Building Vision-Language Models on Solid Foundations with Masked Distillation

Sepehr Sameni[1*] Kushal Kafle[2] Hao Tan[2] Simon Jenni[2]
[1]University of Bern [2]Adobe Research
sepehr.sameni@unibe.ch {kkafle,hatan,jenni}@adobe.com

Abstract

Recent advancements in Vision-Language Models (VLMs) have marked a significant leap in bridging the gap between computer vision and natural language processing. However, traditional VLMs, trained through contrastive learning on limited and noisy image-text pairs, often lack the spatial and linguistic understanding to generalize well to dense vision tasks or less common languages. Our approach, Solid Foundation CLIP (SF-CLIP), circumvents this issue by implicitly building on the solid visual and language understanding of foundational models trained on vast amounts of unimodal data. SF-CLIP integrates contrastive image-text pretraining with a masked knowledge distillation from large foundational text and vision models. This methodology guides our VLM in developing robust text and image representations. As a result, SF-CLIP shows exceptional zero-shot classification accuracy and enhanced image and text retrieval capabilities, setting a new state of the art for ViT-B/16 trained on YFCC15M and CC12M. Moreover, the dense per-patch supervision enhances our zero-shot and linear probe performance in semantic segmentation tasks. A remarkable aspect of our model is its multilingual proficiency, evidenced by strong retrieval results in multiple languages despite being trained predominantly on English data. We achieve all of these improvements without sacrificing the training efficiency through our selective application of masked distillation and the inheritance of teacher word embeddings.

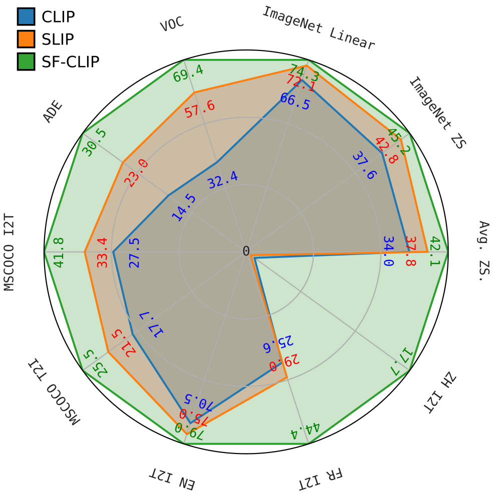

Figure 1. **SF-CLIP's Performance on Vision-Language Tasks.** Our model, SF-CLIP, builds on the knowledge of foundational vision and language models to learn a joint embedding space. As a result, it not only shows improved zero-shot performance but also inherits strong multi-lingual and image segmentation capabilities from its teachers. The plot shows the performance of SF-CLIP compared to SLIP and vanilla CLIP across ten established benchmarks. (all models are pretrained on YFCC-15M)

1. Introduction

The emergence of Vision-Language Models (VLMs), exemplified by pioneering models like CLIP [44] and ALIGN [24], was pivotal in the integration of computer vision and natural language processing. These models foster a unique symbiosis between visual and textual data, opening the door to various applications. For instance, VLMs have been instrumental in enhancing text-guided image retrieval systems [3] and enabling breakthroughs in automated image captioning [29]. Furthermore, they have even extended their impact to creative domains such as the generation of images from text [25, 42, 43, 45]. This diverse range of applications underscores the versatile and transformative nature of VLMs in bridging visual and linguistic domains.

However, despite their success and widespread adoption, VLMs are not without limitations. A fundamental issue is their heavy reliance on alt-text data, which is often readily available but marred by noise and lacks the necessary depth for a nuanced understanding of the visual-textual interplay. This reliance typically leads to models developing representations akin to a "bag of words" approach, where the focus lies predominantly on identifying objects without adequately capturing their compositional context or the intricacies of

[*] Work in part performed during an internship at Adobe Research.

their interactions and attributes [56]. Moreover, the standard contrastive training approach tends to create global representations lacking localized feature sensitivity [57]. Such global representations fail to capture the finer details and the spatial relationships within images, leading to a superficial understanding of the visual content. This problem is even more pronounced in contexts involving low-resource languages, where the scarcity of quality paired data exacerbates the issue. Consequently, this hinders the model's performance in underrepresented languages and raises concerns about the equitable advancement and applicability of VLMs across diverse linguistic landscapes.

Recent works have explored various strategies to address these limitations of VLMs and enhance their understanding of the visual-textual interplay. One prominent trend is the incorporation of extra supervision [9, 37]. While effective, these methods significantly harm training efficiency. The use of cleaner datasets [16] has also been shown to be beneficial. Though this improves model performance, scalability remains a challenge, particularly when bootstrapping from pre-trained VLMs. A third approach involves re-captioning images to refine the alignment between visual and textual data [15, 38]. This method not only necessitates a pre-trained VLM but also tends to reach a performance plateau, indicating a potential saturation point in learning. Despite these advances, a common limitation across these methods is their reliance on existing VLM architectures, and none fully addresses the nuanced understanding of spatial and compositional elements within images.

To overcome these challenges, our approach, SF-CLIP, leverages the solid visual and linguistic understanding captured in foundational vision and language models. These models learned rich visual and textual representations by drawing on vast repositories of unimodal text and image data, allowing them to avoid many of the pathologies resulting from pure weakly-supervised image-text contrastive training. However, foundational vision and language models inherently lack the capacity for direct image-text alignment, an area where models like CLIP thrive. Therefore, we propose building SF-CLIP on the foundations of pre-trained vision and language models by distilling and bridging their knowledge through a combination of masked distillation and contrastive image-text pretraining. Concretely, we leverage the frozen text and image teacher models to provide per-token target latent representations for the text and image encoders during VLM training. This dense per-patch supervision greatly enhances the spatial and compositional understanding of the image encoder and counteracts the tendency to primarily global feature learning of standard VLM training. As a result, SF-CLIP exhibits much-improved performance on tasks that require a deep comprehension of the visual content, such as image segmentation and advanced retrieval scenarios. A standout feature of SF-CLIP is its multilingual proficiency, addressing another limitation of standard VLMs. Typically, achieving such capability would require extensive training on a diverse set of languages [8, 59] or reliance on an off-the-shelf translation models [10]. However, our model achieves this despite being only trained on monolingual data. This is accomplished by designing the VLM text input as a learned projection from the fixed teacher word embeddings, thus inheriting the LLM's multilingual capabilities. The resulting improvements in less common languages are a step towards making VLMs more accessible and applicable globally. Finally, we show that SF-CLIP can be trained efficiently by only selectively applying the distillation on a few examples at each step. As a result, our approach maintains higher training throughput while simultaneously achieving better downstream performance than prior methods using auxiliary training objectives.

