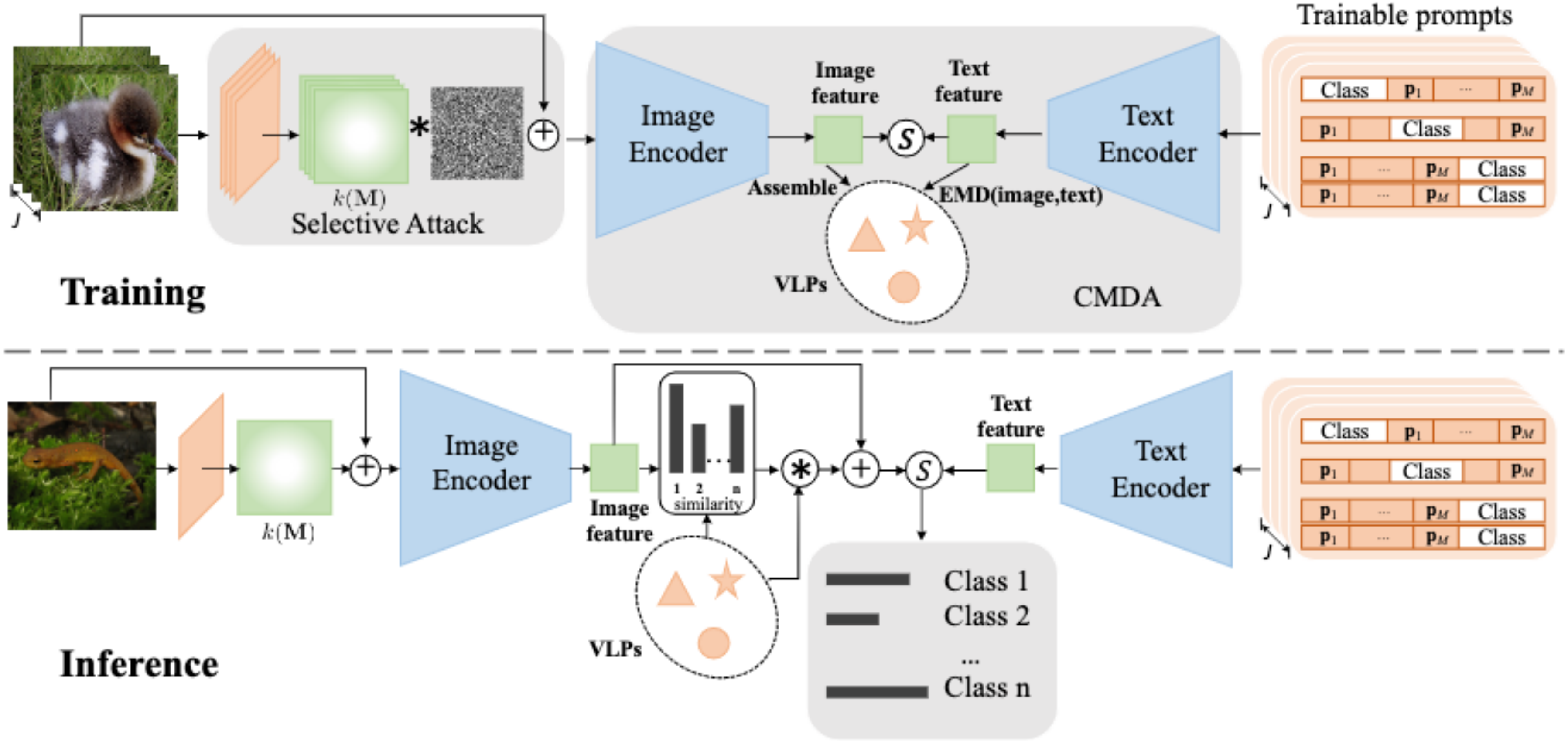
Few-Shot Learning with Visual Distribution Calibration and Cross-Modal Distribution Alignment

Wang Runqi, Zheng Hao, Duan Xiaoyue, Liu Jianzhuang,
Lu Yuning, Wang Tian, Xu Songcen, Zhang Baochang

