FFF: Fixing Flawed Foundations in contrastive pre-training results in very strong Vision-Language models

Adrian Bulat[1,2]   Yassine Ouali[1]   Georgios Tzimiropoulos[1,3]

[1]Samsung AI Center Cambridge, UK   [2] Technical University of Iași, Romania
[3] Queen Mary University of London, UK

Abstract

Despite noise and caption quality having been acknowledged as important factors impacting vision-language contrastive pre-training, in this paper, we show that the full potential of improving the training process by addressing such issues is yet to be realized. Specifically, we firstly study and analyze two issues affecting training: incorrect assignment of negative pairs, and low caption quality and diversity. Then, we devise effective solutions for addressing both problems, which essentially require training with multiple true positive pairs. Finally, we propose training with sigmoid loss to address such a requirement. We show very large gains over the current state-of-the-art for both image recognition ($\sim$ +6% on average over 11 datasets) and image retrieval ($\sim$ +19% on Flickr30k and $\sim$ +15% on MSCOCO).

Figure 1. Our approach, **FFF**, achieves state-of-the-art accuracy across multiple datasets, largely outperforming prior methods.

1. Introduction

Large-scale contrastive image-text pre-training has emerged as the prevalent method for vision-language representation learning [14, 21, 27, 29, 38, 51, 52]. The majority of datasets employed for pre-training are web-collected [4, 10, 34, 40–43, 45]. They offer a varied data distribution and are sufficiently large to effectively train high-performing vision-language models. However, since the raw captions for each image are typically extracted from associated tags or descriptions, they often exhibit low quality, being noisy and suboptimal for training purposes [21, 27]. Although some attempts to fix such issues have been already described, to some extent, in literature (e.g. ALIP [50], BLIP [27]), in this work, we show that the full potential of improving the quality of the training process is far from being fully realized. Specifically, by studying and addressing specific issues related to noise and low data quality, in this work, we show that our improved vision-language training pipeline can achieve massive gains over the current state-of-the-art methods for both image recognition ($\sim$ +6% on
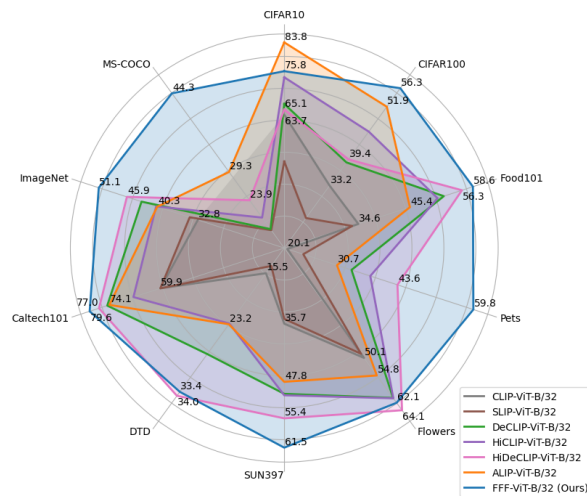
average over 11 datasets) and image retrieval ($\sim$ +19% on Flickr30k [53] and $\sim$ +15% on MSCOCO [30]).

The first issue we study is related to noise impacting contrastive learning: near-duplicate samples which are incorrectly treated as negative pairs. Even within a batch, it is not uncommon to find images and/or captions that are semantically similar or even identical. Since standard contrastive learning assumes one positive pair, this significantly hinders the training process and the quality of the trained models.

The second issue we study is related to low caption quality and diversity. Captions can be short and lacking detail, noisy, or even entirely irrelevant to the image. Moreover, since the mapping process between image and text is one-to-many, more than one caption is needed to provide an approximate description of the image.

To fix issue one, we propose an algorithm that mines new positive pairs based on image-text, image-image, and text-text similarities, aiming to decrease the number of false negatives in the training data arising due to semantically similar images and/or captions.

We fix issue two by firstly generating pseudo-captions for each training image using a state-of-the-art image captioning technique [28] that will act as new true positives for a given image. Then, we propose *batch text augmentation* for training with multiple pseudo-captions (*i.e.* five captions per image selected via beam search) within the same batch to effectively increase caption diversity.

Importantly, after applying the proposed solutions, we end up with a variable number of positive pairs per image *i.e.* newly mined positive pairs and multiple pseudo-captions per image. This implies that we need to train our model with a loss function that accommodates multiple positives and is robust to potential errors in the mining process. Unfortunately, neither contrastive loss [38] nor supervised contrastive loss [22] can be directly applied for this case. To this end, we propose to use the sigmoid loss [54] which allows the number of positives to vary dynamically per sample and per batch at no extra cost and is also robust to noise.

Overall, we make the following **contributions**:

- We study and provide in-depth analyses of two important issues related to vision-language training process/data: false negative pairs due to semantic near-duplicates, and low caption quality and diversity (Sec. 2).
- We provide two simple algorithms for addressing the aforementioned issues: The first one uses text-image, image-image, and text-text similarities for eliminating incorrectly assigned negatives and mining new true positives. The second uses the proposed *batch text augmentation* for training with multiple pseudo-captions per image within the same batch. Both solutions induce multiple new positives per each training image. To address this, we propose to use sigmoid loss for training the model. See Sec. 4.
- We show very large gains over the current state-of-the-art for both image recognition ($\sim +6\%$ on average over 11 datasets) and image retrieval ($\sim +19\%$ on Flickr30k and $\sim +15\%$ on MSCOCO) (Sec. 5). We further ablate the impact of many important components of our method in Sec. 6.

2. Flaws of web-collected datasets & potential solutions

Several observations drawn by analyzing the flaws of a web-collected dataset (CC3M dataset), motivating the proposed approach, are provided below:

Original captions are noisy and repetitive: For example, as illustrated in Fig. 2 for the CC3M dataset, original (raw) captions contain a high number of generic captions that frequently reoccur across the dataset (Fig. 2 (c)), and are often semantically similar (Fig. 2 (a)). Moreover, many raw captions may be unrelated to their associated images and their visual content, as indicated by low CLIP scores (Fig. 2 (b)).

