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

ALIP: Adaptive Language-Image Pre-training with Synthetic Caption

Kaicheng Yang[1], Jiankang Deng[*] [2], Xiang An[1], Jiawei Li[1], Ziyong Feng[1],
Jia Guo[3], Jing Yang[3], Tongliang Liu[4]
[1]DeepGlint [2]Huawei UKRD [3]InsightFace [4]University of Sydney
{kaichengyang,xiangan,jiaweili,ziyongfeng}@deepglint.com, tongliang.liu@sydney.edu.au
{jiankangdeng,guojia,y.jing2016}@gmail.com

Abstract

Contrastive Language-Image Pre-training (CLIP) has significantly boosted the performance of various vision-language tasks by scaling up the dataset with image-text pairs collected from the web. However, the presence of intrinsic noise and unmatched image-text pairs in web data can potentially affect the performance of representation learning. To address this issue, we first utilize the OFA model to generate synthetic captions that focus on the image content. The generated captions contain complementary information that is beneficial for pre-training. Then, we propose an Adaptive Language-Image Pre-training (ALIP), a bi-path model that integrates supervision from both raw text and synthetic caption. As the core components of ALIP, the Language Consistency Gate (LCG) and Description Consistency Gate (DCG) dynamically adjust the weights of samples and image-text/caption pairs during the training process. Meanwhile, the adaptive contrastive loss can effectively reduce the impact of noise data and enhances the efficiency of pre-training data. We validate ALIP with experiments on different scales of models and pre-training datasets. Experiments results show that ALIP achieves state-of-the-art performance on multiple downstream tasks including zero-shot image-text retrieval and linear probe. To facilitate future research, the code and pre-trained models are released at https://github.com/deepglint/ALIP.

1. Introduction

With the development of mobile networks and social platforms, there has been an explosion in the production of image-text pairs on a massive scale [3, 13]. This unprecedented abundance of data has laid a solid foundation for vision-language pre-training [32, 15]. Through image-text alignment on large-scale data, the Contrastive Language–Image Pre-training (CLIP) method [32] has demonstrated huge success in multi-modal learning. Specifically, CLIP learns two separate unimodal encoders for image and text using a contrastive loss, one of the most effective losses for representation learning [36, 14, 5, 7]. Nevertheless, the negative impact of the noise in the web-crawled data has been largely overlooked, shadowed by the performance gain achieved from scaling up the training data [31, 1].

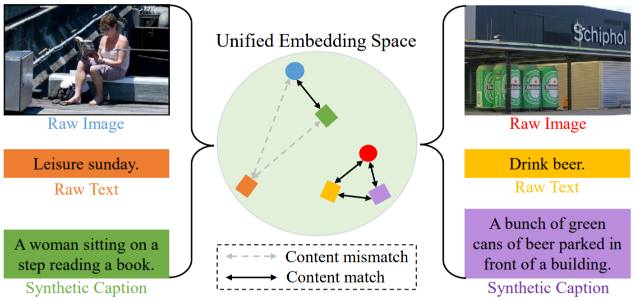


Figure 1. Examples from the YFCC15M dataset to illustrate the mismatched (left) and matched (right) image-text pairs. The synthetic caption is generated from the OFA [37] model. The raw text description "Leisure Sunday" is less aligned with the raw image in the left sample, while the synthetic caption "A woman sitting on a step reading a book" is a more accurate representation.

To facilitate research on large-scale multi-modal models, LAION400M [34] and LAION5B [33] were released, comprising 400 million and 5 billion image-text pairs respectively, which were filtered using the CLIP model. However, the current offline filtering approach results in a substantial loss of training data, as the original dataset contains 5 billion image-text pairs. Furthermore, this approach may introduce biases due to the limited representation power of the pre-trained model used for filtering. Despite efforts to curate data for high-quality image-text pairs (*e.g*., LAION [34, 33] and YFCC100M [35]), noisy and poorly-aligned pairs still exist in existing image-text datasets, which can potentially impact the performance of representation learning.

In Fig. 1, we present two samples from YFCC15M. The raw text of the right sample ("Drink beer") is correct and

[*] corresponding author.

matches the content in the image, while the raw text of the left sample ("Leisure Sunday") is too abstract and does not exactly match the concrete visual signals of the image. To alleviate the influence of the noisy and poorly-aligned image-text pairs, BLIP [19] bootstraps the captions with an online captioner generating synthetic captions and an online filter removing the noisy ones, while momentum-based methods (*e.g*., ALBEF [20] and PSD [2]) employs soft labels computed using embeddings from a moving average momentum model. However, these filtering and momentum-based methods require additional computation costs and memory consumption.

In this paper, we first employ the OFA [37] model to generate synthetic captions that are consistent with the image content. Specifically, the OFA [37] model is guided by the prompt "What does the image describe" to generate synthetic captions. Compared with the raw texts in Fig. 1, the synthetic captions "A woman sitting on a step reading a book" and "A bunch of green cans of beer parked in front of a building" provide additional as well as reliable descriptions, such as object information (book, cans), attribute information (green), action information (sitting, parked), and spatial relationship (in front of), which can be used to enhance the representation learning.

