b) \right), \quad (5)$$

where $\mathcal{K} = \sum_{i=1}^{|B|} K_i$.

Cross-attended concept-aware loss. In addition to SigLIPs attention-pooling in the vision tower, we apply cross-modal attention-pooling to extract information from the visual tokens using the concept embeddings c as queries, as shown in Fig. 1b. We reuse the projected value $\bar{V}$ in Eq. (1) and the MLP in Eq. (3) to project the visual tokens to $\bar{V}' = f_{MLP}(\bar{V})$ in the joint subspace, which are used as key and value in the cross-attention function:

$$\hat{v}(c) = \bar{V}'^T \cdot \text{attn}(c, \bar{V}'). \quad (6)$$

Note that this attention-pooling does not add any additional parameters, and it is only utilized during training.

The **cross-a**ttended **c**oncept-aware (xac) loss is defined similarly to $\mathcal{L}_{npc}$ in Eq. (5), but replacing the unimodal visual latent representation v with the cross-modal concept-attended variant $\hat{v}(c)$, as follows:

$$\mathcal{L}_{xac} = -\frac{1}{\mathcal{K}} \sum_{i=1}^{|B|} \sum_{j=1}^{\mathcal{K}} \log \sigma \left(z'_{ij} (\tau \hat{v}_i(c_j) \cdot c_j + b) \right). \quad (7)$$

Total training loss. Finally, the model is fine-tuned with the combined loss given by

$$\mathcal{L}_{total} = \mathcal{L}_{contrastive} + \lambda_{npc} \mathcal{L}_{npc} + \lambda_{xac} \mathcal{L}_{xac}, \quad (8)$$

where λ_{npc} and λ_{xac} are trade-off hyper-parameters.

4. Experimental results

We detail our experimental settings, baselines, datasets and the evaluation protocols in Sec. 4.1. We present and discuss our quantitative results in Sec. 4.2, provide qualitative results in Sec. 4.3 and ablate the components of our method in Sec. 4.4.

4.1. Experimental setting

We fine-tune the pretrained SigLIP on CC3M [25] using our proposed method. Particularly, we use the new short captions of CC3M introduced in DreamLIP [36], which contain

richer descriptions of objects and attribute bindings. We fine-tune the model for five epochs using Adam optimizer with base learning rate of 1e-5, on eight A40 GPUs with effective batch size of 768. The hyper-parameters λ_{npc} and λ_{xac} in Eq. (8) are empirically chosen as 1 and 0.01, respectively. Our implementation extends from *OpenCLIP* [11].

Data pre-processing. First, we use spaCy for dependency parsing of the original captions, and traverse the dependency trees to extract the noun-phrase concepts.

Baselines. We use a ViT-B/16 in all our experiments. Accordingly, we compare with baselines utilizing the same model, or models with similar parameter count. We select checkpoints of corresponding models trained/fine-tuned on CC3M whenever available, and use any provided checkpoints otherwise. Refer to Tab. 1 for details on the training setup of our baselines.

Our main baseline is the SigLIP model pretrained on WebLI and fine-tuned on the DreamLIP variant of CC3M. This results in a fair baseline that has been trained with exactly the same data, exactly the same training setup and allows a comparison without confounding factors like the effect of a narrower domain of CC3M in comparison to WebLI. Despite that, we also compare performance of our model to that of larger models using a ViT-L encoder, as well as models trained with multi-task objectives.

Evaluation protocol. To ensure fair, consistent performance comparisons and facilitate reproducibility, we evaluate all methods with the *CLIP_Benchmark* tool [3] in the same environment, and report the resulting metrics on a variety of benchmarks:

- *Compositionality*: SugarCrepe and SugarCrepe++
- *Zero-shot classification*: ImageNet1K
- *Zero-shot retrieval*: Flickr30K and MSCOCO
- *Fine-grained retrieval*: DOCCI [21] and ImageInWords (IIW) [7]. Following the experiments of Xiao et al. [30], we split the long captions into single sentences and evaluate retrieval performance on them.

