COSMOS: Cross-Modality Self-Distillation for Vision Language Pre-training

Supplementary Material

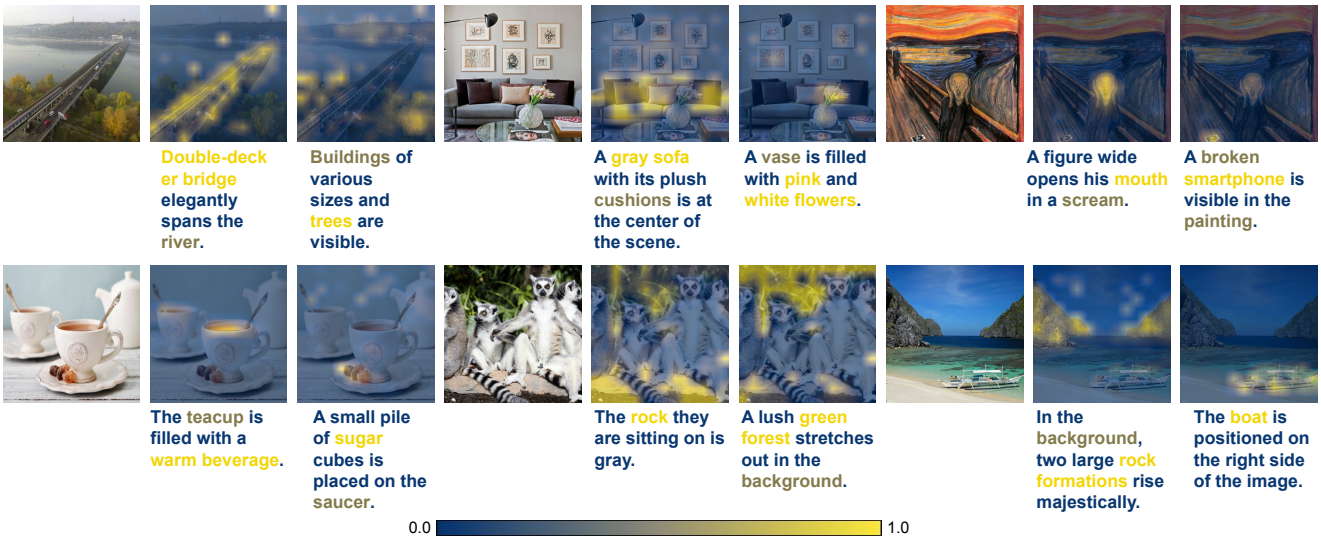

Figure 1. **Visualization of Attention Map.** For different set of captions, we visualize the attention weights of the image and text cross-attention modules. The patch-wise (image) and token-wise (caption) attention weights are both normalized between 0 and 1.

In this supplementary material, we include detailed information about the training procedure of COSMOS. We first provide additional qualitative results in Sec. A. The dataset configurations are described in Sec. B, and the experimental setups are outlined in Sec. C. Additionally, we elaborate on the baseline configurations in Sec. D and present further experiments in Sec. E.

A. Qualitative Results

In Fig. 1, we visualize the attention maps of the cross-attention modules with different captions to illustrate their focus areas. We first normalize the attention weights across all patches and map them back onto the original image. Similarly, tokens in the captions are colored based on the normalized attention weights. The qualitative results show that our cross-attention module enhances the learning of local representations in both modalities, namely image and text. For example, our model is capable of detecting relatively small objects in the image, such as sugar cubes (first example of the second row) or a broken smartphone (third example of the first row). This information is often overlooked by the previous models due to feature suppression.

B. Dataset Configuration

As COSMOS is trained on web datasets, we had to download the images based on the corresponding URLs, but some of the URLs were invalid. Consequently, there is a discrepancy between the number of samples used to train our models and the original dataset sizes. The specific number of samples are reported in Tab. 1. We lost approximately 8.6% (YFCC15M) to 19.4% (CC12M) of samples compared to the original datasets due to expired URLs. This detail should be considered when directly comparing our models with others.

Dataset	Original	Ours	Difference	Percentage
CC3M [44]	3,318,333	2,823,019	500,229	85.1%
CC12M [5]	12,423,374	10,010,225	2,413,149	80.6%
YFCC15M [9]	15,388,848	14,065,827	1,323,021	91.4%
Merged-30M	31,130,555	26,899,071	4,236,399	86.4%
PixelProse [46]	16,896,423	15,037,386	1,859,037	89.0%

Table 1. **Size of the pre-training datasets.** We compare the original dataset sizes and the actual number of samples used to train our models. We also report the percentage of samples that we succeeded to retrieve from the original image URL links.

Additionally, COSMOS is evaluated on SugarCrepe [19], SVO [18], and MMVP [18], to assess the robustness of its multi-modal representations.

The **SugarCrepe** dataset [19] consists of images with positive and negative captions, where the model is required to choose the positive one given an input image (50% is the random chance). Negative captions are slightly different from the positive captions in terms of attributes and objects, with the goal of confusing the model. Since SugarCrepe

is based on the COCO [27] validation set, we were able to download all images without any data loss.

SVO [18] consists of 36,841 data pairs, where each caption is paired with two images (one positive and one negative). Each image has its own triplet (subject, verb, object), and the positive and negative images differ in one component of the triplets. The original task is to match the caption with one of the two images based on the similarity scores given by VLMs (50% by random chance). Since naive CLIP [40] already achieves 80% accuracy, we increased the difficulty of this task by constructing negative captions. Based on the triplet of positive and negative images, we replaced the subject, verb, or object in the positive caption with the negative one. Currently, the model not only has to match the positive image with the positive caption but also match the negative image with the negative caption (25% by random chance). As SVO is released with image URL links, we were able to collect 22,220 data pairs (60.3% of the original size).

MMVP-VLM [18] contains 15 image-text pairs across 9 categories, resulting in a total of 135 image-text pairs. It categorizes 9 challenging visual patterns that most VLMs struggle with. Each data point consists of two images and two captions, requiring the VLMs to match the correct image and caption respectively (25% by random chance). Since Hendricks and Nematzadeh [18] released their dataset including images and captions, we were able to fully download the MMVP-VLM dataset as originally intended.

C. Experiment Configuration

In this section, we detail the configuration of our experiments, including hyperparameters for training (Sec. C.1), the inference process for each downstream task (Sec. C.2), and the pseudocode of our training objective (Sec. C.3).

