Attentive Mask CLIP

Yifan Yang[1*] Weiquan Huang[2†*] Yixuan Wei[3†] Houwen Peng[1] Xinyang Jiang[1]
Huiqiang Jiang[1] Fangyun Wei[1] Yin Wang[2] Han Hu[1] Lili Qiu[1] Yuqing Yang[1]
[1]Microsoft Research Asia [2]Tongji University [3]Tsinghua university
{yifanyang,houwen.peng,xinyangjiang,hjiang,fawe,hanhu,liliqiu,yuqing.yang}@microsoft.com
{weiquanh,yinw}@tongji.edu.cn wei-yx20@mails.tsinghua.edu.cn

Abstract

In vision-language modeling, image token removal is an efficient augmentation technique to reduce the cost of encoding image features. The CLIP-style models, however, have been found to be negatively impacted by this technique. We hypothesize that removing a large portion of image tokens may inadvertently destroy the semantic information associated to a given text description, resulting in misaligned paired data in CLIP training. To address this issue, we propose an attentive token removal approach, which retains a small number of tokens that have a strong semantic correlation to the corresponding text description. The correlation scores are dynamically evaluated through an EMA-updated vision encoder. Our method, termed attentive mask CLIP, outperforms original CLIP and CLIP variant with random token removal while saving the training time. In addition, our approach also enables efficient multi-view contrastive learning. Experimentally, by training ViT-B on YFCC-15M dataset, our approach achieves 43.9% top-1 accuracy on ImageNet-1K zero-shot classification, 62.7/42.1 and 38.0/23.2 I2T/T2I retrieval accuracy on Flickr30K and MS COCO, outperforming SLIP by ***+1.1%****,* ***+5.5/+0.9****, and* ***+4.4/+1.3****, respectively, while being* ***2.30×*** *faster. An efficient version of our approach runs* ***1.16×*** *faster than the plain CLIP model, while achieving significant gains of* ***+5.3%****,* ***+11.3/+8.0****, and* ***+9.5/+4.9*** *on these benchmarks, respectively. Code will be release in* https://github.com/microsoft/A-CLIP.

1. Introduction

Large-scale vision-language pre-training models, such as CLIP [28] and ALIGN [16], have demonstrated remarkable capabilities in zero-shot image classification and multi-modal retrieval. However, these models typically require a large amount of training data, which raises the training

Methods	Training Time	GPU Memory	IN 1K 0-shot	Flickr30K I2T/T2I	MS COCO I2T/T2I
CLIP	1.00×	14G	37.6	51.4/32.6	27.9/17.6
SLIP	2.67×	30G	42.8	57.2/41.2	33.6/21.9
MaskCLIP	1.56×	16G	42.7	60.0/38.8	34.1/21.2
A-CLIP	1.16×	14G	**43.9**	**62.7/42.1**	**38.0/23.2**
A-CLIP-eff	**0.86×**	**13G**	42.9	**62.7**/40.6	37.4/22.5

[1] The full training wall clock time and GPU memory footprint are measured on the same device. We report the training cost relative to the original CLIP [28].

Table 1: We compare our attentive mask CLIP (A-CLIP) with CLIP [28], SLIP [27] and MaskCLIP [7]. A-CLIP outperforms CLIP by +6.3%, +11.3/+9.5 and +10.1/+5.6 on Imagenet-1K [30] zero-shot classification, Flickr30K [39] and MS COCO [23] I2T/T2I retrieval. An efficient variant termed A-CLIP-eff outperforms CLIP by +5.3%, +11.3/+8.0, and +9.5/+4.9 on these benchmarks, while reducing the training time to 0.86×.

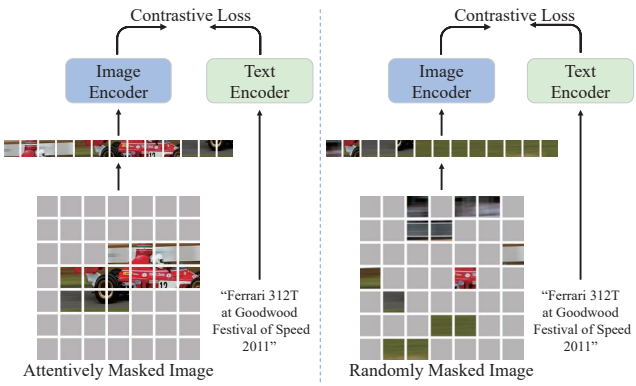

Figure 1: Attentive mask vs. random mask. Left part is the attentive mask applied CLIP training process, and right part is random mask applied. Here we use ViT-B16 model from A-CLIP's visual encoder to generate above masked images with patch size of 32 × 32 and 25% mask ratio. The image and alt-text are sampled from YFCC-100M [32].

*Equal contribution. †This work was done during internship in MSRA.

cost. For instance, CLIP is trained on 400 million image-text pairs, and ALIGN uses more than 1 billion paired data. This raises the need for more efficient language-image pre-training methods.

This paper aims to improve the efficiency of CLIP training by introducing an efficient image augmentation approach called "image token removal". This approach drops a large portion of image tokens, thereby reducing the computation in the image encoder. It has been shown to be effective in masked image modeling [13, 35, 26, 15, 31] when combined with vision Transformer architectures [8]. Thus, we seek to introduce this approach into CLIP training to improve its efficiency.

However, previous work has shown that dropping tokens randomly can harm CLIP performance [9, 19]. This is also evidenced in our own experiments (see Table 2). Specifically, we observe a -2.6% top-1 accuracy drop on ImageNet zero-shot classification when we randomly remove 50% tokens in each image. The underlying reason for this issue is that the process of removing tokens may mistakenly eliminate semantic content that is pertinent to the alt-text, leading to inaccurate image-text pairs for CLIP training.