Evaluation metrics. For evaluation of compositionality, we follow the standard evaluation protocol of SugarCrepe and SugarCrepe++ and report accuracy. These benchmarks consider a sample as correct, if the true caption is assigned a higher similarity to the image than the hard negative caption. The datasets allow to evaluate for different word types separately, namely for *objects*, *attributes* and *relations of concepts*. Accuracy is always reported for a set of subtasks, differing in how the false samples are created: "Add" creates the false caption by adding an unrelated word of a given word type, "Swap" swaps a given word type with the same word type within the caption and "Replace" replaces a word with a new word of the same type. All evaluations on SugarCrepe are image-to-text (I2T).

SugarCrepe++ extends the protocol of SugarCrepe by using two positive captions and requiring both of them to be higher ranked than the negative, to be considered correct. Furthermore, SugarCrepe++ adds the text-to-text tasks (TOT), which probes the text encoders in isolation for compositionality, following the same evaluation pattern.

For the remaining tasks we follow the default evaluation settings of *CLIP_Benchmark* [3]: For ImageNet we report accuracy. For Flickr30k, MSCOCO, DOCCI and IIW we report Recall@5.

Table 1. **Training data of baselines.** Summary of baseline methods and their corresponding training datasets, excluding the original CLIP and SigLIP models. All composition-aware methods rely on training with hard-negatives. The last column indicates whether the provided checkpoints were trained from scratch (✓).

Model	Training data	Train from scratch
Composition-aware		
CE-CLIP [34]	MSCOCO	
NegCLIP [32]	MSCOCO	
CLIC [23]	LAION-1.5B	
DAC [5]	CC3M	
SLVC [4]		
CoN-CLIP [27]		
TripletCLIP [22]		✓
Codebook-based		
Codebook-CLIP [2]		✓
IL-CLIP [35]		✓
Fine-grained training		
DreamLIP-3m [36]		✓
FLAIR-3m [30]		✓
FG-CLIP [31]	LAION-2B + FineHARD	✓
FineCLIP [12]	MSCOCO	
LLIP [14]	Common Crawl 12.8B	✓
Large-sized models		
CLIP-A (ViT-L/14) [18]	LAION-400M	✓
CLIPS (ViT-L/14) [20]	Recap-DataComp-1B	✓
Multi-task training		
BLIP-B [17]	14M samples	-
FLAVA [26]	PMD-80M	-
SigLIP-2 [29]	WebLI	-

4.2. Quantitative evaluation

Comparison with ViT-B baselines. Our main results are shown in Tab. 2. The last two rows show the most important results as they show the most comparable models, both trained by us under the exact same setting on the DreamLIP variant of CC3M with either the standard SigLIP objective or our objectives. It can be seen that fine-tuning on DreamLIP-CC3M alone leads only to marginal improvements with respect to the original SigLIP checkpoint. However, when using our method we observe considerable improvements on all compositionality benchmarks and tasks. Additionally, our model improves on Flickr30k, MSCOCO, and IIW. Thus, it improves both long and short caption retrieval. Only on ImageNet does the performance drop. This

Table 2. **Performance comparison of *ViT-B-based* models**. C^{2}LIP achieves consistently high performance on SugarCrepe and SugarCrepe++ compared to the composition-aware models, despite only relying on regular captions during training. Additionally, our model is able to maintain and improve the retrieval and zero-shot capabilities of the original SigLIP, achieving the best overall score.