Our experiments demonstrate significant improvements in zero-shot and vision-language retrieval tasks using SF-CLIP. For instance, our training procedure enhances zero-shot performance on ImageNet [47] by over 5% compared to standard CLIP training on CC-12M [5]. This improvement is achieved with only a marginal impact on training throughput, maintaining a high rate of 2750 samples per second against 3300 samples per second for traditional training methods.

2. Prior Work

Contrastive Learning in VLMs. Contrastive learning framework [52] associates images with their corresponding textual descriptions via the use of dual encoders to map images and text into a shared embedding space. Contrastive learning has been pivotal in enabling zero-shot learning capabilities in models, as evidenced in numerous studies [24, 44]. However, it predominantly captures high-level associations, often neglecting finer compositional details [56].

Challenges in Spatial and Linguistic Understanding. A significant limitation of the standard contrastive training in VLMs is the insufficient capture of compositional information, object attributes, and relations [48, 60]. This shortcoming stems partly from the nature of web-scraped image-text pairs, which often lack the depth required for understanding complex compositions [15]. The literature has identified this as a critical area for improvement in VLMs [33, 40, 51].

To address these challenges, some researchers have explored data-level interventions, such as data augmentation and hard-negative mining. Techniques include modifying text descriptions by word swapping or replacement [12, 56] and enhancing captions using large language models [13, 15, 35]. While these methods have shown promise, they risk overfitting specific textual modifications, usually ignore the image, and might introduce hallucinations, which is a common problem with LLMs [22].

Another line of research has introduced additional learning objectives to improve VLMs. This includes incorporat-

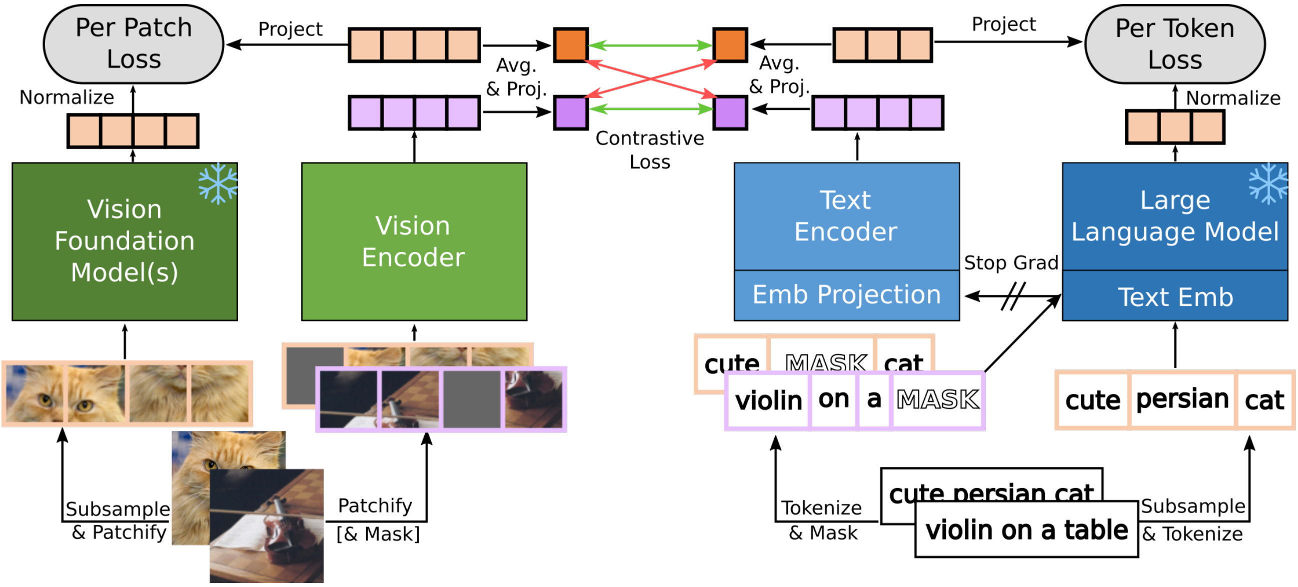

Figure 2. **Model Overview for SF-CLIP.** Our model learns to represent visual and textual data in a shared embedding space through an image and text encoder. We train our model on a dataset comprising image-text pairs, utilizing a combination of optionally masked feature distillation—aimed at inheriting the robust compositional understanding from vision and language foundation models—on a subset of the minibatch (illustrated by the orange sample in this figure) and standard vision-language contrastive learning to align the two modalities.

ing self-supervision in the vision and text branches [9, 30]. These methods, although effective in enhancing model performance, significantly increase the computational overhead of training large-scale VLMs [55]. For example, training MaskCLIP [9] takes 1.75× more than vanilla CLIP, and training SLIP [37] takes 2.67× more.

Leveraging Foundational Models. Recent works have demonstrated the efficacy of vision foundational models in capturing rich spatial features [26, 39] and their robust feature understanding. This insight has guided recent research towards leveraging these models to enhance VLMs. LiT [58] uses a frozen pretrained vision encoder instead of training one from scratch, and Three Towers [27] uses a frozen vision encoder to guide the CLIP's vision encoder. SAM-CLIP [54] starts from a pretrained SAM model and fine-tunes it with both SAM and CLIP objectives via another larger pretrained CLIP model. Despite the success of using pretrained vision models, using LLMs as extra supervision for the text encoder is under-explored.

Multilinguality. There are generally two approaches to train multilingual CLIP models. One way is to simply train the model on multilingual data like mSigLIP [59] trained on WebLI [6] and OpenCLIP [8] trained on LAION5B [49]. The primary advantage of this approach is the direct exposure of the model to a wide range of languages during training, which can lead to more naturally generalized multilingual capabilities. However, it also presents challenges, particularly in terms of the immense computational resources required for training on such large and diverse datasets. An alternative strategy to circumvent the high costs of direct multilingual training is to align CLIP's English text encoder with a pre-trained multilingual text encoder by using parallel sentences [4, 7, 10]. This approach is significantly more resource-efficient, as it leverages existing models and data. However, it may depend on the quality of the parallel sentences and the effectiveness of the alignment process, which can vary based on the languages involved and the quality of the machine translation system used.

In summary, VLM approaches have advanced, but a gap in detailed spatial-textual capture persists compared to uni-modal foundational models. Our SF-CLIP seeks to fill this by blending contrastive pretraining with masked knowledge distillation, leveraging foundational models for improved spatial-textual insight and training efficiency.