To mitigate this issue, we propose an attentive token removal strategy, as shown in Figure 1. The fundamental idea is to retain a small set of image tokens that are more closely related to the corresponding text description while discarding those that are irrelevant. Concretely, we encode all image tokens into a latent space and then calculate the correlation scores between each image-token feature and the text feature extracted by the CLIP's text encoder. We investigate several strategies for selecting image tokens using correlation scores and conclude that retaining these image tokens that are most semantically related to the text description yields the best performance. The correlation scores are computed by using the exponential moving average (EMA) version of the vision encoder. Specifically, we use the attention weights of the $[CLS]$ token of the visual encoder as the correlation scores. We also find that using the averaged attention weights of all layers performs better. Figure 4 shows the selected tokens, which correlate well with the associated text semantics.

The efficiency of the proposed token removal approach allows us to construct multiple masked views from an image while keeping the training as efficient as the original CLIP, e.g., 2 masked views with a token removal ratio of 50%. The random cropping strategy adds certain stochastic effects to different masked views, enabling the application of an auxiliary contrastive loss between the augmented views in addition to the plain CLIP loss. In fact, as has been shown in SLIP [27, 7], the auxiliary task facilitates the CLIP training. We consider both SimCLR [3] and SimSiam [4] methods for the auxiliary contrastive learning task. It is worth noting that the EMA branch can be naturally treated as another view, and we also apply an online-to-EMA contrastive or consistency loss and use the formulation from BYOL [12] for this purpose.

The proposed approach is called A-CLIP, which introduces only 16% computational overhead in comparison to the plain CLIP. It is also 2.30× and 1.34× faster than previous CLIP improvements that also include multiple views and additional self-supervised losses, while being more effective. Using ViT-B and the YFCC-15M [32, 28] dataset, the A-CLIP framework achieves 43.9% top-1 accuracy on ImageNet-1K [30] zero-shot classification(see Table 1). Additionally, it achieves 62.7/42.1 and 38.0/23.2 I2T/T2I retrieval accuracy on Flickr30K [39] and MS COCO [23], respectively, which is **+1.1%**, **+5.5/+0.9**, and **+4.4/+1.3** higher than the SLIP method, and **+1.2%**, **+2.7/+3.3**, and **+3.9/+2.0** higher than the MaskCLIP method.

Also note the training cost of our approach can be further reduced by using a lower resolution input for the EMA network, i.e., 2x lower reduced resolution. This has little affect on the accuracy of correlation score computing, while marginally reduce the efficacy of online-to-EMA contrastive loss. This strategy will reduce the EMA computation by more than 4×, resulting in an efficient variant that is even 1.16× faster than the plain CLIP model, and is significantly more accurate. We refer to this more efficient variant as A-CLIP-eff.

2. Related Work

Contrastive language-image pre-training An important goal of computer vision is to interpret visual signals using language that humans can understand. While the field has long used image classification tasks to learn visual representations which connect visual signals to semantics, recent works, such as CLIP [28] and ALIGN [16], suggest a new way to connect visual signals with linguistic semantics by contrasting image-language pairs. In this visual-language contrastive learning framework, the training data is more scalable as billions of image-alt-text pairs can be easily collected from the internet. At present, CLIP has become a mainstream visual learning method that not only learns transferable representations but also connects visual signals to arbitrary semantics. There have been extensive follow-up studies [20, 11, 38, 27, 7, 34, 37] to improve the effectiveness by sacrificing training efficiency. Our method aims to improve CLIP pre-training while improving both effectiveness and efficiency.

Masking tokens for efficient computation Vision Transformers [8, 25] process images as sequences of patch tokens and perform encoding computation on these tokens. A masking strategy that removes certain tokens from an image can significantly speed up the computation of the image encoder. Dynamic ViTs [29, 33] learn to remove tokens

for efficient image classification. Masked autoencoder [13] randomly masks 75% of tokens, which significantly speeds up self-supervised visual representation pre-training based on masked image modeling.

This paper extends the idea of token masking to speed up the CLIP training process. Instead of using random masking like masked image modeling, we propose an attentive masking method that removes only semantically meaningless tokens, thereby alleviating the issue of noisy pairing between image tokens and text descriptions in the random masking strategy caused by incorrectly discarded semantic content.

Comparison with a concurrent work FLIP [22] Similar to our approach, there is concurrent work called FLIP [22], which also employs token masking to accelerate CLIP training. However, FLIP uses a random masking strategy like MAE [13], and achieves inferior zero-shot accuracy than when trained with full images, while keeping the batch size and using the ViT-B model.

Our approach improves upon FLIP in several ways. First, we propose an attentive masking strategy for CLIP training that significantly outperforms the random masking baseline. This is especially crucial for CLIP, as its target heavily relies on semantic texts, while random masking can work well with MIM that do not have explicit semantic supervision. Second, we introduce multiple masked image views into our framework, enabling us to conveniently incorporate auxiliary pre-text tasks, such as image-to-image contrast learning. By reducing computation through masking, our approach does not increase computation compared to the plain CLIP model using full images, while improving the pre-trained representations.

With these techniques, our approach achieves significant accuracy improvements over the original CLIP model on both zero-shot image classification and multi-modal retrieval. In contrast, FLIP only achieves comparable performance to the original CLIP model, even with additional training tricks. Also note that some of the findings in FLIP, particularly on scaling experiments and hyper-parameter tuning such as larger batch sizes and base learning rate tuning, complement to our approach. Together, these findings provide readers with a more complete view of using masking for CLIP training.

Masking for data augmentation Token masking is not only more efficient, but also serves as a form of data augmentation for visual representation learning. Studies have shown that masking augmentation performs well in instance-based contrastive learning [1, 35, 17, 26]. While not our primary focus, we demonstrate that attentive masking for CLIP training also benefits from its data augmentation property, as reflected in the increasing accuracy with a longer scheduler where saturation is observed using full images as in Table 4.

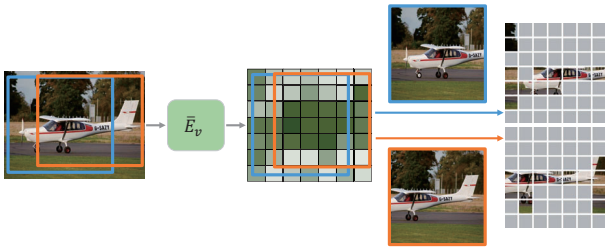

Figure 2: A illustration of computing attentive masks for multiple image views. For different image views, we perform score map computation once using an EMA version of the visual encoder ($\overline{E_v}$) on the original image or the minimum enclosing area of all image views. Then for each image view, a bilinear sampling approach is adopted to generate the selection score for each image token. The masking is performed based on the selection scores.

