A. Additional Implementation Details

Details of compared methods For CLIP and SLIP, we use the publicly available model and checkpoints provided by SLIP[1]. We reproduced MaskCLIP ourselves since it is not open-sourced and does not report results on YFCC15M. We follow the hyperparameters in the MaskCLIP paper: we mask out 75% of tokens in the mask self-distillation branch with a loss weight of 10.0; we start the moment of the EMA updated visual encoder from 0.999 and increase it linearly to 0.9999; we standardize the targets from the EMA encoder by a parameter-free Layer Norm; we set the learning rate to 5e-4, the batch size to 4,096, and the weight decay to 0.5.

We conduct all experiments on 4 nodes with 8 NVIDIA Tesla V100 GPUs each. We perform a speed test of different frameworks using a single node of 8 NVIDIA A100 GPUs to eliminate the effect of network conditions, which has been found to have very stable speed profiles for each framework.

B. Effects of stronger data augmentation

To combine more auxiliary self-superivise learning(SSL) tasks as mentioned in Section 3.2, we considered adding stronger data augmentation to learn robuster representation. In A-CLIP with or without SSL task, we tried with and without stronger data augmentation respectively.

The results in Table A1 demonstrate that the plain A-CLIP, which incorporates an attentive mask strategy into CLIP, benefits from the addition of the classic data augmentations of color+blur, leading to a +1.2% boost in ImageNet-1K zero-shot top-1 accuracy. This suggests that using attentive masking as an efficient data augmentation does not conflict with traditional methods like color and blur. Regarding the addition of SIMCLR, the results show that it relies more heavily on stronger data augmentations, as its performance is weaker when only random crop is added. However, after incorporating the BYOL task, an online-EMA SSL, the model's performance steadily improves regardless of the strength of the data augmentations. We believe that this is because the BYOL task enables the model to learn the output distribution of the complete image from the EMA encoder.

Methods	IN 1K 0-shot	
	crop	crop+color+blur
+attentive mask	41.3	42.5
+attentive mask+SimCLR	39.0	42.8
+attentive mask+SimCLR+BYOL	**41.9**	**43.9**

Table A1. Effects of different data augmentations for A-CLIP. Adding color+blur improves the performance of all settings, but more significantly after adding image self-supervised learning tasks.

[1]https://github.com/facebookresearch/SLIP

C. Effects of using EMA inference

A-CLIP uses an EMA vision encoder for inference in order to generate a stable attentive mask. In Table A2, we find that using EMA for evaluation leads to a stable performance gain, especially for mask training. Without mask, CLIP improves by +0.4% on ImageNet 1K zero-shot classification using EMA; with attentive mask training, it improves by +1.3%. We speculate that EMA alleviates the bias from the mask training.

Methods	IN 1K 0-shot	
	online	EMA
w/o mask	37.6	38.0(+0.4)
+random mask	38.0	39.1(+1.1)
+attentive mask	**40.0**	**41.3(+1.3)**

Table A2. Performance comparison of using online and EMA vision encoders for evaluation on IN 1K 0-shot classification with different mask methods. EMA improves performance more significantly with mask training than without mask.

D. More visualization results for attentive mask

Figure A1 shows more visualization results of attentive mask for A-CLIP. It can be seen that attentive mask always retains the text-related areas, while the deleted areas are more redundant and non-text-related parts, which is what our motivation wants to achieve.

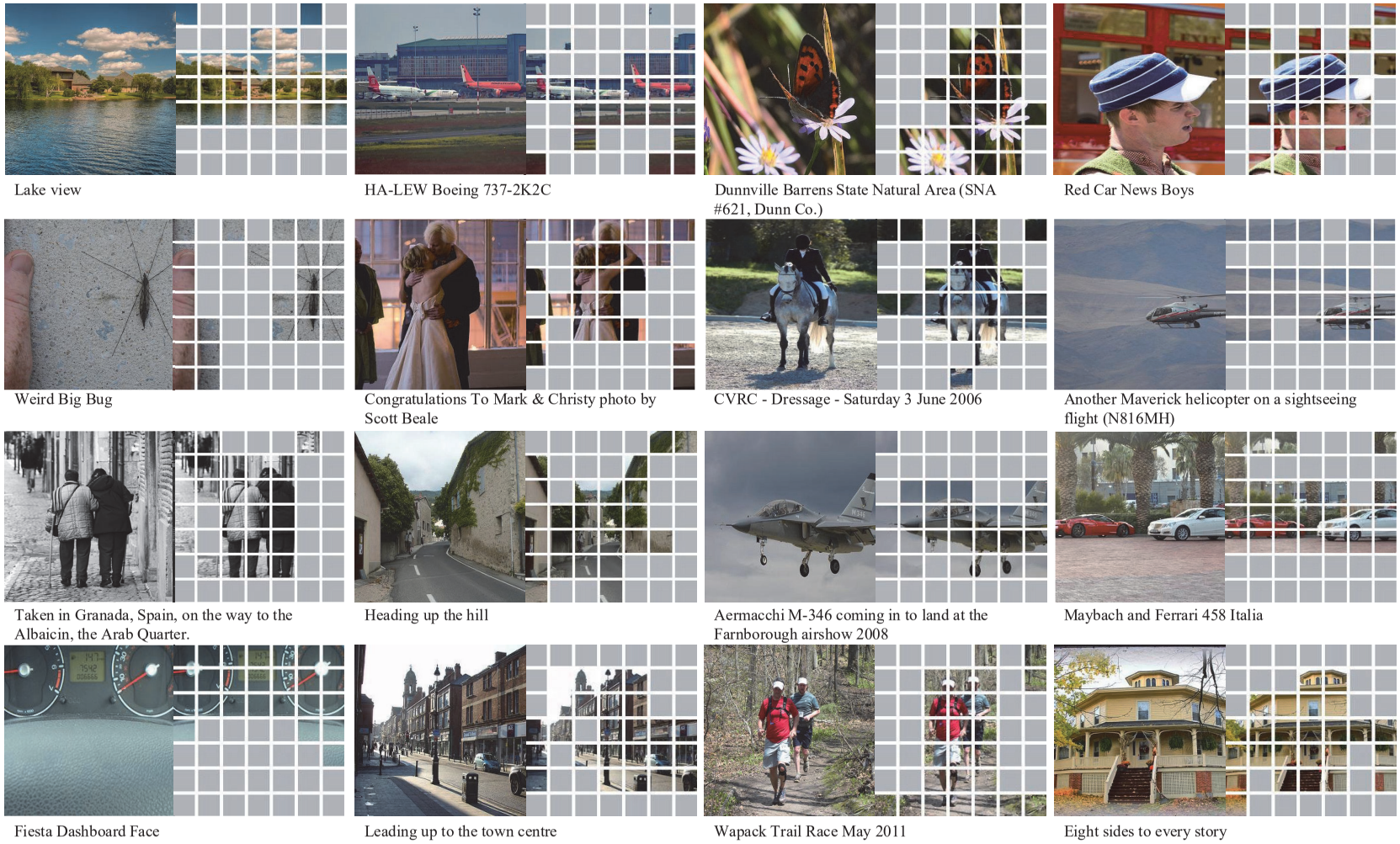

Figure A1. More visualization of attentive mask. Here we use a ViT-B16 model from A-CLIP's EMA vision encoder of to generate a mask with patch size of 32×32 and 50% mask ratio. The image and text used are sampled from YFCC-100M.