Re-captioning enhances quality and diversity: A potential solution to this issue is the use of state-of-the-art image captioning models (*e.g.* BLIP2 [28], OFA [46]) to generate synthetic pseudo-captions, which can enhance the quality and descriptiveness of the captions. When comparing raw and pseudo-captions, it is evident that the latter are more diverse and semantically relevant to their associated images, as shown in Fig. 2.

Multiple pseudo-captions should reduce noise: State-of-the-art image captioning models, despite being capable of generating fluent and diverse captions, are often trained and bootstrapped from the same web-collected data used in training vision-language models. Consequently, as shown in Fig. 4, in some instances, the generated pseudo-captions can be ambiguous and contain hallucinations, errors, and stylistic biases similar to those found in the raw captions. As a result, relying on a single pseudo-caption per image can still introduce a high degree of noise and can hinder the training of an effective vision-language model.

A potential solution to this issue is the use of multiple pseudo-captions or multiple positives per image in the hope that even if individual captions are incorrect, their ensemble is of higher quality and better reflects the content of the associated image. To probe for the possible positive effect of using multiple synthetic captions, in Fig. 3 (a), we show the intra-cosine similarities of 5 pseudo-captions generated using beam search and, respectively, in Fig. 3 (b), the average image-text CLIP score between these synthetic captions and their associated images - contrasted with the score corresponding to a single caption. We observe that: 1) a simple method such as beam search can generate diverse synthetic captions, and, more crucially, 2) using multiple positives per image results in an improved ensemble that better describes the image and helps alleviate the problem of false positives due to incorrect individual instances.

Mining of new positives: As shown in Fig. 3 (c), even for a relatively small batch of $1k$ image-caption pairs, it is common to find captions more similar to the image than the ground-truth caption (*i.e.* higher ranks), and, as displayed in Fig. 5, such high-ranking captions often contain true positives, which are captions that can be considered ground-truth descriptions for the associated image.

A potential solution to this is the use of online mining of new positives based on image and text feature cosine similarities. However, as shown in Fig. 5, text-image pairs with high cosine similarity can still be false positives. To reduce them, we propose to mine the positives based on image-text, image-image, and text-text similarities, aiming to decrease the number of false negatives in the training data arising due to semantically similar images and/or captions.
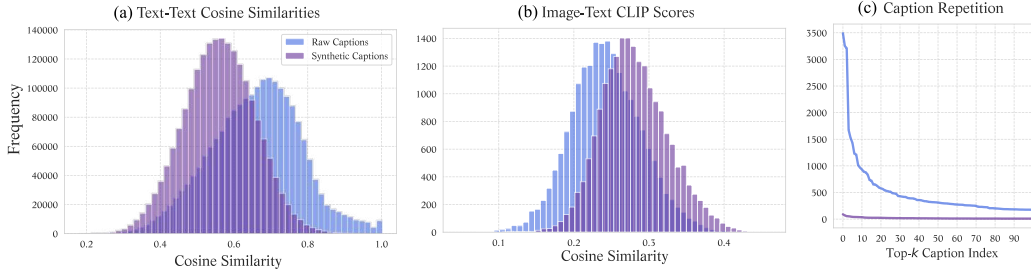
Figure 2. **Semantic and lexical diversity of raw and synthetic pseudo-captions of CC3M:** (a): Average cosine similarities of each caption and its 100 most similar captions using CLIP ViT-L/14 features. (b): Cosine similarities between features of each image and its ground-truth caption. (c): The frequencies of the top-100 most frequent raw and synthetic pseudo-captions (generated using BLIP2). We observe that the raw captions are semantically similar to each other (a), often not well aligned with their associated ground-truth images (b), and contain a high number of basic and redundant captions (c). By swapping them with pseudo-captions, we observe an improved diversity (a,c) and better image-text alignment (b).
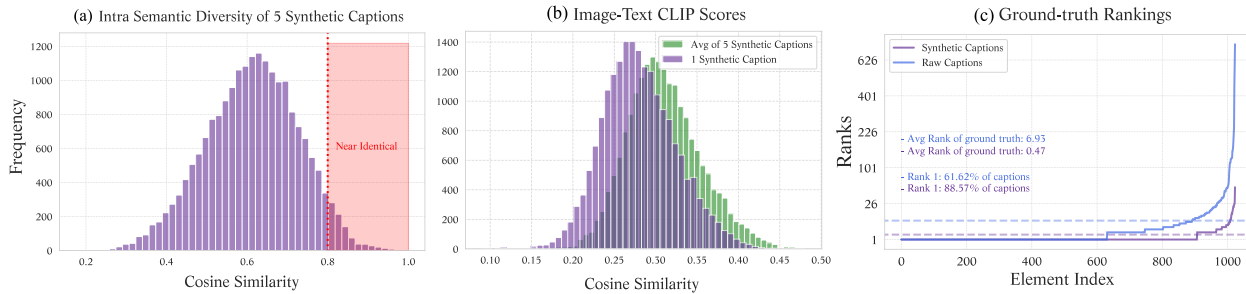

Figure 3. **Quality assessment of synthetic captions of CC3M:** (a) Average intra-cosine similarities between 5 synthetic captions of each image. (b) Cosine similarities between the features of each image and either the features of a single synthetic pseudo-caption or the averaged features of 5 pseudo-captions. In (a) and (b), we observe that using multiple synthetic positives that are diverse (a), possible erroneous captions can be corrected using an ensemble of pseudo-captions that better converge to the ground truth, resulting in text features more aligned with their associated images (b). (c): The rankings of the ground-truth captions for each image in a batch of $1k$ image-caption pairs. This shows that, even with relatively small batches, many negatives are well aligned with some images, and it is very likely that many of these negatives are potentially correct matches for a subset of images, *i.e.* false negatives. Features are computed using CLIP ViT-L/14.

Figure 4. **Qualitative samples of synthetic captions from CC3M:** We show 4 examples featuring original raw and synthetic (BLIP2) pseudo-captions. These examples highlight typical limitations and challenges observed in synthetic captions which, while superior to raw captions, can still be considered noisy.

