I can not believe there is no training!

Paper Reading by Zhiying Lu

2023.06.05

智能多媒体内容计算实验室
Intelligent Multimedia Content Computing Lab

Tip-Adapter: Training-free Adaption of CLIP for Few-shot Classification

Renrui Zhang[*1,2], Wei Zhang[*1], Rongyao Fang[2], Peng Gao[†1], Kunchang Li[1], Jifeng Dai[3], Yu Qiao[1], and Hongsheng Li[2,4]

[1] Shanghai AI Laboratory
[2] The Chinese University of Hong Kong
[3] SenseTime Research
[4] Centre for Perceptual and Interactive Intelligence (CPII)
{zhangrenrui, gaopeng, qiaoyu}@pjlab.org.cn, hsli@ee.cuhk.edu.hk

ECCV 2022

SuS-X: Training-Free Name-Only Transfer of Vision-Language Models

ICCV 2022

Vishaal Udandarao
University of Cambridge
vu214@cam.ac.uk

Ankush Gupta
DeepMind, London
ankushgupta@google.com

Samuel Albanie
University of Cambridge
sma71@cam.ac.uk

- 作者介绍
- 研究背景
- Tip-Adapter
- SuS-X
- 总结

作者介绍

Tip-Adapter: Training-free Adaption of CLIP for Few-shot Classification

Renrui Zhang*[1,2], Wei Zhang*[1], Rongyao Fang[2], Peng Gao†[1], Kunchang Li[1], Jifeng Dai[3], Yu Qiao[1], and Hongsheng Li[2,4]

[1] Shanghai AI Laboratory
[2] The Chinese University of Hong Kong
[3] SenseTime Research
[4] Centre for Perceptual and Interactive Intelligence (CPII)
{zhangrenrui, gaopeng, qiaoyu}@pjlab.org.cn, hsli@ee.cuhk.edu.hk

Renrui Zhang

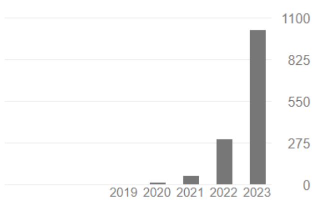
MMLab CUHK & Peking University
在 pku.edu.cn 的电子邮件经过验证 - 首页
Computer Vision　Deep Learning

关注

标题	引用次数	年份
CLIP-Adapter: Better Vision-language Models with Feature Adapters P Gao*, S Geng*, R Zhang*, T Ma, R Fang, Y Zhang, H Li, Y Qiao IJCV 2023	252	2021
Tip-Adapter: Training-free Adaption of CLIP for Few-shot Classification R Zhang, W Zhang, R Fang, P Gao, K Li, J Dai, Y Qiao, H Li ECCV 2022	195 *	2022
End-to-end Object Detection with Adaptive Clustering Transformer M Zheng, P Gao, R Zhang, X Wang, H Li, H Dong BMVC 2021 Oral	162	2020
PointCLIP: Point Cloud Understanding by CLIP R Zhang, Z Guo, W Zhang, K Li, X Miao, B Cui, Y Qiao, P Gao, H Li CVPR 2022	133	2022
LLaMA-Adapter: Efficient Fine-tuning of Language Models with Zero-init Attention R Zhang, J Han, A Zhou, X Hu, S Yan, P Lu, H Li, P Gao, Y Qiao arXiv preprint arXiv:2303.16199	7	

引用次数

	总计	2018 年至今
引用	1400	1397
h 指数	17	17
i10 指数	26	26



开放获取的出版物数量　查看全部

Hongsheng Li (李鸿升)

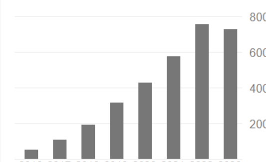
Associate Professor at The Chinese University of Hong Kong
在 ee.cuhk.edu.hk 的电子邮件经过验证 - 首页
Computer Vision　Machine Learning　Medical Image Analysis

关注

标题	引用次数	年份
StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks H Zhang, T Xu, H Li, S Zhang, X Huang, X Wang, D Metaxas IEEE Int. Conf. Comput. Vision (ICCV), 5907-5915	3029	2017
PointRCNN: 3D object proposal generation and detection from point cloud S Shi, X Wang, H Li Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern ...	1911	2019
Cross-scene crowd counting via deep convolutional neural networks C Zhang, H Li, X Wang, X Yang Proceedings of the IEEE Conference on Computer Vision and Pattern ...	1315	2015
PV-RCNN: Point-voxel feature set abstraction for 3D object detection S Shi, C Guo, L Jiang, Z Wang, J Shi, X Wang, H Li Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern ...	1279	2020
StackGAN++: Realistic Image Synthesis with Stacked Generative Adversarial Networks H Zhang, T Xu, H Li, S Zhang, X Wang, X Huang, DN Metaxas IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE 41 (8), 1947-1962	1269	2019