Models	SugarCrepe			SugarCrepe++				ImNet1K	Flickr30k	MSCOCO	DOCCI	IIW	**Average**
	Add	Replace	Swap	Replace		Swap							
				I2T	TOT	I2T	TOT						
SigLIP ViT-B/16	86.5	84.1	65.8	73.8	62.8	48.0	34.7	**76.1**	95.2	78.9	58.9	75.3	70.0
CLIP (OpenAI) ViT-B/32	73.0	80.0	62.7	69.5	60.5	45.7	27.4	63.3	89.0	65.4	47.1	69.5	62.8
CLIP (OpenAI) ViT-B/16	72.7	80.4	62.5	70.1	60.2	43.8	23.9	68.4	90.9	67.6	50.4	71.3	63.5
Fine-grained models													
FG-CLIP	84.7	85.1	69.9	75.8	67.5	51.5	38.2	69.0	95.8	78.4	56.7	75.6	70.7
FineCLIP	85.4	85.3	66.8	72.8	68.7	43.9	26.8	55.8	93.5	79.0	44.9	67.1	65.8
DreamLIP-3m	72.9	77.5	64.2	61.2	51.5	44.4	30.1	31.6	83.1	61.2	58.8	**77.9**	59.5
FLAIR-3m	80.6	81.4	70.9	66.3	57.5	49.5	35.2	33.7	90.4	71.5	47.2	72.8	63.1
LLIP	71.4	78.9	57.8	67.3	56.8	40.8	27.3	60.8	90.2	67.5	46.9	70.7	61.4
Codebook-based models													
Codebook-CLIP	50.6	54.5	47.5	34.3	13.8	28.3	10.1	0.0	0.6	0.1	0.1	0.7	20.0
IL-CLIP	51.8	52.8	55.2	34.4	31.8	36.9	14.7	0.0	0.4	0.1	0.1	0.7	23.2
Composition-aware models													
CE-CLIP	92.9	87.0	74.9	56.5	67.0	34.3	32.5	40.4	87.4	71.9	30.8	50.1	60.5
NegCLIP	85.8	85.0	**75.3**	69.1	70.9	53.4	39.1	55.7	92.4	73.9	45.2	66.4	67.7
CLIC	89.8	85.8	72.5	76.6	58.1	**59.0**	27.0	66.6	91.1	67.4	51.3	73.4	68.2
DAC-SAM	91.6	87.0	73.5	52.4	60.2	30.6	18.4	52.3	84.0	58.8	31.2	50.6	57.5
DAC-LLM	93.7	**89.5**	74.6	53.7	59.6	32.2	18.1	51.1	83.7	59.0	27.1	43.9	57.2
SLVC-R	85.4	78.9	68.8	64.0	68.2	51.3	28.4	58.5	90.1	66.9	42.1	66.1	64.1
SLVC-RL	78.5	75.9	65.5	61.9	70.0	46.0	26.3	59.7	90.0	67.0	41.8	67.3	62.5
CoN-CLIP	82.4	77.4	62.4	68.1	70.1	45.2	28.0	63.7	86.0	61.2	46.0	67.1	63.1
TripletCLIP	88.3	87.0	69.9	69.9	69.4	41.4	28.5	45.9	81.7	54.4	39.8	62.5	61.6
Models fine-tuned by us													
SigLIP ViT-B/16 (ft. CC3M)	87.9	85.6	69.7	73.5	67.9	49.8	36.6	75.9	95.6	80.3	59.8	75.5	71.5
C^{2}LIP	**94.2**	88.3	73.1	**79.7**	**75.3**	55.2	**44.2**	73.5	**97.0**	**82.7**	**60.0**	76.4	**75.0**

phenomenon is observed for most composition-aware models. Compared to these methods, the drop on ImageNet for C^{2}LIP is very small. We attribute the drop in performance to the more scene centric representation encouraged by our training method, which can lead to poorer results on tasks like ImageNet classification, an extremely object centric task, for which it is beneficial to ignore everything except for the most salient and central object.

Compared to previous methods that aimed at improving compositionality (composition-aware models in Tab. 2), C^{2}LIP shows consistently high performance on all benchmarks, making it the new SOTA method among models with similar backbone capacity. While C^{2}LIP is outperformed on some metrics by either NegCLIP or DAC-LLM on the SugarCrepe benchmark (which is known to be hackable), it outperforms or performs on par with both methods on the more trustworthy SugarCrepe++ benchmark. Similarly, CLIC surpasses C^{2}LIP on SugarCrepe++ Swap-I2T task, but a closer look at text-only tasks (TOT) reveals that CLIC performs poorly, exposing major limitations in the compositional representation learned by the CLIC text encoder.

In summary, C^{2}LIP demonstrates overall SOTA performance, achieving the best or second best results on most benchmarks and metrics while remaining among the top models on the remaining ones. C^{2}LIP stands out particularly for its consistently high performance across all benchmarks, whereas other models tend to excel on a single benchmark but perform considerably poorer on others. With this reliable performance, C^{2}LIP is an excellent post-training method to improve compositional comprehension in foundation models, without sacrificing performance on other tasks.