3. Related work

Contrastive pretraining under noisy web-collected data: Current publicly available vision-language datasets are mined from the internet automatically [4, 40, 41, 45] with only basic automatic filtering applied, which results in im-

perfect annotations and duplicate or near-duplicate pairs. A series of papers [1, 11, 15, 48] attempt to alleviate the noise present in annotations by switching from hard to soft labels, akin to knowledge distillation (KD), using various combinations of contrastive loss (i.e. InfoNCE) and KL divergence. The work in [48] constructs the soft labels using an online entropic optimal transport algorithm implemented via the Sinkhorn-Knopp algorithm. The probabilities for each image add up to 1, with 0.5 on the diagonal and the rest distributed. This assumes that, within the batch, there are always some images that are somewhat similar. In our case, we use hard labels, with multiple positives, performing reassignments only when the samples are sufficiently close, instead of forcing a distribution in all cases. Furthermore, we do not require running an optimal transport method, nor rely on a contrastive loss. The work of [1] progressively self-distills soft image-text alignments to more efficiently learned robust representations from noisy data. At every iteration, a subset of labels are "soft" while the rest are kept hard. Similarly, the work of [15] relaxes the

Ground-truths:
Raw: racecar driver steers his car during video game subject .
Synthetic: the red bull racing car is driving down a race track

Captions with higher ranks than the ground-truths:
Raw: seat achieves the record of cars
Synthetic: an older blue van in dozens of other cars
Raw: racecar driver of person steers his car during event
Synthetic: a formula racer in motion during evening session

Ground-truths:
Raw: looking west that 's the old building on the far right .
Synthetic: an old photo of the street in the city

Captions with higher ranks than the ground-truths:
Raw: looking east along a city from 1930s
Synthetic: an old photo of a city street
Raw: view this image of automobile make
Synthetic: a black scion sti parked in the sun with cloudy sky

Ground-truths:
Raw: waiting on a cowboy - live music
Synthetic: two people are singing and playing an acoustic guitar in an empty warehouse

Captions with higher ranks than the ground-truths:
Raw: country artist and pop artist perform on stage during event
Synthetic: two people in musical costumes perform a song in a stage
Raw: young boy and girl singing a song in a choir
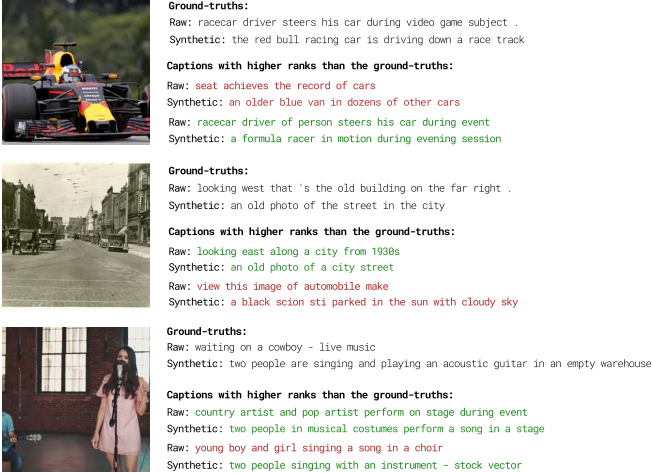Synthetic: two people singing with an instrument - stock vector

Figure 5. **Examples of high-ranking captions from CC3M:** We show 3 examples of raw and synthetic captions ranked higher than the ground-truths from a batch of $1k$ image-caption pairs. In green, we show potential false negatives that can be used as new positives for improved training. However, possible false positives, as shown in red, can still occur. These can be handled by the robust sigmoid loss. Rankings are obtained using CLIP ViT-L/14.

strict one-to-one constraint, transitioning to a soft cross-modal alignment by introducing a softened target, which is generated from the fine-grained intra-modal self-similarity. Additionally, they disentangle the negatives in the distribution to further boost the relation alignment, resulting in a combination of InfoNCE loss performed with hard labels and KL divergence. However, they do not perform batched text augmentations with multiple positives, as in our work, and still use a contrastive loss combined with KL, operating on soft scores. The works of [6, 19, 20] study the effect of removing false negatives in the context of unimodal *i.e.* pure vision models, not considering the case of multi-modal learning. The work of [20] flags (a very small number) of potential negatives using the aggregated score obtained from multiple support views per image, [6] uses a clustering based approach while [19] is based on ranked positives, requiring a known class hierarchy (i.e. a fully supervised case) or known changes/relations (i.e. videos). The works of [6, 20] derive from the Supervised Contrastive Loss, while [19] from InfoNCE. In contrast, our work operates on image-text data, takes into account multi-modal interactions (I2T, T2T, T2I), does not use additional support views, known hierarchies etc. and is easily scalable.

Following a different direction, BLIP [27] and their followup [28] version, use a bootstrapping approach in which the noisy captions are filtered out using the initial model, which is then retrained on the new data. This interplay is performed offline and requires training a multitask model. The work of [39] presents a small-scale study showing that random sampling of pseudo-captions improves CLIP, concluding however that scaling up the number of image-

caption pairs appears to be more effective. Finally, very recently, ALIP [50] adds a synthetic pseudo-caption and a consistency gating mechanism that weights the influence of the samples and image-text pairs on the contrastive loss.

Different from the aforementioned methods, we propose to fix incorrectly assigned negatives and mine for new true positives using text-image, image-image, and text-text similarities. Moreover, to increase caption quality and diversity, we further propose training with multiple pseudo-captions per image within the same batch. As our methods require training with multiple positives per image, we further propose to use the sigmoid loss [54] for training the model.