引用次数　查看全部

	总计	2018 年至今
引用	32678	30112
h 指数	82	78
i10 指数	191	181



开放获取的出版物数量　查看全部

7 篇文章　128 篇文章
无法查看的文章　可查看的文章
根据资助方的强制性开放获取政策

作者介绍

***SuS-X*: Training-Free Name-Only Transfer of Vision-Language Models**

Vishaal Udandarao
University of Cambridge
vu214@cam.ac.uk

Ankush Gupta
DeepMind, London
ankushgupta@google.com

Samuel Albanie
University of Cambridge
sma71@cam.ac.uk

Vishaal Udandarao

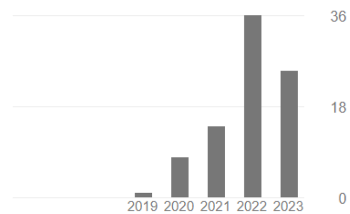
FOLLOW

PhD Student, University of Tübingen & University of Cambridge
Verified email at cam.ac.uk - Homepage
Deep Learning Natural Language Processing Computer Vision

TITLE	CITED BY	YEAR
Cobra: Contrastive bi-modal representation algorithm V Udandarao, A Maiti, D Srivatsav, SR Vyalla, Y Yin, RR Shah arXiv preprint arXiv:2005.03687	23	2020
EDUQA: Educational domain question answering system using conceptual network mapping A Agarwal, N Sachdeva, RK Yadav, V Udandarao, V Mittal, A Gupta, ... ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and ...	23	2019
Sus-x: Training-free name-only transfer of vision-language models V Udandarao, A Gupta, S Albanie arXiv preprint arXiv:2211.16198		
Memeify: A large-scale meme generation system SR Vyalla, V Udandarao Proceedings of the 7th ACM IKDD CoDS and 25th COMAD, 307-311		

Cited by

	All	Since 2018
Citations	85	85
h-index	6	6
i10-index	2	2



Samuel Albanie

University of Cambridge
0000-0003-1732-9198
samuelalbanie.com

Publications	76
h-index	23
Citations	20,217
Highly Influential Citations	1,838

Follow Author...

Author pages are created from data sourced from our academic... show more

Publications | Citing Authors | Referenced Authors | Co-Authors

Search authors, put | Co-Author | Has PDF | More Filters | Sort by Most Influe...

Squeeze-and-Excitation Networks
Jie Hu, Li Shen, Samuel Albanie, Gang Sun, E. Wu · Computer Science ·
IEEE/CVF Conference on Computer Vision and... · 5 September 2017

TLDR This work proposes a novel architectural unit, which is term the "Squeeze-and-Excitation" (SE) block, that adaptively recalibrates channel-wise feature responses by explicitly modelling interdependencies between channels and finds that SE blocks produce significant performance improvements for existing state-of-the-art deep architectures at minimal additional computational cost. Expand

17,124 · 1,480 · PDF · IEEE · Save · Alert · Cite

BLOOM: A 176B-Parameter Open-Access Multilingual Language Model
Teven Le Scao, Angela Fan, +388 authors Thomas Wolf · Computer Science · arXiv.org · 9 November 2022

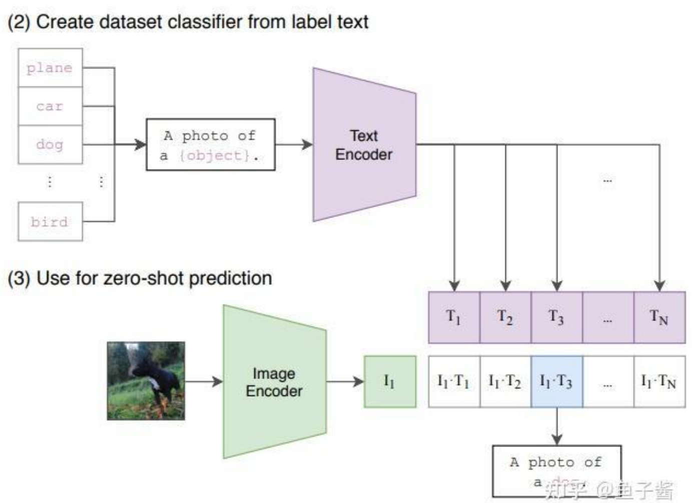
TLDR BLOOM is a 176B-parameter open-access language model designed and built thanks to a collaboration of hundreds of researchers and achieves competitive performance on a wide variety of benchmarks, with stronger results after undergoing multitask prompted finetuning. Expand