C.1. Hyperparameters

As shown in Tab. 2, we primarily follow the hyperparameters described in DreamLIP [58], with the exception of batch size. DreamLIP utilizes a batch size of 1,024 for CC3M and 8,192 for other datasets. Due to computational constraints, we adopt a batch size of 1,024 for CC3M and 4,096 for other datasets to train COSMOS. For reproducing CLIP [40] and SigLIP [57] with our setting, we use a batch size of 1,024 for CC3M and 6,144 for other datasets to establish a strong baseline.

As mentioned in the main paper, the teacher model is updated at each iteration using the exponential moving average (EMA) of the student model. To obtain an effective teacher for self-distillation, we need to determine the momentum parameter λ, which controls the update rule of the teacher parameter θ_t based on the student parameter θ_s (i.e., $\theta_t = \lambda\theta_t + (1-\lambda)\theta_s$). According to Caron et al. [3], a higher batch size requires a lower momentum parameter.

Config	Value
Optimizer	AdamW [31]
Learning rate	5×10^{-4}
Weight decay	0.5
Adam β	$\beta_1, \beta_2 = (0.9, 0.98)$
Adam ϵ	1×10^{-8}
Total epochs	32
Warm up iterations	$2,000$
Learning rate schedule	cosine decay

Table 2. **Hyperparameter configuration.**

Based on their configuration, we choose 0.999 for CC3M and 0.99 for other datasets by default. We did not explicitly perform a grid search for this parameter. Empirically, we found that fixing the momentum parameter obtains a better performance, unlike Caron et al. [3], who used a cosine scheduler for the momentum parameter which eventually converges to 1.0.

C.2. Inference

As COSMOS is applied to various downstream tasks, it is crucial to establish a consistent evaluation protocol for each task. Notably, the cross-attention module is not involved in the inference process. Therefore, we are able to evaluate COSMOS as a standard CLIP-based model, utilizing the class token [cls] and end-of-text token [eot] as image and text embeddings, respectively.

For **zero-shot classification**, we follow the process established by CLIP [40]. First, we construct prompts using the class label names for each dataset, as referenced in ALIP [53]. For each class, text embeddings are generated by the text encoder using different prompts, which are then averaged to obtain the final text embedding. Given an input image, the image encoder extracts the image embedding and calculates the cosine similarity scores between the image embedding and the text embeddings of all predefined classes in each dataset. The class label with the highest score is selected as the prediction.

Zero-shot retrieval is based on the image and text embeddings extracted from the pre-trained models, following the evaluation scheme presented in [53]. For each image-text pair, the pre-trained model generates image and text embeddings, collecting all embeddings for the entire dataset. We then compute a full cosine similarity matrix between all image embeddings and all text embeddings. Captions with the highest similarity scores are selected for each image to calculate the R@1, R@5, and R@10 metrics for image-to-text retrieval. Similarly, for text-to-image retrieval, we choose images with the highest scores for each caption. As each image in the validation or test set is equipped with multiple captions, image-to-text retrieval

scores are generally higher than text-to-image scores.

In **zero-shot semantic segmentation**, we exclude the background category for PASCAL VOC [13] and PASCAL Context [34], following Cha et al. [4] and Wang et al. [50]. To be specific, we denote the original datasets with a background class as VOC21 and Context60, while the variants without the background category are referred to as VOC20 and Context59 in the main table. At inference time, for a given set of classes in the datasets, we obtain the corresponding text embeddings by querying our text encoder with a standard prompt. We compute the cosine similarity between the image patch embeddings (image tokens) and the text features of each class name to generate a segmentation map in a zero-shot manner. We adhere to the evaluation protocol established by SCLIP [50], including specifications for the window size, stride, and other parameters. We believe that the raw segmentation output of a VLM accurately reflects its zero-shot performance; therefore, we do not fine-tune our model or apply any post-refinement techniques such as PAMR [1].

To **evaluate on SugarCrepe [19], SVO [18], and MMVP-VLM [48]**, we primarily referred to their evaluation demo prompts. Image and text embeddings are extracted from the pre-trained models, and then image-text pairs with higher cosine similarity scores are chosen as the final decision.

C.3. Pseudocode

To increase the clarity of our method, we present the pseudocode of the training objective in Algorithm 1. We empirically found that incorporating local image crops on the CLIP loss diminishes zero-shot performance, whereas using local text crops enhances it. Consequently, we compute the CLIP loss between all text crops and global image crops. We infer that integrating diverse captions during training allows the model to learn various objects shown in the images, while including local image crops during training likely leads to the misalignment of image-text pairs.

Algorithm 1 COSMOS: Pseudocode of our loss function

```
# img_g, img_l: Global&local crop of image
# txt_g, txt_l: Global&local crop of text
# Is, Ts: Student image&text encoder
# It, Tt: Teacher image&text encoder

# Generate embeddings of size [batch, seq_len, dim]
s_img_g, s_img_l = Is(img_g, img_l) # Student image
s_txt_g, s_txt_l = Ts(txt_g, txt_l) # Student text
t_img_g = It(img_g) # Teacher image
t_txt_g = Tt(txt_g) # Teacher text

# Split into ([CLS],img_tok) or ([EOT], txt_tok)
s_cls_g, s_img_tok_g = s_img_g
s_eot_g, s_txt_tok_g = s_txt_g
s_cls_l, _ = s_img_l
s_eot_l, _ = s_txt_l
t_cls_g, _ = t_img_g
t_eot_g, _ = t_txt_g

# Calculate CLIP loss
clip_loss = (sym_nce(s_cls_g, s_eot_g) +
             sym_nce(s_cls_g, s_eot_l))/2


# Generate cross-modality embeddings
s_cls = {s_cls_g, s_cls_l} # Combine student crops
s_eot = {s_eot_g, s_eot_l} # Combine student crops
h_img = s_cls + cross-attn(q=s_cls, kv=s_txt_tok_g)
h_txt = s_eot + cross-attn(q=s_eot, kv=s_img_tok_g)

# Calculate cross-modality self-distillation loss
cosmos_loss = (sym_nce(h_img, t_cls_g) +
               sym_nce(h_img, t_eot_g) +
               sym_nce(h_txt, t_cls_g) +
               sym_nce(h_txt, t_eot_g))/4

final_loss = clip_loss + cosmos_loss
```

Notes: We assume one global view and one local view for simplicity. `sym_nce` represents the symmetric InfoNCE loss [38]. `cross-attn` denotes the cross attention module which requires key, value (kv) and query (q) as inputs.