4. Method

This section describes the proposed method, whose aim is to improve vision-language training by denoising and improving the quality of the training process/data. Specifically, Sec. 4.1 addresses the problem of false negative pairs inherent to the noisy nature of large-scale image-text datasets by re-assigning them as true positives[1]. Sec. 4.2 proposes *text batch augmentation* for training the model with *multiple* positives pairs. The effect of Secs. 4.1 and 4.2 is that, for each training image, a *variable* number of positive-negative pairs is formed (Sec. 4.3). Sec. 4.4 proposes a natural way to train the model in this case by using the recently proposed sigmoid loss for vision-language pre-training.
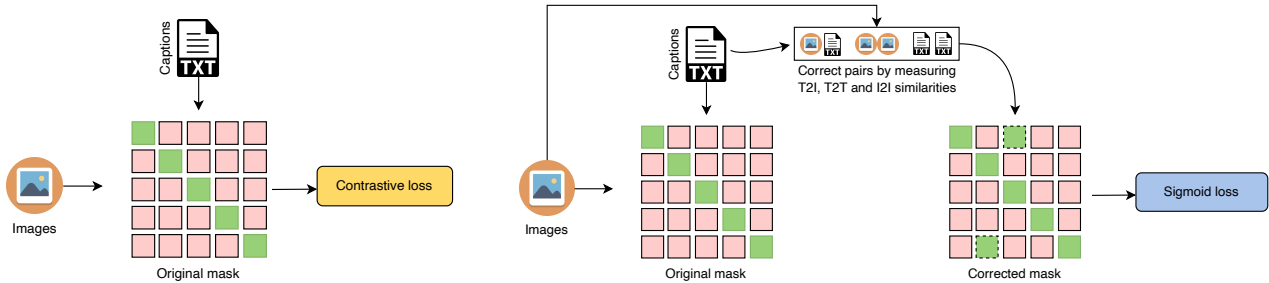
4.1. Fixing incorrect negatives

Let D be a dataset consisting of image-text pairs, with B a batch of randomly selected samples (x_i, t_i), $i = 1, 2, \ldots, N$. In addition to the ground truth positives pairs (x_i, t_i), we seek to identify and correct wrongly co-occurring negative pairs (x_i, t_j) *on-the-fly*. To achieve this, let us first define the image-text, image-image, and text-text cosine similarity matrices $S_{it} = X_f \cdot T_f^T, S_{ii} = X_f \cdot X_f^T$ and $S_{tt} = T_f \cdot T_f^T$, where $S_{it}, S_{ii}, S_{tt} \in \mathbb{R}^{N \times N}$ and $X_f \in \mathbb{R}^{N \times d}$ and $T_f \in \mathbb{R}^{N \times d}$ represent the image and text features, respectively.

Given the similarity score matrices, we define the assignment matrix $M \in \{0, 1\}^{N \times N}$ as follows:

$$M = (S_{it} > p_1) \vee (S_{ii} > p_2) \vee [(S_{tt} > p_3) \wedge (S_{it} > p_1')], \quad (1)$$

where $\vee$ is the logical `or` and $\wedge$ the logical `and` operator, and p_1, p_1', p_2 and p_3 are the thresholds above which a sample is marked as positive, with $p_1' < p_1$. Note that we filter the positives found with text-text matching using image-text similarities (using threshold p_1'), as we observed a high portion of false positives within text-text matching, due to the fact that repeated samples often correlate with poor overall

---

[1] It is possible that such cases can occur in clean datasets too, as multiple captions can describe an image and vice versa, multiple images can be described by one caption.

(a) **Baseline [38]:** Ground truth construction and training loss. Prior work does not take into account that some pairs may be incorrect negatives and is limited to one positive per sample.

(b) **Fixing incorrect negatives:** Our approach analyzes on the fly the image-text, image-image, and text-text similarities, correcting wrong negative pairs. The model is trained using the sigmoid loss (see Sec. 4.4) instead of the standard contrastive loss which is unsuitable for an arbitrary number of positive samples.
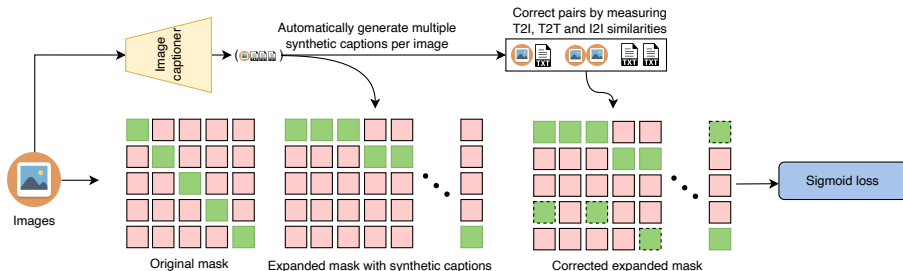
(c) **Overall approach** combining fixing incorrect negatives (Sec. 4.1) with batch text argumentation (Sec. 4.2). The synthetic pseudo-captions are generated offline and packed as part of the dataset.

Figure 6. **Overview of our approach:** Fixing incorrect negatives is shown in (b). In (c) we describe our combined approach (including batch text augmentation) contrasted with the baseline of (a). Green squares denote positive pairs, while pink are negatives. Green squares with a dashed border denote identified false negatives that are corrected to true positives.

image description fidelity. The choice of p_1, p_1', p_2 and p_3 is empirical and generally depends on the characteristics of the model. We ablate the dependency of the method on the threshold values in Sec. 6 where we show little sensitivity. Note that M re-assigns a variable number of positives to each image. Fig. 6b depicts the construction process of M at a high level.

In order to calculate the cosine similarity matrices S_{it}, S_{ii} and S_{tt} required for the construction of M, we use a pre-trained model. This is akin to a form of auto-labeling/auto-filtering, where the pretrained model provides a signal for re-assessing the labeling of the samples. Although one could opt to use an EMA teacher-student approach, we found this simple approach to work sufficiently well. Moreover, some possible errors in M can be handled by the robust sigmoid loss used for training (see Eq. 2).

4.2. Batch text augmentation with multiple positives

The currently available image-text datasets [4, 41, 42] are noisy, with high variability in the quality of the text descriptions among samples. To improve data quality, we use BLIP2 [28], an off-the-shelf image captioner, to generate multiple pseudo-captions for each image in the training set (see supplementary material for visual examples). Inspired by [18], we propose to include *all pseudo-captions* as true positives *within the same batch*, which we call batch text

augmentation. Note that simultaneously training with multiple pseudo-captions within the same batch has not been considered in previous work. We show that this approach enables the training of highly accurate models (see Sec. 5 and our ablation in Sec. 6). In the next section, we also show how batch text augmentation can be integrated with the mask construction process defined in Sec. 4.1. Finally, we note that while batch text augmentation improves the overall performance, it does not address the presence of semantic near duplicates (i.e. false negatives) within the same training batch.