Comparison on attribute-binding performance. As shown in the previous section, C^{2}LIP improves all compositionality related tasks; however, the main gain lies in strengthening attribute-object binding. Therefore, we examine the results on attribute–object binding in more detail. As can be seen in Tab. 3, on average, C^{2}LIP outperforms all prior composition-aware methods for attribute related tasks. C^{2}LIP achieves the best results on almost every metric, being surpassed only by DAC-LLM on the SugarCrepe "Replace" task by 0.3 percentage points. Overall, C^{2}LIP shows consistently high performance, while baselines perform poorly on at least one of the tasks.

Comparison with models that use captioning and/or an LLM component. Contrastive learning offers clear advantages over models trained with captioning losses, pri-

Table 3. **Attribute-binding performance comparison.** We compare the accuracy of C^{2}LIP with composition-aware baselines on ***attribute*** replacement and swapping experiments of SugarCrepe and SugarCrepe++. Our model consistently outperforms the competing baselines on most tasks and attains the best overall score. Moreover, previous methods rely on hard negatives to induce a compositional representation, which often harms generalization to natural data and leads to a considerable performance drop on SugarCrepe++. In contrast, C^{2}LIP maintains high performance on all metrics, further highlighting the strength of our proposed approach.

Model	SugarCrepe		SugarCrepe++				**Average**
	Replace	Swap	Replace-I2T	Replace-TOT	Swap-I2T	Swap-TOT	
SigLIP ViT-B/16	86.7	71.5	75.5	64.2	56.3	46.4	66.8
CE-CLIP	88.8	77.0	52.0	64.2	32.3	36.3	58.4
NegCLIP	85.9	75.4	67.1	70.7	55.0	50.3	67.4
CLIC	86.6	74.0	75.9	52.4	62.2	31.2	63.7
DAC-SAM	85.9	75.1	44.3	56.0	33.5	25.4	53.3
DAC-LLM	**89.5**	74.2	47.7	59.5	32.9	24.8	54.8
SLVC-R	81.4	69.1	61.6	67.4	53.2	36.3	61.5
SLVC-RL	76.8	66.5	57.1	67.0	48.8	34.3	58.4
CoN_CLIP	79.7	66.1	68.1	64.7	50.3	37.2	61.0
TripletCLIP	85.7	69.7	66.0	71.1	44.4	38.1	62.5
C^{2}LIP	89.2	**78.8**	**79.8**	**80.2**	**66.2**	**59.8**	**75.7**

Table 4. **Comparison with SOTA foundation models.** We compare performance of C^{2}LIP with larger contrastive models (ViT-L), and SOTA models trained with multimodal/multi-task objectives (BLIP, FLAVA, SigLIP-2). Despite the substantial disadvantages in model size and training, C^{2}LIP has the best performance on "Add" and "Replace" tasks of SugarCrepe, and maintains close performance on other benchmarks. On average our model is only 1.7 percentage points below the best model, CLIPS, which has 5x more parameters.

Models	SugarCrepe			SugarCrepe++				ImNet1K	Flickr30k	MSCOCO	DOCCI	IIW	**Average**
	Add	Replace	Swap	Replace		Swap							
				I2T	TOT	I2T	TOT						
Multi-task SOTA													
BLIP-B	92.2	84.1	73.6	76.7	**86.0**	52.1	41.3	49.2	97.2	79.1	51.6	72.4	71.3
FLAVA	79.6	84.9	73.8	74.3	75.0	56.6	**51.0**	56.9	91.8	73.9	48.3	72.9	69.9
SigLIP2-ViT-B/16	89.0	86.0	71.9	77.7	69.7	**57.9**	42.0	78.5	96.5	**82.8**	62.2	78.1	74.4
Larger models													
CLIP-A (ViT-L/14)	85.8	82.9	63.8	74.7	74.8	45.8	30.8	79.6	95.2	78.1	59.4	75.4	70.5
CLIPS (ViT-L/14)	86.6	86.1	**77.8**	76.6	66.6	59.4	44.3	76.8	**97.7**	82.1	**66.2**	**81.8**	**75.2**
SigLIP ViT-L/16-res256	86.7	84.0	72.8	75.4	64.9	55.6	42.1	**80.5**	96.7	82.3	62.5	77.6	73.4
C^{2}LIP	**94.2**	88.3	73.1	**79.7**	75.3	55.2	44.2	73.5	97.0	82.7	60.0	76.4	75.0

