Preserving Multi-Modal Capabilities of Pre-trained VLMs for Improving Vision-Linguistic Compositionality

Anonymous ACL submission

Abstract

In this paper, we propose a new method to enhance compositional understanding in pre-trained vision and language models (VLMs) without sacrificing performance in the model's original zero-shot multi-modal tasks. Traditional fine-tuning methods often improve compositional reasoning at the expense of multi-modal capabilities. This drawback stems from the use of global hard negative loss, which contrasts the global representations of images and texts. This can distort multi-modal representations by pushing original texts due to ambiguous global representations. To address this, we propose the Fine-grained Selective Calibrated CLIP (FSC-CLIP). This incorporates local hard negative loss and selective calibrated regularization, designed to provide fine-grained negative supervision while maintaining the integrity of representations. Our extensive evaluation across various benchmarks for compositionality and multi-modal tasks shows that FSC-CLIP not only achieves compositionality on par with state-of-the-art models but also maintains multi-modal capabilities.

1 Introduction

Humans naturally excel at multi-modal understanding, effortlessly perceiving and interpreting different modalities, such as images and text, and forming associations between them. This capability is evident in recognizing novel concepts (Fu et al., 2018), cross-modal retrieval (Kaur et al., 2021), and compositional reasoning (Levesque et al., 2012). To achieve this ability in artificial intelligence, foundational vision and language models (VLMs) have been trained on large-scale image-text datasets (Schuhmann et al., 2022b), significantly bridging the gap between human and machine capabilities in tasks like zero-shot recognition and image-text retrieval (Radford et al., 2021).

Despite these advances, VLMs still face challenges in compositional understanding (Yuksekgonul et al., 2023). Humans intuitively grasp complex compositional language within images, involving spatial reasoning, attributes and relationships in objects, and equivariance between image and text (Wang et al., 2023). In contrast, VLMs often fail to understand these nuanced relationships (Liu et al., 2023a; Ray et al., 2023). This shortfall is attributed to their reliance on single-vector representations (Kamath et al., 2023a) and limited ability to match compositional knowledge (Wang et al., 2024), which restricts effective encoding and utilization of compositional language.

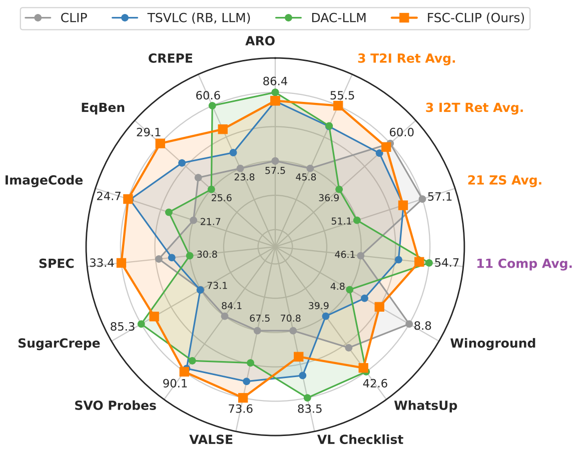

Figure 1: A holistic comparison of fine-tuning methods for visio-linguistic compositionality. Enhancing compositionality often compromises multi-modal task performance in previous approaches. Our FSC-CLIP bridges this gap, minimizing these trade-offs. Full experimental results are provided in Tab. 1.

To improve compositionality in VLMs, both pre-training (Singh et al., 2023; Zheng et al., 2024) and fine-tuning (Zhang et al., 2024; Singh et al., 2024) methods have been proposed. In particular, fine-tuning, which leverages pre-trained knowledge and is cost-effective, is widely adopted in academia. Typically, this involves incorporating hard negative texts (Doveh et al., 2022, 2023; Herzig et al., 2023) into training. However, as shown in Fig. 1, this approach can result in a trade-off, where gains in compositionality come at the expense of performance in the multi-modal tasks: zero-shot classification (`ZS`) and image-to-text retrieval (`I2T Ret`). The hard negative losses in previous methods, which operate on global image and text representations, may disrupt the well-established multi-modal representations due to the ambiguous encoding of original and negative texts (Kamath et al., 2023b).

To this end, we propose a new fine-tuning framework designed to enhance compositional reasoning in pre-trained VLMs while preserving their capabilities in original multi-modal tasks. This approach tackles the degradation of multi-modal representations caused by global hard negative loss on single vector representations, which struggles to capture subtle informational differences between hard negative texts and the original text.

Our framework introduces two key innovations: **(1) Local Hard Negative (LHN) Loss**. We utilize dense alignments between image patches and text tokens to calculate the hard negative loss. This approach, inspired by the dense alignment for vision-language representation (Huang et al., 2021; Bica et al., 2024), aggregates local similarity scores to enhance compositional understanding without undermining multi-modal representations.
(2) Selective Calibrated Regularization (SCR). To mitigate the adverse effects of hard negative losses, which can push original text representations away due to blurred text representations, SCR selectively focuses on challenging hard negative texts. Furthermore, it introduces a slight positive margin for these texts, helping to calibrate the confusion.

The whole framework, dubbed **F**ine-grained and **S**elective **C**alibrated CLIP, offers fine-grained supervision of hard negatives while preserving the integrity of multi-modal representations. As shown in Fig. 1, `FSC-CLIP` not only improves compositionality but also maintains high performance in multi-modal tasks. It outperforms DAC-LLM in `ZS` and `I2T Ret` scores, while achieving similar compositionality (`Comp`) across a wide range of tasks. We summarize our contributions as follows:

- We propose a novel fine-tuning methodology, `FSC-CLIP`, that aims to enhance visio-linguistic compositionality in pre-trained VLMs while maintaining their multi-modal task capabilities.
- We design a local hard negative (LHN) loss and a selective calibrated regularization (SCR) mechanism, effectively capturing subtle differences in hard negative texts and preserving the integrity of multi-modal representations.
- We validate `FSC-CLIP` through an extensive range of experiments, covering 11 compositionality, 21 zero-shot recognition, and 3 image-text retrieval tasks, establishing a comprehensive evaluation of VLMs' multifaceted capabilities.

2 Related Work

Contrastive Vision-Language Models. CLIP (Radford et al., 2021) has revolutionized the multi-modal domain through large-scale pre-training of image-text alignment, showing the remarkable zero-shot capabilities. CLIP utilizes a dual-encoder architecture, which enables versatility across a broad spectrum of vision (Zhou et al., 2022; Liang et al., 2023), and vision-language (Mokady et al., 2021; Kwon and Ye, 2022) downstream tasks. They also serve as the building blocks for modern foundational models in various tasks, including advanced VLMs (Li et al., 2022b), multi-modal language models (MLLMs) (Li et al., 2023; Liu et al., 2023b), and generative models (Podell et al., 2023; Huang et al., 2023). Additionally, these models extend their utility to linking 3D (Sun et al., 2024) or audio (Elizalde et al., 2023) to language, highlighting the essential roles of both multi-modal and compositional tasks in practical applications. We aim to enhance CLIP's compositional understanding while preserving its multi-modal capabilities.
Visio-Linguistic Compositionality. Although vision and language models (VLMs) have promising capabilities like zero-shot classification and retrieval (Radford et al., 2021; Zeng et al., 2022), they still lack compositional reasoning that requires fine-grained understanding (Peng et al., 2024) between image and text. Numerous benchmarks have been proposed, testing various aspects such as attributes, relationships and objects (Zhao et al., 2022; Yuksekgonul et al., 2023), spatial reasoning (Kamath et al., 2023b; Liu et al., 2023a) and linguistic phenomena (Parcalabescu et al., 2022). Meanwhile, incorporating hard negative captions during fine-tuning has become common to enhance compositionality (Zhang et al., 2024), generated through rule-based methods (Doveh et al., 2022; Yuksekgonul et al., 2023), large language models (Doveh et al., 2023), and scene graphs (Singh et al., 2023; Herzig et al., 2023). We comprehensively evaluate the capabilities of VLMs across a broad range of compositionality and multi-modal tasks.