D. Baseline Configuration

In the main tables, we evaluate both the pre-trained weights from the official code repository and our reproduction. For DreamLIP [58], we used the official pre-trained weights with ViT/B-16 as the vision encoder. Since they did not provide weights for the ViT/B-32 vision encoder, we referenced the results from their table as shown in Tab. 5 and Tab. 6. For OpenCLIP [7], we utilized the models described in Tab. 3. Specifically, they trained their CLIP models on LAION-400M [42] until 12.8 billion examples were seen, using a batch size of 33,792 for both ViT-B/16 and ViT-B/32. For DataComp-1B [16], the models were trained until 12.8 billion examples were seen with a batch size of 90,112. Additionally, the models were trained on LAION-2B [43] with a batch size of 88,064 (ViT-B/16) or 79,104 (ViT-B/32) until 34 billion examples were seen.

Architecture	Data	Model Name
ViT-B/16	LAION-400M	laion400m_e32
	DataComp-1B	datacomp_xl_s13b_b90k
	LAION-2B	laion2b_s34b_b88k
ViT-B/32	LAION-400M	laion400m_e32
	DataComp-1B	datacomp_xl_s13b_b90k
	LAION-2B	laion2b_s34b_b79k

Table 3. **OpenCLIP [7] model names.**

E. Extra Experiments

E.1. Comparison to Previous SSL Methods

Previous works, such as SLIP [35] or SILC [36], aim to enhance CLIP through self-supervision by explicitly optimizing local-to-global correspondences. These methods primarily focus on the image encoder. Fig. 2 illustrates how our approach differs from the self-supervised contrastive language-image pre-training methods, SLIP [35]

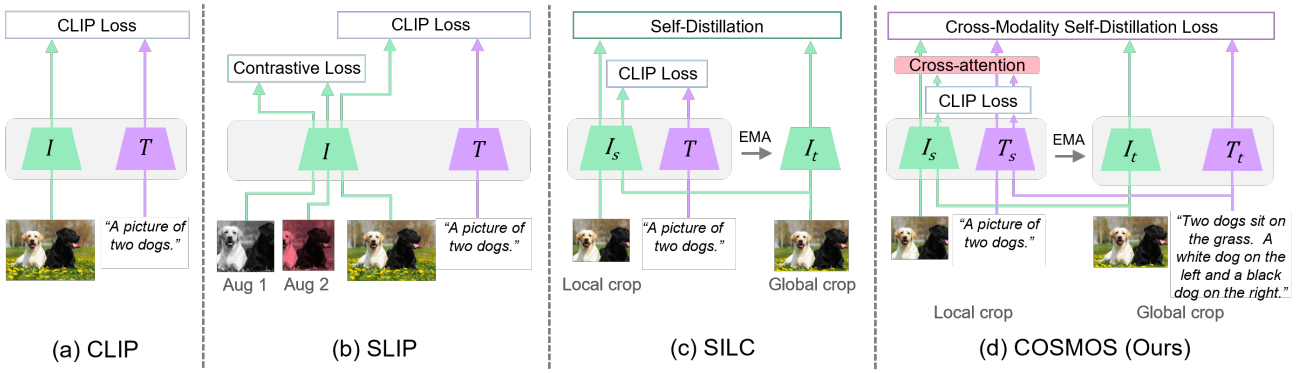

Figure 2. **Illustration of CLIP with self-supervised approaches.** I and T denote the image and text encoders, respectively. I_t (or T_t) and I_s (or T_s) represent the teacher and student image (or text) encoders, where the teacher is an exponential moving average (EMA) of the student. (a) CLIP [40]: image and text embeddings are aligned during training. (b) SLIP [35]: contrastive loss is computed on sets of two different augmentations. (c) SILC [36]: self-distillation loss is obtained between local and global crops of the same image. (d) COSMOS: the cross-attention module is utilized to generate cross-modal representations which are optimized through the cross-modality self-distillation loss. We also design global and local crops of image and text pairs for effective self-supervised learning in VLMs.

Method	ImageNet	MSCOCO		Flickr30k	
	Top-1	I2T@1	T2I@1	I2T@1	T2I@1
(a) CLIP [40]	23.9	40.2	27.2	68.4	52.1
(b) SLIP [35]	26.6	44.4	30.5	75.8	58.7
(c) SILC [36]	30.4	48.6	35.4	79.0	62.1
(d) COSMOS	**37.1**	**53.1**	**40.1**	**84.1**	**68.6**

Table 4. **Comparison to methods described in Fig. 2.** All models are trained on CC3M with long synthetic caption. We use the batch size of 1,024 and ViT-B/16 image encoder.

and SILC [36].

In Tab. 4, we directly compare COSMOS with the previous methods depicted in Fig. 2. For a fair comparison, we re-implemented these methods based on the OpenCLIP [7] code repository and trained them on the CC3M dataset with long captions provided by DreamLIP [58]. The projection layer parameters follow the SLIP configuration, while the optimal temperature and loss scale for SILC were carefully determined. After a grid search, we selected student and teacher temperatures of 0.1 and 0.02, respectively, and used loss scale hyperparameters of (1.9, 0.1) for CLIP loss and self-distillation loss (see Tab. 15). Without extensive hyperparameter tuning, COSMOS significantly outperforms previous methods obtaining an accuracy of 37.1% in zero-shot classification on ImageNet [10], and 53.1% R@1 score in image-to-text retrieval on MSCOCO [27]. As mentioned in the introduction, focusing solely on enhancing image representation leads to sub-optimal results in VLMs, as shown by SILC which reached an accuracy score of only 30.4% on ImageNet and an R@1 score of 48.6% on image-to-text retrieval on MSCOCO, highlighting the importance of self-distillation with cross-modality representations.

E.2. Experiments with ViT-B/32 Architecture

In addition to the experiments with ViT-B/16 presented in the main paper, we conducted the same experiments using the ViT-B/32 vision encoder architecture, as shown in Tab. 5 and Tab. 6. Overall, the improvements achieved with COSMOS are consistent with those observed in the main table using ViT-B/16.