Combining CLIP with other representation learning methods Our method is also related to recent approaches that combine CLIP with other pre-text tasks to improve representation learning [27]. For example, SLIP [27] combines CLIP with image-to-image contrastive learning [14, 3], while MaskCLIP [7] combines CLIP with masked image modeling [2, 13, 36, 24]. These approaches have shown to learn stronger representations, but at the cost of higher training requirements.

In contrast, our method introduces the task of image-to-image contrastive learning without incurring additional training costs. The attentive mask generated by an EMA network on the full image constitutes another view of the image, enabling us to add an auxiliary task of image-to-image contrastive learning with minimal additional cost. Moreover, the masking strategy allows us to introduce more masked views with minimal computation cost, facilitating the application of richer image-to-image tasks in training and further improving representation learning. Our framework can also naturally incorporate masked image modeling, as we use a masked image input, which is a direction for future work.

3. Method

3.1. A Brief Review of CLIP

CLIP [28] is a visual representation learning approach that uses a large amount of image-to-text pairs. The training task treats each image and its associated alt text as a positive pair and pairs of this image to all other alt-texts as negative. For all possible pairs within a batch, the method applies an InfoNCE-like loss to classify them as positive or negative.

The CLIP model shows well connecting images to arbitrary language semantics and achieves remarkably strong zero-shot image recognition accuracy on benchmarks. Its learned representation also performs very well when fine-tuned on various down-stream tasks [21].

Specifically, the CLIP method independently applies a visual encoder $\mathbf{E}_v$ and a language encoder $\mathbf{E}_l$ on each image and each alt-text, respectively. In this paper, the visual and language encoders are instantiated as a Vision Transformer [8] and a standard language Transformer [6]. For the vision Transformer, a learnable $[CLS]$ token is prepended to the image tokens to represent the entire image. For the language encoder, an $[EOS]$ token is appended after the last word to represent the full sentence. The CLIP model projects the $[CLS]$ and $[EOS]$ features into an embedding space, denoted as e^I and e^T respectively, where a vision-language contrastive loss is applied:

$$\mathcal{L}_{vl} = 0.5 * \mathcal{L}_v + 0.5 * \mathcal{L}_l \tag{1}$$

where

$$\mathcal{L}_v = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp\left(\mathrm{sim}\left(e_i^I, e_i^T\right)/\tau\right)}{\sum_{j=1}^{B}\exp\left(\mathrm{sim}\left(e_i^I, e_j^T\right)/\tau\right)} \tag{2}$$

$$\mathcal{L}_l = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp\left(\mathrm{sim}\left(e_i^T, e_i^I\right)/\tau\right)}{\sum_{j=1}^{B}\exp\left(\mathrm{sim}\left(e_i^T, e_j^I\right)/\tau\right)} \tag{3}$$

In the above equations, B denotes the batch size; $\mathrm{sim}(\cdot)$ denotes the cosine similarity function; τ is a learnable temperature to scale the logits.

3.2. Masking for Efficient CLIP Training

We seek to adopt a masking strategy to improve the training efficiency of CLIP.

A naive approach is to randomly select a portion of image tokens to remove, as shown in Figure 1 (right). The random masking method has been shown to be effective for masked image modeling(MIM) [13]. However, we note two main differences between MIM and CLIP tasks: 1) The MIM method is mainly designed for pre-training, while CLIP excels at zero-shot classification and retrieval. In fine-tuning settings, the domain gap between masked-image based pre-training and full-image based fine-tuning can be bridged by adjusting model parameters, which zero-shot evaluation hopes to achieve good accuracy by not changing the original model and thus requires a small domain gap between pre-training and evaluating; 2) MIM pre-training is more of a low-level task without language semantics. Removing highly semantic visuals does not affect training in general. CLIP training is more of a high-level task, aiming to connect images to semantics defined by text. Removing highly semantic objects related to associated alt-text can form noisy pairs. As illustrated in Figure 1 (right), after the Ferrari car is almost discarded from an image, it would be inappropriate to associate the masked image to a textual description of Ferrari at a race.

In fact, as shown in Table 2, using a random ratio with 50% mask ratio, the zero-shot accuracy on ImageNet-1K is 2.6% worse than using the full image for training (35.0% vs. 37.6%). In the following subsection, we will present an attentive masking method that can resolve the issues of random masking described above, and achieves good accuracy for zero-shot and retrieval.

3.3. Attentive Mask

The goal of attentive masking is to retain tokens that are relevant to language description, as illustrated in Figure 1 (left).

To this end, we note that the representation of the $[CLS]$ token after CLIP training corresponds well to the semantics contained in the associated alt-text, and its attention weights to other image tokens can act a good indicator as the relevance measure of each image token to the linguistic semantics. We thus compute the score of the token at location P as:

$$s_P = \frac{1}{HL}\sum_{l=1}^{L}\sum_{h=1}^{H} \mathrm{Softmax}\left(\frac{\mathbf{f}_{lh}^q(CLS)\cdot\mathbf{f}_{lh}^k(P)}{\sqrt{C}}\right), \tag{4}$$

where l denotes the layer index; h denotes the attention head index; $\mathbf{f}_{lh}^q(CLS)$ denotes the query embedding of the $[CLS]$ token at Layer l and Head h; $\mathbf{f}_{lh}^k(P)$ denotes the key embedding of Layer l and Head h for an image token at location P; C is the number of channels for the query and key embedding.

The image tokens to mask are selected based on the scores s_P. We consider three strategies: 1) "Low" strategy. The image tokens with the lowest scores are discarded; 2) "High" strategy. The image tokens with the highest scores are discarded; 3) "Mixed" strategy. A portion of the maintained tokens are the ones with the highest scores, and the other maintained tokens are randomly selected. In the experiments, the "low" strategy performs best, and is set as the default in our method.