marily due to its inexpensive zero-shot capabilities and the strictly separate image and text encoders; properties that, i.e., BLIP [17] does not possess. Nevertheless, captioning based methods remain popular, and a comparison is instructive. Similarly, LLM based approaches such as FLAVA [26] are rapidly gaining traction. We compare C^{2}LIP with models that incorporate a captioning loss and with FLAVA in the top part of Tab. 4. It can be seen that C^{2}LIP is competitive even against these methods.

Comparison with larger models. Next, we compare against contrastively trained models that have larger parameter count. The results are shown in the middle part of Tab. 4. Again, C^{2}LIP performs competitively on compositionality tasks despite being considerably smaller.

4.3. Qualitative Results

To verify that our method behaves as intended and indeed yields better binding at the visual token level, we visualize *the change in the attention map* when training with our approach compared to SigLIP. Fig. 2 shows results for the caption "black sweater" and "green tablecloth" next to the corresponding images. For instance, it can be seen in Fig. 2a that the sign in the background receives considerably less attention from our C^{2}LIP than from SigLIP. At the same time, attention increases or remains high for most patches that cover the sweater. This clearly illustrates that training with our model improves binding at the patch level (i.e., before global pooling). Moreover, the quantitative results presented in Sec. 4.2 further demonstrate that binding improves after global pooling.

Table 5. **Training objective ablation.** Improvements in (green) with respect to the variant fine-tuned on CC3M. ***npc***: noun-phrase loss, ***xac***: cross-attended concept-aware loss. Both components lead to significant gains.

Training Configurations	SugarCrepe								Flickr8K
	Replace			Swap		Add		**Average**	
	Object	Attribute	Relation	Object	Attribute	Object	Attribute		
Original SigLIP ViT-B/16	95.3	86.7	70.3	60.0	71.5	89.1	83.8	79.5	94.9
+ fine-tuned on CC3M	95.8	87.4	73.5	65.3	74.2	88.9	87.0	81.7	95.2
+ npc	**96.7** (+0.9)	**89.7** (+2.3)	78.5 (+5.0)	66.5 (+1.2)	78.5 (+4.3)	93.0 (+4.1)	93.9 (+6.9)	85.3 (+3.6)	96.3 (+1.1)
+ xac (*full*)	**96.7** (+0.9)	89.2 (+1.8)	**78.9** (+5.4)	**67.3** (+2.0)	**78.8** (+4.6)	**93.5** (+4.6)	**94.9** (+7.9)	**85.6** (+3.9)	**96.4** (+1.2)



(a) **Caption:** "Black sweater"



(b) **Caption:** "Green tablecloth"

Figure 2. **Change in attention when using C^{2}LIP** compared to SigLIP. We visualize the *difference in attention* between C^{2}LIP and SigLIP to the visual tokens given a caption and an image. Higher attention for C^{2}LIP is shown as green. Lower attention for C^{2}LIP indicated with violet. White means no change. **(a)** Black regions that are not a sweater get reduced attention, the sweater gets more or unchanged attention. **(b)** The background, the cups and beer cans get attended less after training with our method, while attention to the green tablecloth increases or stays the same.

4.4. Ablation studies

We perform an ablation study of the components of our pipeline and report the results in Tab. 5. As can be seen, we observe large improvements across the benchmark when using the noun-phrase loss (L_{npc}). Adding the cross-modal attention pooling loss (L_{xac}) yields additional gains, particularly for the "Swap" and "Add" tasks. These results demonstrate that focusing on positive examples, i.e., using concept-centric text tokens as extra positives, provides a substantial boost for compositionality. When a cross-modal attention pooling mechanism is also employed, the network can learn more easily, encouraging stronger binding before the global pooling stage. Consequently, performance improves consistently on all benchmarks.

4.5. C^{2}LIP improves visual instruction tuning

Using C^{2}LIP as image encoder for LLaVa [19] leads to improved compositionality compared to using a SigLIP image encoder. We provide details and results in the appendix.