In the zero-shot retrieval tasks (Tab. 5), COSMOS not only surpasses previous strong baselines (CLIP [40], SigLIP [57], and DreamLIP [58]) trained on the same datasets, but also exceeds the performance of OpenCLIP [7] models trained on much larger datasets. Notably, COSMOS trained on Merged-30M achieves 64.3% and 48.4% in image-to-text and text-to-image R@1 scores on MSCOCO, significantly outperforming DreamLIP (58.3% and 41.1%). Furthermore, COSMOS trained on CC12M already demonstrates higher accuracy compared to OpenCLIP trained on LAION-2B dataset. These results highlight the effectiveness of COSMOS in generating more fine-grained and comprehensive multi-modal representations through cross-modality self-distillation.

In the zero-shot classification tasks (Tab. 6), COSMOS improves all metrics compared to CLIP [40] and SigLIP [57] across most datasets, while achieving comparable results to DreamLIP [58] on YFCC15M and Merged-30M. DreamLIP [58] aligns visual patch embeddings with positive text embeddings, promoting the learning of global structure by image tokens, which enhances the performance of classification task where the model focuses on global information. Although COSMOS optimizes local-to-global correspondence to enhance the representation of local information through self-distillation, its performance is similar to DreamLIP. Additionally, increasing the dataset size

Data	Method	Flickr30K						MSCOCO					
		Image-to-text			Text-to-image			Image-to-text			Text-to-image		
		R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
CC3M	CLIP [40]	53.4	77.1	85.0	38.4	65.4	74.6	31.3	55.3	66.9	19.8	42.7	54.8
	SigLIP [57]	56.4	78.9	85.7	40.2	66.5	76.0	30.8	55.5	66.8	20.6	43.9	55.4
	DreamLIP* [58]	57.6	84.4	89.6	42.2	69.0	77.7	33.4	60.7	72.0	23.4	47.2	58.6
	COSMOS	**74.3**	**92.3**	**95.9**	**59.2**	**82.6**	**89.1**	**47.6**	**73.1**	**82.0**	**33.5**	**59.7**	**70.6**
CC12M	CLIP [40]	76.8	92.9	95.8	60.3	83.4	89.4	50.6	76.5	84.3	34.9	61.5	72.3
	SigLIP [57]	75.8	92.7	96.1	60.0	83.1	89.7	51.2	76.5	84.7	35.4	61.9	72.4
	DreamLIP* [58]	78.7	94.6	97.6	61.0	83.9	89.8	53.4	77.1	84.7	36.7	62.3	72.3
	COSMOS	**86.5**	**97.5**	**98.7**	**69.8**	**89.3**	**94.1**	**59.6**	**82.3**	**89.4**	**43.0**	**69.5**	**78.9**
YFCC15M	CLIP [40]	84.6	97.1	98.9	66.0	87.8	92.7	56.7	81.6	88.6	40.1	66.8	76.6
	SigLIP [57]	82.5	97.4	98.5	66.9	87.8	92.3	56.3	82.0	89.0	40.0	66.7	76.9
	DreamLIP* [58]	84.9	97.3	98.7	66.0	86.4	91.4	55.7	80.5	88.2	39.8	66.0	75.5
	COSMOS	**90.2**	**98.7**	**99.4**	**73.3**	**91.6**	**95.4**	**64.5**	**86.1**	**91.8**	**46.0**	**72.2**	**81.0**
Merged-30M	CLIP [40]	85.6	96.7	99.0	69.5	89.9	94.3	59.3	83.1	89.9	42.8	69.3	79.0
	SigLIP [57]	88.4	97.7	99.1	70.7	90.3	94.6	59.5	83.3	90.1	43.8	70.1	79.6
	DreamLIP* [58]	87.2	97.5	98.8	66.4	88.3	93.3	58.3	81.6	88.8	41.1	67.0	76.6
	COSMOS	**89.9**	**98.8**	**99.3**	**76.1**	**92.8**	**96.2**	**64.3**	**86.5**	**92.0**	**48.4**	**74.2**	**82.6**
LAION-400M	OpenCLIP† [7]	79.7	95.0	97.6	60.9	84.8	90.7	51.7	76.6	84.9	33.7	59.4	69.9
DataComp-1B	OpenCLIP† [7]	80.1	94.6	97.2	62.9	85.4	91.1	54.6	78.4	85.8	36.3	62.1	72.6
LAION-2B	OpenCLIP† [7]	85.4	96.2	98.2	68.4	89.0	93.4	56.6	80.3	87.4	38.8	64.8	74.7

Table 5. **Zero-shot image-text retrieval results** in terms of R@1, R@5, and R@10 on the Flickr30K [54] and MSCOCO [27] datasets. The vision encoder architecture is ViT-B/32. The best results are highlighted in **bold**. Results are reproduced with our setup for fair comparison unless otherwise marked. *: Results copied from their work. †: Results obtained using their official pre-trained weights.

generally improves accuracy, suggesting that the total number of images is a crucial factor for achieving high performance in classification. The results obtained with ViT-B/32 are consistent with those of ViT-B/16 from the main paper.

E.3. Experiment on PixelProse Dataset

To demonstrate the adaptability of COSMOS to various captions, we employ PixelProse [46], a synthetic caption dataset similar to DreamLIP. PixelProse filters and combines data from CommonPool [16], CC12M [44], and RedCaps [11] to create over 16 million image and alt-text pairs. They used Gemini 1.0 Pro [46] to generate new captions. As shown in Tab. 7 and Tab. 8, we trained CLIP [40], SigLIP [57], and COSMOS on PixelProse using the same setup as before. The actual number of image-text pairs used to train these models is detailed in Tab. 1.

In Tab. 7, COSMOS consistently outperforms the CLIP and SigLIP models by a significant margin with both ViT-B/16 and ViT-B/32 vision encoders. For instance, COSMOS with ViT-B/16 achieves 62.4% and 43.4% in R@1 scores for image-to-text and text-to-image retrieval on MSCOCO. This performance even surpasses strong baselines from OpenCLIP [7], which achieve 56.5% and 37.9% on LAION-400M, 58.2% and 39.8% on DataComp-1B, and 59.3% and 41.7% on LAION-2B (see Tab. 1 main).

Similarly, Tab. 8 shows that COSMOS outperforms previous baselines in the zero-shot classification tasks. Notably, COSMOS trained on PixelProse achieves the best average accuracy of 60.7% with ViT-B/16 and 58.0% with ViT-B/32 among all models trained on the synthetic caption datasets (e.g., CC3M, CC12M, YFCC15M, and Merged-30M). However, the same models trained on PixelProse do not achieve the best accuracy in retrieval tasks according to Tab. 1 main and Tab. 5. We infer that this might be due to the length and content of the captions used during the training. DreamLIP utilizes three MLLMs to generate long synthetic captions, resulting in longer captions than those generated by PixelProse, which only uses Gemini Pro. Consequently, models trained on longer and more varied captions are better at capturing local information (i.e., better at retrieval task), while models trained on relatively shorter and similar captions are better at focusing on global objects in the image (i.e., better at classification task). Investigating this trade-off based on the nature of these long captions would be an intriguing direction for future work.