3. Method

We aim to learn a unified embedding space of text and images using two separate encoders, a visual encoder V and a text encoder T. We assume access to a paired dataset of images x_i and their noisy captions y_i (alt-text) $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$. Our model also leverages pre-trained teacher models V_{teacher} and T_{teacher} for both modalities. These teacher models are trained on potentially much larger uni-modal datasets (*e.g.*, large amounts of unlabelled images and texts), which are generally easier to obtain than the paired image-text data. All the models in our framework are based on the Transformer architecture [53], which encodes an input into a sequence of feature vectors, *e.g.*, $V(x) \in \mathbb{R}^{n_v \times d_v}$, where n_v is the number of tokens and d_v the latent feature dimension of the visual encoder. At a high level, our proposed training strategy combines the usual contrastive loss between paired data [44] with masked knowledge distillation objective to the teacher in each modality. We name our model SF-CLIP and show an overview in Figure 2. The remainder of this section describes our model and training objective in detail.

Vision-Language Contrastive Objective. We use a standard contrastive learning objective to align our model's vision and language embeddings. Concretely, let $v(x_i) \in \mathbb{R}^d$ be the embedding vector resulting from our visual encoder and $t(y_i) \in \mathbb{R}^d$ be the corresponding text embedding. In our implementation, we calculate $v(x_i)$ and $t(y_i)$ as the average of all the final layer token embeddings in the transformers V and T, followed by a learned linear projection in each modality (*e.g.*, projecting d_v to d for the visual encoder). Finally, the projected embeddings are l2-normalized. We follow prior works [24, 44] and use a symmetric InfoNCE loss [52] formulation

$$\mathcal{L}_{CLIP} = \mathcal{L}_{I \to T} + \mathcal{L}_{T \to I}, \tag{1}$$

with

$$\mathcal{L}_{I \to T} = -\frac{1}{B} \sum_{i}^{B} \log \frac{\exp(v(x_i) \cdot t(y_i)/\tau)}{\sum_{j=1}^{B} \exp(v(x_i) \cdot t(y_j)/\tau)}$$

$$\mathcal{L}_{T \to I} = -\frac{1}{B} \sum_{i}^{B} \log \frac{\exp(v(x_i) \cdot t(y_i)/\tau)}{\sum_{j=1}^{B} \exp(v(x_j) \cdot t(y_i)/\tau)},$$

where τ is a learned temperature parameter, and B is the size of a training mini-batch.

Masked Feature Distillation. We combine the above contrastive vision-text alignment objective with a feature distillation loss in both modalities. These distillation objectives aim to anchor the learned student representations with strong pre-trained visual and textual representations that capture well the structure of visual and textual data (solid foundations). Note that this implies that the student and teacher input tokenizer must result in the same number of tokens for any input. We, therefore, inherit the teacher language tokenizers in practice. Furthermore, we pose feature distillation in a masked setting, where the student only partially observes the input and must recover the latent teacher representation of masked and unmasked input tokens. This masked reconstruction task additionally steers the encoders to learn structural patterns in the inputs. Concretely, given teacher visual encoders V_{teacher} and text encoder T_{teacher}, and their corresponding student models V and T, we pose the distillation losses

$$\mathcal{L}_{VD} = \|V(M_v \odot x) - V_{\text{teacher}}(x)\|_2^2 \tag{2}$$

for the visual encoder and for the text encoder

$$\mathcal{L}_{TD} = \|T(M_t \odot y) - T_{\text{teacher}}(y)\|_2^2, \tag{3}$$

where M_v and M_t are masks that randomly zero out a set of student input tokens. Note that we layer normalize the outputs of both teacher models in the loss calculation and that we include a learned linear projection from the output of V and T to teacher output features (the teacher and student feature dimensions can be different).

Overall Training Objective. Our overall learning objective combines the CLIP loss with the distillation losses

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{CLIP} + \lambda_1 \mathcal{L}_{VD} + \lambda_2 \mathcal{L}_{TD}, \tag{4}$$

where λ_1 and λ_2 weigh the contribution of the distillation terms. In this multitask objective, we can interpret $\mathcal{L}_{CLIP}$ as aligning the two modalities while $\mathcal{L}_{VD}$ and $\mathcal{L}_{TD}$ anchor the visual and textual encoders with strong pre-existing representations of visual and textual data.

Batch Subsampling for Efficient Training. Since our training includes distillation from large teacher networks (*e.g.*, LLMs for the text encoder), a naive implementation would result in much increased computational and memory demands and much lower training throughput. This is due to feeding every example in the mini-batch to the two teacher networks as well. To counteract this, we propose to perform the masked distillation objectives only on a small random subset of each training mini-batch. As our experiments show, this provides the positive influence of masked distillation while preserving high training throughput. Furthermore, it would be possible to pre-compute the teacher representations and avoid the online computation of targets during training, trading off additional storage requirements for virtually no training overhead compared to vanilla CLIP training.

Inheriting Teacher Word Embeddings. A large portion of the text encoder parameters are dedicated to learned word embeddings. Instead of learning these from scratch, we opt for a linear projection from the teacher's frozen word embeddings to T's hidden dimension. Besides accelerating training and enhancing downstream performance, we observe multilingual vision-language understanding capabilities emerging from this design when leveraging a multi-lingual text teacher. This occurs even without seeing any multilingual paired data during training of V and T.

4. Experiments

We conducted extensive experiments to validate our model design and to demonstrate its advantages over the conventional CLIP-style training approach for vision-language models. In these experiments, we employed a vision encoder based on the ViT-B/16 architecture [11] and CLIP's corresponding text encoder architecture [44] but with a non-causal attention (Similar to MaskCLIP [9] and CLIP🚀 [17]). Following the methodology of SLIP [37], we trained our model on YFCC15M (a subset of YFCC100M [50]) for 25 epochs with a batch size of 4096. Our default selection for the visual