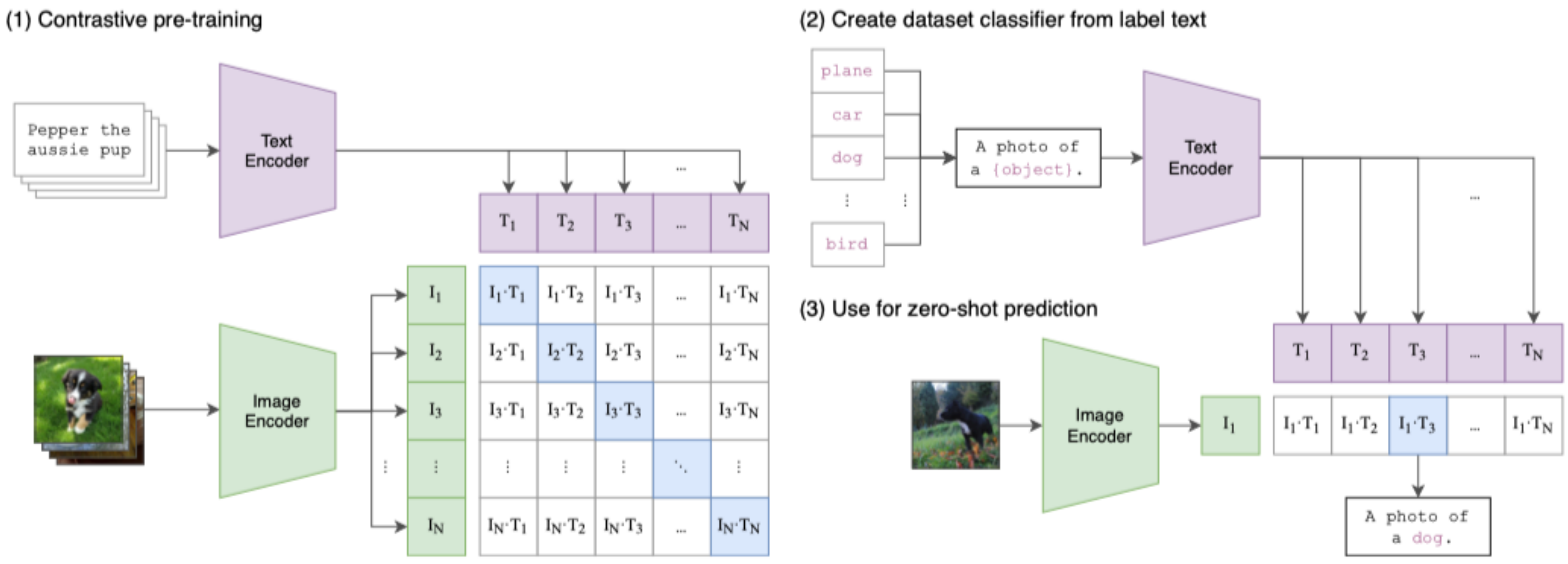
CVPR2023-2990 THU-PM-271

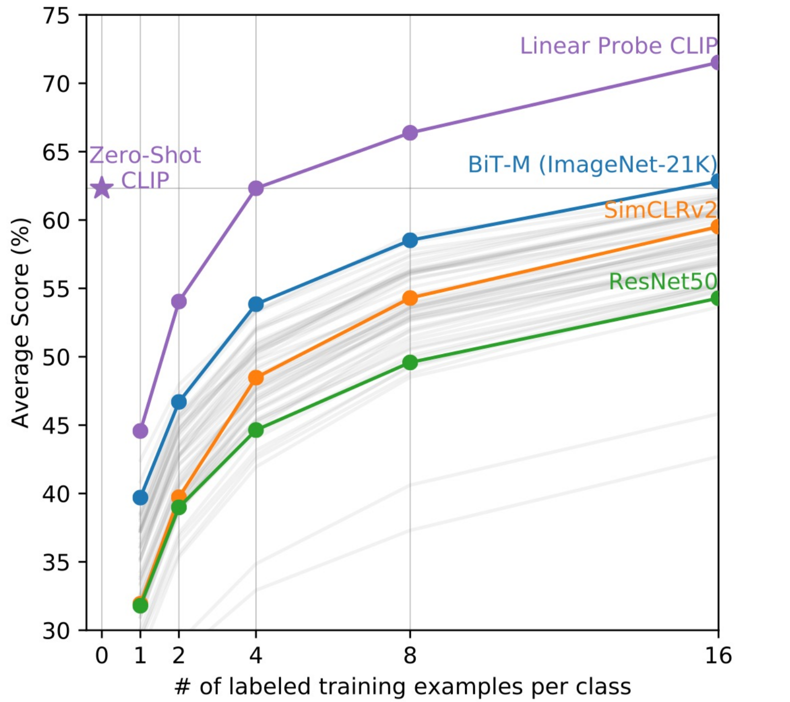
Abstract



Trainable prompts
Selective Attack
$k(\mathrm{M})$
Image Encoder
Image feature
Text feature
Text Encoder
Assemble
EMD(image,text)
VLPs
CMDA
Training
Class
$\mathbf{p}_1$
$\mathbf{p}_M$
Inference
similarity
Class 1
Class 2
...
Class n