Generating scores using an EMA network To obtain the attentive masked images as described above, we need a network that can generate attention scores for all image areas. We introduce an exponential moving average (EMA) version of the CLIP visual encoder and apply this EMA network on the full image to generate the attention scores. Using the EMA network has the following merits: First, the attention scores can be computed online, without a need of an already trained network. Second, the EMA network constitutes another view of the image, so other effective auxiliary tasks such as BYOL [12] can be introduced with little

overhead. Third, the EMA network only involves the inference phase during training, which can save $2/3$ of the computational cost compared to using a trainable network.

In order to further improve efficiency and effectiveness, we propose the following two techniques.

Efficient EMA computation with reduced image resolution The EMA network is primarily used for attention score computation to select image tokens to maintain. It does not need to be very accurate. We thus propose to perform the EMA computation at reduced image resolution, for example, using half the original resolution. By reducing the resolution, the EMA computation cost is saved by more than $4\times$.

Overall, the EMA computation will correspond to be $1/12$ of a regular CLIP visual encoder, which is efficient.

Shared EMA score map for multiple masked views We also consider adding more masked image views for better results. A naive solution would be performing one time of the EMA computation for each view. However, this would waste computational cost. We propose using a shared EMA attention score maps for different image views.

Specifically, we first compute the minimum enclosing rectangle for multiple image views. Accordingly, we crop and resize the original image, where the EMA network is applied to produce an attention score map. The selection scores for each image view are computed from this shared attention score map via bilinear interpolation, and with the selection scores, the masking is performed using the "low" strategy described above. Please see the process in Figure 2.

3.4. Overall Framework

Based on attentive masking, we present an attentive mask CLIP method, dubbed A-CLIP. The overall framework is illustrated in Figure 3. In the following, we describe several key designs and implementations of the A-CLIP framework.

Multiple masked views The A-CLIP framework can take multiple masked views for better results. To make different number of views with roughly the same complexity, we set the token number for each masked view as N/k (N and k are the number of tokens and masked views respectively). The final VL loss is the average of all losses between each masked view and the alt-text. In the experiments, $k = 2, 3$ perform best (see Table 5), and $k = 2$ is set as default.

Auxiliary self-supervised tasks Our framework can conveniently add several auxiliary self-supervised tasks. We consider two tasks: 1) Online-to-EMA contastive task. We adopt the BYOL instantiation [12] for this task, which encourages the feature of a masked view to be the same as the EMA feature. 2) Online-to-online contrastive tasks. When multiple masked views are used, we can introduce contrastive losses between the masked views. We use the SimCLR [3] or SimSiam [4] instantiation for this task. Table 3 shows that both auxiliary self-supervised learning tasks can improve the zero-shot and retrieval accuracy.

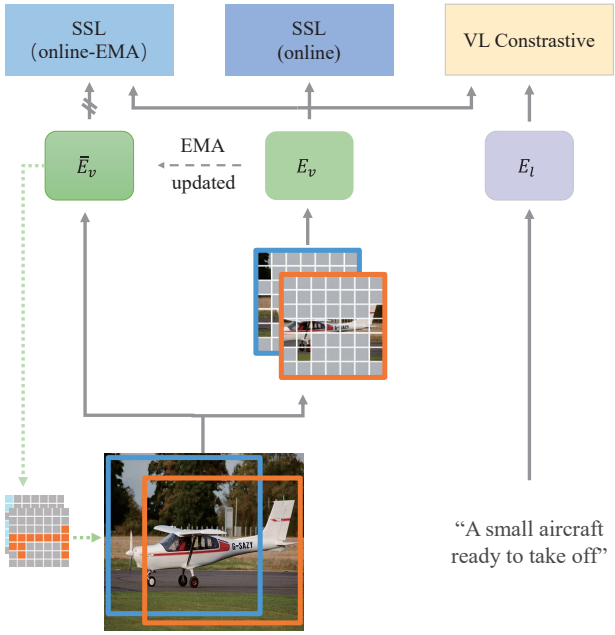

Figure 3: The architecture diagram of A-CLIP. We use masked multi-view input, where each view is computed independently for the vision language(VL) contrastive loss and then averaged. Between multiple online views we introduce auxiliary image self-supervised learning(SSL) to enrich features. An EMA update vision encoder is used to help us generate the attentive mask, which is not involved in the gradient update. Since EMA encoder uses unmasked images, we design an self-distillation based online-EMA SSL task to learn the output distribution of the complete image.

An A-CLIP-eff variant For A-CLIP, we use the full image resolution as the input to the EMA network. We also introduce a variant called A-CLIP-eff, which uses half the resolution for EMA input. The new variant is more efficient than the A-CLIP variant, while being marginally worse in accuracy. Specifically, A-CLIP-eff is $0.86\times$ than that of the original CLIP model in training cost, much faster than the A-CLIP variant ($1.16\times$ of the original CLIP model), and slightly worse than A-CLIP in terms of zero-shot image classification and multi-modal retrieval accuracy (see Table 1).

4. Experiments

4.1. Implementation Details

Dataset and augmentation We train our model on a 15M subset of YFCC100M [32] filtered by Radford et al. [28]. The text data in the subset consists of English-only titles and descriptions. During training, we randomly sample a valid caption (e.g., title or description) for each image, as in SLIP [27]. The data augmentation is also similar to SLIP. Images in the online branches are randomly resized and cropped to between 50% and 100% of the original image size. For the EMA encoder, we choose a larger randomly cropped sub-image than online views to ensure we can get attentive scores by bilinear sampling. When adding auxiliary self-supervised learning tasks such as SimCLR or SimSiam, we used a classic combination of color jitter, grayscale, solarize and blur.

Architecture and training configurations We employ the ViT-B/16[8] architecture as our visual encoder, and a 12-layer, 512-width, and 8-head Transformer as our text encoder following CLIP. The input resolution for the image encoder is 224×224, and the text is encoded into 77 tokens using a vocabulary of 49k tokens with necessary truncations or paddings. To ensure stable training, we use a fixed random patch projection layer to embed image patches, following MoCo v3 [5].