Method	Average	Caltech-101	CIFAR-10	CIFAR-100	Country211	DTD	EuroSAT	FER-2013	Aircraft	Food-101	GTSRB	Memes	KittiDis	MNIST	Flowers	Pets	PatchCam	SST2	RESISC45	Cars	Voc2007
Pretraining on YFCC-15M																					
CLIP [9]	34.0	58.6	68.5	36.9	10.8	21.4	30.5	16.9	5.1	51.6	6.5	51.1	25.9	5.0	52.7	28.6	51.7	**52.5**	22.4	4.5	79.1
SLIP [9]	37.8	70.9	82.6	48.6	11.8	26.6	19.8	18.1	5.6	59.9	**12.6**	51.8	29.4	9.8	56.3	31.4	**55.3**	51.5	28.5	5.4	80.5
MaskCLIP [9]	40.1	72.0	80.2	**57.5**	12.6	27.9	**44.0**	20.3	6.1	64.9	8.5	52.0	34.3	4.9	57.0	34.3	50.1	49.9	35.7	6.7	**82.1**
CLIP🚀 [17]	-	72.8	71.3	38.9	14.6	28.0	12.6	-	9.9	61.5	10.0	52.9	**44.2**	9.4	58.4	30.7	51.1	50.4	37.2	6.7	-
SLIP$_{100ep}$ [17]	40.1	74.0	79.2	50.4	11.5	26.2	20.8	**36.5**	8.4	63.3	11.7	**55.1**	35.2	**17.1**	61.3	34.7	52.1	49.9	27.8	8.1	78.67
CLIP🚀$_{32ep}$ [17]	-	**75.4**	67.1	37.8	**15.6**	30.3	23.2	-	**11.2**	63.0	8.1	54.3	35.6	9.8	**62.8**	35.4	51.6	50.1	36.0	**8.2**	-
SF-CLIP	**42.1**	72.2	**85.0**	53.6	12.0	**35.2**	43.7	30.6	11.0	**65.0**	10.3	49.6	32.9	11.6	59.5	**38.1**	54.1	50.3	**39.7**	**8.2**	80.2
Pretraining on CC-12M																					
CLIP [15]	37.5	77.4	64.9	38.5	5.1	19.4	20.1	30.8	2.4	50.8	7.3	52.1	**36.3**	10.1	33.2	64.1	50.3	47.6	38.9	24.1	77.0
SLIP [15]	-	77.6	80.7	46.3	5.7	25.1	25.8	-	2.3	52.5	6.0	-	-	-	29.2	58.6	-	-	36.6	24.9	-
LaCLIP [15]	41.9	83.3	75.1	43.9	8.9	31.0	27.3	26.7	5.6	60.7	12.7	52.9	16.9	19.2	39.9	72.4	50.6	48.4	44.3	**36.3**	81.9
LaSLIP [15]	-	82.8	82.0	50.2	**9.2**	30.1	20.4	-	4.4	62.9	10.1	-	-	-	37.4	70.6	-	-	45.6	32.2	-
LaSF-CLIP	**46.9**	**84.6**	**86.7**	**57.3**	**9.2**	**42.2**	**35.9**	**34.9**	**7.3**	**65.1**	**18.4**	**53.0**	29.7	**19.3**	**43.7**	**76.3**	**54.8**	**50.3**	**49.1**	35.7	**84.1**
Pretraining on YFCC-15M+CC-3M+CC-12M+ImageNet-21K(ImageNet-1k is removed, around 13M images)																					
MaskCLIP [9]	48.9	86.4	95.3	78.3	11.6	33.0	57.7	18.8	8.0	78.9	17.3	52.8	16.0	7.3	74.2	74.4	52.1	46.2	54.3	26.5	82.3

Table 1. Zero-shot evaluation on ICinW [28] classification benchmarks. Best results in **bold** and second best with underline.

	ImageNet		Flickr30K						MS-COCO					
	Zero	Linear	Image-to-text			Text-to-image			Image-to-text			Text-to-image		
Method	Shot	Probe	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
Pretraining on YFCC-15M														
CLIP [37]	37.6	66.5	52.9	79.6	87.2	32.8	60.8	71.2	27.5	53.5	65.0	17.7	38.8	50.5
SLIP [37]	42.8	72.1	58.6	85.1	91.7	41.3	68.7	78.6	33.4	59.8	70.6	21.5	44.4	56.3
MaskCLIP [9]	44.5	73.7	**70.1**	90.3	**95.3**	45.6	**73.4**	82.1	41.4	67.9	77.5	**25.5**	49.7	**61.3**
SLIP$_{100ep}$ [17]	45.0	73.6	59.7	85.5	91.6	39.6	66.5	76.6	33.8	60.0	71.2	22.9	45.9	57.3
SF-CLIP	**45.2**	**74.3**	68.7	**90.4**	94.8	**46.2**	73.2	**82.7**	**41.8**	**68.3**	**78.4**	**25.5**	**50.2**	**61.3**
Pretraining on CC-12M														
CLIP [15]	40.2	70.3	63.3	86.3	92.4	48.0	73.9	82.5	37.8	65.4	75.7	25.8	51.0	62.5
SLIP [37]	40.7	73.7	62.5	87.2	92.1	46.6	73.3	80.9	37.6	64.9	75.5	26.8	51.4	62.7
LaCLIP [15]	48.4	72.3	63.9	86.5	92.6	51.6	78.8	86.2	38.0	64.8	75.0	26.5	51.2	62.6
LaSF-CLIP	**53.6**	**75.3**	**71.8**	**91.9**	**95.2**	**59.9**	**84.2**	**90.9**	**44.3**	**71.3**	**80.2**	**31.4**	**57.3**	**68.1**

Table 2. Zero-shot and Linear probing accuracies on Imagenet (left) and zero-shot image-text retrieval on Flickr30K [41] and MS-COCO [31] datasets (right). Best results in **bold** and second best with underline.

teachers included SAM-H/16 [26] and DINOv2-L/14 [39], and for the text teacher, we chose XGLM-1.7B [32] with word embedding projection. By default, we masked up to 25% of the text tokens and none of the vision tokens (evaluated in ablations) and employed a subset of 1024 images for the visual teachers and 512 for the text teacher. The values of λ_1 and λ_2 were consistently set to 1. To further demonstrate the broad applicability of our method, we also trained LaSF-CLIP, a language-augmented version of our SF-CLIP, using LaCLIP's language rewrites [15] on CC12M [5] for 35 epochs with a batch size of 8192, employing 1024 images for the vision teachers and 512 for the text teacher. For most of our comparisons, we used the official SLIP[2] and LaCLIP[3] checkpoints. All the models were trained using eight A100 GPUs on a single node using OpenCLIP's codebase [8].