5. Conclusion

We propose a new method to improve the compositionality of CLIP-like models. We process the training captions to extract noun-phrases, i.e., short segments that correspond to a single object and its attributes. Using the caption decomposition we obtain a context embedding for each concept in the caption. By aligning each concept embedding with the global image embedding and subsequently using a parameter-free attention-pooling function to create visual-domain concept embeddings, we enforce alignment between visual and textual concept embeddings via a contrastive loss. Notably, our method **C^{2}LIP** does not rely on any hard negatives, which often yield improvements that fail to generalize to slightly different compositionality tasks. Our approach is simple and incurs no additional computational overhead at inference time; in fact, inference remains identical to the original model. We evaluate **C^{2}LIP** on standard compositionality benchmarks as well as a variety of zero-shot and retrieval tasks. **C^{2}LIP** achieves consistent improvements on all tested compositionality benchmarks and on most standard retrieval tasks. Owing to its consistently strong performance, **C^{2}LIP** attains overall SOTA performance.

Limitations and future work. While our approach leads to consistent improvements without sacrificing general capabilities, we do observe a small drop in performance on ImageNet. Future work should explore whether it is possible to reduce this loss even further. Similarly, our approach clearly improves the compositional representation, but the final pooling layer remains suboptimal. Future work should explore pooling operations that naturally preserve compositionality, and the extensibility of our proposed method to other types of compositional reasoning beyond object-attribute binding.

References

[1] Rim Assouel, Pietro Astolfi, Florian Bordes, Michal Drozdzal, and Adriana Romero-Soriano. Object-centric binding in contrastive language-image pretraining. *arXiv preprint arXiv:2502.14113*, 2025. 1, 3

[2] Yuxiao Chen, Jianbo Yuan, Yu Tian, Shijie Geng, Xinyu Li, Ding Zhou, Dimitris N Metaxas, and Hongxia Yang. Revisiting multimodal representation in contrastive learning: from patch and token embeddings to finite discrete tokens. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2023. 5

[3] Mehdi Cherti and Romain Beaumont. CLIP benchmark, 2025. 5

[4] 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. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2022. 1, 3, 5

[5] 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. In *Neural Information Processing Systems*, 2023. 1, 3, 5

[6] Sri Harsha Dumpala, Aman Jaiswal, Chandramouli Sastry, Evangelos Milios, Sageev Oore, and Hassan Sajjad. SUGARCREPE++ dataset: vision-language model sensitivity to semantic and lexical alterations. In *Neural Information Processing Systems - Datasets and Benchmarks Track*, 2024. 1, 2

[7] Roopal Garg, Andrea Burns, Burcu Karagol Ayan, Yonatan Bitton, Ceslee Montgomery, Yasumasa Onoe, Andrew Bunner, Ranjay Krishna, Jason Baldridge, and Radu Soricut. Imageinwords: Unlocking hyper-detailed image descriptions, 2024. 5

[8] B Gurung, David T Hoffmann, and Thomas Brox. Common data properties limit object-attribute binding in clip. In *German Conference on Pattern Recognition (GCPR)*, 2025. 1, 3

[9] Matthew Honnibal and Ines Montani. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear, 2017. 3, 4

[10] Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. SugarCrepe: Fixing hackable benchmarks for vision-language compositionality. In *Neural Information Processing Systems - Datasets and Benchmarks Track*, 2023. 1, 2, 3

[11] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, 2021. If you use this software, please cite it as below. 5

[12] Dong Jing, Xiaolong He, Yutian Luo, Nanyi Fei, Guoxing Yang, Wei Wei, Huiwen Zhao, and Zhiwu Lu. FineCLIP: Self-distilled region-based CLIP for better fine-grained understanding. In *Neural Information Processing Systems*, 2024. 5

[13] Alexander Kolesnikov, Alexey Dosovitskiy, Dirk Weissenborn, Georg Heigold, Jakob Uszkoreit, Lucas Beyer, Matthias Minderer, Mostafa Dehghani, Neil Houlsby, Sylvain Gelly, Thomas Unterthiner, and Xiaohua Zhai. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 3