The AdamW optimizer with a learning rate of 5e-4 and a weight decay of 0.5 is used, with a batch size of 4,096. The built-in automatic mixed precision library in PyTorch is adopted for training in all experiments. For the exponential moving average (EMA) model which generates the attentive mask, its momentum starts from 0.996 and gradually increases to 1 using a cosine scheduler during training, following [12]. For the A-CLIP-eff variant, we use a halved resolution image as the input of the EMA model, with a bi-cubic interpolation method to get the new position encodings for the lower-resolution images.

Details of Evaluation Tasks For zero-shot retrieval, we perform text-to-image and image-to-text evaluation on two benchmarks: COCO [23] and Flickr30k [39]. Following [28], we retrieve top-k candidates using the similarity scores from the image and text encoder. We evaluate zero-shot transfer capacity of the models on various classification benchmarks, such as ImageNet [30] and Caltech101 [10]. Following [28], we use the same prompt templates and class names for the evaluation.

4.2. Main Results

CLIP with random masking This is the first work to analyze the impact of mask on the zero-shot performance of CLIP. There are two potential effects of mask training on CLIP. One is that it may cause ambiguity or meaninglessness in text and image matching, and the other is that the masked input of the pre-training process differs significantly from the full image during zero-shot evaluation. Our analysis in Table 2a reveals that randomly masking 50% of image tokens in the 1×50% setting leads to a drop of -2.6% in ImageNet-1K zero-shot accuracy compared to the original CLIP model, as well as -2.6/-0.1, +1.0/-1.0 I2T/T2I retrieval accuracy on Flickr30K and MS COCO.

However, the comparison may be unfair due to lower training costs with mask training. To address this, we use two-views for the random masking method, which is still 20% more efficient than the original CLIP model due to the superlinear computation cost reduction for the self-attention layers. The two-view version is now comparable in zero-shot accuracy to the one using full images (38.0% vs. 37.6%), and has shown +3.2/+1.8, +3.2/+1.1 retrieval accuracy improvements on Flickr30K and MS COCO. We believe that in the 2×50% case, two independent random crop samples increase the richness of the input visual signal, alleviate the side effects of image and text description mismatch caused by random mask, and gain the benefits of

Mask Strategy	IN 1K	Flickr30K		MS COCO	
(View × Ratio)	0-shot	I2T	T2I	I2T	T2I
w/o mask					
1×100%	37.6	51.4	32.6	27.9	17.6
+random mask					
1×50%	35.0	48.8	32.5	28.9	16.6
2×50%	38.0	54.6	34.4	31.1	18.7
+attentive mask					
1×50%	39.5	57.6	36.6	34.2	19.8
2×50%	**41.3**	**59.3**	**38.4**	**35.1**	**21.3**

(a) Comparison of different mask strategies.

Methods	IN 1K	Flickr30K		MS COCO	
	0-shot	I2T	T2I	I2T	T2I
Selection					
high	28.5	42.6	29.0	23.5	13.6
mix	40.4	**59.8**	37.6	34.9	20.9
low	**41.3**	59.3	**38.4**	**35.1**	**21.3**
Patch Size					
16	40.8	**61.4**	37.6	**35.1**	20.5
32	**41.3**	59.3	**38.4**	**35.1**	**21.3**
Layers					
last	40.4	**59.4**	36.8	34.9	20.0
all	**41.3**	59.3	**38.4**	**35.1**	**21.3**
EMA					
eff	41.0	56.8	37.5	**35.1**	20.4
full	**41.3**	**59.3**	**38.4**	**35.1**	**21.3**

(b) Ablations for adding attentive mask to CLIP training.

Table 2: Comparison of different mask strategies and ablation experiments of attentive mask. The indigo background highlights our default setting.

mask as an effective data augmentation at the same time.

The effects of attentive mask As described in Section 3.3, we developed attentive mask strategies for CLIP training that allow us to selectively sample tokens based on their potential relevance to the text, rather than random masking.

In the 1×50% setting, attentive mask improved performance by +4.5%, +8.8/+4.1 and +5.3/+3.2 compared to random mask in ImageNet-1K, Flickr30K and MS COCO I2T/T2I. In the 2×50% setting, improvements to random mask were +4.7%, +4.7/+4.0 and +4.0/+2.6. It can be found that the attentive mask solves the non-negligible performance degradation caused by random mask in single masked view setting, and is even much higher than full token CLIP. These results demonstrated that attentive mask can effectively integrate mask learning into CLIP, providing the benefits of data augmentation without introducing bias to the vision language contrastive task.

We also conducted experiments on different selection strategies, as shown in Table 2b. The *Low* strategy, which retained the top 50% of the most relevant tokens, achieving the best results. It increased ImageNet-1K zero-shot performance by +3.7%, and improved Flickr30K and MS COCO I2T/T2I retrieval accuracy by +7.9/+5.8 and +7.2/+3.7. This indicates that attentive selection can provide a more informative and valuable input distribution. In contrast, the *High* strategy, which masks off the most relevant 50% of tokens and is a good strategy for AttnMask [18] in MIM, resulted in a drastic performance collapse, while the *Mix* strategy, which combines *Low* with 25% random selection, did not match the performance of *Low*. These results suggest that for CLIP's VL contrastive task, only tokens with greater relevance are needed, which is consistent with the assumption of MaskCLIP [7] that CLIP's training mainly relates to the language-described region and plays a low role.

Although masking brings efficiency gains, the additional cost of attentive selection is the need to pre-infer with the EMA encoder. Although back-propagation is not required, this takes close to 30% of the training time. To further reduce the cost, we use half-resolution images as input to the EMA encoder, which reduces the cost to 5%. As shown in the results of EMA-eff in Table 2b, this step achieves competitive results compared to EMA-full and resulted in a +3.4%, +5.4/+4.9, +7.2/+2.8 improvement on ImageNet-1K zero-shot classification, Flickr30K and MS COCO I2T/T2I retrieval, respectively.