Zero-shot Classification on Small Datasets. We use the 20 datasets of the Image Classification in the Wild (ICinW) challenge [28] to assess our zero-shot classification accuracies in Table 1. Other than the datasets that most methods perform poorly on, like MNIST and Aircraft (likely due to the domain gap between YFCC15M/CC12M and these datasets [9]), our method outperforms the others on average

[2] https://github.com/facebookresearch/SLIP#vit-base

[3] https://github.com/LijieFan/LaCLIP#pre-trained-models

Method	Pascal-Context	ADE-20K
CLIP [9]	13.5	7.2
MaskCLIP [9]	17.2	10.2
SF-CLIP	**25.9**	**11.6**

Table 3. Zero-shot semantic segmentation (mIoU%) using models trained on YFCC-15M.

by 2% on YFCC15M and by 5% on CC-12M. Not only does our model outperform SLIP$_{100ep}$ (which was trained for 100 epochs, instead of 25), it even gets close to MaskCLIP [9] trained on significantly more data.

ImageNet Classification. We evaluate both zero-shot and linear probing accuracy of our model on ImageNet-1k [47]. We use the 7 prompts in SLIP [37] for zero-shot and 90 epochs of training with added batch normalization [23] without affine parameters [19] for the linear probing. As can be seen on the left of Table 2, SF-CLIP performs better than all the other models with both pretraining datasets. Most notably, on CC-12M, SF-CLIP gets a whopping 5.2% improvement over LaCLIP in the zeroshot setting.

Zero-shot Text/Image Retrieval. We report the zero-shot text-image retrieval results on two benchmark datasets, Flicr30K [41] and MS-COCO [31] on the right side of Table 2. Overall, in the case of YFCC-15M, we see on-par performance with MaskCLIP, and on CC-12M we see significant improvements over LaCLIP.

Zero-shot Semantic Segmentation. Even though CLIP was trained on whole images, DenseCLIP [62] showed that one can still get per-patch classifications from CLIP. Since our model inherits additional spatial understanding through our visual distillation objective, we expect to see improved performance compared to vanilla CLIP on zero-shot semantic segmentation. Following DenseCLIP, we use the final attention keys of the vision encoder and project them into the joint embedding space, where we apply a per patch zero-shot classification on Pascal-Context [36] and ADE-20K [61]. Table 3 shows that SF-CLIP performs much better than CLIP and MaskCLIP (which also has a per patch loss).

Linear Probing for Semantic Segmentation. Following the previous experiment, we also conducted linear probing on our per-patch representations for the task of semantic segmentation on Pascal-VOC [14], Pascal-Context [36], ADE-20K [61], and COCO-Stuff [2] datasets. Results in Table 4 show that SF-CLIP indeed has a richer spatial representation than CLIP and significantly outperforms it in this task. Most notably on Pascal-VOC, SF-CLIP is almost performing as well as CLIP trained on 1B images [18].

Method	VOC	Context	ADE	COCO
Pretraining on YFCC-15M				
CLIP [37]	32.4	29.6	14.5	22.6
SLIP [37]	57.6	41.8	23.0	32.1
SF-CLIP	**69.4**	**47.9**	**30.5**	**35.1**
Pretraining on CC-12M				
CLIP [15]	35.2	30.1	18.0	24.4
LaCLIP [15]	33.7	30.1	17.5	24.5
LaSF-CLIP	**69.1**	**47.0**	**31.6**	**34.9**
Larger Dataset Pretraining, 448×448 Evaluation				
SAM$_{SA-1B}$ [26]	46.6	-	26.6	-
CLIP$_{DataComp-1B}$ [18]	70.7	-	36.4	-

Table 4. Linear head probing evaluations (mIoU%) on semantic segmentation datasets with ViT-B/16.

Method	Training Data	mAP	mAR
SAM [26]	SA-1B	57.8	60.8
SF-CLIP	YFCC15M	45.0	54.6

Table 5. Zero-shot instance segmentation (using frozen SAM decoder) with ViT-B/16 on the COCO dataset, both models are evaluated with 1024×1024 images using ground truth bounding boxes.

Zero-shot Instance Segmentation. Since SF-CLIP was trained to mimic SAM's [26] final representations, we can use SAM's decoder on top of our model "out of the box" and get better than chance results. In Table 5, we use the ground-truth bounding boxes of MS-COCO to get instance segmentations. Note that SF-CLIP was only trained to predict SAM's features on YFCC-15M and was not trained to process 1024×1024 images, but still manages to learn something useful and compatible with SAM's decoder.

Compositional Understanding Benchmark. Because of the noisy training data and the coarse contrastive loss, VLMs mostly act like a bag of words [56] and lack a deeper compositional understanding of the images. Previous benchmarks to assess compositional understanding like Winoground [51], VL-CheckList [60], ARO [56], CREPE [34], and Cola [46], were found to have shortcomings and be gameable in many cases. The SugarCREPE [21] benchmark aims to address those shortcomings and provides a more reliable metric to measure compositional understanding of the VLMs. Results in Table 6 (left) show that SF-CLIP gets a decent improvement over prior dual encoder joint embedding VLM approaches on YFCC-15M but gets worse performance than CLIP with language rewrites [15]. This result shows that even though naively sampling from an LLM for text data augmentation can be useful for many tasks (as was shown in other experiments), the LLM hallucinations might make the model worse in some aspects at the end and just using the hidden representations of an LLM, like as in SF-CLIP is a more reliable way than sampling from an LLM without looking at the image.

	SugarCREPE				SVO			
Method	Replace	Swap	Add	Average	Subject	Verb	Object	All
Pretraining on YFCC-15M								
CLIP [37]	73.3	59.4	74.0	68.9	79.3	70.5	87.8	75.4
SLIP [37]	75.2	58.6	73.7	69.2	80.3	72.8	**89.5**	77.4
SF-CLIP	**77.3**	**61.6**	**74.8**	**71.2**	**81.0**	**74.7**	87.1	**78.2**
Pretraining on CC-12M								
CLIP [15]	**77.5**	61.8	**73.5**	**70.9**	80.8	76.9	89.5	80.0
LaCLIP [15]	75.1	60.6	71.2	69.0	85.6	80.7	91.8	83.8
LaSF-CLIP	76.7	**63.3**	72.0	70.7	**87.8**	**84.0**	**94.2**	**86.7**

Table 6. Benchmarks on the shortcomings of VLMs. SugarCREPE [21] (compositional understanding), and SVO [20](verb understanding).

