A Implementation Details

Model Architecture. We employ the ViT-B/16 version of the Segment Anything Model (SAM) as our base architecture [20], comprising 12 transformer layers. To integrate CLIP capabilities, we append a lightweight CLIP head consisting of 3 transformer layers to the SAM backbone. The patch token outputs from this CLIP head undergo a pooling layer to produce an image-level embedding, akin to the role of the CLS token output in ViT models. We adopt max-pooling since we observe that it can lead to better zero-shot classification and semantic segmentation performance of `SAM-CLIP` than average pooling. It is noteworthy that max-pooling has been found to be able to encourage the learning of spatial visual features [38]. With the pooling layer, the CLIP head can output an embedding for the whole image, which can be aligned with a text embedding just like the original CLIP model [37].

Dataset Preparation. For the CLIP distillation, we merge images from several datasets: CC3M [44], CC12M [4], YFCC-15M [37] (a curated subset of YFCC-100M [47] by OpenAI) and ImageNet-21k [41]. This forms our $\mathcal{D}_{\texttt{CLIP}}$ containing 40.6M unlabeled images. For the SAM self-distillation, we sample 5.7% subset from the SA-1B dataset to form $\mathcal{D}_{\texttt{SAM}}$, which originally comprises 11M images and 1.1B masks. We randomly select 1% of $\mathcal{D}_{\texttt{CLIP}}$ and $\mathcal{D}_{\texttt{SAM}}$ as validation sets. Overall, we have 40.8M images for training, which we term as Merged-41M in this work.

Training. As we discussed in Sec. 2, the training is conducted in two phases to optimize convergence, in a "*probing then full finetuning*" style. The first stage of CLIP-head probing takes 20 epochs on $\mathcal{D}_{\texttt{CLIP}}$, while the backbone is kept frozen. Here, the teacher model is the OpenCLIP [18] ViT-L/14 trained on the DataComp-1B dataset [12]. In the second stage (16 epochs), we unfreeze the backbone $\mathrm{Enc}_{\texttt{SAM-CLIP}}$ and proceed with joint fine-tuning together with $\mathrm{Head}_{\texttt{CLIP}}$ and $\mathrm{Head}_{\texttt{SAM}}$, incorporating both CLIP and SAM distillation losses at the ratio of 1:10. The original SAM ViT-B model serves as the teacher in SAM loss. Further, the learning rates applied to $\mathrm{Enc}_{\texttt{SAM-CLIP}}$ and $\mathrm{Head}_{\texttt{SAM}}$ are 10 times smaller than that of $\mathrm{Head}_{\texttt{CLIP}}$ in order to reduce the forgetting of the original SAM abilities. Besides, we adopt a mixed input resolution strategy for training. A notable difference between SAM and CLIP is their pretraining resolution. SAM is trained and works best on 1024px resolution while often lower resolutions (e.g., 224/336/448px) are adopted for CLIP training and inference [37, 7, 45]. Hence, we employ variable resolutions of 224/448px for the CLIP distillation via the variable batch sampler approach of [31], while SAM distillation utilizes a 1024px resolution in accordance with SAM's original training guidelines [20]. In every optimization step, we form a batch of 2048 images from $\mathcal{D}_{\texttt{CLIP}}$ and 32 images (each with 32 mask annotations) from $\mathcal{D}_{\texttt{SAM}}$ and perform training in a multi-task fashion.

Resolution Adaption. After the two training stages, `SAM-CLIP` can accomplish CLIP tasks (e.g., zero-shot classification) using the CLIP-head under 224/336/448px, and run inference with the SAM-head under 1024px. However, if one wants to apply the two heads together on a single input image for certain tasks (we present a demo of this in Sec. A.3), it would be inefficient to pass the image twice to the image encoder with two resolutions for the two heads respectively. To remedy this issue, we adapt the CLIP head for 1024px input using a very short and efficient stage of fine-tuning: freezing the image encoder and only finetuning the CLIP-head with $\mathcal{L}_{\texttt{CLIP}}$ for 3 epochs (it is the same as the first stage of training, which is also CLIP-head probing) under variable resolutions of 224/448/1024px. *Note:* resolution upscaling strategies are prevalent in CLIP training: [37, 45, 22] show it is more efficient than training with high resolution from the beginning.