E.4. Ablation Study on Training Efficiency

In Tab. 9, we report the GPU memory requirements and training time based on the number of global and local crops, along with the corresponding zero-shot performance. Compared to CLIP, COSMOS requires 10-20% more training time and GPU memory due to its teacher-student setup (row 2 vs row 4, and row 3 vs row 5). As no gradients flow into the teacher during training, COSMOS does not add too much computational overhead while outperforming CLIP.

For COSMOS, we fixed the number of global crops for

Data	Model	Food-101	CIFAR-10	CIFAR-100	SUN397	Cars	Aircraft	DTD	Pets	Caltech-101	Flowers	ImageNet	Average
CC3M	CLIP [40]	14.2	73.5	37.8	39.3	2.4	1.4	14.4	16.9	66.4	10.7	19.7	27.0
	SigLIP [57]	13.5	74.2	40.1	39.8	2.8	1.3	12.9	16.3	67.3	11.3	21.2	27.3
	DreamLIP* [58]	16.1	**82.0**	45.4	41.3	2.5	1.0	13.9	18.8	64.4	14.1	25.9	29.6
	COSMOS	**24.9**	80.0	**51.9**	**51.5**	**3.8**	**1.5**	**25.2**	**29.9**	**77.2**	**21.9**	**33.0**	**36.4**
CC12M	CLIP [40]	41.6	88.5	57.5	54.5	11.3	2.3	23.4	43.3	80.4	15.7	37.4	41.4
	SigLIP [57]	41.9	87.5	59.3	55.2	11.8	1.6	27.3	41.8	80.9	18.4	37.8	42.1
	DreamLIP* [58]	48.9	86.4	63.0	55.7	17.9	1.9	23.5	41.9	83.2	25.8	44.2	44.8
	COSMOS	**52.9**	**91.2**	**67.5**	**61.0**	**23.8**	**3.7**	**32.1**	**54.2**	**85.5**	**30.6**	**46.7**	**49.9**
YFCC15M	CLIP [40]	38.9	86.2	58.2	53.3	7.1	4.0	23.9	27.6	76.8	38.0	38.9	41.2
	SigLIP [57]	37.7	86.1	57.1	53.2	6.4	4.3	25.3	30.4	77.4	35.3	38.6	41.1
	DreamLIP* [58]	**51.7**	**87.9**	**60.7**	**54.8**	9.4	**7.1**	26.8	36.3	79.6	**48.6**	46.6	**46.3**
	COSMOS	40.3	84.9	57.0	54.6	**13.5**	5.9	**31.3**	**38.6**	**82.1**	47.8	**48.1**	45.8
Merged-30M	CLIP [40]	54.8	90.0	67.1	62.0	13.0	3.6	27.6	49.4	83.8	41.4	45.6	48.9
	SigLIP [57]	52.8	90.8	66.1	63.4	15.0	5.4	29.9	47.8	84.4	35.7	46.5	48.9
	DreamLIP* [58]	**68.2**	**91.8**	69.2	62.2	20.7	**8.0**	32.1	**62.8**	**86.1**	48.5	**55.7**	55.0
	COSMOS	65.9	91.5	**70.8**	**64.6**	**23.4**	7.6	**37.6**	57.3	**86.1**	**52.2**	53.4	**55.5**
LAION-400M	OpenCLIP† [7]	78.2	88.4	68.3	65.0	74.5	14.6	52.4	84.9	88.4	65.9	60.2	67.3
DataComp-1B	OpenCLIP† [7]	86.3	95.6	80.4	67.3	87.3	24.8	57.2	90.2	91.6	73.2	69.2	74.8
LAION-2B	OpenCLIP† [7]	82.7	93.6	75.8	68.7	86.1	24.5	55.8	90.4	90.5	71.6	66.5	73.3
2.5B	MetaCLIP* [52]	82.7	95.2	77.7	66.8	77.4	27.0	58.9	90.9	92.8	69.9	67.6	73.4
	Llip* [24]	84.1	95.5	80.8	68.6	82.2	34.9	58.8	92.3	92.9	74.8	67.5	75.6

Table 6. **Zero-shot classification results** in terms of top-1 accuracy on the ImageNet [10] and Food101 [2], CIFAR-10 [22], CIFAR-100 [22], SUN397 [51], Stanford Cars [21], FGVCAircraft [33], DTD [8], Oxford Pets [39], Caltech101 [15], Flowers102 [37], and ImageNet [10] datasets. The vision encoder architecture is ViT-B/32. The best results are highlighted in **bold**. Results are reproduced with our setup for fair comparison unless otherwise marked. *: Results copied from their work. †: Results obtained using their official pre-trained weights.

Method	Flickr30K						MSCOCO					
	Image-to-text			Text-to-image			Image-to-text			Text-to-image		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
Model Architecture: ViT-B/32												
CLIP [40]	76.1	92.1	95.7	57.5	81.8	88.6	49.4	73.7	82.2	32.0	57.1	68.2
SigLIP [57]	74.4	92.2	96.1	56.8	81.3	87.7	48.9	74.4	82.2	32.0	57.2	68.0
COSMOS	**85.6**	**96.9**	**98.3**	**66.3**	**87.6**	**92.7**	**57.2**	**80.9**	**88.0**	**38.9**	**64.8**	**74.6**
Model Architecture: ViT-B/16												
CLIP [40]	85.2	96.2	98.6	66.3	87.8	93.2	56.9	79.7	87.4	38.0	64.1	74.5
SigLIP [57]	85.4	96.7	97.8	66.6	88.4	93.5	57.5	80.5	87.6	38.9	64.5	74.4
COSMOS	**89.9**	**98.5**	**99.5**	**73.6**	**92.0**	**95.4**	**62.4**	**84.0**	**89.7**	**43.4**	**69.3**	**78.8**

Table 7. **PixelProse: Zero-shot image-text retrieval results** in terms of R@1, R@5, and R@10 on the Flickr30K [54] and MSCOCO [27] datasets. The best results are highlighted in **bold**. Results are reproduced with our setup for fair comparison.

both image and text to two, while varying the number of local crops included during training (row 4-row 7). As a result, increasing the number of local crops improves zero-shot classification and retrieval tasks, achieving 37.1% accuracy on ImageNet and 53.1% and 40.1% R@1 scores on image-to-text and text-to-image retrieval on the MSCOCO validation set with six local crops. The improvement is most significant between zero and two local crops (row 4 vs row 5), with diminishing returns as more local crops are added. GPU memory usage and training time also increase with the number of local crops. Therefore, one could determine the optimal number of local crops based on the available computational resources.