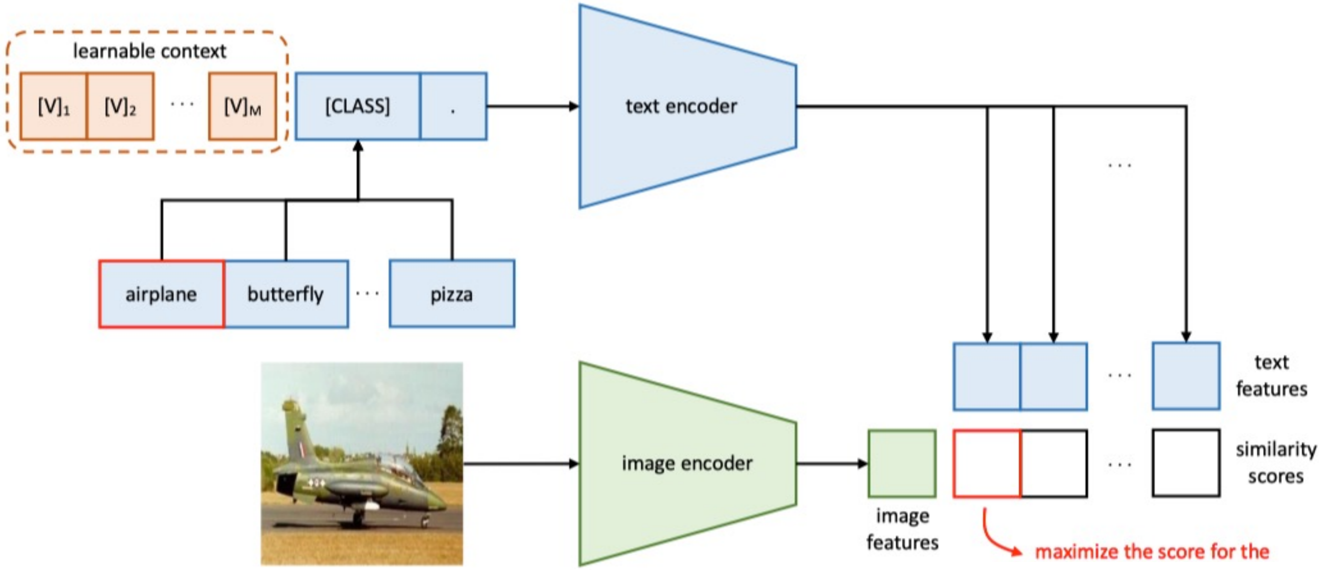
Background and Challenges



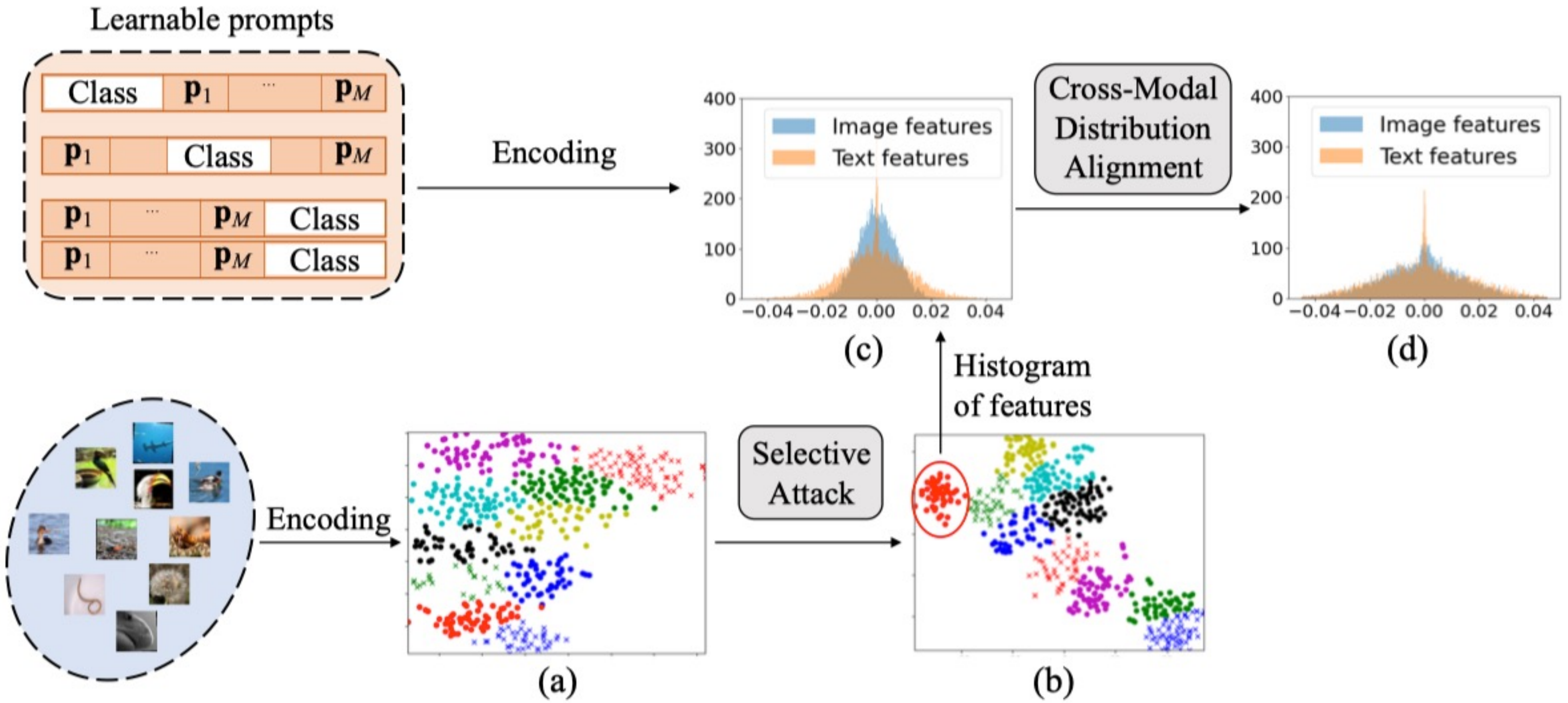


- The frame of CLIP are shown on the left and middle figures. The right figure is a comparison of the CLIP pre-training model with the few-shot learning results of the public models (the gray lines are other models in eval).
- The Zero-shot CLIP nearly matches the best results of the 16-shot linear classifier in the common model. The linear layer fine-tuned by only 16-shot can give the best results over BiT-M and SimCLRv2 by a wide margin.
- Twenty datasets, including ImageNet, were used in this analysis 。