4.3. Combined approach

Our approach for fixing incorrect negative pairs (Sec. 4.1) and batch text augmentation (Sec. 4.2) can be naturally combined in order to define the total number of true positives per image.

To this end, and without loss of generality, we assume k captions per image (original caption plus pseudo-captions), hence the total number of captions and images are related by $N_{txt} = kN_{img}$. Given this, the image-text similarity matrix has now (by construction) size $S_{it} \in \mathbb{R}^{N_{img} \times N_{txt}}$. Hence, the computation of S_{ii} and S_{tt} needs to be adjusted to reflect this change. For the image-image case, and as $N_{img} < N_{txt}$, to make the image-image similarity matrix S_{ii} have the same dimensions as S_{it}, i.e. $S_{ii} \in \mathbb{R}^{N_{img} \times N_{txt}}$, we

replicate the scores k times. In other words, a given image x_i will share the score with each group of captions belonging to image x_j, $\forall i, j \in N_{img}$. For the text-text case, the similarity matrix is now of size $S_{tt} \in \mathbb{R}^{N_{txt} \times N_{txt}}$. Analogously, to make the S_{tt} have the same dimensions as S_{it}, we take the average score between each caption of image x_i and all k captions of image x_j.

Overall, we end up with similarity matrices of the same dimensions $S_{it}, S_{ii}, S_{tt} \in \mathbb{R}^{N_{img} \times N_{text}}$ and hence the assignment matrix M can be again constructed by applying Eq. 1. The overall process is depicted in Fig. 6c.

4.4. Loss function

The symmetrical contrastive loss (i.e. *text → image* and *image → text*) used in CLIP [38] supports only one positive pair per sample (see Fig. 6a), being in discordance with the requirement of training with a variable number of positive pairs per image set by the proposed methods in Secs. 4.1 and 4.2. A solution to this problem could be given by the Supervised Contrastive Loss [22], originally introduced to enable multi-view training of supervised image recognition. However, this loss is prone to noise [2], with the harder positive pairs dominating the signal and hindering, in part, the effect from the rest of the positive samples. This is especially problematic in the context of web-collected datasets, which are notoriously noisy. Finally, it is memory intensive and computationally demanding. In practice, we observe a $1.9\times$ slowdown for a batch size of 8,096 samples.

A natural alternative is the BCE loss, shown to outperform cross-entropy for image classification [47], and also shown to be a viable alternative for image-text representation learning [54]. Such formulation is particularly advantageous for the proposed approach, as the BCE loss natively supports an arbitrary number of positives per sample per batch, with the ground truth being provided simply as a binary mask. Moreover, the loss is more robust to noise in general, and hence to false negatives and positives [54]. Finally, the initial negative bias prevents the model from being forced to learn incorrect assignments early one. Hence, we propose to use the following loss:

$$\ell_{mp} = -\frac{1}{N_{txt}} \sum_{i=1}^{N_{img}} \sum_{j=1}^{N_{txt}} \log \frac{1}{1 + \exp(m_{ij}(-s_{ij}/\tau + \beta))},$$
(2)

where m_{ij} is the i, j element of M (-1 for negative and 1 for positive pairs), and respectively, s_{ij} the i, j element of the similarity matrix S_{it}.

As the negative pairs considerably outnumber the positive ones, to ensure that we start from a low initial loss (making the same observation as in [54]), we add a learnable scalar β, set initially to a negative value. However, as the number of positive pairs is dynamic and is typically tied to both the specifics of the dataset and the threshold used to define a positive sample, different from [54], we propose to estimate β at the beginning of the training process. Specifically, given the randomly initialized model, we sample b batches out of the training set, and then compute and store the cosine similarities. Then, given the scores and the corresponding labels, we search for β such that the initial loss is minimized (everything else is kept frozen). The value of β can be found either by gradient descent or alternatively, by performing a grid search.

5. Results

Pretraining Datasets: To allow for fair comparisons with prior work, we pre-train our approach on YFCC15M-v2 [26], a subset of YFCC100M [45] containing approximately 15M image-text pairs. To cover different dataset sizes, we also conduct experiments on CC3M [42] and CC12M [4], and in the supplementary material, on Open30M and Open70M datasets, further showcasing our method's scalability with respect to the dataset size.

Implementation details: Architecturally, we use the same model topology and setting as in CLIP [38], specifically, using AdamW [31], learning rate of $1e-3$ and weight decay of 0.1, except for CC3M where we set the weight decay to 0.5, as in prior work [29]. In terms of augmentations, we follow [29], randomly resizing and cropping the image to 224×224px, applying random flipping, random Gaussian blur (between 0.1 and 2.0) and color jittering (0.4, 0.4, 0.4, 0.1). For text, the data is truncated to 77 tokens. Note, that the branch used to construct the assignment matrix M uses no augmentations (i.e. resize to 256×256px, followed by center crop, resulting in a 224×224px image). The thresholds were set to $p_1 = 0.27$, $p_2 = 0.92$, $p_3 = 0.99$, $p_1' = 0.24$. Unless otherwise specified, the models are trained for 32 epochs with a batch size of 8,096 on 8 NVIDIA A100 GPUs. All of our models and training code are implemented using PyTorch [37].

5.1. Comparison with state-of-the-art

Following recent work on vision-language pretraining [16, 33, 50], we compare our method with state-of-the-art approaches for zero-shot classification and zero-shot retrieval. See supplementary material for linear probe evaluation.

Zero-shot classification: For zero-shot classification evaluation, for the main setting, we select the common subset of datasets that facilitate a direct comparison with prior state-of-the-art. In particular, we evaluate our approach on the following datasets: CIFAR-10 [24], CIFAR-100 [24], Food101 [3], Pets [36], Flowers [35], SUN397 [49], Stanford Cars [23], DTD [7], Caltech101 [12], FGVC-Aircraft [32] and ImageNet [9]. The evaluation is performed using the same prompt templates and class names as in prior work [16, 33, 50].