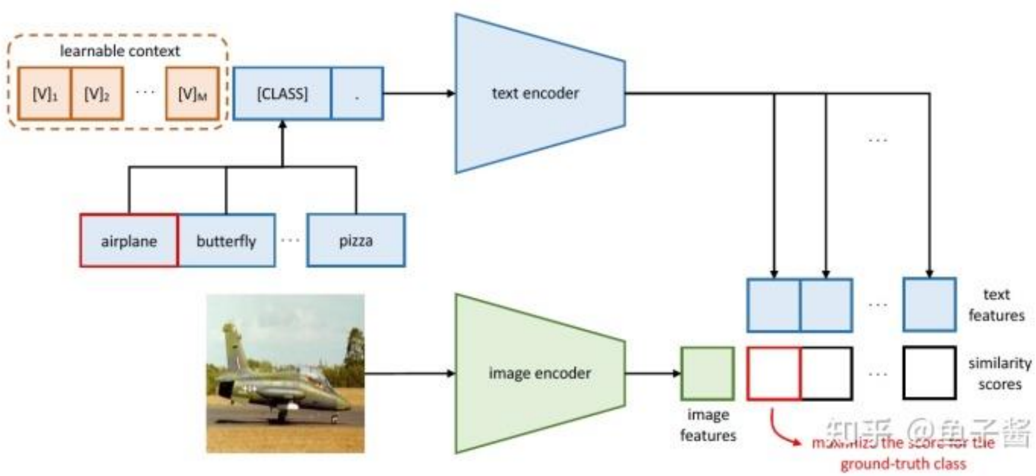
615 · 91 · PDF · arXiv · Save · Alert · Cite

