1 Training Loss & Valication Accuracy

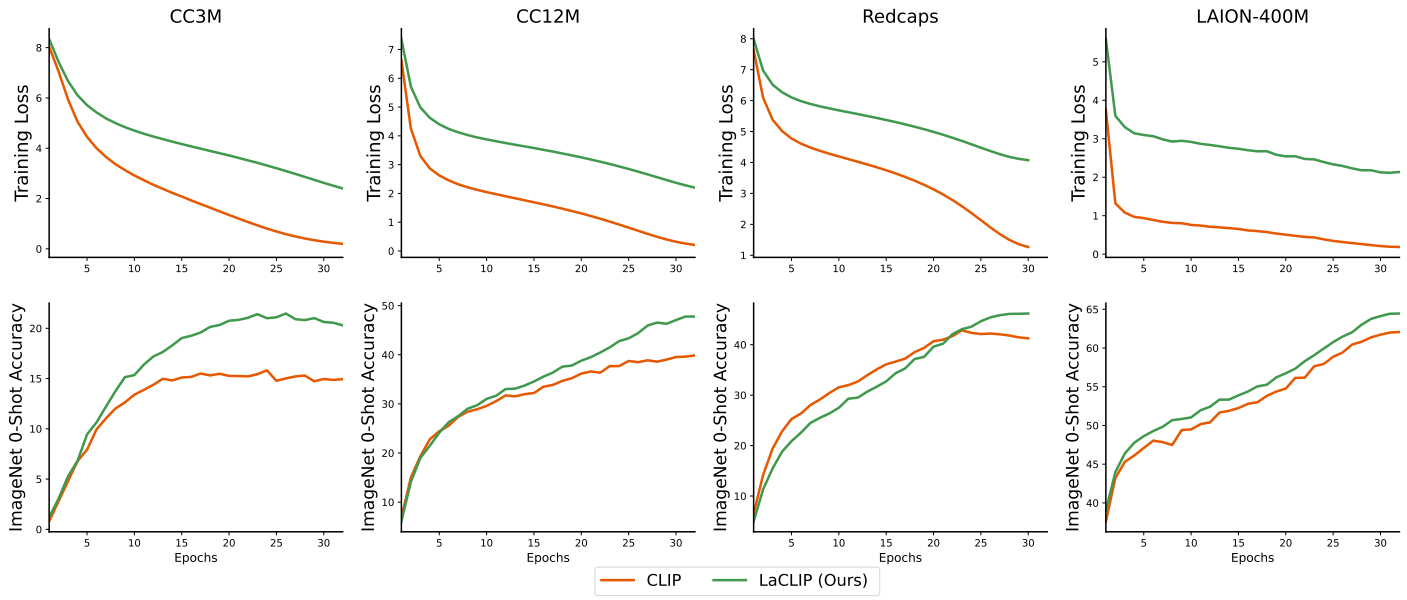

Figure 1: Training loss and validation accuracy for CLIP and LaCLIP trained on CC3M, CC12M, RedCaps and LAION-400M datasets. Top row is the training loss curve, bottom row is the validation accuracy, measured by zero-shot accuracy on ImageNet. Each column corresponds to a specific training dataset. X-axis is the training epoch for each figure. This result shows LaCLIP achieves higher validation accuracy and higher training loss, indicating LaCLIP improves CLIP generalization instead of optimization.

2 Text Encoder Sizes

Config	Layers	Width	Heads
Small	12	384	6
Base	12	512	8
Large	12	768	12

Table 1: Detailed text encoder configurations for different sizes. The one used in our main paper is Base. The text encoder size design follow the exact same set up as OpenCLIP.

Text Encoder	**Small**		**Base**		**Large**	
	DownStream	ImageNet	DownStream	ImageNet	DownStream	ImageNet
CLIP	39.5	40.7	38.8	40.2	39.0	40.1
LaCLIP	45.0	47.6	46.2	48.4	45.5	48.6

Table 2: Zero-shot performance comparison of CLIP and LaCLIP trained with different text encoder sizes on CC12M dataset. The vision encoder size is ViT-B/16. The results indicates changing the text encoder alone will not lead to significant performance changes, and LaCLIP outperforms CLIP significantly under all circumstances.

3 Visualizations of correct examples

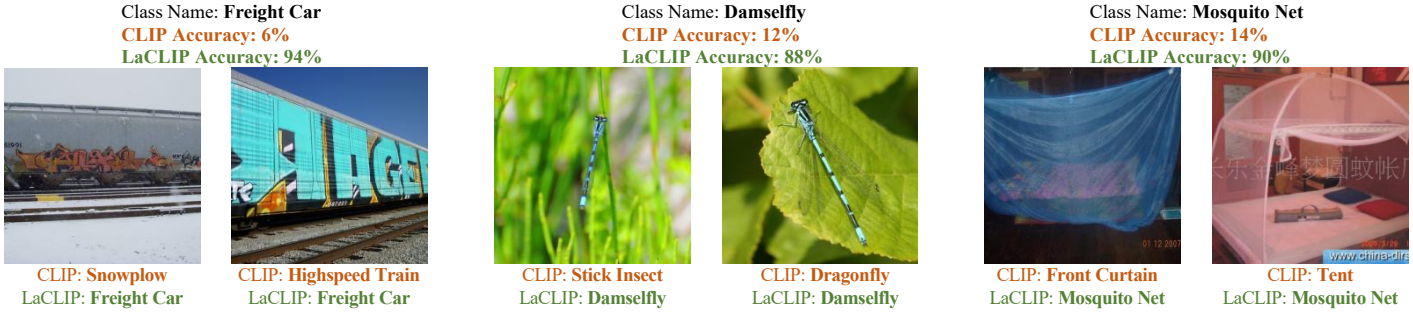

Figure 2: Visualization of examples corrected by LaCLIP on ImageNet validation set. We illustrate examples from the three categories with the most significant accuracy improvements.