Method	EN	ES	FR	IT	DE	RU	ZH	TR	JP	PL	KO
Pretraining on YFCC-12M											
CLIP [37]	70.5	23.3	25.6	23.4	21.4	1.1	0.9	3.6	0.7	6.6	0.7
SLIP [37]	75.0	26.8	29.0	22.1	21.7	0.3	0.5	3.8	0.7	7.5	0.6
SF-CLIP	79.0	48.7	44.4	43.1	41.3	32.5	17.7	14.8	10.4	9.4	6.5
Pretraining on CC-12M											
CLIP [15]	78.9	4.3	10.8	8.5	7.2	0.7	0.4	2.3	1.0	4.2	0.5
LaCLIP [15]	80.1	8.4	16.1	12.9	14.0	1.0	1.6	3.5	0.4	7.1	0.8
LaSF-CLIP	84.0	34.2	38.1	33.2	33.5	40.3	47.3	13.9	27.5	9.1	12.1

Table 7. Comparison of I2T.R@5 (T2I.R@5 follows the same patterns) performance across different languages on the XTD10 [1] benchmark.

Verb Understanding Benchmark. VLMs often fail at identifying image-text pairs that show a mismatch concerning subjects, verbs, and objects. The SVO [20] benchmark aims to quantify this problem and identified that VLMs perform worse on verb understanding, likely due to the noisy training data. We see a clear improvement using SF-CLIP in this task (Table 6, middle), but we note that verbs remain challenging even for SF-CLIP compared to subject and object, which are mostly nouns.

Multilingual Capabilities. YFCC15M, a subset of YFCC100M [50], primarily features English captions. Therefore, training a standard CLIP model on this data will not permit image/text retrieval with other languages. However, thanks to using a large multi-lingual language model as our text teacher (XGLM-1.7B [32]) and by inheriting its word embedding through a learned projection, SF-CLIP shows out-of-the-box multilingual capabilities even though it was never explicitly trained on non-English data. Table 7 demonstrates that SF-CLIP performs drastically better than baselines on XTD10 [1] for all languages. In contrast, we observe CLIP's and SLIP's performance drop strongly even for languages similar to English and nearing random levels for distant languages. SF-CLIP, on the other hand, performs significantly better than chance even in very distant languages like Japanese and Korean. We believe it to be remarkable that just by learning a projection of input tokens and matching the outputs of the LLM in one language (*i.e.*, English in our case), the model can generalize well to various languages. We observe the same behavior on the models trained on CC-12M but with different languages. Based on our initial investigations, language rewrites [15] sometimes output sentences in Russian and Chinese, explaining the performance difference in these languages. This also shows that a small set of sentences in other languages can greatly benefit multilingual capabilities through teacher distillation.

5. Ablations

We perform extensive ablation experiments to verify the various design choices in our model and training algorithm. All the models in this section are trained on YFCC-15M for 8 epochs with a batch size of 4096. Unless stated otherwise we use a small 564M version of XGLM [32] and only DINO-L/14 as our vision teacher. For all of the evaluations we measured ImageNet-ZeroShot accuracy, MSCOCO text-to-image and image-to-text retrieval performance, image-to-text top 5 accuracies on a close to English language (ES) and distant language (RU), and finally zero shot semantic segmentation accuracy on Pascal-Context to have a full picture and compare models on many aspects.

	$\text{Teacher}_{\text{txt}}$	Mask_{txt}	Emb. Proj.	$\text{Teacher}_{\text{img}}$	Mask_{img}	Subset Ratio	IN-ZS	MSC-T2I	MSC-I2T	ES-I2T	RU-I2T	Context
1	✓	✓	✓	✓	✗	12.5%	33.9	18.9	33.3	39.1	34.2	25.4
2	✗	✓	N/A	✓	✗	12.5%	33.2	18.6	31.5	18.4	1.6	25.3
3	✓	✗	✓	✓	✗	12.5%	33.6	18.9	32.5	38.0	31.1	25.2
4	✓	✓	✗	✓	✗	12.5%	35.2	18.8	31.9	22.6	0.8	23.8
5	✓	✓	✓	✗	✗	12.5%	31.3	16.6	29.0	35.5	33.8	22.7
6	✓	✓	✓	✓	✓	12.5%	34.3	18.7	32.3	37.4	33.6	25.6
7	✓	✓	✓	✓	✗	6.25%	33.9	18.8	32.9	38.9	34.1	25.2
8	✓	✓	✓	✓	✗	25%	34.1	18.7	31.5	39.0	32.4	25.2

Table 8. Importance of the Different Components: We study the different components that matter for different evaluation metrics. For the different variants, we highlight the differences from the default SF-CLIP setting.

Text	Vision	IN-ZS	MSC-T2I	MSC-I2T	ES-I2T	RU-I2T	Context
-	-	31.5	15.5	28.0	16.9	0.8	21.6
-	DINO	33.2	18.6	31.5	18.4	1.6	25.3
564M	-	31.3	16.6	29.0	35.5	33.8	22.7
564M	DINO	33.9	18.9	33.3	39.1	34.2	25.4
1.7B	DINO	34.6	19.5	33.4	41.7	34.9	25.7
1.7B	DINO+SAM	36.2	20.6	36.0	41.7	36.2	25.7

Table 9. We study the effects of different teachers of varying sizes on VLM performance.

Setting	Training Time
CLIP	1.00×
CLIP+SimCLR	2.67×
MaskCLIP	1.75×
SF-CLIP $_{\text{Full Batch}}$	2.28×
SF-CLIP $_{\text{Subsampled}}$	1.20×

Table 10. Training Time.

Importance of Different Components. In Table 8, we report different model variants as we add or remove components. First, we can see that removing the text teacher or not inheriting the word embedding removes the multilingual capabilities (rows 2 and 4). If we remove the projected word embedding, we see better performance for ImageNet-ZeroShot but worse performance on everything else which indicates trade-offs between different benchmarks. Second, removing the image teacher leads to a drop in performance on all metrics (row 5). Next, we see consistent benefits for masking with text (row 3), while image masking (row 6) provides mixed results with ViT-B. Although we observe improvements from image masking for larger ViT architectures in initial exploration, we disable it by default for ViT-B. Lastly, adjusting the distillation rate to either double or half (rows 7 and 8) reveals an optimal performance at the default rate and indications of overfitting at higher rates.

On the Choice of Teachers. Our model supports various teacher combinations in both modalities. Table 9 compares these combinations, utilizing XGLM models (564M and 1.7B parameters) for text and DINO-L/14 and SAM-H/16 for vision. Generally, we note improved performance with an increased number and size of the teachers.

Training Efficiency. We compare training speeds with and without the proposed batch subsampling (12.5%) for the distillation objective in Table 10. SF-CLIP trains at a high speed, only slightly slower than standard CLIP, but significantly faster than SLIP(CLIP+SimCLR) [37] and MaskCLIP [9]. It also shows improved performance in other experiments, indicating greater computational efficiency.