Given the normalized embedding features of the image, raw text, and synthetic caption, we propose an Adaptive Language-Image Pre-training (ALIP) method, a bi-path model that integrates raw text supervision and synthetic caption supervision. As the core components of ALIP, the Language Consistency Gate (LCG) and the Description Consistency Gate (DCG) are designed to dynamically adjust the weights of samples and image-text/caption pairs during the training process. The LCG considers the consistency between raw text and synthetic caption to identify the high-quality sample, while the DCG considers the consistency of image-text or image-caption to adjust the contrastive pair loss. The adaptive contrastive loss influenced by the above weights substantially reduces the impact of noise data and enhances the efficiency of pre-training data. Extensive experiments show that ALIP achieves state-of-the-art performance on multiple downstream tasks including zero-shot image-text retrieval and linear probe. Experiment results on different model sizes and pre-training datasets also prove the strong robustness of the ALIP. The main contributions of this paper are summarized as follows:

- We propose a bi-path model that integrates raw text supervision and synthetic caption supervision. Based on the similarity triplet between image, text, and caption, the proposed method can dynamically adjust the weights of samples and image-text/caption pairs through the language consistency gate and description consistency gate.
- Based on the adaptive weights, we design the adaptive contrastive loss, which can effectively reduce the impact of noise data and improve the pre-training data efficiency.
- We conduct extensive experiments and prove that ALIP achieves state-of-the-art performance on multiple downstream tasks including zero-shot image-text retrieval and linear probe task.

2. Related Work

Image-Language Pre-training. Image-language pre-training aims to improve the performance of downstream vision and language tasks by pre-training the model on large-scale image-text pairs. The milestone work CLIP [32] has attracted unprecedented attention for its impressive zero-shot recognition ability and excellent transfer ability. Recently, a number of improved methods based on CLIP have been proposed. For more effective training, SLIP [25] significantly improves performance by combining self-supervised learning and CLIP pre-training. DeCLIP [21] explores self-supervision and cross-modal multi-view supervision in the million-scale vision-language pre-training. FILIP [39] learns fine-grained representation for patches in the images and words in the sentences. UniCLIP [18] improves data efficiency by integrating contrastive losses defined across multiple domains into a single universal space. HiCLIP [12] equips both the visual and language branches in CLIP with hierarchy-aware attention which significantly improves the cross-modal alignment. In this paper, we propose an Adaptive Image-Language Pre-training (ALIP) method to effectively utilize raw text supervision guided by synthetic captions.

Noise Alleviation for Contrastive Pre-training. Large-scale contrastive pre-training [32, 15, 34, 33] typically requires dataset sizes of hundreds of millions to billions level. Despite the performance gain obtained by scaling up the training data, the noisy web text is sub-optimal for image-language pre-training. Nevertheless, the cleaning strategy applied to these large-scale data is primitive (*e.g*., removing samples with short or non-English captions) or biased (*e.g*., filtering samples based on alignment scores from existing models) [6]. To reduce the adverse effects of noisy image-text pairs in the training data, ALBEF [20] and PSD [2] use soft labels computed using embeddings from a moving average momentum model. However, momentum-based approaches are infeasible for large-scale training due to the increased computation and memory consumption. LiT [40] shows that when a well pre-trained vision encoder is adopted, it is better to lock the vision encoder to protect vision representations from being corrupted by noisy language supervision. However, LiT lacks the ability to align complex text to a fully-trained image encoder, thus underperforming on the multi-modal task, cross-modal retrieval. BLIP [19] uses the bootstrapped image-grounded text encoder to filter out noisy captions, but the captioner and filter models need to be finetuned on the COCO dataset beforehand, and these models also increase the overall number of parameters of the model.

In contrast to the previous work, we introduce synthetic captions to alleviate the influence of noise in large-scale vision-language pre-training. Our method can dynamically adjust the weights of samples and image-text pairs through the language consistency gate and description consistency gate. Meanwhile, the adaptive contrastive loss can effectively reduce the impact of noise data and improve the pre-training data efficiency. Our approach is a cheaper alternative since it does not require us to run the expensive online model throughout the training as synthetic captions can be pre-computed and stored offline. In addition, we take advantage of all training samples with adaptive weights instead of directly filtering out image-text pairs.

3. Methodology

In this section, we first introduce the model architecture of the proposed method (Sec. 3.1). Then, we delineate the Language Consistency Gate (LCG) and Description Consistency Gate (DCG) in Sec. 3.2 and 3.3. Lastly, we explain the training objectives of the proposed adaptive contrastive loss for vision-language representation learning (Sec. 3.4).

3.1. ALIP Architecture

The primary focus of this paper is on the task of contrastive image-text pre-training. Different from the image-text pairs used in CLIP [32], we adopt the off-the-shelf OFA$_{base}$ [37] model to generate a synthetic caption for each image by applying the prompt "What does the image describe?". This method results in a dataset $D = \{(X_i, T_i, C_i)\}_{i=1}^{N}$ comprising of image-text-caption triplets. Next, we train a dual encoder model $\Phi = \{\Phi_{\text{image}}, \Phi_{\text{text/caption}}\}$, where Φ_{image} represents the image encoder, and $\Phi_{\text{text/caption}}$ denotes the shared text/caption encoder. We use the shorthand $\mathbf{x} = \Phi_{\text{image}}(X)/\left\|\Phi_{\text{image}}(X)\right\|$, $\mathbf{t} = \Phi_{\text{text}}(T)/\left\|\Phi_{\text{text}}(T)\right\|$, and $\mathbf{c} = \Phi_{\text{caption}}(C)/\left\|\Phi_{\text{caption}}(C)\right\|$ to denote the l_2 normalized embeddings of image, text, and caption, respectively, for an image-text-caption triplet (X, T, C).

Large vision and language datasets such as YFCC100M [35] and LAION [21] have collected a large number of image-caption pairs from the web, which makes them a good fit for large-scale contrastive pre-training. However, these datasets lack semantic-based curation and can contain unilateral or irrelevant raw texts. Moreover, the automatically generated captions can be also noisy or lack fine granularity[19]. Fig. 3 shows some examples of the web raw text T and the synthetic caption C. Each box in the figure represents a sample that includes an image along with its corresponding text and caption descriptions. To mitigate the negative impact of such noise during training, our approach takes into account the similarities between the image, text, and caption triplet.