Method	Pre-train dataset	CIFAR10	CIFAR100	Food101	Pets	Flowers	SUN397	Cars	DTD	Caltech101	Aircraft	ImageNet	Average
CLIP-ViT-B/32[38]	YFCC15M	63.7	33.2	34.6	20.1	50.1	35.7	2.6	15.5	59.9	1.2	32.8	31.8
SLIP-ViT-B/32 [33]	YFCC15M	50.7	25.5	33.3	23.5	49.0	34.7	2.8	14.4	59.9	1.7	34.3	30.0
FILIP-ViT-B/32 [51]	YFCC15M	65.5	33.5	43.1	24.1	52.7	50.7	3.3	24.3	68.8	3.2	39.5	37.2
DeCLIP-ViT-B/32 [29]	YFCC15M	66.7	38.7	52.5	33.8	60.8	50.3	3.8	27.7	74.7	2.1	43.2	41.3
DeFILIP-ViT-B/32 [8]	YFCC15M	70.1	46.8	54.5	40.3	63.7	52.4	4.6	30.2	75.0	3.3	45.0	44.2
HiCLIP-ViT-B/32 [16]	YFCC15M	74.1	46.0	51.2	37.8	60.9	50.6	4.5	23.1	67.4	3.6	40.5	41.8
HiDeCLIP-ViT-B/32 [16]	YFCC15M	65.1	39.4	56.3	43.6	**64.1**	55.4	5.4	**34.0**	77.0	**4.6**	45.9	44.6
ALIP-ViT-B/32 [50]	YFCC15M	**83.8**	51.9	45.4	30.7	54.8	47.8	3.4	23.2	74.1	2.7	40.3	41.7
FFF-ViT-B/32 (Ours)	YFCC15M	75.8	**56.3**	**58.6**	**59.8**	62.1	**61.5**	**16.3**	33.4	**79.6**	4.6	**51.1**	**50.8**

Table 1. Zero-shot classification performance on 11 downstream datasets. All models were pre-trained on YFCC15M.

Method	Text retrieval						Image retrieval					
	Flickr30k			MSCOCO			Flickr30k			MSCOCO		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
CLIP-ViT-B/32[38]	34.9	63.9	75.9	20.8	43.9	55.7	23.4	47.2	58.9	13.0	31.7	42.7
SLIP-ViT-B/32 [33]	47.8	76.5	85.9	27.7	52.6	63.9	32.3	58.7	68.8	18.2	39.2	51.0
DeCLIP-ViT-B/32 [29]	51.4	80.2	88.9	28.3	53.2	64.5	34.3	60.3	70.7	18.4	39.6	51.4
UniCLIP-ViT-B/32 [25]	52.3	81.6	89.0	32.0	57.7	69.2	34.8	62.0	72.0	20.2	43.2	54.4
HiCLIP-ViT-B/32 [16]	-	-	-	34.2	60.3	70.9	-	-	-	20.6	43.8	55.3
HiDeCLIP-ViT-B/32 [16]	-	-	-	38.7	64.4	74.8	-	-	-	23.9	48.2	60.1
ALIP-ViT-B/32 [50]	70.5	91.9	95.7	46.8	72.4	81.8	48.9	75.1	82.9	29.3	54.4	65.4
FFF-ViT-B/32 (Ours)	**85.3**	**97.5**	**99.4**	**61.7**	**84.5**	**90.4**	**67.6**	**89.1**	**93.3**	**44.3**	**70.9**	**80.1**

Table 2. Zero-shot image-text retrieval on the test splits of Flickr30k and MSCOCO. All models were pre-trained on YFCC15M.

As the results from Tab. 1 show, our approach outperforms all prior methods, improving by 6.2% in absolute terms on top of the previous best result of HiDeCLIP [16] (which benefits from a better architecture) when aggregated across 11 datasets. Notably, we set a new state-of-the-art result on ImageNet, too (51.1%). Finally, we significantly improve upon ALIP [50], which also makes use of synthetic captions, outperforming it by 9.1%.

Method	CC3M	CC12M
CLIP [38]	20.6	36.5
ProtoCLIP [5]	21.5	-
CyCLIP [17]	22.1	-
CLOOB [14]	24.0	-
SoftCLIP [15]	24.2	43.2
DeCLIP [29]	27.2	41.0
CLIP-Rocket [13]	27.4	44.4
BoW [44]	30.3	-
FFF (Ours)	**33.4**	**47.4**

Table 3. Zero-shot evaluation on Imagenet in terms of Top-1 (%) accuracy for a ResNet-50 model pre-trained on CC3M/CC12M.

For completeness, we also adhere to the protocol of pre-training a ResNet-50 on CC3M, and respectively, CC12M

and then evaluating it for zero-shot classification on ImageNet. As the results from Tab. 3 show, the same conclusions hold. Our method outperforms the previous best result by 3.1% on CC3M (30.3% vs 33.4%) and 3.0% on CC12M (44.4% vs 47.4%). See supplementary material for results on Open30M and Open70M.

Zero-shot retrieval: Consistent with prior work, we evaluate our approach for zero-shot retrieval on Flickr-30k [53] and MS-COCO [30] reporting results in terms of R@{1,5,10} for both text and image retrieval. The results are summarized in Tab. 2. As it can be observed, our approach offers significant gains across all metrics and datasets used, improving on top of the prior state-of-the-art ALIP [50] by 14.8% and 18.7% in terms of R@1 on Flickr30k for text, and respectively, image retrieval. Similarly, we outperform the previous best result by 14.9% and 15.0% in terms of R@1 on MSCOCO for text and image retrieval. This highlights that our approach results in representations that can capture subtle and fine-grained details.

6. Ablation studies

For our ablation studies, the results reported are produced using a ViT-B/16 model pretrained on CC3M dataset.