Background and Challenges



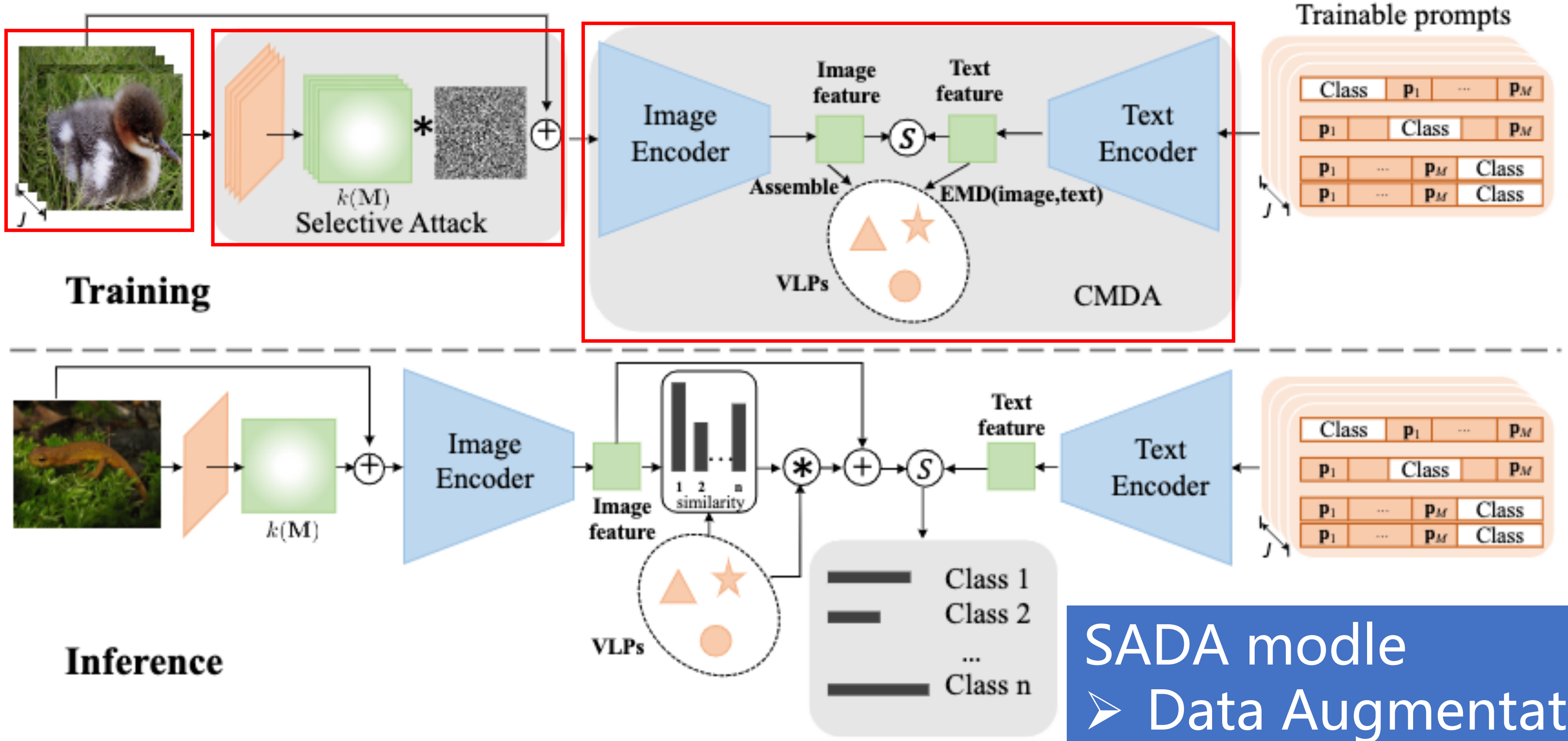
- Prompt tuning with few-hot samples. (CoOp,2022)



There are still deficiencies on CoOp:

- Few-shot learning tends to overfit.
- There are distribution gap between the modals