6. Conclusion & Limitations

In this paper, we introduced SF-CLIP, a novel Vision Language Model (VLM) approach that harnesses the robust foundations of large-scale unimodal models to enhance the VLM's visual and linguistic capabilities. By integrating contrastive image-text pretraining with masked knowledge distillation from unimodal teachers, SF-CLIP effectively absorbs and aligns the strengths of both techniques. This approach results in notable improvement in zero-shot classification accuracy and image-text retrieval performance, achieving new state-of-the-art results without sacrificing training efficiency. SF-CLIP also displays promising multilingual retrieval performance, suggesting its applicability in various linguistic contexts despite being primarily trained in English data.

However, this multilingual capability necessitates using the same tokenizer as the text teacher, which can be potentially restrictive. One important limitation is that this method cannot be trivially modified to finetune an existing pre-trained CLIP model. Moreover, in pursuit of simplicity and lightness, SF-CLIP avoids using transformer decoders for distillation in favor of linear projections. While this benefits computational efficiency, it might limit the model's learning capacity compared to architectures that utilize more complex decoding mechanisms [9, 54].

References

[1] Pranav Aggarwal and Ajinkya Kale. Towards zero-shot cross-lingual image retrieval. *arXiv preprint arXiv:2012.05107*, 2020. 7

[2] Holger Caesar, Jasper R. R. Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1209–1218, 2016. 6

[3] Min Cao, Shiping Li, Juntao Li, Liqiang Nie, and Min Zhang. Image-text retrieval: A survey on recent research and development. *ArXiv*, abs/2203.14713, 2022. 1

[4] Fredrik Carlsson, Philipp Eisen, Faton Rekathati, and Magnus Sahlgren. Cross-lingual and multilingual clip. In *International Conference on Language Resources and Evaluation*, 2022. 3

[5] Soravit Changpinyo, Piyush Kumar Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3557–3567, 2021. 2, 5

[6] Xi Chen, Xiao Wang, Soravit Changpinyo, A. J. Piergiovanni, Piotr Padlewski, Daniel M. Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Nan Ding, Keran Rong, Hassan Akbari, Gaurav Mishra, Linting Xue, Ashish V. Thapliyal, James Bradbury, Weicheng Kuo, Mojtaba Seyedhosseini, Chao Jia, Burcu Karagol Ayan, Carlos Riquelme, Andreas Steiner, Anelia Angelova, Xiaohua Zhai, Neil Houlsby, and Radu Soricut. Pali: A jointly-scaled multilingual language-image model. *ArXiv*, abs/2209.06794, 2022. 3

[7] Zhongzhi Chen, Guangyi Liu, Bo Zhang, Fulong Ye, Qinghong Yang, and Ledell Yu Wu. Altclip: Altering the language encoder in clip for extended language capabilities. *ArXiv*, abs/2211.06679, 2022. 3

[8] 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. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2829, 2022. 2, 3, 5

[9] 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. *ArXiv*, abs/2208.12262, 2022. 2, 3, 4, 5, 6, 8

[10] Gabriel Oliveira dos Santos, Diego A. B. Moreira, Alef Iury Ferreira, Jhessica Silva, Luiz Pereira, Pedro Bueno, Thiago Sousa, Helena de Almeida Maia, N'adia Da Silva, Esther Colombini, Helio Pedrini, and Sandra Avila. Capivara: Cost-efficient approach for improving multilingual clip performance on low-resource languages. *ArXiv*, abs/2310.13683, 2023. 2, 3

[11] 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. *ArXiv*, abs/2010.11929, 2020. 4

[12] Sivan Doveh, Assaf Arbelle, Sivan Harary, Rameswar Panda, Roei Herzig, Eli Schwartz, Donghyun Kim, Raja Giryes, Rogério Schmidt Feris, Shimon Ullman, and Leonid Karlinsky. Teaching structured vision&language concepts to vision&language models. *ArXiv*, abs/2211.11733, 2022. 2

[13] Sivan Doveh, Assaf Arbelle, Sivan Harary, Roei Herzig, Donghyun Kim, Paola Cascante-Bonilla, Amit Alfassy, Rameswar Panda, Raja Giryes, Rogério Schmidt Feris, Shimon Ullman, and Leonid Karlinsky. Dense and aligned captions (dac) promote compositional reasoning in vl models. *ArXiv*, abs/2305.19595, 2023. 2

[14] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2): 303–338, 2010. 6

[15] Lijie Fan, Dilip Krishnan, Phillip Isola, Dina Katabi, and Yonglong Tian. Improving clip training with language rewrites. *ArXiv*, abs/2305.20088, 2023. 2, 5, 6, 7

[16] Alex Fang, Albin Madappally Jose, Amit Jain, Ludwig Schmidt, Alexander Toshev, and Vaishaal Shankar. Data filtering networks. *ArXiv*, abs/2309.17425, 2023. 2

[17] Enrico Fini, Pietro Astolfi, Adriana Romero-Soriano, Jakob Verbeek, and Michal Drozdzal. Improved baselines for vision-language pre-training. *ArXiv*, abs/2305.08675, 2023. 4, 5

[18] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, Eyal Orgad, Rahim Entezari, Giannis Daras, Sarah Pratt, Vivek Ramanujan, Yonatan Bitton, Kalyani Marathe, Stephen Mussmann, Richard Vencu, Mehdi Cherti, Ranjay Krishna, Pang Wei Koh, Olga Saukh, Alexander J. Ratner, Shuran Song, Hannaneh Hajishirzi, Ali Farhadi, Romain Beaumont, Sewoong Oh, Alexandros G. Dimakis, Jenia Jitsev, Yair Carmon, Vaishaal Shankar, and Ludwig Schmidt. Datacomp: In search of the next generation of multimodal datasets. *ArXiv*, abs/2304.14108, 2023. 6

[19] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Doll'ar, and Ross B. Girshick. Masked autoencoders are scalable vision learners. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15979–15988, 2021. 6

[20] Lisa Anne Hendricks and Aida Nematzadeh. Probing image-language transformers for verb understanding. *arXiv preprint arXiv:2106.09141*, 2021. 7

[21] Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. Sugarcrepe: Fixing hackable benchmarks for vision-language compositionality. *arXiv preprint arXiv:2306.14610*, 2023. 6, 7

[22] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*, 2023. 2

[23] Sergey Ioffe and Christian Szegedy. Batch normalization:

Accelerating deep network training by reducing internal covariate shift. *ArXiv*, abs/1502.03167, 2015. 6

[24] 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 *International Conference on Machine Learning*, 2021. 1, 2, 4