Effect of fixing incorrect negatives: herein, we analyze the effectiveness of the proposed algorithm of Sec. 4.1. By analyzing the result from Tab. 4, we can observe consistent gains for all 3 cases of interest: a) when using the web-collected captions (+2.7% gain), b) when using one pseudo-caption (+3.5% improvement) and c) when all available pseudo-captions at once (+1.8%). Overall, compared to the baseline accuracy of 18.6%, our approach improves by +14.3% (top-1 accuracy of 32.9%). The results show that our approach provides gains across all options considered.

Fix incorrect negatives	Num. captions	Top-1 (%)
✗	0	18.6
✓	0	21.3
✗	1	23.3
✓	1	26.8
✗	5	31.1
✓	5	**32.9**

Table 4. **Effect of fixing incorrect negatives:** Zero-shot evaluation on ImageNet in terms of Top-1 (%) accuracy.

Effect of different components in Eq. 1: In Eq. 1, the constructed assignment matrix M is computed from three feature similarity matrices S_{it}, S_{ii} and S_{tt}. Herein, we evaluate the impact of each of these components. As the results from Tab. 5 show, viewed independently, the S_{it} is the most impactful, as it has a dual effect, both in terms of filtering incorrect pairs and of adjusting for semantically similar samples. Moreover, the results hold for both ground truth captions and pseudo-captions.

Assign. Matrix M	Num. captions	Top-1 (%)
None	0	18.6
$S_{tt} > p_3$	0	18.8
$S_{ii} > p_2$	0	21.3
$S_{it} > p_1$	0	21.4
Eq. 1 (all)	0	22.0
None	1	23.1
$S_{tt} > p_3$	1	23.6
$S_{ii} > p_2$	1	24.6
$S_{it} > p_1$	1	26.0
Eq. 1 (all)	1	26.8

Table 5. **Effect of different components in Eq. 1:** Zero-shot evaluation on ImageNet in terms of Top-1 (%) accuracy.

Effect of batch text augmentation: Herein, we assess the impact of training with multiple pseudo-captions within the same batch, as described in Sec. 4.2. Tab. 6 shows accuracy vs number of pseudo-captions used during training. As we can observe, increasing the number of captions increases the accuracy of the model, inline with the expectations.

As an additional baseline, we compare against a model trained by randomly sampling 1 out of 5 captions (as opposed to using them jointly as proposed in our work) on CC3M and YFCC-15M. On CC3M the performance drops by 1.5%, from 32.9% to 31.4%, while on YFCC-v2 from 51.1% to 44.1%. This further highlights the importance of the proposed batch text augmentation.

Effect of image captioner: We also compare the effect of using two different state-of-the-art image captioners, OFA [46] and BLIP-2 [28]. As the results from Tab. 7 show, both captioners lead to identical performance.

Num. captions	0	1	3	5
Top-1 (%)	18.6	23.3	30.2	31.1

Table 6. **Effect of batch text augmentation:** Zero-shot evaluation on ImageNet in terms of Top-1 (%) accuracy.

Image captioner	Top-1 (%)
OFA [46]	32.9
BLIP-2 [28]	32.9

Table 7. **Effect of the image captioner:** Zero-shot evaluation on ImageNet in terms of Top-1 (%) accuracy.

Comparison with the supervised contrastive loss: To further validate the loss choice, we compare against a model trained with the supervised contrastive loss [22]. For a fair comparison, both models were trained using the same settings on CC3M. When evaluated for zero-shot classification on ImageNet, the supervised contrastive model achieved only 19.0% accuracy vs 21.3% achieved by our model. Note, that similar results are obtained using a InfoNCE based loss. This result empirically solidifies the arguments made in Sec. 4.4.

7. Conclusions

In this work, we propose a new approach to vision-language pretraining based on multi-positive sample pairing that fixes incorrect negatives and addresses low caption quality. The latter is tackled by a newly introduced batch text augmentation strategy, in which multiple new positive pairs are concomitantly added via synthetic recaptioning. Departing from the typical contrastive loss, to enable efficient training under an arbitrary number of positives per sample, we propose to train the model with a sigmoid loss. In the process, we highlight the crucial role of noise and caption quality in vision-language pre-training, offering an in-depth analysis. All in all, we show large improvements over the current state-of-the-art method for both zero-shot image recognition ($\sim +6\%$ on average of 11 datasets) and retrieval ($\sim +19\%$ on Flickr30k and $\sim +15\%$ on MSCOCO).

References

[1] Alex Andonian, Shixing Chen, and Raffay Hamid. Robust cross-modal representation learning with progressive self-distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16430–16441, 2022. 3

[2] Lucas Beyer, Olivier J Hénaff, Alexander Kolesnikov, Xiaohua Zhai, and Aäron van den Oord. Are we done with imagenet? *arXiv preprint arXiv:2006.07159*, 2020. 6

[3] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VI 13*, pages 446–461. Springer, 2014. 6

[4] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12M: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *CVPR*, 2021. 1, 3, 5, 6

[5] Delong Chen, Zhao Wu, Fan Liu, Zaiquan Yang, Yixiang Huang, Yiping Bao, and Erjin Zhou. Prototypical contrastive language image pretraining. *arXiv preprint arXiv:2206.10996*, 2022. 7

[6] Tsai-Shien Chen, Wei-Chih Hung, Hung-Yu Tseng, Shao-Yi Chien, and Ming-Hsuan Yang. Incremental false negative detection for contrastive learning. *arXiv preprint arXiv:2106.03719*, 2021. 4

[7] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3606–3613, 2014. 6

[8] Yufeng Cui, Lichen Zhao, Feng Liang, Yangguang Li, and Jing Shao. Democratizing contrastive language-image pre-training: A clip benchmark of data, model, and supervision. *arXiv preprint arXiv:2203.05796*, 2022. 7

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 6

[10] Karan Desai, Gaurav Kaul, Zubin Aysola, and Justin Johnson. Redcaps: Web-curated image-text data created by the people, for the people. *arXiv preprint arXiv:2111.11431*, 2021. 1

[11] Lijie Fan, Dilip Krishnan, Phillip Isola, Dina Katabi, and Yonglong Tian. Improving clip training with language rewrites. *arXiv preprint arXiv:2305.20088*, 2023. 3

[12] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *2004 conference on computer vision and pattern recognition workshop*, pages 178–178. IEEE, 2004. 6