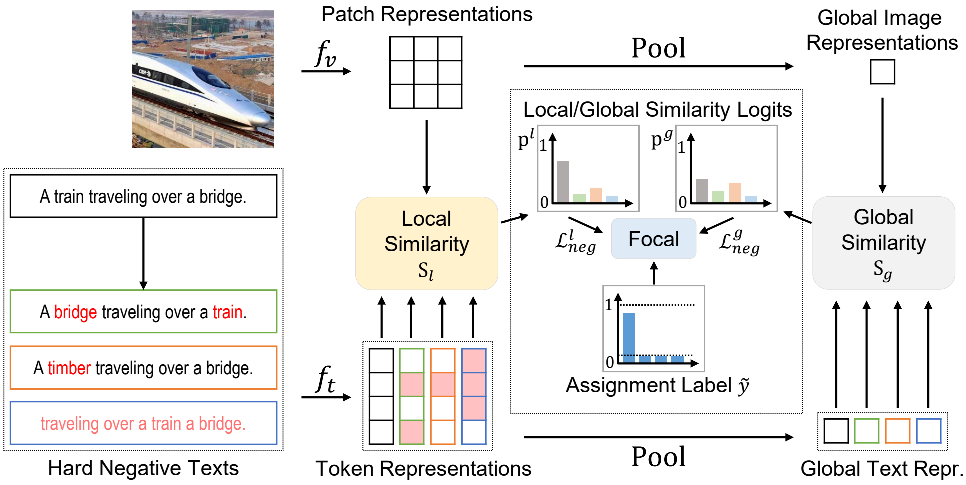

Figure 2: A complete `FSC-CLIP` framework that incorporates Local Hard Negative (LHN) Loss with Selective Calibrated Regularization (SCR), alongside a global HN loss. The LHN loss measures similarity between an image and a text at the patch and token levels to more accurately identify subtle differences between original and HN texts. SCR combines focal loss with label smoothing to mitigate the adverse effects of using hard negative losses.

3 Methodology

We first outline the fine-tuning setting of CLIP in Sec. 3.1. We then introduce `FSC-CLIP`, which includes **Local Hard Negative (LHN) Loss** and **Selective Calibrated Regularization (SCR)** in Secs. 3.2 and 3.3. The training objective for `FSC-CLIP` is detailed in Sec. 3.4. We illustrate the `FSC-CLIP` framework, which integrates both global and local HN losses with SCR as shown in Fig. 2.

3.1 CLIP with Global Contrastive Loss

CLIP objective. Consider a mini-batch $\mathcal{B} = \{(I_i, T_i)\}_{i=1}^{B}$ of size B, consisting of image and text pairs (I_i, T_i). Using CLIP's visual and language encoders, $f_v(\cdot)$ (*e.g.*, ViT (Dosovitskiy et al., 2021)) and $f_t(\cdot)$ (*e.g.*, Transformers (Vaswani et al., 2017)), each image I_i is encoded into a sequence of visual tokens $\mathbf{V}_i = f_v(I_i)$, and each text T_i into a sequence of textual tokens $\mathbf{T}_i = f_t(T_i)$. These sequences are represented in a shared multi-modal space, with $\mathbf{V}_i = \{\mathrm{v}_{p,i}\}_{p=1}^{P}$ comprising P local patch embeddings and $\mathbf{T}_i = \{\mathrm{t}_{w,i}\}_{w=1}^{W}$ consisting of W token embeddings, both in the shared embedding dimension d. The global representations of image and text v_i and $t_i \in \mathbb{R}^d$ can be obtained by pooling the local representations: $v_i = \mathrm{Pool}(\mathbf{V}_i)$ and $t_i = \mathrm{Pool}(\mathbf{T}_i)$, respectively. For example, $\mathrm{Pool}(\cdot)$ corresponds to `avgpool` and `argmax` for images and texts in (Radford et al., 2021)).

CLIP aligns the corresponding images and texts by measuring the global-level similarity:

$$\mathrm{S}_g(I_i, T_i) = \exp\left(\cos(v_i, t_i)/\tau\right), \tag{1}$$

where $\cos(v, t) = \frac{v^T t}{\|v\| \cdot \|t\|}$. The image to text loss $\mathcal{L}_{i2t}$ of CLIP maximizes $\mathrm{S}_g(I_i, T_i)$, while minimizing $\mathrm{S}_g(I_i, T_j)$ for all non-matching texts $j \neq i$:

$$\mathcal{L}_{i2t} = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\mathrm{S}_g(I_i, T_i)}{\sum_{j=1}^{B} \mathrm{S}_g(I_i, T_j)}, \tag{2}$$

and the text to image loss $\mathcal{L}_{t2i}$ is the reciprocal of $\mathcal{L}_{i2t}$ which aligns the matching image per text. The final CLIP loss is $\mathcal{L}_{\text{clip}} = \frac{1}{2}(\mathcal{L}_{i2t} + \mathcal{L}_{t2i})$.

Incorporating hard negative texts. To enhance the compositional reasoning of CLIP, hard negative (HN) texts are commonly incorporated into training, whether they are rule-based (Yuksekgonul et al., 2023) or generated by language models (Doveh et al., 2023). Consider a set of K different HN texts $\tilde{\mathrm{T}}_i = \{\tilde{T}_i^k\}_{k=1}^{K}$ originated from T_i. We introduce a separate hard negative loss added to $\mathcal{L}_{\text{clip}}$, similar to (Doveh et al., 2022). First, we compute a similarity prediction probability p_i^g, assigned to the original caption T_i as follows:

$$p_i^g = \frac{\mathrm{S}_g(I_i, T_i)}{\mathrm{S}_g(I_i, T_i) + \sum_{k=1}^{K} \mathrm{S}_g\left(I_i, \tilde{T}_i^k\right)}. \tag{3}$$

Here, g represents the global representation, and the hard negative (HN) loss applied to this similarity assignment is formulated as cross entropy:

$$\mathcal{L}_{neg}^{g} = -\frac{1}{B}\sum_{i=1}^{B} \log p_i^g. \tag{4}$$

However, incorporating such global HN loss can inadvertently harm the multi-modal representations due to the similarly encoded global text representations between original and HN texts.

3.2 Local Hard Negative (LHN) Loss

To address the issue, we propose a novel Local Hard Negative (LHN) loss that utilizes a local similarity score $\mathrm{S}_l(I,T)$. This score focuses on the local alignment between text tokens and sub-image regions (Huang et al., 2021; Bica et al., 2024), better capturing subtle differences between the original text and HN texts. Replacing the global similarity S_g with S_l, the LHN loss is formulated as:

$$\mathcal{L}_{neg}^{l} = \frac{-1}{B}\sum_{i=1}^{B}\log \underbrace{\frac{\mathrm{S}_l\left(I_i,T_i\right)}{\mathrm{S}_l\left(I_i,T_i\right)+\sum_{k=1}^{K}\mathrm{S}_l\left(I_i,\tilde{T}_i^k\right)}}_{p_i^l}, \tag{5}$$

where p_i^l represents the local similarity prediction, and the LHN loss is calculated in the same manner as $\mathcal{L}_{neg}^{g}$ in Eq. (4). We further describe the process for obtaining the local similarity $\mathrm{S}_l(I,T)$.

Textual-aligned Visual Patches. $\mathrm{S}_l(I,T)$ measures the similarity between token and patch embeddings for each token in the given text T. From the patch representations $\mathbf{V} = \{\mathrm{v}_p\}_{p=1}^{P}$, we first derive the textual-aligned patch embeddings $\hat{\mathbf{V}} = \{\hat{\mathrm{v}}_w\}_{w=1}^{W}$, corresponding to each textual token feature t_w in $\mathbf{T} \in \mathbb{R}^{W,d}$. This is achieved by performing a weighted average of patches $\mathbf{V}$ using attention weights $\mathrm{a} \in \mathbb{R}^{W,P}$ derived from normalizing the similarity map $\mathrm{s} \in \mathbb{R}^{W,P}$ between token and patch embeddings. This process assigns a patch embedding to each token, enabling similarity measurement on a per-token basis. We denote the similarity map as $\mathrm{s} = \mathbf{T}^T\mathbf{V} \in \mathbb{R}^{W,P}$, where $\mathrm{s}_{w,p} = \mathrm{t}_w^T \mathrm{v}_p$. To relate multiple similar patches for a single token, we min-max normalize s to obtain a:

$$\mathrm{a}_{w,p} = \frac{\mathrm{s}_{w,p} - \min_k \mathrm{s}_{w,k}}{\max_k \mathrm{s}_{w,k} - \min_k \mathrm{s}_{w,k}}, \tag{6}$$

and use the attention weights a to aggregate $\mathbf{V}$, obtaining textual-aligned patches $\hat{\mathbf{V}} = \{\hat{\mathrm{v}}_w\}_{w=1}^{W}$:

$$\hat{\mathrm{v}}_w = \frac{1}{\sum_{p=1}^{P}\mathrm{a}_{w,p}} \cdot \sum_{p=1}^{P}\mathrm{a}_{w,p}\cdot \mathrm{v}_p. \tag{7}$$

Token-level Similarity. Having obtained the textual-aligned visual tokens $\hat{\mathbf{V}}$, we aggregate the per-token similarities between $\hat{\mathbf{V}}$ and $\mathbf{T}$:

$$\mathrm{S}_l\left(I,T\right) = \sum_{w=1}^{W}\exp\left(\cos\left(\hat{\mathrm{v}}_w,\mathrm{t}_w\right)/\tau\right), \tag{8}$$

where $\hat{\mathrm{v}}_w \in \hat{\mathbf{V}}$ and $\mathrm{t}_w \in \mathbf{T}$. Unlike $\mathrm{S}_g(I,T)$, $\mathrm{S}_l(I,T)$ focuses on the local alignment between image and text, better distinguishing features between correct and HN texts, thereby reducing the negative impact on the multi-modal representations by the hard negative loss, as illustrated in Fig. 2.