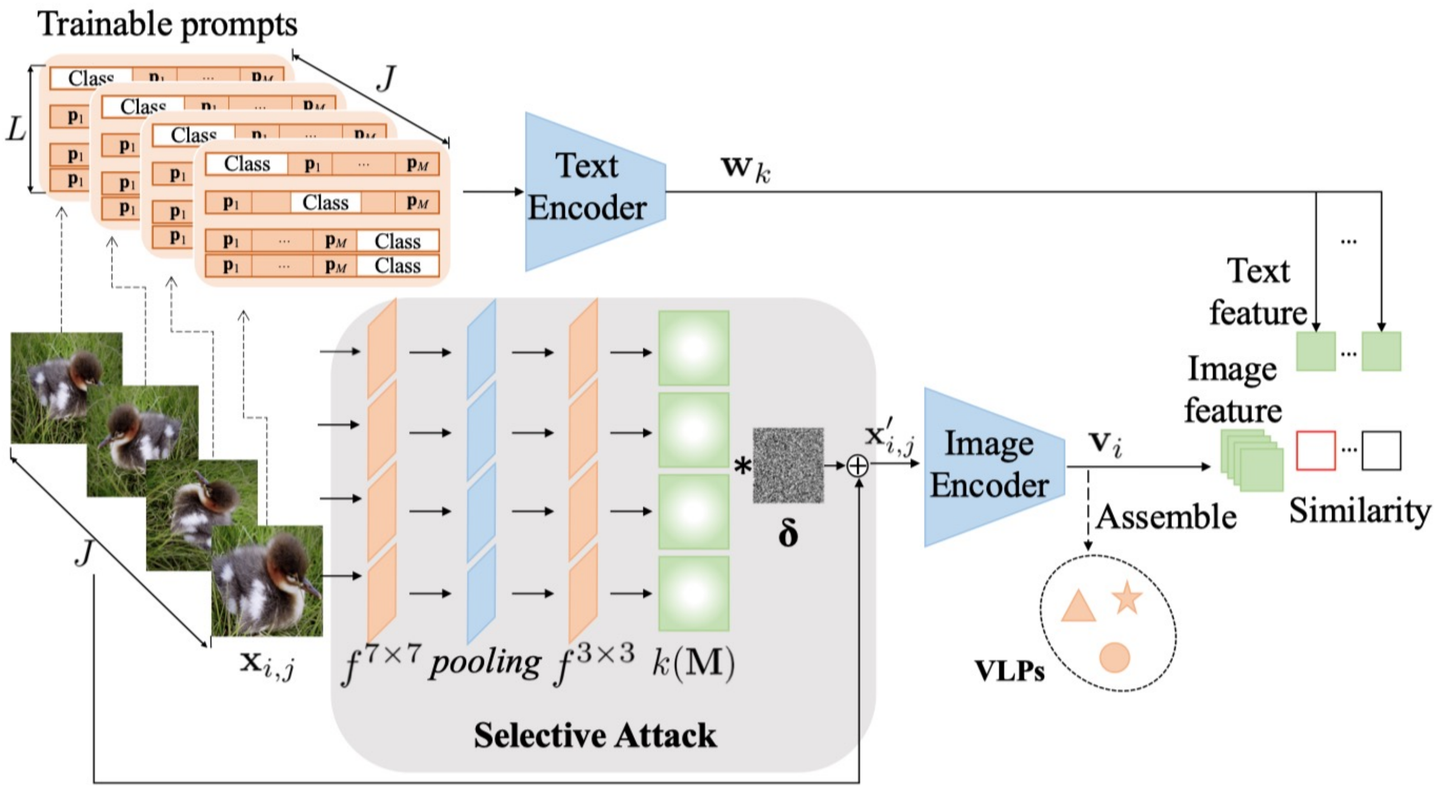
Method



SADA modle

- Data Augmentation
- Selective Attack (SA)
- Cross-Modal Distribution Alignment (CMDA)

Method——Augmentation Strategy



Step1

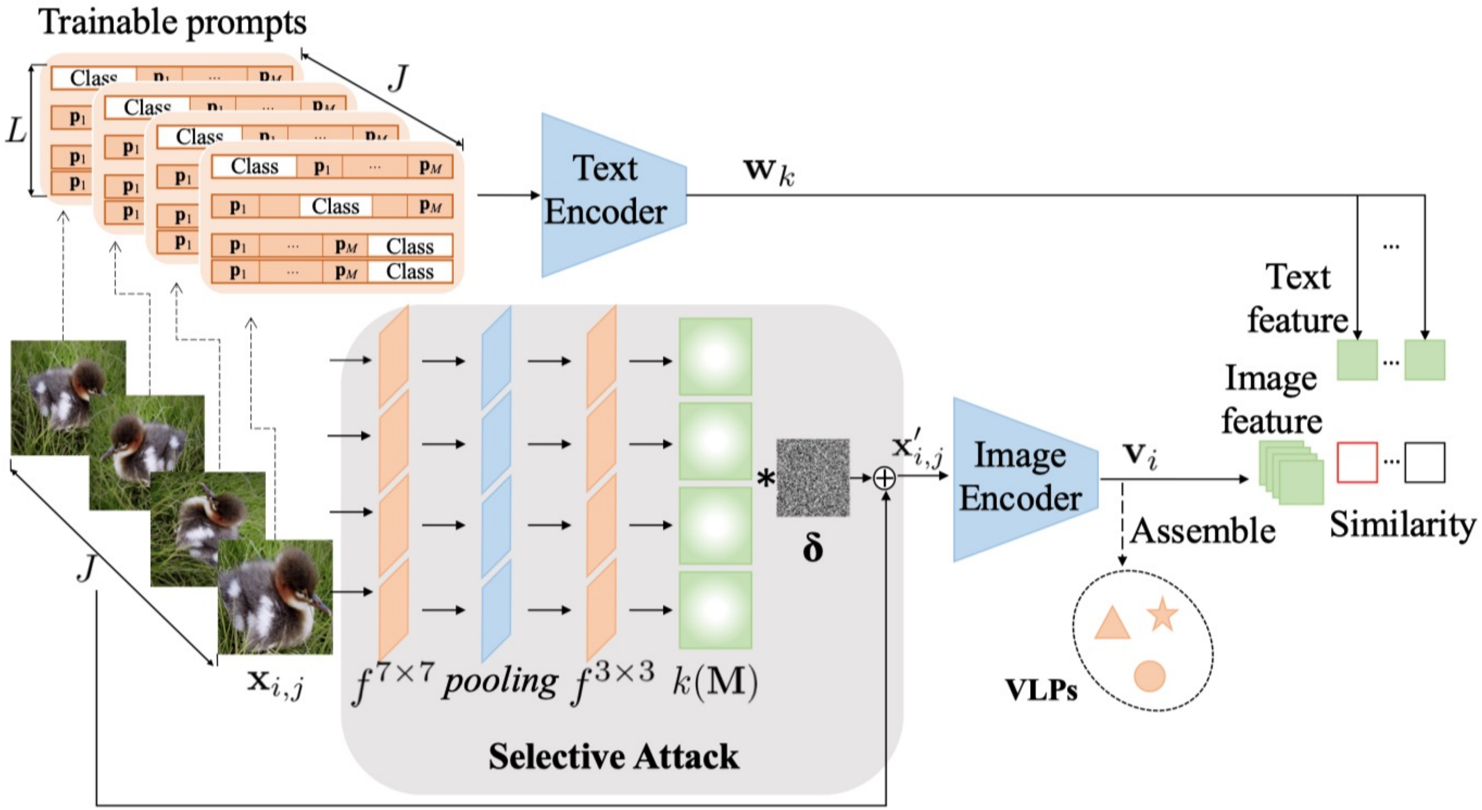
- ➢ The images and prompt were augmented in J groups, each containing L prompt and one image.