E.5. Experiment on the Winoground Dataset

In addition to the visual perception and contextual understanding tasks presented in the main paper, we also evaluate COSMOS on the Winoground dataset [47]. Each entry in

Model	Food-101	CIFAR-10	CIFAR-100	SUN397	Cars	Aircraft	DTD	Pets	Caltech-101	Flowers	ImageNet	Average
Model Architecture: ViT-B/32												
CLIP [40]	52.4	85.2	56.1	54.1	26.0	1.9	32.0	59.0	85.0	34.4	44.9	48.3
SigLIP [57]	53.8	85.7	59.4	55.6	25.2	3.1	32.1	64.1	84.8	37.1	44.6	49.6
COSMOS	**66.3**	**89.9**	**67.4**	**60.4**	**47.0**	**4.4**	**37.3**	**76.6**	**89.5**	**45.1**	**54.3**	**58.0**
Model Architecture: ViT-B/16												
CLIP [40]	63.2	87.9	60.2	59.1	37.0	2.3	40.1	68.8	86.6	42.1	50.8	54.4
SigLIP [57]	63.4	87.5	58.6	60.5	36.2	3.5	37.8	63.5	87.0	43.4	51.1	53.9
COSMOS	**72.1**	**92.1**	**71.9**	**60.8**	**53.0**	**3.6**	**43.4**	**74.4**	**90.1**	**46.8**	**59.6**	**60.7**

Table 8. **PixelProse: Zero-shot classification results** in terms of top-1 accuracy on ImageNet [10] and 10 common downstream datasets. The best results are highlighted in **bold**. Results are reproduced with our setup for fair comparison.

	Method	ImageNet	MSCOCO		Flickr30k		Training	
		Top-1	I2T@1	T2I@1	I2T@1	T2I@1	Time	Mem.
1	CLIP: 1 globals + 0 locals	24.5	38.9	26.8	69.1	51.2	3.8h	10.5G
2	CLIP: 2 globals + 0 locals	29.7	47.6	33.8	78.1	60.7	12.2h	16.3G
3	CLIP: 2 globals + 2 locals	31.3	49.6	35.9	78.4	62.1	17.6h	20.6G
4	COSMOS: 2 globals + 0 locals	31.0	50.1	35.4	80.9	63.2	14.6h	18.2G
5	COSMOS: 2 globals + 2 locals	34.8	51.5	38.5	82.7	65.9	21.4h	23.4G
6	COSMOS: 2 globals + 4 locals	36.3	52.3	39.6	83.5	67.4	26.5h	28.2G
7	COSMOS: 2 globals + 6 locals	37.1	53.1	40.1	84.1	68.6	31.7h	32.6G

Table 9. **Ablation on the training efficiency of COSMOS.** Total running time and peak memory usage per GPU are shown with different configuration of global and locap crops. We also report zero-shot classification results on ImageNet [10] and zero-shot retrieval results on Flickr30K [54] and MSCOCO [27]. We train our model (ViT-B/16) on four 4-GPU machines using the CC3M dataset with the batch size of 1,024.

Method	Data Size	Text	Image	Group
CLIP [40]	30M	29.3	13.0	8.5
SigLIP [57]	30M	29.0	11.0	8.3
DreamLIP [58]	30M	27.8	15.8	**11.5**
COSMOS	30M	**30.8**	**16.5**	11.3
OpenCLIP [7]	400M	25.5	11.5	7.8
	1B	29.8	8.8	7.3
	2B	28.0	10.8	8.3

Table 10. **Evaluation on Winoground [47].**

the dataset consists of two images and two captions, where the task is to correctly match the image-text pairs (25% by random chance). Both captions contain an identical set of words, but in different orders, assessing whether the model possesses sufficient compositional understanding for challenging image-text pairs.

In Tab. 10, we compare COSMOS with three baselines (CLIP [40], SigLIP [57], and DreamLIP [58]) as well as OpenCLIP [7] models trained on larger datasets. The results of the OpenCLIP models indicate that scaling up the size of pre-training datasets does not necessarily improve the performance. In other words, large datasets alone cannot ensure high compositional knowledge in VLMs. However, COSMOS outperforms all OpenCLIP models, suggesting that our self-distillation and cross-modality learning effectively enhance the model's contextual understanding. COSMOS achieves 30.8% and 16.5% in text and image scores, respectively, surpassing the previous best baselines of 29.3% text score from CLIP and 15.8% image score from DreamLIP, while achieving comparable results in terms of group score.

E.6. Experiment on MLLM setting

In order to evaluate COSMOS on MLLM setting, we adapt our vision encoder (ViT-B/16) to LLaVA framework [28]. We follow the training process of LLaVA v1.5 which consists of feature alignment pre-training and visual instruction tuning. Various benchmarks are selected for comprehensive evaluation including ScienceQA [32], POPE [26],

Method	Data Size	ScienceQA	POPE	GQA	TextVQA	MMMU
CLIP [40]	30M	65.2	80.1	58.7	53.5	33.9
OpenCLIP [7]	400M	67.2	81.3	59.8	54.4	36.5
COSMOS	30M	**67.8**	**83.2**	**60.4**	**55.3**	**36.8**

Table 11. **LLaVA v1.5 experiment**. CLIP and COSMOS trained on Merged-30M and OpenCLIP trained on LAION-400M.

GQA [20], TextVQA [45], and MMMU [56]. In Tab. 11, COSMOS outperforms both CLIP trained on the same data (Merged-30M) and OpenCLIP trained on LAION-400M, which demonstrates the effectiveness of COSMOS on visual reasoning and compositional question answering.