We observe that $\mathcal{L}_{neg}^{l}$ maintains multi-modal task performance close to that of the pre-trained representations, while significantly boosting compositionality. Additionally, we highlight this process does not introduce any additional model parameters for heavy modality interaction layers (*e.g.*, cross attention) (Li et al., 2022b; Yu et al., 2022). It also maintains the efficient inference pipeline of CLIP without relying on text-dependent image embeddings during inference (Lavoie et al., 2024).

3.3 Selective Calibrated Regularization (SCR)

Reliance on the HN losses can adversely affect multi-modal representations. To counteract this, we propose a Selective Calibrated Regularization (SCR) mechanism applicable to both global and local HN losses. SCR comprises two complementary components: one regulates the prediction of image-text similarity, while the other adjusts the assignment labels. Our experimental validation confirms that both components are crucial for preserving the integrity of the representations.

Focal Loss to Target Challenging HN Texts. We intend to focus selectively on *challenging* HN texts, *i.e.*, those with higher similarity to the image than positive texts. This strategy is aligned with the concept behind focal loss (Lin et al., 2017). Formally, let the similarity prediction logit vector of the i-th batch item along with K generated HN texts be $\mathrm{p}_i \in \mathbb{R}^{1+K}$, where the first element corresponds to the original text. Depending on whether using global or local representations, the logit vector is further represented as either p_i^g or p_i^l, similar to Eqs. (4) and (5). The respective HN losses can be re-formulated in a vector representation with p_i as $\mathrm{CE}(\mathrm{p}_i, \mathrm{y}_i) = \sum_{k=0}^{K} l_{i,k}$, where $l_{i,k} = -\mathrm{y}_{i,k}\log \mathrm{p}_{i,k}$ and $\mathrm{y}_i = \mathbb{1}_{[k=0]} \in \mathbb{R}^{1+K}$ indicates the assignment label between an image and all texts. To reduce the negative impact caused by the confidently correct associations, we apply confidence-based weighting to CE loss:

$$\mathrm{Focal}\left(\mathrm{p}_i,\mathrm{y}_i\right) = \sum_{k=0}^{K}\left(1-\mathrm{p}_{i,k}\right)^{\gamma} l_{i,k}, \tag{9}$$

where γ is the modulation parameter. This strategy prioritizes challenging image-text associations, which are crucial for learning compositionality.

Label Smoothing to Calibrate HN Text Assignments. From the HN losses in Eqs. (4) and (5), the label vector y_i assigns a value of 1 exclusively to the single positive text, while assigning a value of 0 to all HN texts, thereby producing a binary label vector. This treats HN texts as certainly negative. Given that the original text and its hard negative (HN) texts exhibit similar representations from a global perspective, we assign a slight positive margin to the HN texts instead of categorizing them as entirely negative. Specifically, we adopt label smoothing (Guo et al., 2017) to the assignment label vector y_i, using a smoothing parameter β:

$$\tilde{\mathrm{y}}_{i,k} = (1-\beta) \cdot \mathrm{y}_{i,k} + \frac{\beta}{1+K}, \quad (10)$$

where $\tilde{\mathrm{y}}_i$ provides such non-binary label for the global and local HN losses, *i.e.*, $\mathrm{Focal}(\mathrm{p}_i, \tilde{\mathrm{y}}_i)$. This accommodates similar representations in the HN texts, preserving the original representations.

3.4 Overall Training Objective

Our framework incorporates two hard negative losses, $\mathcal{L}^g_{neg}$ and $\mathcal{L}^l_{neg}$, representing global and local HN losses respectively, into the CLIP training loss $\mathcal{L}_{\text{clip}}$ with additional hard negative texts:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{clip}} + \lambda_g \mathcal{L}^g_{neg} + \lambda_l \mathcal{L}^l_{neg}, \quad (11)$$

where λ_g and λ_l are the weighting factors for the global and local HN losses, respectively. Training with $\mathcal{L}_{\text{total}}$ neither modifies the architecture of CLIP nor introduces additional model parameters.

4 Experiments

For reproducibility, we will release our codes for training and evaluation, along with the checkpoints.

Training Datasets. We consider two image-text datasets for fine-tuning: LAION-COCO (Schuhmann et al., 2022a) and CC-3M (Sharma et al., 2018), each with a 100K randomly sampled subset, consistent with the literature (Singh et al., 2023; Zhang et al., 2024). For training, we use synthetic captions generated by an image captioning model from paired images instead of raw captions. Specifically, LAION-COCO captions are generated using BLIP (Li et al., 2022b) with ViT-L/14, applied to LAION-2B (Schuhmann et al., 2022b). For the CC-3M subset, we generated synthetic captions using CoCa (Yu et al., 2022) with ViT-L/14. Importantly, we avoid using COCO 100K subset (Yuksekgonul et al., 2023) for fine-tuning as it shares data with several evaluation benchmarks, which could inadvertently influence the results, as also noted by (Singh et al., 2023).

Hard Negative (HN) Texts. We adopt a simple rule-based methods for generating hard negative texts that do not rely on external language models such as (Le Scao et al., 2023) adopted in (Doveh et al., 2023). Consequently, rule-based approach enables online text augmentation at each training step, ensuring variations in each iteration. For each caption, we apply three distinct negative augmentations in an online version: `negclip` (Yuksekgonul et al., 2023), `replace` (Hsieh et al., 2023), and `bi-gram shuffle`. This process results in a total of four captions, including the original one, paired with an image for every batch item. We provide further details on these augmentations, along with corresponding examples, in Appendix A.1.

Training Setup. Consistent with previous methods (Yuksekgonul et al., 2023; Zhang et al., 2024; Singh et al., 2023), we trained our models during 5 epochs with batch size 256, using OpenCLIP repository (Ilharco et al., 2021). The learning rate is set to 5e-6 and decayed by a cosine schedule, with a warmup of 50 steps. Models are optimized using AdamW with a weight decay of 0.1. We use a single Quadro RTX 8000 GPU with 48GB memory for training. Images are re-scaled to 224, and the context length is 77 for texts. We set the weighting factors $\lambda_g = 0.5$ and $\lambda_l = 0.2$. For SCR, we set $\gamma = 2.0$ and $\beta = 0.02$ for focal loss and label smoothing, respectively. We also explore fine-tuning with LoRA (Hu et al., 2022) setting the rank to 4 as in (Doveh et al., 2022, 2023). Training our model takes less than one hour for 100K samples.

Evaluation Setup. We use an *extensive* range of compositionality and multi-modal task benchmarks for a comprehensive evaluation, far surpassing the scope of previous works. For compositionality, we employ 11 benchmarks in total: ARO (Yuksekgonul et al., 2023), CREPE-Productivity (Ma et al., 2023), EqBen (Wang et al., 2023), ImageCoDe (Krojer et al., 2022), SPEC (Peng et al., 2024), SugarCrepe (Hsieh et al., 2023), SVO Probes (Hendricks and Nematzadeh, 2021), VALSE (Parcalabescu et al., 2022), VL-Checklist (Zhao et al., 2022), WhatsUp (Kamath et al., 2023b), Winoground (Thrush et al., 2022), testing a diverse array of aspects