A-CLIP provides an efficient paradigm for combining SSL with CLIP SLIP [27] and MaskCLIP [7] both add image self-supervised learning (SSL) to CLIP using a separate branch, whereas our proposed A-CLIP integrates different SSL tasks naturally and efficiently through masked multi-view input. In Table 3, with the same SimCLR task added, A-CLIP can achieve comparable ImageNet [30] zero-shot performance to SLIP [27], +6.4/-0.2 in Flickr30K retrieval and +2.4/+0.7 in MS COCO retrieval, but the reduced branches accelerate the training time by 2.3×. We tried SimSiam without using negative samples as a SSL task and had similar performance gains, demonstrating that our approach is not dependent on specific SSL tasks.

Online-EMA SSL can further improve performance. To obtain the attentive mask through the EMA encoder, we use a self-distillation based BYOL head to learn the output distribution of the full image. Table 3 shows that Online-EMA

Methods	IN 1K	Flickr30K		MS COCO	
	0-shot	I2T	T2I	I2T	T2I
CLIP					
plain	37.6	51.4	32.6	27.9	17.6
+SimCLR(SLIP)	42.8	57.2	41.2	33.6	21.9
+MAE(MaskCLIP)	42.7	60.0	38.8	34.1	21.2
A-CLIP					
plain	41.3	59.3	38.4	35.1	21.3
+SimCLR	42.8	63.6	41.0	36.0	22.6
+SimCLR+BYOL	**43.9**	62.7	**42.1**	38.0	23.2
+SimSiam	43.1	62.5	41.3	37.6	22.6
+SimSiam+BYOL	43.4	**64.1**	41.5	**38.1**	**23.3**

Table 3: Ablation for auxiliary self-supervised tasks.

Methods	IN 1K	Flickr30K		MS COCO	
	0-shot	I2T	T2I	I2T	T2I
CLIP(25ep)	37.6	51.4	32.6	27.9	17.6
SLIP(25ep)	42.8	57.2	41.2	33.6	21.9
A-CLIP(25ep)	**43.9**	**62.7**	**42.1**	**38.0**	**23.2**
CLIP(50ep)	39.4	53.9	35.8	30.2	19.2
SLIP(50ep)	44.1	60.6	41.1	33.2	22.3
A-CLIP(50ep)	**46.3**	**66.7**	**43.2**	**39.8**	**24.4**
CLIP(100ep)	42.7	61.0	37.9	34.4	20.9
SLIP(100ep)	45.0	59.3	41.4	34.6	22.7
A-CLIP(100ep)	**48.0**	**66.3**	**45.7**	**40.7**	**25.1**
CLIP(VIT-L)	40.4	51.4	35.2	28.9	18.5
SLIP(VIT-L)	46.2	60.6	43.7	35.3	23.5
A-CLIP(VIT-L)	**48.9**	**64.1**	**48.2**	**39.1**	**26.9**

Table 4: The results using longer training schedulers and bigger model size.

Methods	tokens per view × k views			
	196 × 1	98 × 2	65 × 3	48 × 4
+attentive mask	37.6	**41.3**	**41.3**	38.9

Table 5: Effects of more attentive mask views. In this table, we tried more views, i.e., 3 and 4 views, while keeping the total image tokens of multiple views to be the same (196 tokens) so as to keep the computation overheads roughly the same. Results evaluated by zero-shot classification accuracy on ImageNet-1K validation set.

Epochs	Methods	Food-101	CIFAR-10	CIFAR-100	CUB	SUN397	Cars	Aircraft	DTD	Pets	Caltech-101	Flowers	MNIST	FER-2013	STL-10	EuroSAT	RESISC45	GTSRB	KITTI	Country211	PCAM	UCF101	Kinetics700	CLEVR	HatefulMemes	SST2	ImageNet	Average
25	CLIP	50.6	66.0	34.5	38.8	51.1	4.0	5.4	21.2	28.5	60.9	53.3	8.4	17.3	90.5	30.2	21.5	6.1	35.1	10.5	53.5	28.5	22.1	10.8	52.4	50.7	37.6	34.2
	SLIP	59.5	78.6	45.2	38.7	53.4	**5.4**	5.7	**26.1**	31.1	71.0	56.6	9.8	19.6	94.4	20.3	28.9	**14.5**	34.0	11.6	**55.4**	37.7	26.9	**17.5**	**52.8**	**51.1**	42.8	38.0
	MaskCLIP	**60.6**	70.1	41.6	**43.3**	54.0	4.9	**8.2**	25.5	**36.8**	68.9	53.6	**11.2**	22.4	93.9	**35.1**	24.8	10.1	30.5	12.5	51.2	37.0	28.1	12.9	**52.8**	50.0	42.7	37.8
	A-CLIP	58.3	**82.8**	**51.0**	43.0	**57.0**	**5.4**	7.6	26.0	32.0	**71.6**	**57.7**	9.8	**29.7**	**95.4**	29.3	**30.3**	13.1	**35.2**	**13.5**	51.6	**38.3**	**29.6**	14.1	**52.8**	49.9	**43.9**	**39.6**
50	CLIP	55.2	77.0	43.8	38.9	49.0	4.7	6.3	23.5	27.2	63.5	56.2	**12.5**	30.2	92.1	21.0	31.9	7.4	33.6	10.9	50.8	35.5	24.8	14.0	49.9	50.1	39.4	36.5
	SLIP	61.9	76.8	48.9	39.2	54.8	7.3	9.0	**29.8**	31.9	**75.0**	57.7	9.8	24.9	**95.6**	37.8	32.5	9.0	**35.1**	12.7	54.4	41.1	30.3	13.8	49.5	49.9	44.1	39.7
	A-CLIP	**62.2**	**81.5**	**53.7**	**48.2**	**58.7**	**8.3**	**10.2**	27.7	**40.5**	73.3	**61.0**	11.3	**32.9**	95.5	**39.7**	**37.5**	**9.4**	23.3	**14.4**	**63.7**	**42.5**	**31.6**	**19.6**	**50.8**	**52.3**	**46.3**	**42.2**
100	CLIP	60.4	79.4	44.6	43.3	53.0	8.5	8.2	26.2	34.7	68.9	59.2	11.4	20.4	93.2	23.3	27.3	10.3	23.1	12.0	54.0	36.7	27.7	13.0	50.9	50.1	42.7	37.8
	SLIP	63.0	83.1	50.4	43.0	52.0	8.3	8.3	26.2	34.0	74.6	61.1	**16.1**	32.4	95.1	22.6	28.5	10.5	**34.8**	11.5	52.1	37.3	28.3	13.7	**55.2**	49.9	45.0	39.9
	A-CLIP	**66.7**	**86.6**	**58.6**	**51.4**	**58.6**	**10.5**	**11.9**	**33.1**	**48.5**	**74.9**	**64.3**	7.8	**31.2**	**96.7**	**35.6**	**35.8**	**12.9**	30.5	**15.7**	**57.1**	**44.1**	**33.1**	**22.9**	52.7	**50.7**	**48.1**	**43.8**