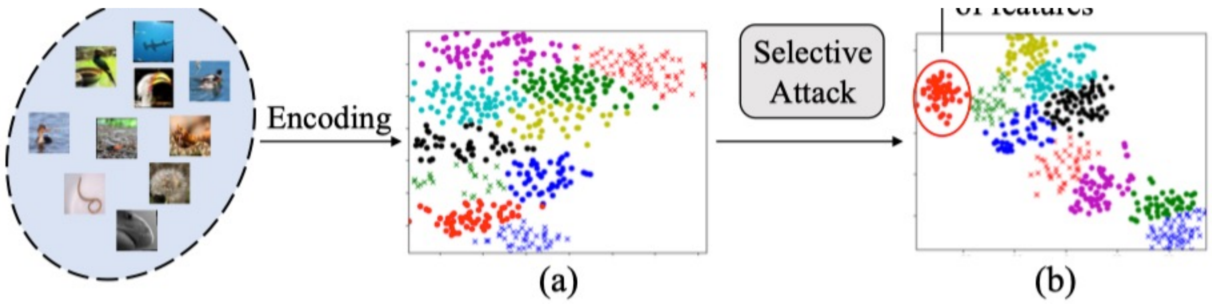
Step2

- ➢ Rotating,
- ➢ Flipping,
- ➢ Random gray scaling,
- ➢ Random cropping+Resizing,
- ➢ None,
- ➢ Color jittering,
- ➢ Gaussian blurring

From the above 7 augmentations, the most suitable J augmentations is searched

Method——Selective Attack（SA）





Step1

- ➢ Constructing spatial attention.

Step2

- ➢ Spatial attention guides perturbation to selective attack.

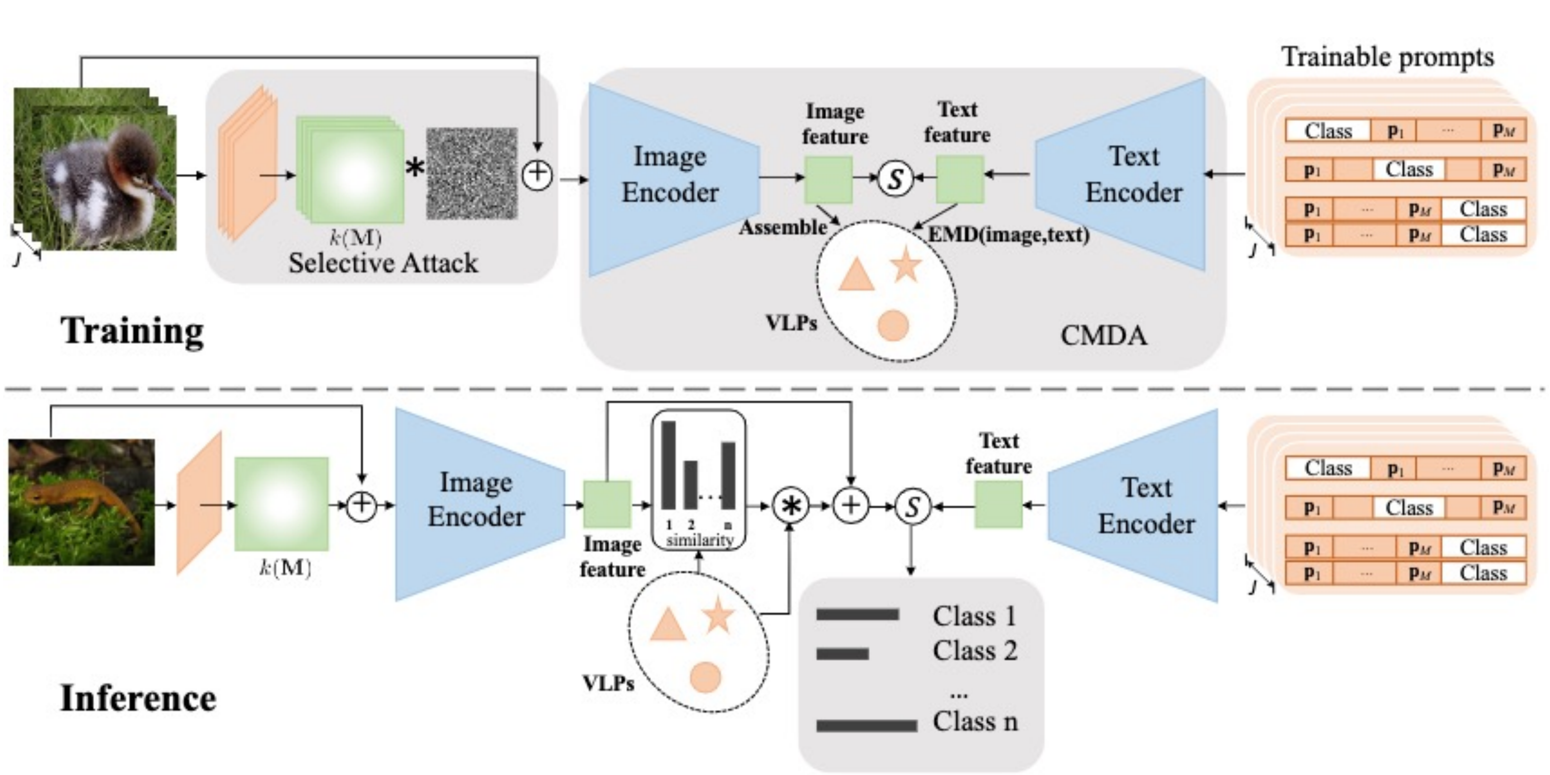
$$p(y_i|\mathbf{x}_{i,j}) = \frac{e^{\langle \mathbf{z}_{i,j},\ \sum_l g(\mathbf{t}_{y_i}(\mathbf{P}_{l,j}))/L\rangle/\tau}}{\sum_{k=1}^{K} e^{\langle \mathbf{z}_{i,j},\ \sum_l g(\mathbf{t}_k(\mathbf{P}_{l,j}))/L\rangle/\tau}},$$

$$\mathbf{F}_{i,j} = \varphi(f_j^{7\times7}(\mathbf{x}_{i,j}))$$

$$\mathbf{M}_{i,j} = \varphi(f_j^{3\times3}([\mathbf{F}_{i,j}^{avg}, \mathbf{F}_{i,j}^{max}]))$$

$$\begin{aligned}\mathbf{x}'_{i,j} &= \mathbf{x}_{i,j} + k(\mathbf{M}_{i,j}) \circ \boldsymbol{\delta} \\ &= \mathbf{x}_{i,j} + (1 - \mathbf{M}_{i,j} \circ \mathbf{M}_{i,j}) \circ \boldsymbol{\delta}\end{aligned}$$