Method	LoRA	ARO	CREPE	EqBen	ImageCoDe	SugarCrepe	SVO Probes	VALSE	VL-Checklist	WhatsUp	Winoground	SPEC	`Comp`	`ZS`	`I2T Ret`	`T2I Ret`
CLIP (ViT-B/32)		57.5	23.8	26.5	21.7	73.1	84.1	67.5	70.8	41.5	8.8	31.9	46.1	57.1	60.0	45.8
Fine-tuned: MS-COCO, 100K Samples																
NegCLIP[1]		80.9	30.3	**30.3**	**26.4**	83.7	**90.8**	73.7	74.9	42.9	8.0	**34.6**	52.4	55.9	66.8	58.4
CE-CLIP[2]		76.3	34.7	26.8	24.5	**85.7**	90.1	**76.7**	76.9	41.7	5.2	33.0	52.0	49.9	59.2	57.4
GNM-CLIP[3]		57.1	17.4	28.3	25.0	78.7	89.2	71.1	70.6	42.1	**10.2**	33.1	47.5	**56.3**	66.1	55.5
MosaiCLIP[†,4]		82.6	-	-	-	-	90.7	-	76.8	-	-	-	-	-	-	-
Fine-tuned: Conceptual Captions – 3M (CC-3M), 100K Samples																
MosaiCLIP[†,*,4]		78.6	-	-	-	-	88.7	-	77.6	-	-	-	-	-	-	-
NegCLIP[‡]		**86.5**	50.5	25.8	24.6	83.4	88.6	72.4	79.0	**43.2**	7.0	32.8	54.0	52.6	51.8	54.1
`FSC-CLIP` (Ours)		78.8	44.0	28.5	25.2	84.3	88.2	74.9	77.4	42.6	6.8	34.2	53.2	53.5	55.8	54.6
`FSC-CLIP` (Ours)	✓	84.4	50.6	27.7	24.5	82.3	88.8	74.5	80.3	42.1	5.0	32.2	53.9	53.6	56.1	54.0
Fine-tuned: Conceptual Captions - 3M (CC-3M), 3M Samples																
TSVLC[5] (RB)	✓	83.5	36.1	27.4	24.0	76.9	89.8	69.3	77.5	40.9	6.8	31.6	51.2	54.9	54.9	52.1
TSVLC[5] (RB+LLM)	✓	82.7	33.1	27.6	24.6	73.2	89.7	72.2	79.2	39.9	5.8	31.4	50.9	55.4	55.1	52.3
DAC-LLM[6]	✓	86.4	60.6	25.6	22.8	85.3	88.9	70.5	83.5	42.6	4.8	30.8	54.7	51.1	36.9	52.4
DAC-SAM[6]	✓	83.3	**63.7**	25.3	24.3	83.8	88.5	70.2	**84.7**	42.4	8.5	29.9	**55.0**	51.9	41.1	49.0
MosaiCLIP[†,4]		80.4	-	-	-	-	-	-	77.3	-	-	-	-	53.5	-	-
Fine-tuned: LAION-COCO, 600M Samples																
CLoVe[7]		83.0	41.7	26.9	25.3	84.6	87.9	71.8	66.6	41.8	6.5	31.7	51.6	51.0	53.1	**56.0**
Fine-tuned: LAION-COCO, 100K Samples																
NegCLIP[‡]		86.4	48.7	27.2	25.3	80.9	89.6	70.9	76.0	43.0	7.8	32.3	53.5	54.1	52.3	54.1
`FSC-CLIP` (Ours)		82.8	46.8	29.1	24.7	82.6	90.1	73.6	75.7	42.4	6.8	33.4	53.5	55.3	**58.2**	55.5
`FSC-CLIP` (Ours)	✓	85.5	54.4	29.1	24.9	80.6	89.7	72.6	78.4	42.8	5.8	32.5	54.2	55.9	57.3	54.3

[†]Numbers taken from the original paper. [*]Fine-tuned on 100K subset of CC12M. [‡]Our implementation, without additional image batch.

References: [1](Yuksekgonul et al., 2023) [2](Zhang et al., 2024) [3](Sahin et al., 2024) [4](Singh et al., 2023) [5,6](Doveh et al., 2022, 2023) [7](Castro et al., 2024)

Table 1: A comprehensive comparison of various fine-tuning methods applied to the pre-trained CLIP ViT-B/32 model across 11 compositionality, 21 zero-shot classification, and 3 retrieval tasks, including their meta averages: `Comp`, `ZS`, and `I2T/T2I Ret`. `FSC-CLIP` achieves superior compositionality scores while maintaining strong multi-modal task performance. The best numbers are **bold**, and the second-best numbers are underlined for each metric.

for compositional reasoning. For the multi-modal tasks, we consider 21 zero-shot classification tasks, combining ImageNet (Deng et al., 2009) and 20 datasets from the ELEVATER toolkit (Li et al., 2022a). We also evaluate on COCO (Chen et al., 2015), Flickr30k (Young et al., 2014), and COCO-Counterfactuals (Le et al., 2023) for retrieval.

We report a single aggregated number, which is the average of sub-tasks for each compositionality benchmark. We also provide the meta-average across all compositionality benchmarks (`Comp`), the average performance over 21 zero-shot classification tasks (`ZS`), and the average Recall@1 for three image-to-text (`I2T Ret`) and text-to-image (`T2I Ret`) retrieval tasks, as shown in Tab. 1. For a fair and consistent comparison, we run evaluations for all the models including previous methods with available checkpoints, across all the benchmarks.

4.1 Main Results

We compare our `FSC-CLIP` to previous fine-tuning methods for compositionality. We report both compositionality and multi-modal task performance as shown in Tab. 1. In Fig. 3, we visualize the trade-off trajectory between `Comp` and `ZS` through the robust fine-tuning method (Wortsman et al., 2022). Here, CLIP ViT-B/32 from OpenAI (Radford et al., 2021) is fine-tuned on the respective datasets.

Compositionality while Sacrificing Multi-Modal Tasks. We introduce our baseline, NegCLIP[‡], directly comparable to our `FSC-CLIP`. Unlike the original implementation (Yuksekgonul et al., 2023), we utilize an online version of hard negatives generation (*e.g.*, `negclip`) and omit additional similar image batches. This baseline will be further used in our ablation study. As indicated in Tab. 1, NegCLIP, fine-tuned with subsets of CC-3M and LAION-COCO, demonstrates competitive `Comp` scores compared to methods like TSVLC[5], and CLoVe[7]. However, both NegCLIP and other methods experience a significant decline in `ZS` and `I2T Ret` scores relative to the pre-trained CLIP. For instance, CE-CLIP[2] increases the meta-average of compositionality scores, `Comp`, by 5.9 but the `ZS`

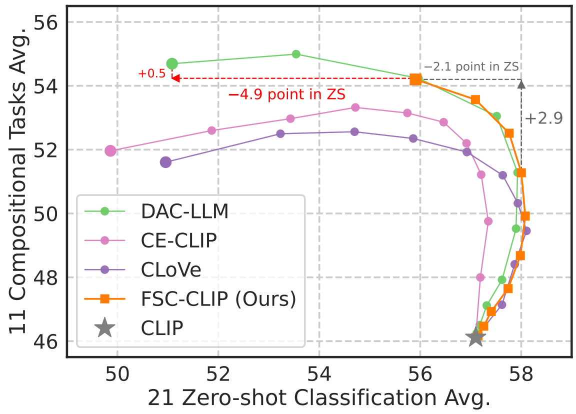

Figure 3: Fine-tuning trajectories between compositionality (Comp) and zero-shot classification (ZS) via robust fine-tuning method (Wortsman et al., 2022). Each point represents the interpolated model between the pre-trained and each fine-tuned version, at varying ratios. FSC-CLIP offers better trade-offs between Comp and ZS, maintaining ZS scores in the fully fine-tuned model.

score drops drastically by 7.2, compared to the pre-trained CLIP. Similarly, DAC-LLM[6], despite strong Comp score aided by LLM-augmented captions, shows marked declines in both ZS and I2T Ret by 6.0 and 23.1, respectively. Meanwhile, GNM-CLIP[3] maintains a ZS score close to that of the pre-trained model, but shows only a modest increase in Comp. These methods apply hard negative (HN) loss to global-level representations, potentially causing the observed performance drops. As note, we have grayed out the retrieval scores of models fine-tuned on COCO due to the influence of overlapping data on these tasks.

Preserving Multi-Modal Tasks. FSC-CLIP stands out by achieving Comp scores higher than previous models and comparable to DAC-LLM, while also maintaining robust multi-modal task performance. Specifically, when fine-tuned on the 100K subset of LAION-COCO, our model attains a Comp score of 53.5 – significantly surpassing its pre-trained counterpart – and a ZS score of 55.9, nearly matching the pre-trained CLIP. It also reaches an I2T Ret score of 58.2, the highest among models not fine-tuned on COCO. Further improvements are observed with using LoRA (Hu et al., 2022) for fine-tuning, which boosts the Comp score to 54.2 while maintaining the ZS score. Similar positive trends are evident when we fine-tune FSC-CLIP on the 100K subset of CC3M. Remarkably, these results are achieved by our innovative Local HN loss and Selective Calibrated Regularization design. We further analyze these contributions in Sec. 4.2.

id	$\mathcal{L}^g_{neg}$	$\mathcal{L}^l_{neg}$	Focal	LS	Comp	ZS	I2T Ret	T2I Ret
1	✓	-	-	-	54.0	53.6	47.4	53.7
2	-	✓	-	-	51.7	55.7	61.6	54.5
3	✓	✓	-	-	54.4	52.6	46.9	53.8
4	✓	✓	✓	-	54.2	54.2	53.1	54.8
5	✓	✓	-	✓	53.9	53.8	51.7	54.9
6	✓	✓	✓	✓	53.5	55.3	58.2	55.5
7	✓	-	✓	✓	52.8	55.3	57.1	55.6
8	-	✓	✓	✓	50.2	55.9	63.2	55.1

Table 2: Impact by individual component. The local HN loss preserves multi-modal task performance. In addition, focal loss and label smoothing (LS) in SCR complement each other, improving the decreased multi-modal task performance caused by the HN losses.

Robust Fine-tuning on Compositionality and Zero-shot Tasks. As depicted in Fig. 3, we utilize the weight-space ensembling technique, WiSE-FT (Wortsman et al., 2022), to compare different fine-tuning methods and their trajectories, specifically in terms of Comp and ZS scores. We create intermediate models by interpolating between each fine-tuned model and the pre-trained one. The blending ratio increases from 0.0 (*e.g.*, pre-trained) to 1.0 (*e.g.*, fully fine-tuned), in increments of 0.1.