A.1 Zero-Shot Evaluations

CLIP Task: Zero-Shot Image Classification. To examine the CLIP-related capabilities of `SAM-CLIP`, we evaluate it with zero-shot image classification on ImageNet [8], ImageNet-v2 [39] and Places365 [54], under image resolution of 224x. We use the text templates as CLIP [37] utilizing the textual embeddings from the text encoder of `SAM-CLIP` (which is kept frozen from our CLIP teacher) to perform zero-shot classification without any finetuning. The evaluation results are presented in Table 1. Employing a ViT-B architecture, our model achieves zero-shot accuracy comparable to the state-of-the-art CLIP ViT-B models pretrained on LAION-2B [43] and DataComp-1B [12] (both released by [18]), over the three datasets. These results validate the efficacy of our merging approach in inheriting CLIP's capabilities.

SAM Task: Zero-Shot Instance Segmentation. For the SAM component of `SAM-CLIP`, we evaluate its performance in instance segmentation, a task at which the original SAM model excels [20], with

(a) Input image (b) Ground-Truth (c) CLIP-head prediction (d) SAM-head refined

Figure 5: Demo on zero-shot semantic segmentation. Passing an input image through the image encoder, $\text{Head}_{\texttt{CLIP}}$ can predict a semantic segmentation mask, and $\text{Head}_{\texttt{SAM}}$ can refine it to a more fine-grained mask with auto-generated geometric prompts.

Table 1: Zero-shot evaluations on classification and instance segmentation tasks, comparing `SAM-CLIP` with state-of-the-art models that use the ViT-B architecture. `SAM-CLIP` demonstrates minimal forgetting compared to the baseline FMs on their original tasks.

Model	Training Data	0-Shot Classification (%)			0-Shot Instance Seg. (mAP)	
		ImageNet	**ImageNet-v2**	**Places-365**	**COCO**	**LVIS**
SAM [20]	SA-1B	-	-	-	41.2	36.8
CLIP [37]	OpenAI-400M	68.3	62.6	42.2	-	-
CLIP [7]	LAION-2B	71.1	61.7	43.4	-	-
CLIP [12]	DataComp-1B	73.5	65.6	43.0	-	-
`SAM-CLIP` (Ours)	Merged-41M	72.4	63.2	43.6	40.9	35.0

COCO [26] and LVIS [14] datasets. Following the original practices of [20], we first generate object detection bounding boxes using a ViT-Det model (ViT-B version) [23]. These bounding boxes act as geometric prompts for SAM's prompt encoder, which then predicts masks for each object instance. The evaluation results of `SAM-CLIP` and the original SAM ViT-B are provided in Table 1 (both under 1024px resolution), showing that `SAM-CLIP` is very close to SAM on the two benchmarks, not suffering from catastrophic forgetting during training.

Zero-Shot Transfer to Semantic Segmentation. We extend our evaluation to (text-prompted) zero-shot semantic segmentation over 5 datasets, Pascal VOC [10], Pascacl Context [33], ADE20k [55], COCO-Stuff [2] and COCO-Panoptic [19, 26]. We adopt a common evaluation protocol for this task: i) each input image is resized to 448×448px and pass to the image encoder and CLIP-head of `SAM-CLIP` to obtain 28×28 patch features; ii) OpenAI's 80 pre-defined CLIP text templates are employed to generate textual embeddings for each semantic class, and these embeddings act as mask prediction classifiers and operate on the patch features from the CLIP head; iii) we linearly upscale the mask prediction logits to match the dimensions of the input image. Evaluation results of `SAM-CLIP` and previous zero-shot models over the five datasets are demonstrated in Fig. 2. Notably, `SAM-CLIP` establishes new state-of-the-art performance on all 5 datasets, with a significant margin over past works.