Method——Cross-Modal Distribution Alignment（CMDA）



Step1

➢ The Vision-Language Prototype (VLP) was built through EMD.

$$\mathcal{L}_{\text{EMD}} \triangleq \sum_k \left(\|\boldsymbol{\mu}_{\text{v}}^k - \boldsymbol{\mu}_{\text{w}}^k\|^2 + \|\boldsymbol{\Sigma}_{\text{v}}^{k\,\frac{1}{2}} - \boldsymbol{\Sigma}_{\text{w}}^{k\,\frac{1}{2}}\|^2\right)$$

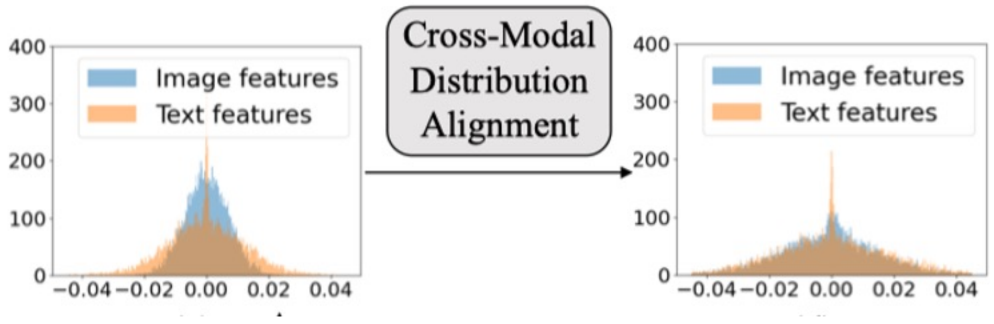
Step2

➢ Inference period: VLPs corrects image feature to realize CMDA.

$$\mathbf{d} = (d_1, d_2, \ldots, d_K)^T,\ d_k = \frac{1}{\|\mathbf{z}_i - \mathbf{v}_k\|},$$

$$\bar{\mathbf{d}} = (\bar{d_1}, \bar{d_2}, \ldots, \bar{d_K})^T,\ \bar{d_k} = \frac{d_k}{\sum_{m=1}^{K} d_m},$$

$$p(y_i|\mathbf{x}_i) = \frac{e^{\langle (1-\alpha)\mathbf{z}_i + \alpha(\bar{\mathbf{d}}^T \mathbf{VLP})^T,\ \sum_l g(\mathbf{t}_{y_i}(\mathbf{P}_{l,j}))/L\rangle/\tau}}{\sum_{k=1}^{K} e^{\langle (1-\alpha)\mathbf{z}_i + \alpha(\bar{\mathbf{d}}^T \mathbf{VLP})^T,\ \sum_l g(\mathbf{t}_k(\mathbf{P}_{l,j}))/L\rangle/\tau}}$$