[25] Minguk Kang, Jun-Yan Zhu, Richard Zhang, Jaesik Park, Eli Shechtman, Sylvain Paris, and Taesung Park. Scaling up gans for text-to-image synthesis. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10124–10134, 2023. 1

[26] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross B. Girshick. Segment anything. *ArXiv*, abs/2304.02643, 2023. 3, 5, 6

[27] Jannik Kossen, Mark Collier, Basil Mustafa, Xiao Wang, Xiaohua Zhai, Lucas Beyer, Andreas Steiner, Jesse Berent, Rodolphe Jenatton, and Efi Kokiopoulou. Three towers: Flexible contrastive learning with pretrained image models. *ArXiv*, abs/2305.16999, 2023. 3

[28] Chunyuan Li, Haotian Liu, Liunian Harold Li, Pengchuan Zhang, Jyoti Aneja, Jianwei Yang, Ping Jin, Yong Jae Lee, Houdong Hu, Zicheng Liu, and Jianfeng Gao. Elevater: A benchmark and toolkit for evaluating language-augmented visual models. *Neural Information Processing Systems*, 2022. 5

[29] Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *ArXiv*, abs/2301.12597, 2023. 1

[30] Yangguang Li, Feng Liang, Lichen Zhao, Yufeng Cui, Wanli Ouyang, Jing Shao, Fengwei Yu, and Junjie Yan. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. *ArXiv*, abs/2110.05208, 2021. 3

[31] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, 2014. 5, 6

[32] Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona T. Diab, Ves Stoyanov, and Xian Li. Few-shot learning with multilingual language models. *ArXiv*, abs/2112.10668, 2021. 5, 7

[33] Fangyu Liu, Guy Edward Toh Emerson, and Nigel Collier. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 11:635–651, 2022. 2

[34] Zixian Ma, Jerry Hong, Mustafa Omer Gul, Mona Gandhi, Irena Gao, and Ranjay Krishna. Crepe: Can vision-language foundation models reason compositionally? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10910–10921, 2023. 6

[35] Liliane Momeni, Mathilde Caron, Arsha Nagrani, Andrew Zisserman, and Cordelia Schmid. Verbs in action: Improving verb understanding in video-language models. *ArXiv*, abs/2304.06708, 2023. 2

[36] 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 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 6

[37] Norman Mu, Alexander Kirillov, David A. Wagner, and Saining Xie. Slip: Self-supervision meets language-image pretraining. *ArXiv*, abs/2112.12750, 2021. 2, 3, 4, 5, 6, 7, 8

[38] Thao Nguyen, Samir Yitzhak Gadre, Gabriel Ilharco, Sewoong Oh, and Ludwig Schmidt. Improving multimodal datasets with image captioning. *ArXiv*, abs/2307.10350, 2023. 2

[39] Maxime Oquab, Timoth'ee Darcet, Théo Moutakanni, Huy Q. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russ Howes, Po-Yao (Bernie) Huang, Shang-Wen Li, Ishan Misra, Michael G. Rabbat, Vasu Sharma, Gabriel Synnaeve, Huijiao Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision. *ArXiv*, abs/2304.07193, 2023. 3, 5

[40] Letitia Parcalabescu, Michele Cafagna, Lilitta Muradjan, Anette Frank, Iacer Calixto, and Albert Gatt. VALSE: A task-independent benchmark for vision and language models centered on linguistic phenomena. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8253–8280, Dublin, Ireland, 2022. Association for Computational Linguistics. 2

[41] Bryan A. Plummer, Liwei Wang, Christopher M. Cervantes, Juan C. Caicedo, J. Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. *International Journal of Computer Vision*, 123:74 – 93, 2015. 5, 6

[42] Dustin Podell, Zion English, Kyle Lacey, A. Blattmann, Tim Dockhorn, Jonas Muller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *ArXiv*, abs/2307.01952, 2023. 1

[43] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *ArXiv*, abs/2209.14988, 2022. 1

[44] 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 *International Conference on Machine Learning*, 2021. 1, 2, 4

[45] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *ArXiv*, abs/2204.06125, 2022. 1

[46] Arijit Ray, Filip Radenovic, Abhimanyu Dubey, Bryan A Plummer, Ranjay Krishna, and Kate Saenko. Cola: How to adapt vision-language models to compose objects localized with attributes? *arXiv preprint arXiv:2305.03689*, 2023. 6

[47] 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*, 115:211 – 252, 2014. 2, 6

[48] Madeline Chantry Schiappa, Michael Cogswell, Ajay Divakaran, and Yogesh Singh Rawat. Probing conceptual understanding of large visual-language models. *ArXiv*, abs/2304.03659, 2023. 2

[49] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. Laion-5b: An open large-scale dataset for training next generation image-text models. *ArXiv*, abs/2210.08402, 2022. 3

[50] Bart Thomee, David A. Shamma, Gerald Friedland, Benjamin Elizalde, Karl S. Ni, Douglas N. Poland, Damian Borth, and Li-Jia Li. The new data and new challenges in multimedia research. *ArXiv*, abs/1503.01817, 2015. 4, 7

[51] 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. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5228–5238, 2022. 2, 6

[52] Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *ArXiv*, abs/1807.03748, 2018. 2, 4

[53] Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, 2017. 3

[54] Haoxiang Wang, Pavan Kumar Anasosalu Vasu, Fartash Faghri, Raviteja Vemulapalli, Mehrdad Farajtabar, Sachin Mehta, Mohammad Rastegari, Oncel Tuzel, and Hadi Pouransari. Sam-clip: Merging vision foundation models towards semantic and spatial understanding. *ArXiv*, abs/2310.15308, 2023. 3, 8

[55] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022. 3

[56] Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? In *The Eleventh International Conference on Learning Representations*, 2022. 2, 6

[57] Yan Zeng, Xinsong Zhang, and Hang Li. Multi-grained vision language pre-training: Aligning texts with visual concepts. *ArXiv*, abs/2111.08276, 2021. 2

[58] Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. Lit: Zero-shot transfer with locked-image text tuning. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18102–18112, 2021. 3

[59] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. *ArXiv*, abs/2303.15343, 2023. 2, 3

[60] Tiancheng Zhao, Tianqi Zhang, Mingwei Zhu, Haozhan Shen, Kyusong Lee, Xiaopeng Lu, and Jianwei Yin. Vl-checklist: Evaluating pre-trained vision-language models with objects, attributes and relations. *ArXiv*, abs/2207.00221, 2022. 2, 6

[61] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 6

[62] Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from clip. In *European Conference on Computer Vision*, 2021. 6