- 作者介绍
- 研究背景
- Tip-Adapter
- SuS-X
- 总结

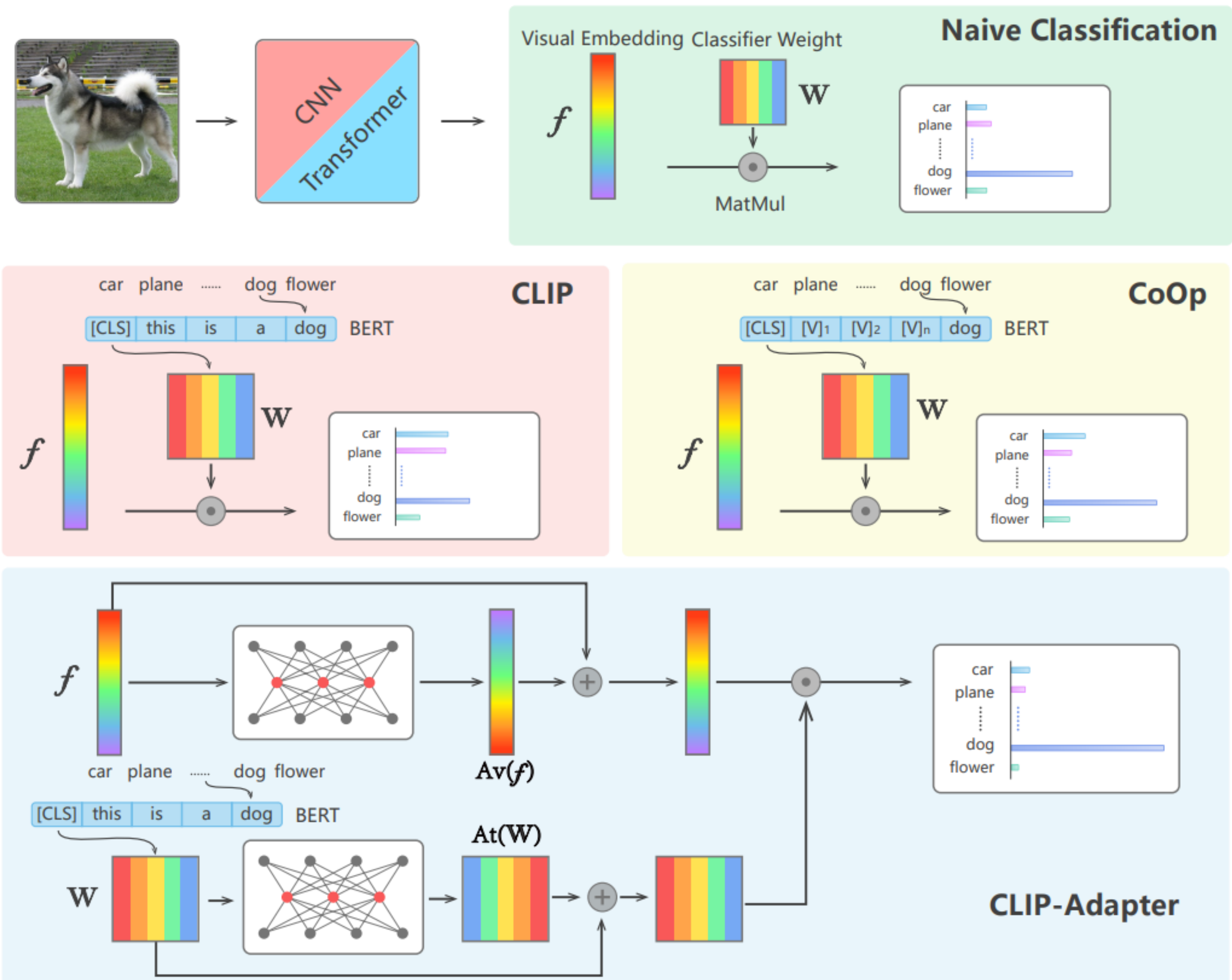
研究背景

- 具有零样本识别潜力的多模态CLIP模型成为主流
- Parameter-Efficient-Finetuning 成为利用大模型适配下游任务的范式





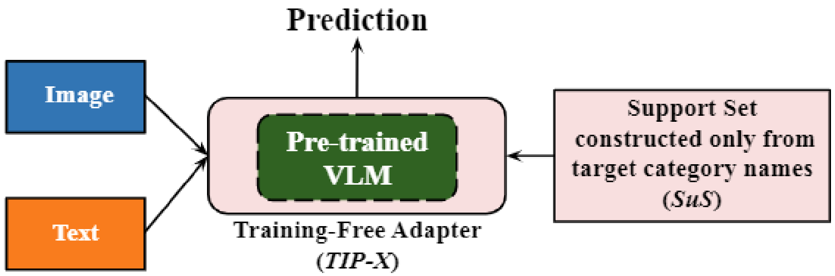
CLIP-Adapter



Tip-Adapter

Models	Training	Epochs	Time	Accuracy	Gain	Infer. Speed	GPU Mem.
Zero-shot CLIP [48]	Free	0	0	60.33	0	10.22ms	2227MiB
Linear-probe CLIP [48]	Required	-	13min	56.13	−4.20	-	-
CoOp [73]	Required	200	14h 40min	62.95	+2.62	299.64ms	7193MiB
CLIP-Adapter [16]	Required	200	50min	63.59	+3.26	10.59ms	2227MiB
Tip-Adapter	**Free**	**0**	**0**	62.03	+1.70	10.42ms	2227MiB
Tip-Adapter-F	Required	20	5min	**65.51**	+5.18	10.53ms	2227MiB