E.7. Ablation Study on Text Cropping Strategy

Method	ImageNet	MSCOCO		Flickr30K	
	Top-1	I2T@1	T2I@1	I2T@1	T2I@1
Masked text [25]	28.5	46.0	32.8	76.2	59.2
Summarized text [17]	33.9	51.7	37.8	81.0	65.3
Local within global	34.6	52.1	37.8	81.2	66.1
COSMOS	**35.1**	**52.6**	**38.9**	**83.0**	**66.5**

Table 12. **Ablation on text cropping strategies.** We compare various text cropping methods in terms of zero-shot classification results on ImageNet [10] and zero-shot retrieval results on Flickr30K [54] and MSCOCO [27]. All models are trained on CC3M with batch size 1,024 and ViT-B/16 image encoder.

In our paper, both global and local crops of captions are randomly sampled from long synthetic captions. Typically, global texts (1-5 sentences) are longer than local texts (1 sentence). The sampling processes of global and local crops are entirely independent. Therefore, global captions may or may not include local captions, similar to the image crop method in SimCLR [6] and DINO [3]. While mismatches can occur between global and local captions via random sampling, as with the global and local crops of images, the model can learn conceptual correspondences. For example, for an image of a park, the global text may describe the park, while the local text describes a dog. The model then learns that "park" and "dog" are related in the text.

To validate our text cropping method, we compare various cropping strategies in Tab. 12, referring to previous works. In the *masked text* setting, we randomly select one to five sentences from a long caption and set it as the global caption. Then, 15% of text tokens are replaced with the [mask] token, which is used as the local caption. This setting is similar to text self-supervised learning in DeCLIP [25]. In the *summarized text* setting, we sample one summary sentence as the global caption and one detailed sentence as the local caption, similar to PyramidCLIP [17]. As the synthetic captions of DreamLIP [58] already distinguish between short and long captions, which generally describe the summary and details respectively, we directly use their categories to construct global and local crops. In the *local within global* setting, we ensure that local captions are always included in global captions, while the rest remains the same as our text cropping method. The results show that our text-cropping strategy, inspired by image cropping, performs the best, as it learns various conceptual similarities via independent random sampling.

E.8. Additional Results on Semantic Segmentation

Method	VOC20	City.	Context59	ADE	COCO-Stf.
CLIP [40]	11.3	5.0	4.5	1.3	2.8
SigLIP [57]	14.5	5.5	5.8	2.2	3.8
DreamLIP [58]	1.8	0.9	0.4	0.1	0.1
COSMOS	**53.6**	**13.9**	**15.7**	**8.5**	**10.7**

Table 13. **Zero-shot semantic segmentation results** in terms of mean Intersection over Union (mIoU). The vision encoder architecture is ViT-B/16 and all models are trained on Merged-30M.

In addition to Table 3 in main paper, we compare segmentation performance of COSMOS to CLIP [40], SigLIP [57], and DreamLIP [58] in Tab. 13. Surprisingly, DreamLIP performs poorly in every semantic segmentation benchmarks, likely because its loss function weakens local image representation by matching local visual patches with global text. Ours is by far the best performing method, likely due to its cross-modality embedding and local-to-global matching, alleviating local feature suppression.

E.9. Additional SOTA Comparison

Method	Data Size	Batch Size	MSCOCO		Flickr30K	
			I2T	T2I	I2T	T2I
VeCLIP [23]	300M	32k	67.8	48.9	91.2	76.3
MobileCLIP-B [49]	1B	65k	**68.8**	50.6	91.4	77.3
COSMOS	30M	4k	68.0	**52.5**	**92.9**	**80.3**

Table 14. **Comparison to VeCLIP [23] and MobileCLIP [49]** in terms of zero-shot retrieval results on Flickr30K [54] and MSCOCO [27].

In Table 1 and 2 in the main paper, COSMOS is compared to other methods using multi-modal data augmentation including DreamLIP [58], LaCLIP [14], and MLLM-A [30]. We additionally report the results of VeCLIP [23] and MobileCLIP [49] for comparison. VeCLIP exploits LLaVA [29] to generate detailed captions while MobileCLIP utilizes CoCa [55] to generate multiple synthetic cations. Although COSMOS is trained with much smaller pre-training set and batch size, it outperforms other methods, demonstrating the efficient usage of synthetic captions within our framework.

Method	ImageNet	MSCOCO		Flickr30k	
	Top-1	I2T@1	T2I@1	I2T@1	T2I@1
CLIP w/ Aug. EMA	20.0	19.2	14.7	38.7	29.3
SILC (1.0, 1.0)	14.7	9.2	13.4	26.8	18.7
SILC (1.5, 0.5)	19.0	17.4	12.9	33.8	25.2
SILC (1.8, 0.2)	21.0	20.8	14.8	40.2	29.3
SILC (1.9, 0.1)	**21.4**	**21.1**	**15.3**	**42.0**	**29.7**
SILC (1.95, 0.05)	**21.4**	20.4	15.1	40.2	29.5

Table 15. **Ablation on the loss scale of SILC [36].** SILC with different loss scale (a,b) where the total loss is calculated as $\mathcal{L}_{\text{total}} = a\mathcal{L}_{\text{CLIP}} + b\mathcal{L}_{\text{self-distill}}$. Models are trained on CC3M with the batch size of 1,024 and one global and one local crops are used as augmentation.

E.10. Rescaling Loss in Previous Methods

In Tab. 15, we conduct an experiment to demonstrate the effect of loss scaling in SILC [36], which also utilizes self-supervision in contrastive vision-language pre-training. We adjust the scale of the CLIP loss (a) and the self-distillation loss (b), while maintaining their sum constant (i.e., $a + b = 2$). For comparison, we also include CLIP [40] with the same augmentation and EMA. Interestingly, a naive summation of the two losses (i.e., $a = 1.0, b = 1.0$) results in a worse performance compared to CLIP, highlighting the importance of selecting appropriate scaling parameters for optimal performance. Consequently, $a = 1.9$ and $b = 0.1$ yield the best results, which we used to reproduce their results in Tab. 4.

Similarly, other works [12, 25, 35, 41] that integrated self-supervised learning in VLM training adopt loss scaling parameters to balance the learning speed between the contrastive objective and the self-distillation objective. This is due to the different scales of the loss functions, as previous works often employ the symmetric InfoNCE loss [38] for contrastive learning while using the cross-entropy loss between masked inputs or global and local crops for self-supervised learning. We unify the loss function with the InfoNCE loss, eliminating the need for a scaling factor, since the CLIP loss and the COSMOS loss are already updated on the same scale.