As illustrated in Fig. 2, we propose an Adaptive Language-Image Pre-Training (ALIP) approach to make full use of data and reduce the impact of noise. Using the l_2 normalized triplet embeddings $\mathbf{x}$, $\mathbf{t}$, and $\mathbf{c}$ of image, text, and caption, we can calculate three types of similarities: (1) the similarity between raw text and synthetic caption $S_{tc} = \mathbf{t} * \mathbf{c}$, (2) the similarity between image and raw text $S_{xt} = \mathbf{x} * \mathbf{t}$, and (3) the similarity between synthetic caption and image $S_{xc} = \mathbf{x} * \mathbf{c}$. Based on the triplet similarities, we design two gate functions, Language Consistency Gate (LCG) and Description Consistency Gate (DCG). More specifically, the LCG predicts a sample weight based on the similarity between raw text embedding and synthetic caption embedding (S_{tc}). Besides, the DCG computes the image-description weights based on the consistency between the image and text/caption (S_{xt} and S_{xc}). Finally, these weights are fed into the adaptive contrastive loss to reduce the impact of noise.

3.2. Language Consistency Gate

In ensemble learning, confidence in the prediction can be increased when two independent inferences have arrived at the same prediction [22]. Inspired by this, we boost the label confidence of a training sample when the similarity between the raw text and synthetic caption is high. To facilitate the accurate assessment of language labels, we introduce a historical average similarity metric H_{tc}, which is dynamically updated during the training process as follows:

$$H_{tc} = m * H_{tc} + (1 - m) * \bar{S}_{tc}, \tag{1}$$

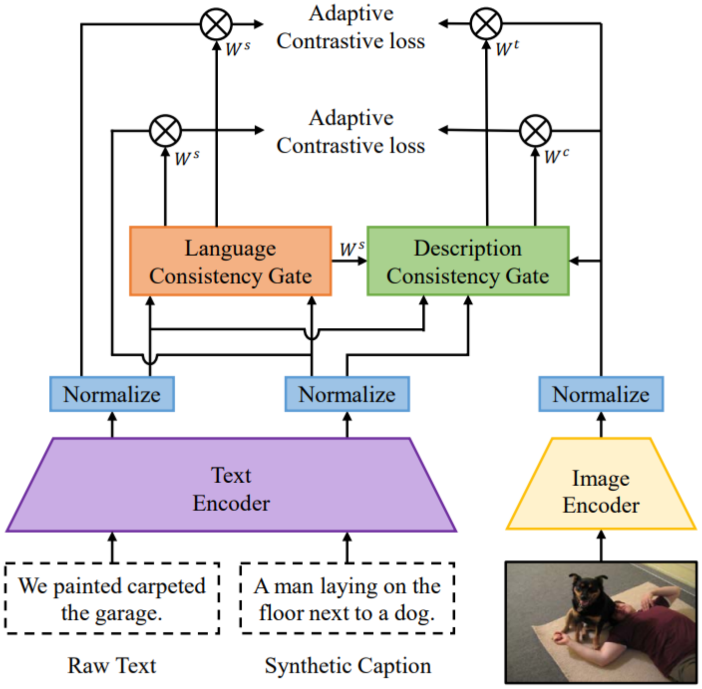
where m is the momentum and $\bar{S}_{tc}$ denotes the average of S_{tc}. As both raw text and synthetic caption are explaining the same image, samples with a similarity score S_{tc} higher than the historical average similarity threshold H_{tc} will be considered as high-quality samples with reliable language labels. By contrast, samples with a lower similarity score are considered as low-quality samples with unreliable language labels. To distinguish these two kinds of training samples, we design a sample weight W^s and the calculation of W^s is given by the following equation:

$$W^s = \begin{cases} \exp((S_{tc} - H_{tc}) * \gamma_s), & S_{tc} \le H_{tc} \\ 1, & S_{tc} > H_{tc} \end{cases}, \tag{2}$$

where γ_s is a hyper-parameter, and W^s is constrained to the range of $(0, 1]$. Consequently, the LCG assigns a lower weight to low-quality samples, reducing the influence of unmatched image-text or image-caption pairs.

3.3. Description Consistency Gate

While the language consistency gate can identify high-quality training samples, it is important to note that some

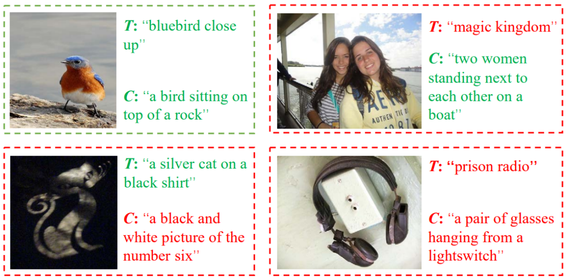


Adaptive Contrastive loss
W^s
W^t
Adaptive Contrastive loss
W^s
W^c
Language Consistency Gate
W^s
Description Consistency Gate
Normalize
Normalize
Normalize
Text Encoder
Image Encoder
We painted carpeted the garage.
A man laying on the floor next to a dog.
Raw Text
Synthetic Caption

```
Algorithm 1: Language Consistency Gate
Input: Text and caption embeddings $t, c$,
       historical average similarity $H_{ct}$
Start:
  $S_{tc} \leftarrow \cos(t, c)$
  $H_{tc} \leftarrow m * H_{tc} + (1 - m) * \bar{S}_{tc}$
  $W^s \leftarrow \exp((S_{tc} - H_{tc}) * \gamma_s)$
  $W^s \leftarrow \max(W^s, 1)$
End
Output: $W^s$
```

```
Algorithm 2: Description Consistency Gate
Input: Image, text, and caption embeddings $x, t, c$,
       historical average similarity $H_{xt}$, $H_{xc}$,
       sample weight $W^s$
Start:
  $S_{xt}, S_{xc} \leftarrow \cos(x, t), \cos(x, c)$
  $H_{xt} \leftarrow m * H_{xt} + (1 - m) * \bar{S}_{xt}$
  $H_{xc} \leftarrow m * H_{xc} + (1 - m) * \bar{S}_{xc}$
  if $W^s < 1$:
      $W^t \leftarrow \exp((H_{xt} - S_{xt}) * \gamma_p)$
      $W^c \leftarrow \exp((H_{xc} - S_{xc}) * \gamma_p)$
  else:
      $W^t, W^c \leftarrow 1, 1$
End
Output: $W^t, W^c$
```