A.2 Head-Probing Evaluations on Learned Representations

By merging the SAM and CLIP models, we anticipate that the resultant model will inherit advantages at the representation level from both parent models. Specifically, SAM excels at capturing low-level spatial visual details pertinent to segmentation tasks, while CLIP specializes in high-level semantic visual information encompassing the entire image. We hypothesize that the merged model combines these strengths, thereby enhancing its utility in broad range of downstream vision tasks. To

Table 2: Zero-shot semantic segmentation performance comparison with recent works. (†SegCLIP is trained on COCO data, so it is not zero-shot transferred to COCO-Stuff.)

Model	Arch	Training Data	0-Shot Semantic Segmentation (mIoU %)				
			Pascal VOC	Pascal-Context	ADE20k	COCO-Stuff	COCO-Panoptic
GroupViT [49]	ViT-S	Merged-26M	52.3	22.4	-	24.3	-
ViewCo [40]	ViT-S	Merged-26M	52.4	23.0	-	23.5	-
ViL-Seg [27]	ViT-B	CC12M	37.3	18.9	-	18.0	-
OVS [50]	ViT-B	CC4M	53.8	20.4	-	25.1	-
CLIPpy [38]	ViT-B	HQITP-134M	52.2	-	13.5	-	25.5
TCL [3]	ViT-B	CC3M+CC12M	51.2	24.3	14.9	19.6	-
SegCLIP [28]	ViT-B	CC3M+COCO	52.6	24.7	8.7	26.5†	-
SAM-CLIP	ViT-B	Merged-41M	**60.6**	**29.2**	**17.1**	**31.5**	**28.8**

Table 3: Head probing evaluations on semantic segmentation datasets, comparing our model with SAM and CLIP that use the ViT-B architecture. Avg is the average evaluation results of three heads.

Model	Training Data	Pascal VOC				ADE20k			
		Linear	DeepLabv3	PSPNet	Avg	Linear	DeepLabv3	PSPNet	Avg
SAM	SA-1B	46.6	69.9	71.2	62.6	26.6	32.8	36.2	31.9
CLIP	DataComp-1B	70.7	78.9	79.7	76.4	36.4	39.4	40.7	38.8
SAM-CLIP	Merged-41M	**75.0**	**80.3**	**81.3**	**78.8**	**38.4**	**41.1**	**41.7**	**40.4**

Table 4: Composing both CLIP and SAM heads of SAM-CLIP for zero-shot semantic segmentation on Pascal VOC.

Method	Resolution	mIoU
CLIP head only	448px	60.6
CLIP+SAM heads	1024px	**66.0**

Table 5: Linear probing evaluations on image classification datasets with ViT-B models.

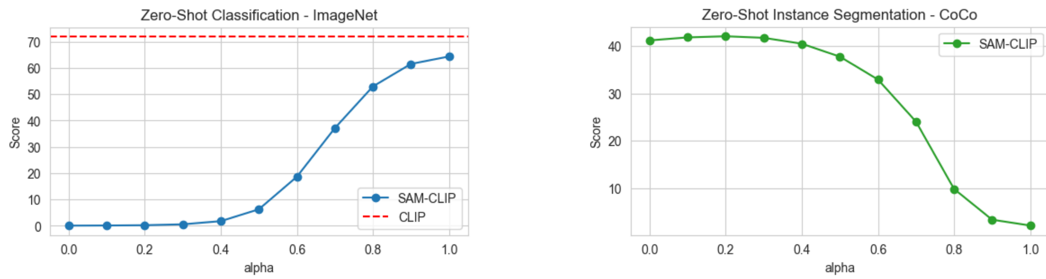
Model	**Linear Probing**	
	ImageNet	Places365
SAM	41.2	41.5
CLIP (DataComp1B)	**81.3**	55.1
CLIP (LAION-2B)	79.6	55.2
SAM-CLIP	80.5	**55.3**