Table 6: Zero-shot evaluation on a variety of classification benchmarks. The *Epochs* indicates the number of training rounds. A-CLIP significantly outperforms other methods at all epochs setting, both in terms of average accuracy and number of winning tracks of above 25 downstream tasks.

task BYOL improves performance in both tasks, and it is complementary to online SSL.

Zero-shot performance on 25 benchmarks We tested the zero-shot classification performance of the proposed A-CLIP approach on a larger set of validation benchmarks, which included 25 classification tasks, following SLIP's [27] evaluation setting. In testing, we also followed SLIP's prompts for evaluation for a fair comparison.

Table 6 reports the results under different training lengths. A-CLIP achieves +1.6%, +2.5%, and +3.9% gains over the counterpart SLIP method regarding the average accuracy of this 25-dataset suite, using 25-epoch, 50-epoch, and 100-epoch training, respectively. Also, note that A-CLIP is 2.3× faster than SLIP. These results further prove the efficacy and efficiency of the proposed A-CLIP.

Comparison with other methods In Table 1, we compare the differences with other methods. Using ViT-B and YFCC-15M dataset, our approach achieves 43.9% top-1 accuracy on ImageNet-1K zero-shot classification, as well as 62.7/42.1 and 38.0/23.2 I2T/T2I retrieval accuracy on Flickr30K and MS COCO, which are +1.1% higher, +5.5/+0.9 higher, and +4.4/+1.3 higher than the SLIP method, respectively, while being 2.30× faster and requiring less GPU memory. A-CLIP-eff, an efficient version that is even 1.16× faster than the plain CLIP model, achieves significant gains of +5.3%, +11.3/+8.0, and +9.5/+4.9 on these benchmarks over that of the plain CLIP model.

4.3. Ablation Study and Analysis

Effects of training longer Table 4 ablates the effects of longer training for different frameworks. Using 2× and 4× training schedulers, i.e., 50 and 100 epochs, the accuracy gaps between the A-CLIP and SLIP models are larger on benchmarks than that using 25-epoch training: +1.1%, **+2.2%**, and **+3.0%** gains ImageNet-1K zero-shot classification, and +4.4/+1.3, **+6.6/+2.1**, and **+6.1/+2.4** on MS COCO [23] I2T/T2I retrieval, by 25, 50, and 100 epochs training, respectively. On Fickr30K [39], longer training also overall benefits more for A-CLIP: +5.5/+0.9, **+6.1/+2.1**, and **+7.0/+4.3** gains on I2T/T2I retrieval by 25, 50, and 100 epochs training, respectively.

We hypothesize that the attentive mask input play as strong augmentation to the input images, which can greatly alleviate the over-fitting issue and thus perform better when training is longer.

Effects of larger model size In Table 4, we have included a comparison with ViT-L/16 on 25 epochs setting. A-CLIP benefits more with larger model size.

Effects of more attentive mask views Our proposed A-CLIP is not limited to two views, and can be extended for multiple views such as 3 and 4. Table 5 shows the results of using different views. In this comparison, we keep the total number of tokens the same for different k views settings. $k = 1$ means the the original CLIP model which keeps all image tokens. For other k, each view selects the most relevant $196/k$ tokens. So the computation overheads for different k are roughly the same.

It can be seen that all $k = 2, 3, 4$ performs significantly better than that of $k = 1$. $k = 2, 3$ both perform the best, and are higher than $k = 4$ (+2.4%).

Use alt-text to select the image patches? We tried using alt-text to select image regions, yielding high training but low evaluation accuracy due to information leakage. Using the text to directly select its corresponding image alters the distance between every positive pairs, rendering the contrastive learning trivial. To address this, we employ EMA attention weights in Eq.(4) as a form of regularization, effectively avoiding direct information leakage.

Figure 4: Visualization of attentive mask. Here we use a ViT-B16 model from A-CLIP's EMA vision encoder of to generate a mask with patch size of 32×32 and 50% mask ratio. Attentive mask can be found to magically preserve the content of text descriptions and filter out redundant backgrounds. The image and text used are sampled from YFCC-100M [32].

Visualization We have done some visualization of the attentive mask. Figure 4 showes the original image and the two views inputted in A-CLIP. It can be found that we basically keep the significant part of the image. The token that is closer to the language semantics makes the input of A-CLIP effective and valuable.

Effects of masked patch size Referring to the conclusion of SimMIM [36], we experimented with different mask patch sizes. As can be seen in Table 2, it obtained a +0.5% ImageNet-1K zero-shot improvement from change mask size to 32. Due to the continuity and redundancy of images, greater granularity of filtering may lead to better performance.

Using all layers works better for attentive token selection Table 2 ablates the use of different layers for attentive token selection. Using all layers performs +0.9% higher on ImageNet-1K zero-shot than that using the last layer.

5. Conclusion

In this paper, we present the attentive mask CLIP framework, or A-CLIP for short, which achieves more efficient and more effective CLIP training by introducing attentive masks for the image branch. Compared to a baseline method using random masking, the attentive mask approach maintains image tokens that are more relevant to the alt-text description and performs much better. The A-CLIP framework is also flexible in incorporating multiple masked views as well as several auxiliary self-supervised tasks, which can further boost efficacy. The attentive mask also serves as good data augmentation, which benefits more from longer training than the original CLIP method.