Figure 2. The overall architecture of the proposed Adaptive Language-Image Pre-training (ALIP) method, a bi-path model with triplet input of raw text, synthetic caption, and image. The language consistency gate and description consistency gate are designed to dynamically adjust the weights of samples and image-text/caption pairs during training.



T: "bluebird close up"
C: "a bird sitting on top of a rock"
T: "magic kingdom"
C: "two women standing next to each other on a boat"
T: "a silver cat on a black shirt"
C: "a black and white picture of the number six"
T: "prison radio"
C: "a pair of glasses hanging from a lightswitch"

Figure 3. Examples of the web raw text T and the synthetic caption C can be categorized into four situations. Green and red dotted boxes are used to indicate high-quality and low-quality samples, where green descriptions are considered correct and red descriptions are considered incorrect.

low-quality samples, as illustrated in Fig. 3, can still have well-matched image-text or image-caption pairs that are beneficial for representation learning. To fully utilize the pre-training data, we propose the description consistency gate, which computes the image-text pair weight W^t and the image-caption pair weight W^c for each training image.

Due to the considerable discrepancy between raw text and synthetic caption (which is discussed in Sec. 4.3), we separately record the historical image-text pair similarity H_{xt} and the historical image-caption pair similarity H_{xc}, which are updated dynamically as follows:

$$
\begin{aligned}
H_{xt} &= m * H_{xt} + (1 - m) * \bar{S}_{xt}, \\
H_{xc} &= m * H_{xc} + (1 - m) * \bar{S}_{xc},
\end{aligned}
\tag{3}
$$

where the $\bar{S}_{xt}$ and $\bar{S}_{xc}$ denote the average similarity of the image-text and image-caption pairs. Based on the similarity scores and historical image-text or image-caption pair similarity, the description consistency gate computes the pair weights W^t and W^c.

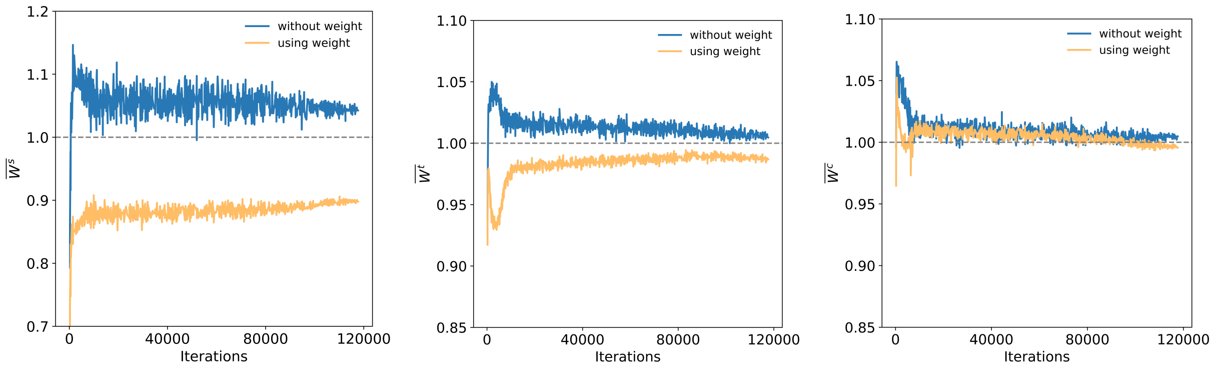
$$
\begin{aligned}
W^t &= \begin{cases} \exp((S_{xt} - H_{xt}) * \gamma_p), & W^s < 1 \\ 1, & W^s = 1 \end{cases} \\
W^c &= \begin{cases} \exp((S_{xc} - H_{xc}) * \gamma_p), & W^s < 1 \\ 1, & W^s = 1 \end{cases}
\end{aligned}
\tag{4}
$$

The pair weight W^t and W^c share a common hyper-parameter γ_p. When $W^s = 1$, the training sample is considered to be high-quality and both W^t and W^c are set to 1. However, when $W^s < 1$, W^t will be larger than 1 if $S_{xt} > H_{xt}$, and W^c will be larger than 1 if $S_{xc} > H_{xc}$. One noteworthy benefit of introducing W^t and W^c is that they are capable of precisely exploiting high-quality image-text or image-caption pairs from low-quality samples.

3.4. Adaptive Contrastive Loss

CLIP [32] utilizes the InfoNCE loss [27] for multi-modal alignment. Given a mini-batch $D = \{(x_i, t_i)\}_{i=1}^{N}$ of image-text feature embeddings, the multi-modal InfoNCE objective is defined as,

$$
L_{\text{NCE}} = -\sum_{i=1}^{N} \left[\log \frac{e^{x_i^\top t_i / \tau}}{\sum_j e^{x_i^\top t_j / \tau}} + \log \frac{e^{x_i^\top t_i / \tau}}{\sum_j e^{x_j^\top t_i / \tau}} \right], \tag{5}
$$



(a) W^s (b) W^t (c) W^c

Figure 4. Visualization of different weights during training. Sample weight W^s and raw image-text pair weight W^t exhibit an obvious decrease. The reduction in image-caption pair weight W^c is relatively minor due to the superior consistency in generated captions.

where τ is the temperature parameter. Even though the InfoNCE loss has achieved huge success in language-vision pre-training, when learning from large-scale noisy web data, the uniform weighting of all training samples can lead to adverse effects on representation learning.