investigate this hypothesis, we conduct head-probing (i.e., learn a task specific head with a frozen image backbone) evaluations on SAM, CLIP, and SAM-CLIP, utilizing different segmentation head structures (linear head, DeepLab-v3 [5] and PSPNet [53]) across two semantic segmentation datasets, Pascal VOC and ADE20k. The results are presented in Table 3. We observe that SAM representations do not perform as well as those of CLIP for tasks that require semantic understanding, even for semantic segmentation task. However, SAM-CLIP outperforms both SAM and CLIP across different head structures and datasets, thereby confirming its superior visual feature representation capabilities.

Besides, we apply linear probing to these models for image classification tasks on two datasets, ImageNet and Places365. Results in Table 5 show that SAM-CLIP attains comparable performance with CLIP, implying that the image-level representation of SAM-CLIP is also well-learned. All head probing evaluation results are visualized in Figure 3 to deliver messages more intuitively.

A.3 Composing Both CLIP and SAM Heads for Better Segmentation

Given that SAM-CLIP is a multi-task model with SAM and CLIP heads, one would naturally ask if the two heads can work together towards better performance on some tasks. Here, we showcase that a simple composition of the CLIP and SAM heads can lead to better zero-shot semantic segmentation. Specifically, we resize the input image to 1024px and pass it through $\mathrm{Enc}_{\texttt{SAM-CLIP}}$, and use the CLIP head to generate low-resolution mask prediction (32×32) using text prompts. Then, we generate some point prompts from the mask prediction (importance sampling based on the mask prediction confidence), and pass the mask prediction and point prompts together to the prompt encoder module as geometric prompts. Finally, $\mathrm{Head}_{\texttt{SAM}}$ takes embeddings from both the prompt encoder and the image encoder to generate high-resolution mask predictions (256×256) as shown in Figure 2 (right). Examples of this pipline are shown in Figure 5. One can clearly observe that the refined segmentation by the SAM-head is more fine-grained.



(a) Zero-Shot Accuracy (%)

(b) Zero-Shot Instance Segmentation (mAP)

Figure 6: Wise-FT [48] to a CLIP-distilled SAM ViT-B model. The red dashed line marks the performance of the CLIP teacher model.

Note that this pipeline requires *only one forward pass* on $\mathrm{Enc}_{\texttt{SAM-CLIP}}$ with 1024px resolution. For fair comparison, in Table 1 and Figure 1 we report `SAM-CLIP` zero-shot segmentation performance with 448px resolution using $\mathrm{Head}_{\texttt{CLIP}}$ only. Using our high-resolution pipeline we obtain further gain in zero-shot semantic segmentation as shown in Table 4.

B Weight Averaging

Weight averaging is a straightforward post-processing method proven to mitigate forgetting across a variety of fine-tuning tasks. Specifically, Wise-FT [48] proposes linearly interpolating the pretrained and fine-tuned parameters using a coefficient α. In this study, we explore the application of Wise-FT in our setup. We focus exclusively on CLIP distillation applied to SAM ViT-B (serving as the student model), with a CLIP ViT-B/16 model acting as the teacher model. The model is trained on ImageNet-21k for 20 epochs. It is evident that the fine-tuned student model ($\alpha = 1$) gains zero-shot classification capabilities at the expense of forgetting its original zero-shot instance segmentation abilities. Upon applying Wise-FT to the fine-tuned model, we observe an inherent tradeoff between learning and forgetting. Notably, no optimal point exists where both high classification accuracy ($> 60\%$ on ImageNet) and a high mAP (> 35 mAP on COCO) are achieved simultaneously.