FSC-CLIP attains a ZS score of 58 at the intermediate, surpassing the scores of other models, while improving Comp to 50. When fully fine-tuned, it attains superior Comp score and offers better trade-offs than CLoVe and CE-CLIP, without the significant loss in ZS. In contrast, DAC-LLM sees a significant drop in ZS, gaining only 0.5 point in Comp, as highlighted by the red marker. Meanwhile, FSC-CLIP not only matches but exceeds the ZS score by 4.9 in the fully fine-tuned model.

4.2 Analysis

We further present an in-depth analysis on our FSC-CLIP including ablation study, as follows:

Impact of Individual Components. From Tab. 2, we observe that applying the local HN loss alone (row 2) surprisingly preserves the multi-modal scores. However, when both global and local HN losses are activated (row 3), Comp is further boosted but at the cost of ZS and I2T Ret scores, likely due to the complicated adverse effects of the losses. The proposed SCR effectively addresses this degradation. Both focal loss (row 4) and label smoothing (row 5) are effective and, when combined, complementarily boost all the ZS, I2T Ret, and T2I Ret scores. Notably, I2T Ret increases by 11.3 (rows 3 to 6) with only a relatively mild drop in Comp. We also note that comparing rows 7 and 8 with rows 1

id	λ_l	Comp	ZS	I2T Ret	T2I Ret
1	-	52.9	55.8	57.5	55.5
2	0.1	53.0	55.7	57.4	55.4
3	0.2	53.5	55.3	58.2	55.5
4	0.5	53.5	55.7	57.3	55.4

(a) Sensitivity to the weighting factor λ_l of the local HN loss.

id	γ	Comp	ZS	I2T Ret	T2I Ret
1	-	53.9	53.8	51.7	54.9
2	1.0	53.4	54.9	54.7	55.1
3	2.0	53.5	55.3	58.2	55.5
4	5.0	52.3	55.6	60.2	55.5

(b) Sensitivity to the modulation factor γ of focal loss.

id	β	Comp	ZS	I2T Ret	T2I Ret
1	-	54.2	54.2	53.1	54.8
2	0.02	53.5	55.3	58.2	55.5
3	0.05	53.1	55.2	59.0	55.1
4	0.10	52.3	55.2	58.7	55.3

(c) Sensitivity to the label smoothing factor β.

Table 3: Sensitivity analysis of each component in our FSC-CLIP framework. **(a):** With the global HN loss applied, applying the local HN loss benefits the compositionality while preserving the multi-modal task scores. **(b)** and **(c)**: Both focal loss and label smoothing, the two components of our Selective Calibrated Regularization (SCR), mutually enhance multi-modal task performance but may compromise compositionality when applied too strongly. We highlight the cells corresponding to our design choices in the final FSC-CLIP model.

CLIP[1]	LoRA	Comp	ZS	I2T Ret	T2I Ret
ViT-B/16		46.2	60.3	62.9	49.0
+ NegCLIP		54.1	55.9	53.8	58.1
+ **FSC-CLIP**		54.1	57.0	59.7	59.3
+ **FSC-CLIP**	✓	54.6	57.4	59.9	58.8

[1]Pre-trained: 400M OpenAI, Fine-tuned: LAION-COCO 100K subset.

Table 4: Fine-tuning results of CLIP with a ViT-B/16 encoder, pre-trained on 400M samples of OpenAI data.

CLIP[2]	LoRA	Comp	ZS	I2T Ret	T2I Ret
ViT-B/32		44.3	63.0	63.8	51.2
+ NegCLIP		53.5	59.2	52.1	52.3
+ **FSC-CLIP**		52.9	61.1	56.8	53.8
+ **FSC-CLIP**	✓	54.0	60.7	56.8	53.1

[2]Pre-trained: DataComp-XL, Fine-tuned: LAION-COCO 100K subset.

Table 5: Fine-tuning results of CLIP with a ViT-B/32 encoder, pre-trained on 12.8B DataComp-XL.

and 2, SCR significantly boosts multi-modal task scores. Furthermore, as shown in row 6, applying both global and local HN losses is essential for achieving better Comp and I2T Ret scores.

Sensitivity Analysis. We explore the impact of individually varying each component's parameters in the final model, as detailed in Tab. 3. From Tab. 3a, we find that increasing the local HN loss parameter λ_l improves Comp score while preserving multi-modal task scores. Tab. 3b shows that enhancing the modulation parameter γ boosts multi-modal tasks; however, beyond a certain point, it starts to diminish compositionality by weakening the learning signal from HN texts. Similarly, Tab. 3c indicates that label smoothing benefits multi-modal tasks, particularly I2T Ret. Yet, assigning too much positive margin with β to negative samples can impede the learning of compositionality.

Fine-tuning CLIP with ViT-B/16. We also fine-tuned CLIP with a ViT-B/16 encoder from OpenAI for comparison, as detailed in Tab. 4. This model uses more image patches in training, showing better multi-modal capabilities. However, no gains are observed in Comp compared to the ViT-B/32 model from Tab. 1. After fine-tuning, NegCLIP decreases ZS and I2T Ret scores. In contrast, FSC-CLIP maintains its Comp score and significantly enhances multi-modal task performances. With fine-tuning using LoRA, it achieves a higher Comp score, along with improved ZS and I2T Ret scores.

Scaling Pre-training Data for Fine-tuning. We explore the effect of large-scale pre-training data when fine-tuned. From Tab. 5, we fine-tuned a CLIP model with a ViT-B/32 encoder, pre-trained on 12.8B DataComp-XL dataset (Gadre et al., 2023), far exceeding the 400M samples from OpenAI (Radford et al., 2021). Despite the larger scale pre-training yielding a promising ZS score of 63.0, it underperforms in compositionality compared to OpenAI's pre-trained ViT-B/32 model. For fine-tuning, NegCLIP results in a notable drop in multi-modal task performance. In contrast, FSC-CLIP with LoRA not only counters this degradation but also achieves a higher Comp score than NegCLIP.

5 Conclusion

In this paper, we introduce Fine-grained and Selective Calibrated CLIP (FSC-CLIP), a new fine-tuning framework for visio-linguistic compositionality. It aims to preserve multi-modal capabilities and address the limitations of existing methods relying on global representations. We achieve this by employing dense representations between images and texts and refining the calibration of hard negative losses, thereby facilitating the introduction of Local Hard Negative Loss and Selective Calibrated Regularization. Our extensive validation shows improved compositional reasoning and promising performance in standard multi-modal tasks.

Limitations

Hard Negative Texts. In our approach, we specifically focused on enhancing existing hard negative losses rather than creating new hard negative texts. We utilized rule-based hard negative texts readily available within the existing data, simplifying the process and eliminating the need for external sources. However, this rule-based method may limit the inherent diversity and complexity of negative examples. Additionally, employing hard negative images alongside texts could provide extra learning signals, such as the concept of equivariance (Goel et al., 2022; Wang et al., 2023). However, generating such counterfactual image-text pairs is not as straightforward as rule-based hard negative text generation. Integrating richer, more diverse negative samples through external means remains an intriguing avenue.

Short captions. Our methodology, like prior approaches, relies on short captions for both training and evaluation benchmarks. This practice constrains the models' exposure to and understanding of longer contexts, which are essential for a genuine visio-linguistic compositional understanding. Longer and detailed captions involve more complex associations and contextual nuances that are essential for advanced compositionality in visual and language models. Moving forward, there is a compelling need within the community to develop training and evaluation protocols that incorporate longer captions to better address compositionality.

References

Romain Beaumont. 2021. img2dataset: Easily turn large sets of image urls to an image dataset. https://github.com/rom1504/img2dataset.

Ioana Bica, Anastasija Ilić, Matthias Bauer, Goker Erdogan, Matko Bošnjak, Christos Kaplanis, Alexey A Gritsenko, Matthias Minderer, Charles Blundell, Razvan Pascanu, et al. 2024. Improving fine-grained understanding in image-text pre-training. *arXiv preprint arXiv:2401.09865*.

Santiago Castro, Amir Ziai, Avneesh Saluja, Zhuoning Yuan, and Rada Mihalcea. 2024. Clove: Encoding compositional language in contrastive vision-language models. *arXiv preprint arXiv:2402.15021*.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

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

Sivan Doveh, Assaf Arbelle, Sivan Harary, Roei Herzig, Donghyun Kim, Paola Cascante-Bonilla, Amit Alfassy, Rameswar Panda, Raja Giryes, Rogerio Feris, et al. 2023. Dense and aligned captions (dac) promote compositional reasoning in vl models. *Advances in Neural Information Processing Systems*, 36.