In this paper, we propose an adaptive contrastive loss that incorporates additional sample weight and pair weight into the InfoNCE loss. By dynamically adjusting the sample and pair weights during the training process, the adaptive loss can significantly reduce the impact of noise. Specifically, given a mini-batch $D = \{(x_i, t_i, c_i)\}_{i=1}^N$ of image-text-caption feature embeddings, the adaptive contrastive loss L_{xt} and L_{xc} between the image-text pair and image-caption pair are defined by the following formula:

$$\begin{aligned} L_{\mathrm{xt}} &= -\sum_{i=1}^{N} W_i^s W_i^t \left[\log \frac{e^{x_i^\top t_i/\tau}}{\sum_j e^{x_i^\top t_j/\tau}} + \log \frac{e^{x_i^\top t_i/\tau}}{\sum_j e^{x_j^\top t_i/\tau}} \right], \\ L_{\mathrm{xc}} &= -\sum_{i=1}^{N} W_i^s W_i^c \left[\log \frac{e^{x_i^\top c_i/\tau}}{\sum_j e^{x_i^\top c_j/\tau}} + \log \frac{e^{x_i^\top c_i/\tau}}{\sum_j e^{x_j^\top c_i/\tau}} \right], \end{aligned} \tag{6}$$

where W_i^s is calculated by the language consistency gate, and W_i^t and W_i^c are computed by the description consistency gate. Finally, the overall loss function of our ALIP is defined by combining the bi-path contrastive loss $L_{ALIP} = L_{xt} + L_{xc}$. Fig. 4 illustrates the variation of weights during the training process. ALIP is capable of effectively adjusting the weights to mitigate the impact of noise.

4. Experiments

4.1. Experimental Settings

Pre-training Datasets. We train our model on the YFCC15M dataset, which is a subset of YFCC100M [35] filtered by DeCLIP [21]. To further verify the effectiveness and generalizability of ALIP, we randomly select subsets of 10M and 30M from the LAION400M dataset [34] and conduct a series of experiments with different model sizes and pre-training data scales.

Downstream Datasets. Following recent works [25, 39, 18], we evaluate the effectiveness of our approach in zero-shot image-text retrieval tasks on the Flickr30K [29] and MSCOCO [30] benchmarks. Besides, consistent with HiCLIP [12], we report the linear probe performance over 10 datasets and the zero-shot classification performance over 11 datasets, including CIFAR10 & CIFAR100 [17], Food101 [4], Oxford Pets [28], Flowers102 [26], SUN397 [38], Stanford Cars [16], DTD [8], Caltech101 [11], FGVC Aircraft [24], and ImageNet [10].

Implementation Details. We employ OFA$_{base}$ to generate synthetic captions. The image encoder and text encoder in ALIP follow the same architecture as in CLIP [32]. We use AdamW [23] as the optimizer with an initial learning rate of 0.001 and a weight decay of 0.2. Consistent with CLIP, we set β_1 to 0.9 and β_2 to 0.98 to improve training stability. The input image size is 224 × 224, and the input text sequence length is truncated or padded to 77. The temperature parameter τ is initialized to 0.07. To ensure a fair comparison with baselines, we train ALIP for 32 epochs with a batch size of 4096 on 16 NVIDIA V100 GPUs.

4.2. Experimental Results

We compare ALIP with state-of-the-art approaches by using YFCC15M. Following HiCLIP [12], We report the performance of ALIP in zero-shot text-image retrieval, linear probe, and zero-shot classification, respectively.

Zero-shot Image-text Retrieval. In Tab. 1, we present a comparison of our method with state-of-the-art approaches in zero-shot image-text retrieval on Flickr30k and MSCOCO. Our proposed ALIP achieves new state-of-the-art results on all evaluation metrics. Specifically, ALIP achieves 46.8%/29.3% I2T/T2I retrieval Recall@1 on MSCOCO, which is 8.1%/5.4% higher than HiDeCLIP and 12.6%/8.7% higher than HiCLIP. Similarly, ALIP demonstrates significant improvements of 18.2% to 35.6% and

Table 1. Zero-shot image-text retrieval on the test splits of Flickr30k and MSCOCO. All models are pre-trained on YFCC15M, and ALIP creates new state-of-the-art results on all the metrics.

	Text retrieval						Image retrieval					
	Flickr30k			MSCOCO			Flickr30k			MSCOCO		
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
CLIP-ViT-B/32[32]	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 [25]	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 [21]	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 [18]	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 [12]	-	-	-	34.2	60.3	70.9	-	-	-	20.6	43.8	55.3
HiDeCLIP-ViT-B/32 [12]	-	-	-	38.7	64.4	74.8	-	-	-	23.9	48.2	60.1
ALIP-ViT-B/32	**70.5**	**91.9**	**95.7**	**46.8**	**72.4**	**81.8**	**48.9**	**75.1**	**82.9**	**29.3**	**54.4**	**65.4**

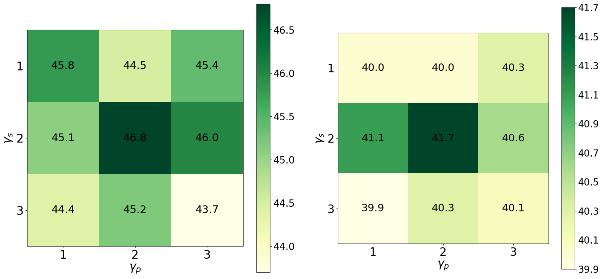
Table 2. Linear probe performance on 10 downstream datasets. ALIP achieves higher average accuracy with an improvement of 1.4∼9.2%.