Due to these advantages, the A-CLIP framework performs much better in both efficiency and efficacy than other CLIP improvements such as SLIP and MaskCLIP. An efficient variant of A-CLIP-eff is even more efficient than the original CLIP method, while being significantly better in terms of zero-shot and multi-modal retrieval accuracy.

References

[1] Mahmoud Assran, Mathilde Caron, Ishan Misra, Piotr Bojanowski, Florian Bordes, Pascal Vincent, Armand Joulin, Michael Rabbat, and Nicolas Ballas. Masked siamese networks for label-efficient learning. 2022. 3

[2] Hangbo Bao, Li Dong, and Furu Wei. Beit: Bert pre-training of image transformers. *Learning*, 2021. 3

[3] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR, 2020. 2, 3, 5

[4] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. *computer vision and pattern recognition*, 2020. 2, 5

[5] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9640–9649, 2021. 6

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, 2019. Association for Computational Linguistics. 4

[7] Xiaoyi Dong, Yinglin Zheng, Jianmin Bao, Ting Zhang, Dongdong Chen, Hao Yang, Ming Zeng, Weiming Zhang, Lu Yuan, Dong Chen, Fang Wen, and Nenghai Yu. Maskclip: Masked self-distillation advances contrastive language-image pretraining. 2022. 1, 2, 3, 7

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. 2, 4, 6

[9] Zi-Yi Dou, Yichong Xu, Zhe Gan, Jianfeng Wang, Shuohang Wang, Lijuan Wang, Chenguang Zhu, Pengchuan Zhang, Lu Yuan, Nanyun Peng, Zicheng Liu, and Michael Zeng. An empirical study of training end-to-end vision-and-language transformers. *arXiv: Computer Vision and Pattern Recognition*, 2021. 2

[10] Li Fei-Fei, Rob Fergus, and Pietro Perona. One-shot learning of object categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2006. 6

[11] Yuting Gao, Jinfeng Liu, Zihan Xu, Jun Zhang, Ke Li, and Chunhua Shen. Pyramidclip: Hierarchical feature alignment for vision-language model pretraining. 2022. 2

[12] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Ávila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent - A new approach to self-supervised learning. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. 2, 4, 5, 6

[13] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. *arXiv: Computer Vision and Pattern Recognition*, 2021. 2, 3, 4

[14] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. Momentum contrast for unsupervised visual representation learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 9726–9735. IEEE, 2020. 3

[15] Zhicheng Huang, Xiaojie Jin, Chengze Lu, Qibin Hou, Ming-Ming Cheng, Dongmei Fu, Xiaohui Shen, and Jiashi Feng. Contrastive masked autoencoders are stronger vision learners. 2022. 2

[16] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 4904–4916. PMLR, 2021. 1, 2

[17] Li Jing, Jiachen Zhu, and Yann LeCun. Masked siamese convnets. 2022. 3

[18] Ioannis Kakogeorgiou, Spyros Gidaris, Bill Psomas, Yannis Avrithis, Andrei Bursuc, Konstantinos Karantzalos, and Nikos Komodakis. What to hide from your students: Attention-guided masked image modeling. 2022. 7

[19] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 5583–5594. PMLR, 2021. 2

[20] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *international conference on machine learning*, 2022. 2

[21] 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. 4

[22] Yanghao Li, Haoqi Fan, Ronghang Hu, Christoph Feichtenhofer, and Kaiming He. Scaling language-image pre-training via masking. 2022. 3

[23] 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. *european conference on computer vision*, 2014. 1, 2, 6, 8

[24] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, et al. Swin transformer v2: Scaling up capacity and resolution. *ArXiv preprint*, abs/2111.09883, 2021. 3

[25] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *ArXiv preprint*, abs/2103.14030, 2021. 2

[26] Shlok Mishra, Joshua Robinson, Huiwen Chang, David Jacobs, Aaron Sarna, Aaron Maschinot, and Dilip Krishnan. A simple, efficient and scalable contrastive masked autoencoder for learning visual representations. 2022. 2, 3

[27] Norman Mu, Alexander Kirillov, David Wagner, Saining Xie, and U Berkeley. Slip: Self-supervision meets language-image pre-training. 2022. 1, 2, 3, 6, 7, 8

[28] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 2021. 1, 2, 3, 6

[29] Yongming Rao, Wenliang Zhao, Benlin Liu, Jiwen Lu, Jie Zhou, and Cho-Jui Hsieh. Dynamicvit: Efficient vision

transformers with dynamic token sparsification. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 2

[30] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 2014. 1, 2, 6, 7

[31] Chenxin Tao, Xizhou Zhu, Gao Huang, Yu Qiao, Xiaogang Wang, Jifeng Dai, and Kong. Siamese image modeling for self-supervised vision representation learning. 2022. 2

[32] 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. 2022. 1, 2, 6, 9

[33] Yulin Wang, Rui Huang, Shiji Song, Zeyi Huang, and Gao Huang. Not all images are worth 16x16 words: Dynamic transformers for efficient image recognition, 2021. 2

[34] Yixuan Wei, Yue Cao, Zheng Zhang, Zhuliang Yao, Zhenda Xie, Han Hu, and Baining Guo. icar: Bridging image classification and image-text alignment for visual recognition, 2022. 2

[35] Zhirong Wu, Zihang Lai, Xiao Sun, and Stephen Lin. Extreme masking for learning instance and distributed visual representations. 2022. 2, 3

[36] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. *arXiv: Computer Vision and Pattern Recognition*, 2021. 3, 9

[37] Jianwei Yang, Chunyuan Li, Pengchuan Zhang, Bin Xiao, Ce Liu, Lu Yuan, and Jianfeng Gao. Unified contrastive learning in image-text-label space, 2022. 2

[38] 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. *Learning*, 2021. 2

[39] 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, 2, 6, 8