Sivan Doveh, Assaf Arbelle, Sivan Harary, Rameswar Panda, Roei Herzig, Eli Schwartz, Donghyun Kim, Raja Giryes, Rogério Schmidt Feris, Shimon Ullman, et al. 2022. Teaching structured vision & language concepts to vision & language models. 2023 ieee. In *CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2657–2668.

Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. 2023. Clap learning audio concepts from natural language supervision. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Christiane Fellbaum. 2010. Wordnet. In *Theory and applications of ontology: computer applications*, pages 231–243. Springer.

Yanwei Fu, Tao Xiang, Yu-Gang Jiang, Xiangyang Xue, Leonid Sigal, and Shaogang Gong. 2018. Recent advances in zero-shot recognition: Toward data-efficient understanding of visual content. *IEEE Signal Processing Magazine*, 35(1):112–125.

Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. 2023. Datacomp: In search of the next generation of multimodal datasets. *Advances in Neural Information Processing Systems*, 36.

Shashank Goel, Hritik Bansal, Sumit Bhatia, Ryan Rossi, Vishwa Vinay, and Aditya Grover. 2022. Cyclip: Cyclic contrastive language-image pretraining. *Advances in Neural Information Processing Systems*, 35:6704–6719.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR.

Lisa Anne Hendricks and Aida Nematzadeh. 2021. Probing image-language transformers for verb understanding. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3635–3644, Online. Association for Computational Linguistics.

Roei Herzig, Alon Mendelson, Leonid Karlinsky, Assaf Arbelle, Rogerio Feris, Trevor Darrell, and Amir Globerson. 2023. Incorporating structured representations into pretrained vision & language models using scene graphs. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14077–14098, Singapore. Association for Computational Linguistics.

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

Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. 2023. Sugarcrepe: Fixing hackable benchmarks for vision-language compositionality. *Advances in Neural Information Processing Systems*, 36.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. 2023. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *Advances in Neural Information Processing Systems*, 36:78723–78747.

Shih-Cheng Huang, Liyue Shen, Matthew P Lungren, and Serena Yeung. 2021. Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3942–3951.

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. 2021. Openclip. If you use this software, please cite it as below.

Amita Kamath, Jack Hessel, and Kai-Wei Chang. 2023a. Text encoders bottleneck compositionality in contrastive vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4933–4944, Singapore. Association for Computational Linguistics.

Amita Kamath, Jack Hessel, and Kai-Wei Chang. 2023b. What's "up" with vision-language models? investigating their struggle with spatial reasoning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9161–9175, Singapore. Association for Computational Linguistics.

Parminder Kaur, Husanbir Singh Pannu, and Avleen Kaur Malhi. 2021. Comparative analysis on cross-modal information retrieval: A review. *Computer Science Review*, 39:100336.

Benno Krojer, Vaibhav Adlakha, Vibhav Vineet, Yash Goyal, Edoardo Ponti, and Siva Reddy. 2022. Image retrieval from contextual descriptions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3426–3440, Dublin, Ireland. Association for Computational Linguistics.

Gihyun Kwon and Jong Chul Ye. 2022. Clipstyler: Image style transfer with a single text condition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18062–18071.

Samuel Lavoie, Polina Kirichenko, Mark Ibrahim, Mahmoud Assran, Andrew Gordon Wildon, Aaron Courville, and Nicolas Ballas. 2024. Modeling caption diversity in contrastive vision-language pretraining. *arXiv preprint arXiv:2405.00740*.

Tiep Le, Vasudev Lal, and Phillip Howard. 2023. Coco-counterfactuals: Automatically constructed counterfactual examples for image-text pairs. *Advances in Neural Information Processing Systems*, 36.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2023. Bloom: A 176b-parameter open-access multilingual language model. *arxiv preprint arXiv:2211.05100*.

Hector Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. In *Thirteenth international conference on the principles of knowledge representation and reasoning*.

Chunyuan Li, Haotian Liu, Liunian Li, Pengchuan Zhang, Jyoti Aneja, Jianwei Yang, Ping Jin, Houdong Hu, Zicheng Liu, Yong Jae Lee, et al. 2022a. Elevater: A benchmark and toolkit for evaluating language-augmented visual models. *Advances in Neural Information Processing Systems*, 35:9287–9301.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022b. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR.

Feng Liang, Bichen Wu, Xiaoliang Dai, Kunpeng Li, Yinan Zhao, Hang Zhang, Peizhao Zhang, Peter Vajda, and Diana Marculescu. 2023. Open-vocabulary semantic segmentation with mask-adapted clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7061–7070.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988.

Fangyu Liu, Guy Emerson, and Nigel Collier. 2023a. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 11:635–651.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Zixian Ma, Jerry Hong, Mustafa Omer Gul, Mona Gandhi, Irena Gao, and Ranjay Krishna. 2023. Crepe: Can vision-language foundation models reason compositionally? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10910–10921.

Ron Mokady, Amir Hertz, and Amit H Bermano. 2021. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*.

Letitia Parcalabescu, Michele Cafagna, Lilitta Muradjan, Anette Frank, Iacer Calixto, and Albert Gatt. 2022. 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. Association for Computational Linguistics.

Wujian Peng, Sicheng Xie, Zuyao You, Shiyi Lan, and Zuxuan Wu. 2024. Synthesize diagnose and optimize: Towards fine-grained vision-language understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13279–13288.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2023. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*.

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

Arijit Ray, Filip Radenovic, Abhimanyu Dubey, Bryan Plummer, Ranjay Krishna, and Kate Saenko. 2023. cola: A benchmark for compositional text-to-image retrieval. *Advances in Neural Information Processing Systems*, 36.

Ugur Sahin, Hang Li, Qadeer Khan, Daniel Cremers, and Volker Tresp. 2024. Enhancing multimodal compositional reasoning of visual language models with generative negative mining. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5563–5573.

Christoph Schuhmann, Andreas , Köpf, Richard Vencu, Theo Coombes, and Romain Beaumont. 2022a. Laion coco: 600m synthetic captions from laion2b-en. https://laion.ai/blog/laion-coco/.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. 2022b. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. 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, Melbourne, Australia. Association for Computational Linguistics.

Harman Singh, Pengchuan Zhang, Qifan Wang, Mengjiao Wang, Wenhan Xiong, Jingfei Du, and Yu Chen. 2023. Coarse-to-fine contrastive learning in image-text-graph space for improved vision-language compositionality. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 869–893, Singapore. Association for Computational Linguistics.

Jaisidh Singh, Ishaan Shrivastava, Mayank Vatsa, Richa Singh, and Aparna Bharati. 2024. Learn" no" to say" yes" better: Improving vision-language models via negations. *arXiv preprint arXiv:2403.20312*.

Zeyi Sun, Ye Fang, Tong Wu, Pan Zhang, Yuhang Zang, Shu Kong, Yuanjun Xiong, Dahua Lin, and Jiaqi Wang. 2024. Alpha-clip: A clip model focusing on wherever you want. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13019–13029.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. 2022. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5238–5248.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Jin Wang, Shichao Dong, Yapeng Zhu, Kelu Yao, Weidong Zhao, Chao Li, and Ping Luo. 2024. Diagnosing the compositional knowledge of vision language models from a game-theoretic view. *arXiv preprint arXiv:2405.17201*.

Tan Wang, Kevin Lin, Linjie Li, Chung-Ching Lin, Zhengyuan Yang, Hanwang Zhang, Zicheng Liu, and

Lijuan Wang. 2023. Equivariant similarity for vision-language foundation models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11998–12008.

Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, et al. 2022. Robust fine-tuning of zero-shot models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7959–7971.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. 2022. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*.

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. 2023. When and why vision-language models behave like bags-of-words, and what to do about it? In *The Eleventh International Conference on Learning Representations*.

Yan Zeng, Xinsong Zhang, and Hang Li. 2022. Multi-grained vision language pre-training: Aligning texts with visual concepts. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 25994–26009. PMLR.

Le Zhang, Rabiul Awal, and Aishwarya Agrawal. 2024. Contrasting intra-modal and ranking cross-modal hard negatives to enhance visio-linguistic compositional understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13774–13784.

Tiancheng Zhao, Tianqi Zhang, Mingwei Zhu, Haozhan Shen, Kyusong Lee, Xiaopeng Lu, and Jianwei Yin. 2022. An explainable toolbox for evaluating pre-trained vision-language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 30–37, Abu Dhabi, UAE. Association for Computational Linguistics.

Chenhao Zheng, Jieyu Zhang, Aniruddha Kembhavi, and Ranjay Krishna. 2024. Iterated learning improves compositionality in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13785–13795.

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

A Additional Details

A.1 Rule-based Hard Negative Texts

We provide details on the generation process of hard negative texts adopted in our model. We employ three types of rule-based methods for generating hard negative texts: `negclip` (Yuksekgonul et al., 2023), `replace` (Hsieh et al., 2023), and `bi-gram shuffle`. Each method is implemented in an online version and applied to the original text at every training step, resulting in total of four texts including the original caption for every batch as illustrated in Fig. 2. In the online augmentation process, some captions do not yield a hard negative counterpart; these are masked out and excluded from the hard negative loss calculation.