References

[1] Nikita Araslanov and Stefan Roth. Single-stage semantic segmentation from image labels. In *CVPR*, 2020. 3

[2] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *ECCV*, 2014. 6

[3] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021. 2, 8

[4] Junbum Cha, Jonghwan Mun, and Byungseok Roh. Learning to generate text-grounded mask for open-world semantic segmentation from only image-text pairs. In *CVPR*, 2023. 3

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

[6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020. 8

[7] 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. In *CVPR*, 2023. 3, 4, 5, 6, 7, 8

[8] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *CVPR*, 2014. 6

[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. *ICML Workshop*, 2022. 1

[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. 4, 6, 7, 8

[11] Karan Desai, Gaurav Kaul, Zubin Aysola, and Justin Johnson. Redcaps: Web-curated image-text data created by the people, for the people. *NeurIPS*, 2021. 5

[12] Xiaoyi Dong, Jianmin Bao, Yinglin Zheng, Ting Zhang, Dongdong Chen, Hao Yang, Ming Zeng, Weiming Zhang, Lu Yuan, Dong Chen, et al. Maskclip: Masked self-distillation advances contrastive language-image pretraining. In *CVPR*, 2023. 9

[13] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International journal of computer vision*, 2015. 3

[14] Lijie Fan, Dilip Krishnan, Phillip Isola, Dina Katabi, and Yonglong Tian. Improving clip training with language rewrites. *NeurIPS*, 2023. 8

[15] 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. 6

[16] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. Datacomp: In search of the next generation of multimodal datasets. *NeurIPS*, 2024. 3, 5

[17] Yuting Gao, Jinfeng Liu, Zihan Xu, Jun Zhang, Ke Li, Rongrong Ji, and Chunhua Shen. Pyramidclip: Hierarchical feature alignment for vision-language model pretraining. *NeurIPS*, 2022. 8

[18] Lisa Anne Hendricks and Aida Nematzadeh. Probing image-language transformers for verb understanding. In *ACL*, 2021. 1, 2, 3

[19] Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. Sugarcrepe: Fixing hackable

benchmarks for vision-language compositionality. *NeurIPS*, 2024. 1, 3

[20] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, 2019. 8

[21] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *ICCV Workshop*, 2013. 6

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

[23] Zhengfeng Lai, Haotian Zhang, Bowen Zhang, Wentao Wu, Haoping Bai, Aleksei Timofeev, Xianzhi Du, Zhe Gan, Jiulong Shan, Chen-Nee Chuah, et al. Veclip: Improving clip training via visual-enriched captions. In *ECCV*, 2024. 8

[24] Samuel Lavoie, Polina Kirichenko, Mark Ibrahim, Mido Assran, Andrew Gordon Wilson, Aaron Courville, and Nicolas Ballas. Modeling caption diversity in contrastive vision-language pretraining. In *ICML*, 2024. 6

[25] 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. In *ICLR*, 2021. 8, 9

[26] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *EMNLP*, 2023. 7

[27] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 2, 4, 5, 6, 7, 8

[28] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *CVPR*, 2024. 7

[29] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *NeurIPS*, 2024. 8

[30] Yanqing Liu, Kai Wang, Wenqi Shao, Ping Luo, Yu Qiao, Mike Zheng Shou, Kaipeng Zhang, and Yang You. Mllms-augmented visual-language representation learning. *arXiv preprint arXiv:2311.18765*, 2023. 8

[31] I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 2

[32] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *NeurIPS*, 2022. 7

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

[34] Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. The role of context for object detection and semantic segmentation in the wild. In *CVPR*, 2014. 3

[35] Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. Slip: Self-supervision meets language-image pre-training. In *ECCV*, 2022. 3, 4, 9

[36] Muhammad Ferjad Naeem, Yongqin Xian, Xiaohua Zhai, Lukas Hoyer, Luc Van Gool, and Federico Tombari. Silc: Improving vision language pretraining with self-distillation. In *ECCV*, 2024. 3, 4, 9

[37] 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. 6

[38] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 3, 9

[39] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *ICCV*, 2012. 6

[40] 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. 2, 4, 5, 6, 7, 8, 9

[41] Sepehr Sameni, Kushal Kafle, Hao Tan, and Simon Jenni. Building vision-language models on solid foundations with masked distillation. In *CVPR*, 2024. 9

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

[43] 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. *NeurIPS*, 2022. 3

[44] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*, 2018. 1, 5

[45] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *CVPR*, 2019. 8

[46] Vasu Singla, Kaiyu Yue, Sukriti Paul, Reza Shirkavand, Mayuka Jayawardhana, Alireza Ganjdanesh, Heng Huang, Abhinav Bhatele, Gowthami Somepalli, and Tom Goldstein. From pixels to prose: A large dataset of dense image captions. *arXiv preprint arXiv:2406.10328*, 2024. 1, 5

[47] Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *CVPR*, 2022. 6, 7

[48] Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *CVPR*, 2024. 3

[49] Pavan Kumar Anasosalu Vasu, Hadi Pouransari, Fartash Faghri, Raviteja Vemulapalli, and Oncel Tuzel. Mobileclip: Fast image-text models through multi-modal reinforced training. In *CVPR*, 2024. 8

[50] Feng Wang, Jieru Mei, and Alan Yuille. Sclip: Rethinking self-attention for dense vision-language inference. *ECCV*, 2024. 3

[51] 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. 6

[52] Hu Xu, Saining Xie, Xiaoqing Tan, Po-Yao Huang, Russell Howes, Vasu Sharma, Shang-Wen Li, Gargi Ghosh, Luke Zettlemoyer, and Christoph Feichtenhofer. Demystifying clip data. In *ICLR*, 2023. 6

[53] Kaicheng Yang, Jiankang Deng, Xiang An, Jiawei Li, Ziyong Feng, Jia Guo, Jing Yang, and Tongliang Liu. Alip: Adaptive language-image pre-training with synthetic caption. In *ICCV*, 2023. 2

[54] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *ACL*, 2014. 5, 6, 7, 8

[55] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *TMLR*, 2022. 8

[56] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *CVPR*, 2024. 8

[57] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *ICCV*, 2023. 2, 4, 5, 6, 7, 8

[58] Kecheng Zheng, Yifei Zhang, Wei Wu, Fan Lu, Shuailei Ma, Xin Jin, Wei Chen, and Yujun Shen. Dreamlip: Language-image pre-training with long captions. In *ECCV*, 2024. 2, 3, 4, 5, 6, 7, 8