Method	Pre-train dataset	CIFAR10	CIFAR100	Food101	Pets	Flowers	SUN397	Cars	DTD	Caltech101	Aircraft	Average
CLIP-ViT-B/32[32]	YFCC15M	86.5	64.7	69.2	64.6	90.6	66.0	24.9	61.3	79.1	23.1	63.0
DeCLIP-ViT-B/32 [21]	YFCC15M	89.2	69.0	75.4	72.2	94.4	71.6	31.0	68.8	87.9	27.6	68.7
HiCLIP-ViT-B/32 [12]	YFCC15M	89.5	71.1	73.5	70.6	91.9	68.8	30.8	63.9	84.8	27.4	67.2
HiDeCLIP-ViT-B/32 [12]	YFCC15M	88.1	70.7	**77.6**	75.5	**95.6**	72.2	**36.0**	70.1	**90.0**	32.6	70.8
ALIP-ViT-B/32	YFCC15M	**94.3**	**77.8**	75.8	**76.0**	95.1	**73.3**	33.6	**71.7**	88.5	**36.1**	**72.2**

14.1% to 25.5% on Flickr30K. The performance improvement is mainly attributed to the more robust image description supervision as ALIP can dynamically adjust the weights of samples and image-text/caption pairs to reduce the impact of noise.

Linear Probe. Following the same evaluation setting as CLIP, we freeze the ALIP model and only train a logistic regression classifier. In Tab. 2, we report our linear probe performance on 10 downstream datasets by referring to HiCLIP [12]. Compared with the baseline methods, our ALIP yields an improvement of 1.4% to 9.2% on average, and it surpasses HiCLIP on all datasets and HiDeCLIP on 5 out of 10 datasets. Although ALIP does not exhibit superior performance to HiDeCLIP in half of the datasets, the performance gaps are marginal. Remarkably, compared with HiDeCLIP, ALIP observes a significant performance boost of 6.2%, 7.1%, and 3.5% on the CIFAR10, CIFAR100, and Aircraft datasets, respectively. The performance improvement demonstrates that ALIP can effectively enhance the representation power in instance discrimination.

Zero-shot Classification. We also present our performance on 11 zero-shot classification benchmarks. The prompt templates and class names used for evaluation are consistent with HiCLIP [12] and SLIP [25]. As shown in Tab 3, ALIP achieves substantial improvement only on CIFAR10 and CIFAR100, but it still lags behind HiDeCLIP in terms of zero-shot accuracy. This performance gap is mainly due to the coarse-grained synthetic captions generated by the OFA model. For instance, the OFA model can only recognize the presence of flowers in an image and can not identify the specific species of flower. Besides, ALIP aims to reduce the impact of noisy image-text pairs, and it does not fully account for the hierarchical nature of fine-grained semantics, as does HiDeCLIP.



(a) Zero-shot text retrieval (b) Zero-shot classification

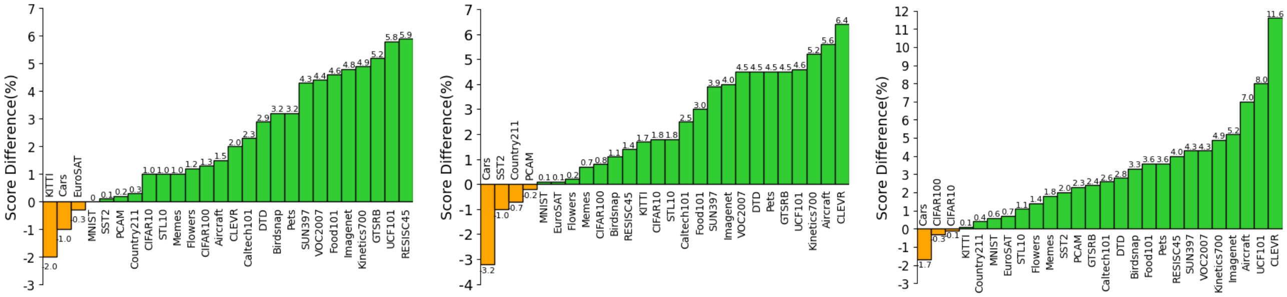
Figure 5. Ablation on the parameters γ_s and γ_p. (a) is the zero-shot text retrieval Recall@1 on MSCOCO. (b) is the average zero-shot classification accuracy on 11 datasets.

4.3. Ablation Study

Ablation on Adaptive Weights. To further explore the effectiveness of the sample weight and image-text/caption pair weight, we perform ablation experiments based on the zero-shot image-text retrieval task. The retrieval Recall@1 for I2T/T2I on Flickr30K and MSCOCO is presented in Tab. 4. Our results indicate that both the sample weight

Table 3. zero-shot classification performance on 11 downstream datasets.

Method	Pre-train dataset	CIFAR10	CIFAR100	Food101	Pets	Flowers	SUN397	Cars	DTD	Caltech101	Aircraft	ImageNet	Average
CLIP-ViT-B/32[32]	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 [25]	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 [39]	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 [21]	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 [9]	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 [12]	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 [12]	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	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



(a) (b) (c)

Figure 6. Linear probe performance comparison between ALIP and CLIP on 27 downstream datasets. (a) ALIP-ViT-B/32 vs. CLIP-ViT-B/32 on LAION10M; (b) ALIP-ViT-B/16 vs. CLIP-ViT-B/16 on LAION10M; (c) ALIP-ViT-B/32 vs. CLIP-ViT-B/32 on LAION30M.