[14] Samuel Lavoie, Polina Kirichenko, Mark Ibrahim, Mido Assran, Andrew Gordon Wilson, Aaron Courville, and Nicolas Ballas. Modeling caption diversity in contrastive vision-language pretraining. In *International Conference on Machine Learning*, 2024. 5

[15] Martha Lewis, Nihal V Nayak, Peilin Yu, Qinan Yu, Jack Merullo, Stephen H Bach, and Ellie Pavlick. Does clip bind concepts? probing compositionality in large image models. *arXiv preprint arXiv:2212.10537*, 2022. 1, 3

[16] Haoxin Li and Boyang Li. Enhancing vision-language compositional understanding with multimodal synthetic data. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2025. 3

[17] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, 2022. 5, 7

[18] Xianhang Li, Zeyu Wang, and Cihang Xie. An inverse scaling law for CLIP training. In *Neural Information Processing Systems*, 2023. 5

[19] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2024. 8

[20] Yanqing Liu, Xianhang Li, Zeyu Wang, Bingchen Zhao, and Cihang Xie. CLIPS: An enhanced CLIP framework for learning with synthetic captions. *arXiv [cs.CV]*, 2024. 3, 5

[21] Yasumasa Onoe, Sunayana Rane, Zachary Berger, Yonatan Bitton, Jaemin Cho, Roopal Garg, Alexander Ku, Zarana Parekh, Jordi Pont-Tuset, Garrett Tanzer, Su Wang, and Jason Baldridge. DOCCI: Descriptions of Connected and Contrasting Images. In *European Conference on Computer Vision*, 2024. 5

[22] Maitreya Patel, Abhiram Kusumba, Sheng Cheng, Changhoon Kim, Tejas Gokhale, Chitta Baral, and Yezhou Yang. TripletCLIP: Improving compositional reasoning of CLIP via synthetic vision-language negatives. In *Neural Information Processing Systems*, 2024. 1, 3, 5

[23] Amit Peleg, Naman Deep Singh, and Matthias Hein. Advancing compositional awareness in CLIP with efficient fine-tuning. In *Neural Information Processing Systems*, 2025. 3, 5

[24] Alec Radford, Jong Wook Kim, Chris Hallacy, A. Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021. 1, 3

[25] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, 2018. 4

[26] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. FLAVA: A foundational language and vision alignment model. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2022. 5, 7

[27] Jaisidh Singh, Ishaan Shrivastava, Mayank Vatsa, Richa Singh, and Aparna Bharati. Learn "no" to say "yes" better: Improving vision-language models via negations. In *Winter Conference on Applications of Computer Vision*, 2025. 1, 3, 5

[28] Yingtian Tang, Yutaro Yamada, Yoyo Zhang, and Ilker Yildirim. When are lemons purple? the concept association bias of vision-language models. In *Empirical Methods in Natural Language Processing*, 2023. 1, 3

[29] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025. 5

[30] Rui Xiao, Sanghwan Kim, Mariana-Iuliana Georgescu, Zeynep Akata, and Stephan Alaniz. FLAIR: Vlm with fine-grained language-informed image representations. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2025. 3, 5

[31] Chunyu Xie, Bin Wang, Fanjing Kong, Jincheng Li, Dawei Liang, Gengshen Zhang, Dawei Leng, and Yuhui Yin. FG-CLIP: Fine-grained visual and textual alignment. In *International Conference on Machine Learning*, 2025. 5

[32] Mert Yüksekgönül, 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? In *International Conference on Learning Representations*, 2023. 1, 3, 5

[33] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *IEEE International Conference on Computer Vision*, 2023. 3

[34] Le Zhang, Rabiul Awal, and Aishwarya Agrawal. Contrasting intra-modal and ranking cross-modal hard negatives to enhance visio-linguistic compositional understanding. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2024. 1, 3, 5

[35] Chenhao Zheng, Jieyu Zhang, Aniruddha Kembhavi, and Ranjay Krishna. Iterated learning improves compositionality in large vision-language models. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2024. 5

[36] Kecheng Zheng, Yifei Zhang, Wei Wu, Fan Lu, Shuailei Ma, Xin Jin, Wei Chen, and Yujun Shen. Dreamlip: Language-image pre-training with long captions. In *European Conference on Computer Vision*, 2024. 3, 4, 5