The `negclip` method rearranges words within captions by swapping similar phrase types – such as nouns, verbs, or adjectives – within the text.

The `replace` method generates hard negative texts by replacing specific elements in the caption – entities, relations, or attributes – using antonyms or co-hyponyms from WordNet (Fellbaum, 2010).

The `bi-gram shuffle` rearranges text by shuffling bi-grams (*e.g.*, pairs of adjacent words), within a sentence. It varies the sentence structure, ensuring the generated texts serve as challenging negatives to the original.

All the augmentation methods above utilize the SpaCy (Honnibal and Montani, 2017) package. We implemented `bi-gram shuffle`, while for `negclip` and `replace`, we adopted the implementations from CLoVe (Castro et al., 2024). For illustrative purposes, we provide examples of each method applied to image-caption pairs, in Fig. 4.

A.2 Details on Evaluation Benchmark

Compositionality. VLMs are presented with either an image or text query and must identify the correct match from a set of candidates, which includes subtly altered incorrect options of texts and images.

Depending on the given query modality types, compositionality benchmarks are classified into three categories, as presented in Tab. 6 with corresponding licenses. (1) `Image-to-Text`, where the objective is to choose the correct textual description for a presented image: ARO (Yuksekgonul et al., 2023), CREPE (Ma et al., 2023), SugarCrepe (Hsieh et al., 2023), VALSE (Parcalabescu et al., 2022), VL-Checklist (Zhao et al., 2022), and WhatsUp (Kamath et al., 2023b).

(2) `Text-to-Image` requires the selection of the

Image-Text Pair	negclip	replace	bi-gram shuffle
 Three statues of elephants on the steps in front of an old building.	Three statues of steps on the elephants in front of an old building. Three statues of elephants on the steps in building of an old front. Three elephants of statues on the steps in front of an old building.	Three statues of pikas on the steps in front of an old building. Three statues of elephants into the steps in front of an old building. Three statues of megatherian mammal on the steps in front of an old building.	on the an old steps in building. Three statues front of of elephants Three statues building. of elephants an old steps in front of on the steps in on the front of an old building. Three statues of elephants
 Four different types of sandals with laces.	Four different sandals of types with laces. Four different laces of sandals with types. Four different types of laces with sandals.	Four different types of slingbacks with laces. Four inactive types of sandals with laces. Four different types of sandals with arms.	Four different laces. types of sandals with sandals with types of Four different laces. laces. types of Four different sandals with
 The small blue van is parked in front of a fence.	The blue small van is parked in front of a fence. The small blue van is parked in fence of a front.	The small blue regiment is parked in front of a fence. The small large van is parked in front of a fence. The small average van is parked in front of a fence.	is parked of a The small in front fence. blue van blue van in front of a is parked fence. The small The small in front blue van fence. is parked of a

Figure 4: Example results of rule-based hard negative texts used for training our model. Image-text pairs were randomly sampled from LAION-COCO (Schuhmann et al., 2022a). For `negclip` (Yuksekgonul et al., 2023) and `replace` (Hsieh et al., 2023), differences from the original captions are highlighted in red.

correct image that matches a given text query: ImageCoDE (Krojer et al., 2022) and SVO Probes (Hendricks and Nematzadeh, 2021).
(3) Involving two counterfactual image-text pairs, where the challenge is to pair each image with its corresponding text and the vice versa: Winoground (Thrush et al., 2022), EqBen (Wang et al., 2023), and SPEC (Peng et al., 2024).

For the `Image-to-Text` and `Text-to-Image` tasks, top-1 accuracy is used. For the last group tasks, group accuracy measures whether VLMs correctly match all the associated image-text pairs.

To elaborate on details in specific benchmarks, for EqBen, we cap the evaluation sample size at 20,000. This is because the subtasks `eqbenag` and `eqbenyoucook2` contain 195,872 and 45,849 samples respectively, and evaluating all samples would be excessively time-consuming. Limiting the number of samples does not significantly alter the evaluation results. We do not use the official repository's 10% evaluation split because it does not support sub-task-specific evaluations.

For SVO-Probes, we downloaded images and corresponding captions using the img2dataset (Beaumont, 2021) package from the provided URL list[1], as they are not available as physical files. Out of the original 36.8k samples, 22,162 were successfully downloaded, with 3,728 for the `subj_neg`, 13,523 for the `verb_neg`, and 4,911 for the `obj_neg` subtasks, respectively.

Zero-shot Classification. We leverage ELEVATER toolkit (Li et al., 2022a) for 21 zero-shot classification tasks, licensed under MIT License.

Image-Text Retrieval. We utilize COCO (Chen et al., 2015), Flickr30k (Young et al., 2014), and COCO-CounterFactuals (Le et al., 2023) for the retrieval task, which are licensed under BSD-3-Clause, CC0: Public Domain, CC-BY-4.0, respectively. For COCO-CounterFactuals, we randomly selected 30% of the total 17,410 samples for evaluation, resulting in 5,223 samples. This number is comparable to the scale of the COCO retrieval evaluation dataset.

A.3 Train Dataset

We used a subset of LAION-COCO (Schuhmann et al., 2022a) which is licensed under CC-BY-4.0, and the CC-3M (Sharma et al., 2018)[2] datasets.

A.4 Baseline Methods

From the comparisons to previous methods in Tab. 1, we evaluated previous methods using the same protocol as ours to ensure a fair and consistent evaluation. As such, we obtained the corre-

[1] https://huggingface.co/datasets/MichiganNLP/svo_probes

[2] https://github.com/google-research-datasets/conceptual-captions/blob/master/LICENSE

Benchmark	License	Image source	Tasks and Subtasks
ARO (Yuksekgonul et al., 2023)	MIT	COCO, Visual Genome, Flickr30k	VG_Relation, VG_Attribution, Flickr30k_Order, COCO_Order
CREPE-Productivity (Ma et al., 2023)	*unspecified*	Visual Genome	Atomic Foils, Negate, Swap
SugarCrepe (Hsieh et al., 2023)	MIT	COCO	Add_{object, attribute}, Replace_{object, attribute, relation}, Swap_{object, attribute}
VALSE (Parcalabescu et al., 2022)	MIT	Visual7w, COCO, SWiG, VisDial_v1.0, FOIL-it	Actions_{swap, replacement}, Coreference_{hard, standard}, Counting_{adversarial, hard, small}, Existence, Foil-it, Plurals, Relations
VL-Checklist (Zhao et al., 2022)	*unspecified*	Visual Genome, SWiG, COCO, HAKE, HICO_Det, Pic, HCVRD, OpenImages	Object_Location_{center, margin, mid}, Object_Size_{large, medium, small}, Attribute_{action, color, material, size, state}, Relation_{action, spatial}
WhatsUp (Kamath et al., 2023b)	MIT	Controlled_Images (*self-captured*), COCO, GQA	Controlled_Images_{A, B}, COCO_QA_{One, Two}, VG_QA_{One, Two}
ImageCoDe (Krojer et al., 2022)	MIT	OpenImages, MSRVTT, Video-Storytelling, YouCook	Static (*e.g.*, images), Video (*e.g.*, videos)
SVO Probes (Hendricks and Nematzadeh, 2021)	Apache-2.0	Google Image Search API	Subject, Verb, Object
Winoground (Thrush et al., 2022)	META IMAGES RESEARCH LICENSE	Getty Images	-
EqBen (Wang et al., 2023)	Apache-2.0	Action Genome (AG), GEBC, YouCook2, Kubric, StableDiffusion (SD)	EQ-AG, EQ-GEBC, EQ-YouCook2, EQ-Kubric_{location, counting, attribute}, EQ-SD
SPEC (Peng et al., 2024)	*unspecified*	Stable-Diffusion-XL 1.0 (Podell et al., 2023)	Absolute_size, Absolute_position, Count, Relative_size, Relative_position, Existence

Table 6: A complete list of compositionality benchmarks in our work. The table is further sub-divided depending on the given query types for a single test.

sponding checkpoints from each official repository and loaded with open_clip package (Ilharco et al., 2021). When loading the previous models' checkpoints, also including the other models, we explicitly set `quick_gelu` to True in open_clip, for NegCLIP (Yuksekgonul et al., 2023), CE-CLIP (Zhang et al., 2024), and GNM-CLIP (Sahin et al., 2024). This adjustment aligns with the original OpenAI models, which were pre-trained and also fine-tuned with this option activated, though it was omitted in their implementations.

We list the previous methods with corresponding licenses. NegCLIP (Yuksekgonul et al., 2023): MIT License, CE-CLIP (Zhang et al., 2024): MIT License, GNM-CLIP (Sahin et al., 2024): Apache-2.0 License, TSVLC[3] and DAC[4] (Doveh et al., 2022, 2023): *unspecified*, CLoVe (Castro et al., 2024): MIT License.