Expeiments

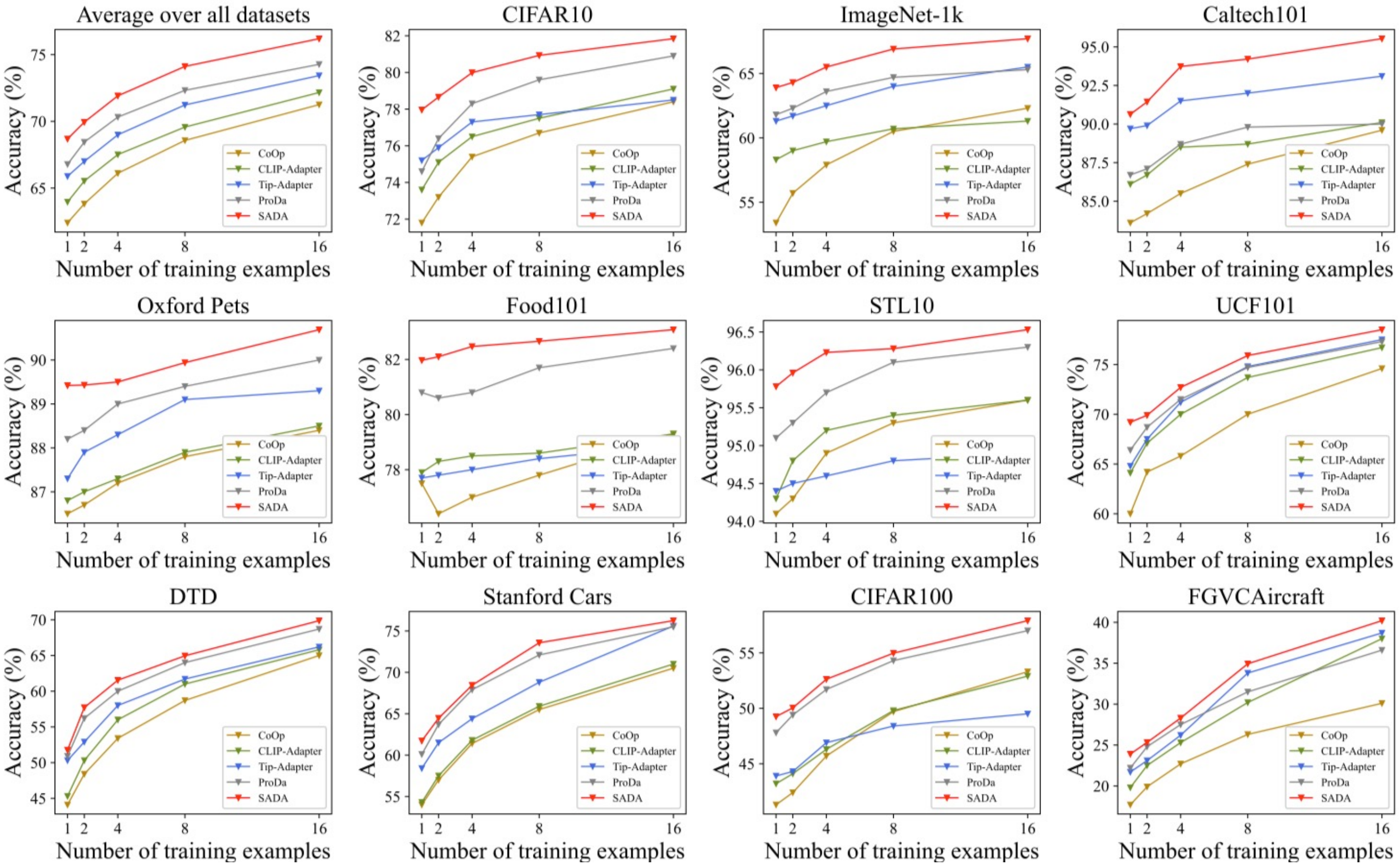


Figure 1. Main results of few-shot learning on 11 datasets. Our SADA consistently shows better performance than prior arts across different number of training samples.

Table 1. Effect of the augmentation on prompt diversity.

Group j	Mean				Std
	1	2	3	4	
SADA w/o Aug	0.0954	0.0923	0.0892	0.0998	0.0045
SADA	0.1035	0.1441	0.0855	0.1173	**0.0247**

Table 2. Ablation of SA and CMDA on CIFAR10.

#Shots	1	2	4	8	16
Baseline	74.61%	76.40%	78.34%	79.63%	80.90%
Baseline w SA	77.61%	78.2%	79.63%	80.53%	81.38%
Baseline w CMDA	76.79%	77.37%	79.02%	80.15%	81.31%

Table 3. Ablation on the objective function to optimize the VLPs.

#Shots	1	2	4	8	16
EMD	**76.7%**	**77.3%**	**79.0%**	**80.1%**	**81.3%**
MMD	73.5%	76.1%	77.4%	79.7%	80.5%
JS-Divergence	74.3%	75.9%	77.6%	79.2%	80.1%

Ablations

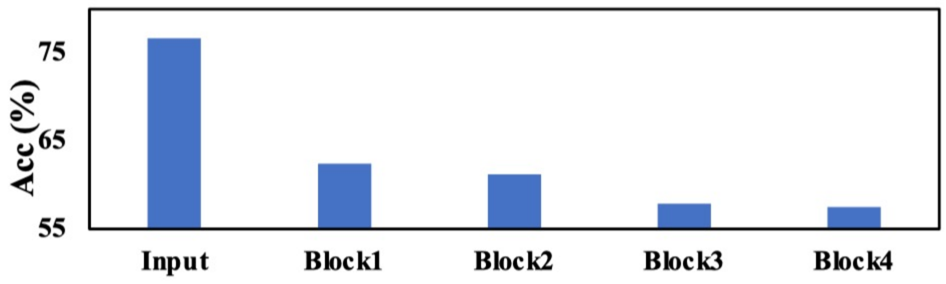


Figure 7. 1-shot accuracy (%) on CIFAR10 when SA is at different layers of the image encoder.

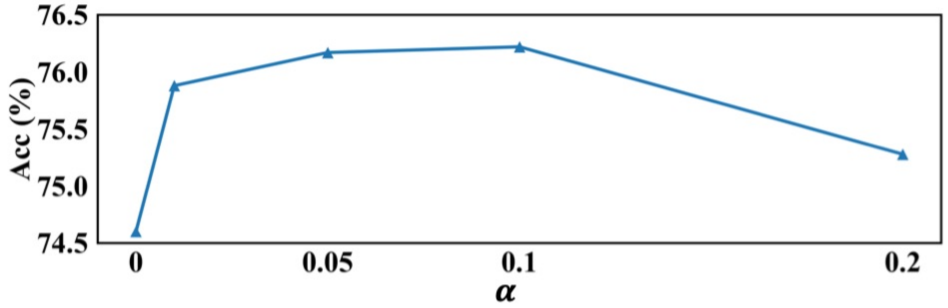


Figure 8. 1-shot accuracy (%) of different calibration ratio α.

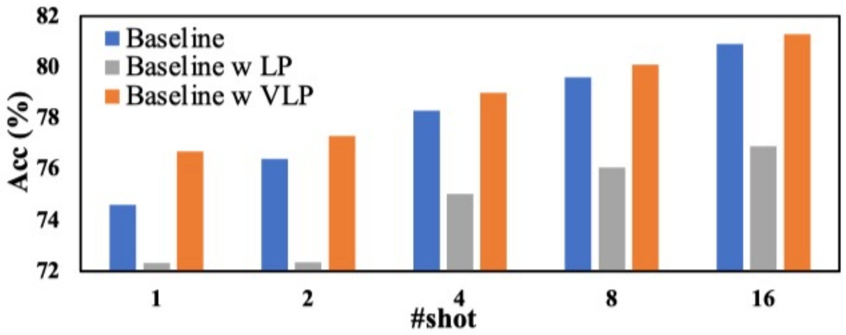


Figure 9. Effect of VLPs on CIFAR10.

Figure 10. Visualization of attacked areas (in red) guided by $1 - \mathbf{M} \circ \mathbf{M}$. The images are from ImageNet-1k.