[13] Enrico Fini, Pietro Astolfi, Adriana Romero-Soriano, Jakob Verbeek, and Michal Drozdzal. Improved baselines for vision-language pre-training. *arXiv preprint arXiv:2305.08675*, 2023. 7

[14] Andreas Fürst, Elisabeth Rumetshofer, Johannes Lehner, Viet T Tran, Fei Tang, Hubert Ramsauer, David Kreil, Michael Kopp, Günter Klambauer, Angela Bitto, et al. Cloob: Modern hopfield networks with infoloob outperform clip. *Advances in neural information processing systems*, 35: 20450–20468, 2022. 1, 7

[15] Yuting Gao, Jinfeng Liu, Zihan Xu, Tong Wu, Wei Liu, Jie Yang, Ke Li, and Xing Sun. Softclip: Softer cross-modal alignment makes clip stronger. *arXiv preprint arXiv:2303.17561*, 2023. 3, 7

[16] Shijie Geng, Jianbo Yuan, Yu Tian, Yuxiao Chen, and Yongfeng Zhang. Hiclip: Contrastive language-image pre-training with hierarchy-aware attention. *arXiv preprint arXiv:2303.02995*, 2023. 6, 7

[17] Shashank Goel, Hritik Bansal, Sumit Bhatia, Ryan Rossi, Vishwa Vinay, and Aditya Grover. Cyclip: Cyclic contrastive language-image pretraining. *Advances in Neural Information Processing Systems*, 35:6704–6719, 2022. 7

[18] Elad Hoffer, Tal Ben-Nun, Itay Hubara, Niv Giladi, Torsten Hoefler, and Daniel Soudry. Augment your batch: Improving generalization through instance repetition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8129–8138, 2020. 5

[19] David T Hoffmann, Nadine Behrmann, Juergen Gall, Thomas Brox, and Mehdi Noroozi. Ranking info noise contrastive estimation: Boosting contrastive learning via ranked positives. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 897–905, 2022. 4

[20] Tri Huynh, Simon Kornblith, Matthew R Walter, Michael Maire, and Maryam Khademi. Boosting contrastive self-supervised learning with false negative cancellation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2785–2795, 2022. 4

[21] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021. 1

[22] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020. 2, 6, 8

[23] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 554–561, 2013. 6

[24] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 6

[25] Janghyeon Lee, Jongsuk Kim, Hyounguk Shon, Bumsoo Kim, Seung Hwan Kim, Honglak Lee, and Junmo Kim. Uniclip: Unified framework for contrastive language-image pretraining. *Advances in Neural Information Processing Systems*, 35:1008–1019, 2022. 7

[26] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021. 6

[27] 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*, pages 12888–12900. PMLR, 2022. 1, 4

[28] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 2, 4, 5, 8

[29] Yangguang Li, Feng Liang, Lichen Zhao, Yufeng Cui, Wanli Ouyang, Jing Shao, Fengwei Yu, and Junjie Yan. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. *arXiv preprint arXiv:2110.05208*, 2021. 1, 6, 7

[30] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 1, 7

[31] Ilya Loshchilov and Frank Hutter. Fixing weight decay regularization in adam. 2018. 6

[32] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. 6

[33] Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. Slip: Self-supervision meets language-image pre-training. In *European Conference on Computer Vision*, pages 529–544. Springer, 2022. 6, 7

[34] Thao Nguyen, Gabriel Ilharco, Mitchell Wortsman, Sewoong Oh, and Ludwig Schmidt. Quality not quantity: On the interaction between dataset design and robustness of clip. *Advances in Neural Information Processing Systems*, 35:21455–21469, 2022. 1

[35] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian conference on computer vision, graphics & image processing*, pages 722–729. IEEE, 2008. 6

[36] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3498–3505. IEEE, 2012. 6

[37] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017. 6

[38] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1, 2, 5, 6, 7

[39] Shibani Santurkar, Yann Dubois, Rohan Taori, Percy Liang, and Tatsunori Hashimoto. Is a caption worth a thousand images? a study on representation learning. In *The Eleventh International Conference on Learning Representations*, 2022. 4

[40] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021. 1, 3

[41] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022. 3, 5

[42] 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. 5, 6

[43] Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2443–2449, 2021. 1

[44] Ajinkya Tejankar, Maziar Sanjabi, Bichen Wu, Saining Xie, Madian Khabsa, Hamed Pirsiavash, and Hamed Firooz. A fistful of words: Learning transferable visual models from bag-of-words supervision. *arXiv preprint arXiv:2112.13884*, 2021. 7

[45] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016. 1, 3, 6

[46] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International Conference on Machine Learning*, pages 23318–23340. PMLR, 2022. 2, 8

[47] Ross Wightman, Hugo Touvron, and Hervé Jégou. Resnet strikes back: An improved training procedure in timm. *arXiv preprint arXiv:2110.00476*, 2021. 6

[48] Bichen Wu, Ruizhe Cheng, Peizhao Zhang, Peter Vajda, and Joseph E Gonzalez. Data efficient language-supervised zeroshot recognition with optimal transport distillation. *arXiv preprint arXiv:2112.09445*, 2021. 3

[49] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 3485–3492. IEEE, 2010. 6

[50] Kaicheng Yang, Jiankang Deng, Xiang An, Jiawei Li, Ziyong Feng, Jia Guo, Jing Yang, and Tongliang Liu. Alip: Adaptive language-image pre-training with synthetic caption. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2922–2931, 2023. 1, 4, 6, 7

[51] Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. Filip: Fine-grained interactive language-image pre-training. *arXiv preprint arXiv:2111.07783*, 2021. 1, 7

[52] Haoxuan You, Luowei Zhou, Bin Xiao, Noel Codella, Yu Cheng, Ruochen Xu, Shih-Fu Chang, and Lu Yuan. Learning visual representation from modality-shared contrastive language-image pre-training. In *European Conference on Computer Vision*, pages 69–87. Springer, 2022. 1

[53] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 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, 2014. 1, 7

[54] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. *arXiv preprint arXiv:2303.15343*, 2023. 2, 4, 6