B Additional Results

For thoroughness, we include additional results not featured in the main paper. Note that all models were fine-tuned using the CLIP ViT-B/32 encoder from OpenAI (Radford et al., 2021).

B.1 Multiple Runs

In Tab. 7, we report the mean and standard deviation for our models across all tasks listed in Tab. 1, using three distinct seeds: 0, 1, and 2 for training each model.

B.2 Zero-shot Classification

We report the results for each benchmark within the 21 zero-shot classification tasks in Tab. 8.

B.3 Image-Text Retrieval

We present the results for each benchmark included in the three image-text retrieval tasks in Tab. 9.

[3] `https://github.com/SivanDoveh/TSVLC`
[4] `https://github.com/SivanDoveh/DAC`

Method	LoRA	ARO	CREPE	EqBen	ImageCoDe	SugarCrepe	SVO Probes	VALSE	VL-Checklist	WhatsUp	Winoground	SPEC	Comp	ZS	I2T Ret	T2I Ret
								Fine-tuned: LAION-COCO, 100K Samples								
FSC-CLIP		$82.7_{0.10}$	$46.6_{0.35}$	$29.3_{0.17}$	$24.6_{0.94}$	$82.6_{0.14}$	$90.1_{0.03}$	$73.5_{0.15}$	$75.7_{0.33}$	$42.1_{0.25}$	$6.2_{0.63}$	$33.5_{0.17}$	$53.4_{0.09}$	$55.6_{0.32}$	$57.8_{0.52}$	$55.3_{0.20}$
FSC-CLIP	✓	$85.3_{0.14}$	$52.9_{1.28}$	$28.9_{0.17}$	$24.9_{0.11}$	$80.5_{0.11}$	$89.7_{0.05}$	$72.4_{0.17}$	$78.7_{0.20}$	$42.9_{0.05}$	$5.4_{0.38}$	$32.4_{0.11}$	$54.0_{0.17}$	$56.1_{0.18}$	$57.3_{0.13}$	$54.4_{0.08}$

Table 7: Evaluation across three training runs of our model using different seeds. We report the mean and standard deviation obtained from the evaluation results of the models across three trials.

Method	caltech101	cifar10	cifar100	country211	dtd	eurosat-clip	fer2013	fgvc-aircraft-2013b	flower102	food101	gtsrb	hateful-memes	imagenet-1k	kitti-distance	mnist	oxford-iiit-pets	patchcamelyon	rendered-sst2	resisc45-clip	stanfordcar	voc2007classification	Average
CLIP-ViT-B/32	88.3	89.8	65.1	17.2	44.4	45.5	42.3	19.7	66.7	84.0	32.6	55.9	63.3	27.4	48.3	87.1	60.6	58.6	60.0	59.7	82.6	57.1
										Fine-tuned: MS-COCO, 100K Samples												
NegCLIP	88.2	88.9	63.2	15.0	43.1	47.3	47.6	16.8	62.3	79.4	30.2	54.3	60.9	27.6	49.7	85.4	59.7	58.8	56.9	54.0	84.4	55.9
CE-CLIP	82.2	85.9	60.2	9.6	35.2	44.9	39.7	10.0	47.2	70.1	28.0	53.5	49.9	34.6	40.6	66.0	58.8	61.1	51.5	35.3	83.1	49.9
GNM-CLIP	86.8	88.4	65.7	15.2	42.0	50.1	46.6	17.3	62.4	81.8	30.2	54.9	61.4	25.2	54.4	86.3	59.0	58.5	58.7	53.1	84.0	56.3
										Fine-tuned: Conceptual Captions – 3M (CC-3M), 100K Samples												
TSVLC (RB)	83.7	92.3	66.0	16.2	39.5	52.1	43.6	14.7	58.2	81.2	24.2	57.8	58.5	30.4	46.9	85.5	50.0	59.8	58.6	49.2	84.7	54.9
TSVLC (RB+LLM)	84.6	92.0	66.8	16.2	40.3	56.5	46.8	13.8	58.5	81.6	27.1	56.9	59.7	27.8	43.9	84.7	50.5	60.1	59.5	50.5	84.7	55.4
DAC-LLM	82.6	90.4	64.1	14.3	38.4	52.5	50.7	10.5	49.7	74.1	24.2	56.3	51.0	16.3	42.1	74.4	50.0	54.5	52.2	39.4	85.1	51.1
DAC-SAM	81.3	89.9	64.1	14.8	40.4	49.8	48.0	8.9	48.9	72.3	24.9	55.7	52.3	18.7	45.2	76.7	58.9	60.0	54.7	39.8	84.1	51.9
										Fine-tuned: LAION-COCO, 600M Samples												
CLoVe	85.5	85.8	66.2	12.6	37.7	49.1	38.0	9.0	44.6	71.9	22.6	54.6	53.1	34.9	36.4	74.2	56.7	51.3	55.2	48.7	81.9	51.0
										Fine-tuned: LAION-COCO, 100K Samples												
FSC-CLIP (Ours)	86.5	87.5	65.7	15.3	42.4	43.9	48.9	14.9	55.5	80.5	31.6	55.9	58.1	29.1	52.4	84.2	61.0	56.0	56.9	52.0	83.6	55.3
FSC-CLIP (Ours, LoRA)	85.9	88.5	66.3	15.8	39.8	52.8	48.2	14.2	57.0	81.0	27.9	56.3	57.4	33.9	54.3	82.7	59.8	57.2	58.7	52.6	83.7	55.9

Table 8: Expanded results for the 21 zero-shot classification tasks from ELEVATER (Li et al., 2022a).

	COCO Retrieval						Flickr30k Retrieval						COCO-CounterFactuals Retrieval						Avg.	
	Image to text (I2T)			Text to image (T2I)			Image to text (I2T)			Text to image (T2I)			Image to text (I2T)			Text to image (T2I)			I2T	T2I
Method	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@1
CLIP-ViT-B-32	50.1	74.9	83.5	30.4	56.0	66.8	78.8	94.9	98.3	58.7	83.5	90.0	51.0	79.3	86.7	48.1	77.4	85.9	60.0	45.8
									Fine-tuned: MS-COCO, 100K Samples											
NegCLIP	59.3	82.8	89.4	45.2	72.1	81.7	85.7	96.4	98.8	71.6	91.8	95.7	55.3	82.5	89.2	58.3	84.9	91.3	66.8	58.4
CE-CLIP	56.0	81.6	89.0	47.1	74.1	83.1	75.3	93.2	96.9	68.9	89.6	94.2	46.3	75.7	84.5	56.2	83.6	90.5	59.2	57.4
GNM-CLIP	58.1	81.4	88.8	41.1	67.5	77.8	82.9	96.2	98.6	68.8	89.9	94.1	57.2	84.5	90.5	56.7	84.5	91.1	66.1	55.5
									Fine-tuned: Conceptual Captions – 3M (CC-3M), 100K Samples											
TSVLC (RB)	46.1	71.7	80.4	36.3	62.0	72.4	74.0	93.2	96.4	64.9	87.2	92.7	44.6	72.0	80.2	55.0	83.3	90.0	54.9	52.1
TSVLC (RB+LLM)	46.4	71.8	80.8	36.6	62.2	72.7	74.8	92.6	96.8	65.1	87.6	92.7	44.1	71.5	80.1	55.1	83.3	90.4	55.1	52.3
DAC-LLM	29.9	54.5	65.6	37.3	63.5	73.8	52.9	79.8	87.9	64.6	88.0	93.0	28.1	53.6	64.4	55.2	83.0	90.0	36.9	52.4
DAC-SAM	33.1	57.9	68.8	34.0	59.7	70.0	59.8	82.7	89.0	61.7	85.7	91.2	30.4	55.2	64.8	51.5	79.9	87.3	41.1	49.0
									Fine-tuned: LAION-COCO, 600M Samples											
CLoVe	48.3	73.9	82.8	42.7	68.7	78.2	69.5	90.4	95.6	68.7	90.0	94.5	41.5	69.1	78.3	56.5	84.2	90.8	53.1	56.0
									Fine-tuned: LAION-COCO, 100K Samples											
FSC-CLIP (Ours)	49.7	73.6	82.4	40.4	66.4	76.4	75.6	93.3	97.4	68.2	90.0	94.3	49.2	77.5	85.8	57.9	85.4	91.4	58.2	55.5
FSC-CLIP (Ours, LoRA)	48.2	73.6	81.8	39.0	64.9	75.0	75.1	93.2	96.4	66.9	88.6	93.6	48.5	76.0	84.4	57.1	84.7	91.0	57.3	54.3

Table 9: Expanded results for the three zero-shot image-text retrieval tasks, including COCO (Chen et al., 2015), Flickr30k (Young et al., 2014), and COCO-CounterFactuals (Le et al., 2023).