Table 4. Ablation on the sample weight W^s and image-text/caption pair weights W^t and W^c. All models are pre-trained on YFCC15M.

Methods	Weight			Flickr30K		MSCOCO	
	W^s	W^t	W^c	I2T	T2I	I2T	T2I
ALIP-ViT-B/32	×	×	×	68.7	48.1	45.1	27.9
ALIP-ViT-B/32	✓	×	×	69.8	49.1	45.8	29.1
ALIP-ViT-B/32	×	✓	×	69.5	**49.4**	45.6	29.3
ALIP-ViT-B/32	×	×	✓	68.9	48.3	45.3	28.7
ALIP-ViT-B/32	✓	✓	✓	**70.5**	48.9	**46.8**	**29.3**

Table 5. The influence of the caption model in the linear probe and zero-shot classification tasks.

Method	Caption Model	Linear probe AVG	Zero-shot AVG
ALIP-ViT-B/32	OFA$_{base}$	72.2	41.7
ALIP-ViT-B/32	OFA$_{large}$	72.3	42.0

and image-text pair weight improved retrieval performance. Additionally, the improvements are more significant when applying W^t alone than applying W^c, indicating that raw texts have weaker description consistency.

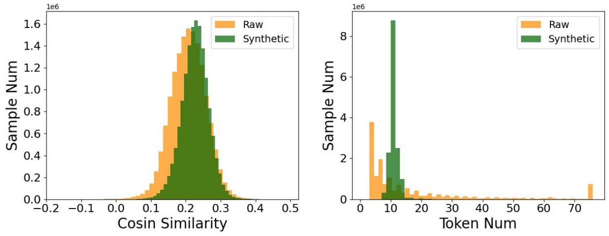
Ablation on the Parameters γ_s and γ_p. The parameters γ_s and γ_p directly affect the sample weight and pair weight. In Fig. 5, we show the zero-shot text retrieval Recall@1 on MSCOCO and the average zero-shot classification accuracy on 11 downstream datasets under different parameter settings. When $\gamma_s = 2$ and $\gamma_p = 2$, ALIP achieves the best performance on both zero-shot text retrieval and zero-shot classification tasks.

Ablation on Different Capacity of Caption Model. Given the significance of synthetic captions in this study, we investigate the impact of captions generated by different sizes of the OFA model on downstream tasks. Specifically, in addition to OFA$_{base}$, we employ the OFA$_{large}$ model which has 470M parameters to generate synthetic captions on the YFCC15M dataset. Then, we train the ALIP-ViT-B/32 and evaluate the average accuracy of linear probe and zero-shot classification on 10 and 11 datasets. The experiment results are presented in Tab. 5, it is worth noting that despite the OFA$_{large}$ model having 2.5 times the parameters of OFA$_{base}$, it only yields a marginal improvement 0.1% on linear probe and 0.4% on zero-shot classification. Additionally, we provide some examples of synthetic captions generated by OFA$_{base}$ and OFA$_{large}$ in the supplementary material. While the synthetic captions generated by OFA$_{large}$ are of higher quality, they still remain coarse-grained descriptions.

Analysis of Raw Text and Synthetic Caption. To examine the distinctions between synthetic captions and raw texts, we conduct a statistical analysis of the token counts and use the CLIP ViT-L/14 model to compute the distribution

Table 6. The linear probe and zero-shot classification performance of CLIP-ViT-B/32 trained on LAION10M.

Method	Pre-train data	CIFAR10	CIFAR100	Food101	Pets	Flowers	SUN397	Cars	DTD	Caltech101	Aircraft	Average
Linear probe:												
CLIP-ViT-B/32	*Raw text-image*	91.2	74.8	66.9	71.0	63.0	89.5	71.1	68.5	40.3	84.7	72.1
CLIP-ViT-B/32	*Synthetic caption-image*	90.7	71.9	65.1	68.6	63.8	88.2	39.5	68.3	40.3	85.5	68.2
zero-shot classification:												
CLIP-ViT-B/32	*Raw text-image*	78.5	49.3	42.0	42.5	28.6	40.8	39.9	23.7	73.2	4.4	42.3
CLIP-ViT-B/32	*Synthetic caption-image*	57.1	21.1	9.9	8.3	4.8	10.8	2.8	9.2	39.5	1.0	16.5



(a) Distribution of similarity (b) Distribution of token num

Figure 7. We conduct a statistical analysis of raw text and synthetic caption on YFCC15M. (a) is the image-text/caption similarity distribution; (b) is the token number distribution of the raw texts and synthetic captions.

of similarity between raw and synthetic image-text pairs. As illustrated in Fig. 7, in comparison to raw texts, synthetic captions demonstrate a higher average similarity and more compact similarity distribution. Additionally, we observe that the number of tokens in the synthetic caption is predominantly concentrated between 10 and 15, which is significantly lower than in raw text.

To better investigate the performance disparities between synthetic caption and raw text in downstream tasks, based on LAION10M, we train CLIP-B/32 on raw and synthetic image-text pairs respectively. We present the linear probe and zero-shot classification performance in Tab. 6. Compared with the CLIP-B/32 trained on the raw image-text pairs, the CLIP-B/32 trained on the synthetic caption-image pairs achieves similar or better linear probe performance on all the datasets except Cars. However, the zero-shot results reveal a significant deficiency of synthetic captions in zero-shot classification task. This is mainly due to the coarse granularity of the synthetic captions, which also explains the inferior performance of ALIP in Tab. 3.

Effectiveness across Different Pre-training Datasets. In addition to YFCC15M, we conduct experiments on randomly selected subsets of 10M and 30M from LAION400M. For a more comprehensive comparison, we report the linear probe performance on 27 downstream datasets. As illustrated in Fig. 6, ALIP significantly improves the performance on different models and pre-training datasets. Specifically, the ALIP-ViT-B/32 models pre-trained on LAION10M and LAION30M outperform the CLIP-ViT-B/32 models on 23 and 24 datasets, respectively. Additionally, when training a larger model, the ALIP-ViT-B/16 model surpasses the CLIP-ViT-B/16 model on 23 datasets. The experimental results demonstrate that ALIP exhibits robustness and extensibility. Please refer to the supplementary material for more detailed experimental results.

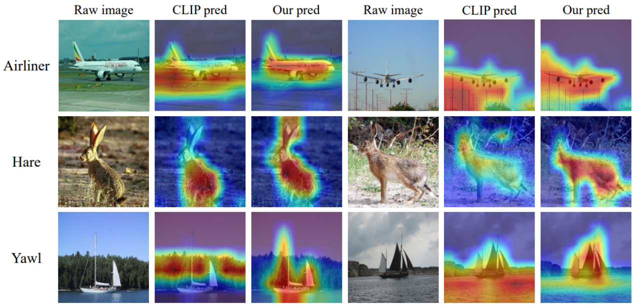


Figure 8. Class activation maps for ALIP and CLIP on different classes from ImageNet.

As shown in Fig. 8, we compare the class activation maps of ALIP and CLIP on different classes from ImageNet. Here, we use the class label as the textual tokens. As can be seen, ALIP is superior in aligning the image patches and textual tokens. For instance, CLIP only focuses on the body of the rabbit, but ALIP is able to also capture the ears. These results highlight the potential of ALIP to enhance the performance of image-text retrieval tasks.

5. Conclusion

In this paper, we introduce a bi-path adaptive contrastive learning model, which includes the language consistency gate and description consistency gate. Specifically, LCG and DCG can adjust the weights of samples and image-text/caption pairs during training, thus effectively reducing the impact of the noisy or unaligned language description. Our method shows superior performance with different models and pre-training datasets on different downstream tasks. We hope our work could bring insights into exploiting the language-image pre-training model.

References

[1] Xiang An, Jiankang Deng, Kaicheng Yang, Jaiwei Li, Ziyong Feng, Jia Guo, Jing Yang, and Tongliang Liu. Unicom: Universal and compact representation learning for image retrieval. In *ICLR*, 2023. 1

[2] Alex Andonian, Shixing Chen, and Raffay Hamid. Robust cross-modal representation learning with progressive self-distillation. In *CVPR*, 2022. 2

[3] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *TPAMI*, 2018. 1

[4] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *ECCV*, 2014. 5

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020. 1

[6] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. *arXiv:2212.07143*, 2022. 2

[7] Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 1, pages 539–546. IEEE, 2005. 1

[8] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *CVPR*, 2014. 5

[9] 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:2203.05796*, 2022. 7

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 5

[11] 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 *CVPR*, 2004. 5

[12] Shijie Geng, Jianbo Yuan, Yu Tian, Yuxiao Chen, and Yongfeng Zhang. HiCLIP: Contrastive language-image pretraining with hierarchy-aware attention. In *ICLR*, 2023. 2, 5, 6, 7

[13] Wenzhong Guo, Jianwen Wang, and Shiping Wang. Deep multimodal representation learning: A survey. *IEEE Access*, 2019. 1

[14] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020. 1

[15] 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 *ICML*, 2021. 1, 2

[16] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *ICCVW*, 2013. 5

[17] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 5

[18] Janghyeon Lee, Jongsuk Kim, Hyounguk Shon, Bumsoo Kim, Seung Hwan Kim, Honglak Lee, and Junmo Kim. Uniclip: Unified framework for contrastive language-image pretraining. *arXiv:2209.13430*, 2022. 2, 5, 6

[19] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022. 2, 3

[20] 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. In *NeurIPS*, 2021. 2

[21] 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:2110.05208*, 2021. 2, 3, 5, 6, 7

[22] Zhining Liu, Pengfei Wei, Jing Jiang, Wei Cao, Jiang Bian, and Yi Chang. Mesa: boost ensemble imbalanced learning with meta-sampler. *NeurIPS*, 2020. 3

[23] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv:1711.05101*, 2017. 5

[24] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv:1306.5151*, 2013. 5

[25] Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. Slip: Self-supervision meets language-image pretraining. In *ECCV*, 2022. 2, 5, 6, 7

[26] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, 2008. 5

[27] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv:1807.03748*, 2018. 4

[28] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *ICCV*, 2012. 5

[29] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *ICCV*, 2015. 5

[30] Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari. Connecting vision and language with localized narratives. In *ECCV*, 2020. 5

[31] Filip Radenovic, Abhimanyu Dubey, Abhishek Kadian, Todor Mihaylov, Simon Vandenhende, Yash Patel, Yi Wen, Vignesh Ramanathan, and Dhruv Mahajan. Filtering, distillation, and hard negatives for vision-language pre-training. *arXiv:2301.02280*, 2023. 1

[32] 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 *ICML*, 2021. 1, 2, 3, 4, 5, 6, 7

[33] 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. *arXiv:2210.08402*, 2022. 1, 2

[34] 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:2111.02114*, 2021. 1, 2, 5

[35] 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*, 2016. 1, 3, 5

[36] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *ECCV*, 2020. 1

[37] 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 *ICML*, 2022. 1, 2, 3

[38] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *ICCV*, 2010. 5

[39] 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:2111.07783*, 2021. 2, 5, 7

[40] Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. Lit: Zero-shot transfer with locked-image text tuning. In *CVPR*, 2022. 2