- 以往的方法还需要大量资源进行微调
- Tip-Adapter方法无需进行任务微调，直接即插即用



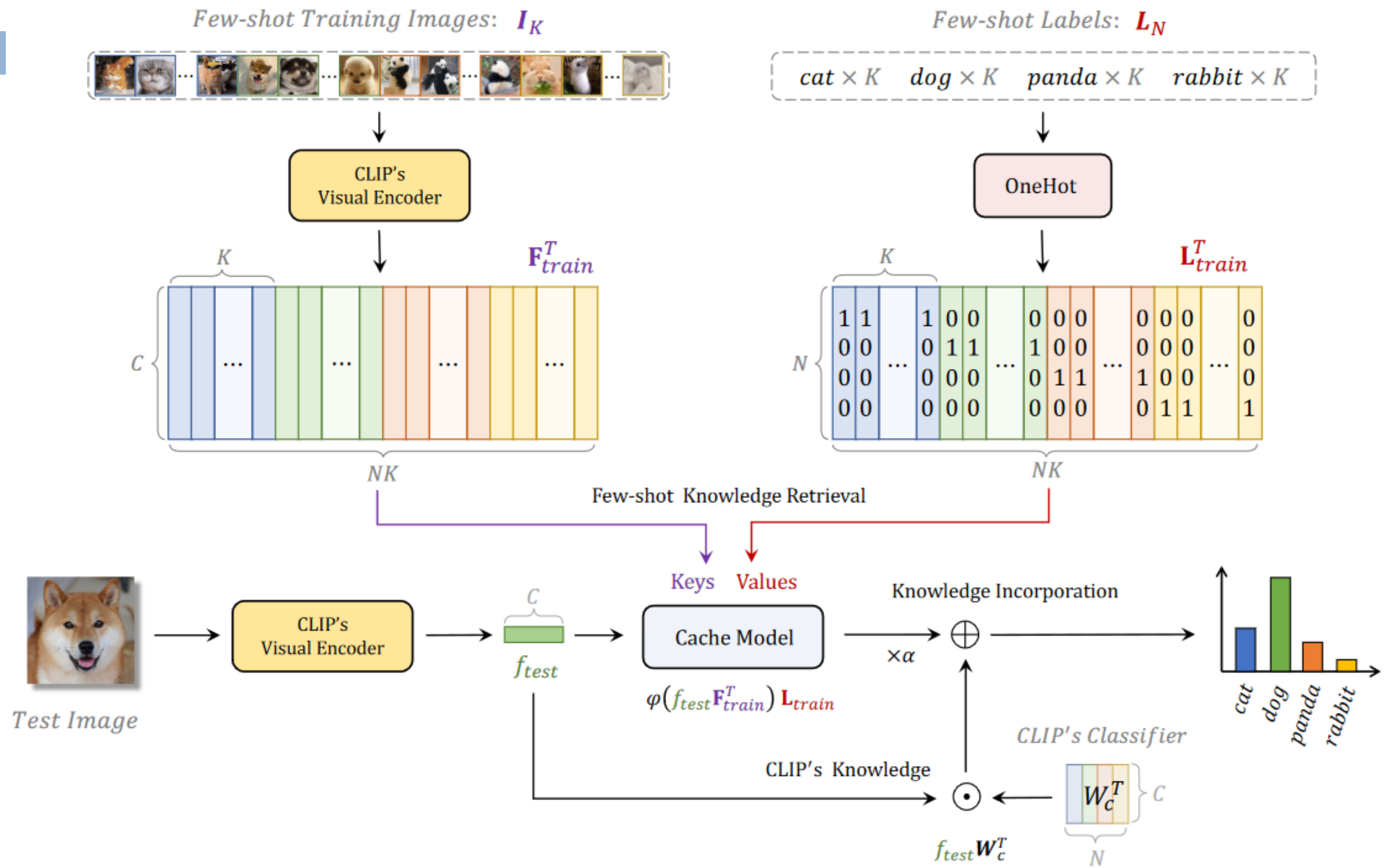
SuS-X

Table 1: **Taxonomy of CLIP adaptation methods for downstream classification.** We underline the Zero-Shot CLIP model to signify that it is the base model that all others build on top of. *This method considers access to all test-set samples simultaneously, hence we still consider it zero-shot. †This method additionally uses class hierarchy maps.

	Method	Does not require training	Does not require labelled data	Does not require target data distribution
Few-shot fine-tuning methods	LP-CLIP [61]	✗	✗	✗
	CoOp [88]	✗	✗	✗
	PLOT [12]	✗	✗	✗
	LASP [10]	✗	✗	✗
	SoftCPT [21]	✗	✗	✗
	VT-CLIP [83]	✗	✗	✗
	VPT [19]	✗	✗	✗
	ProDA [49]	✗	✗	✗
	CoCoOp [87]	✗	✗	✗
	CLIP-Adapter [28]	✗	✗	✗
Intermediate methods	TIP-Adapter [84]	✓	✗	✗
	UPL [40]	✗	✓	✗
	SVL-Adapter [58]	✗	✓	✗
	TPT [52]	✗	✓	✓
	CLIP+SYN [36]	✗	✓	✓
	CaFo [82]	✗	✓	✓
Zero-shot methods	<u>Zero-Shot CLIP</u> [61]	✓	✓	✓
	CALIP [34]	✓	✓	✓
	CLIP+DN [89]*	✓	✓	✓
Training-free name-only transfer methods	CuPL [60]	✓	✓	✓
	VisDesc [53]	✓	✓	✓
	CHiLS [57]†	✓	✓	✓
	SuS-X (ours)	✓	✓	✓

- 作者介绍
- 研究背景
- Tip-Adapter
- SuS-X
- 总结

Tip-Adapter



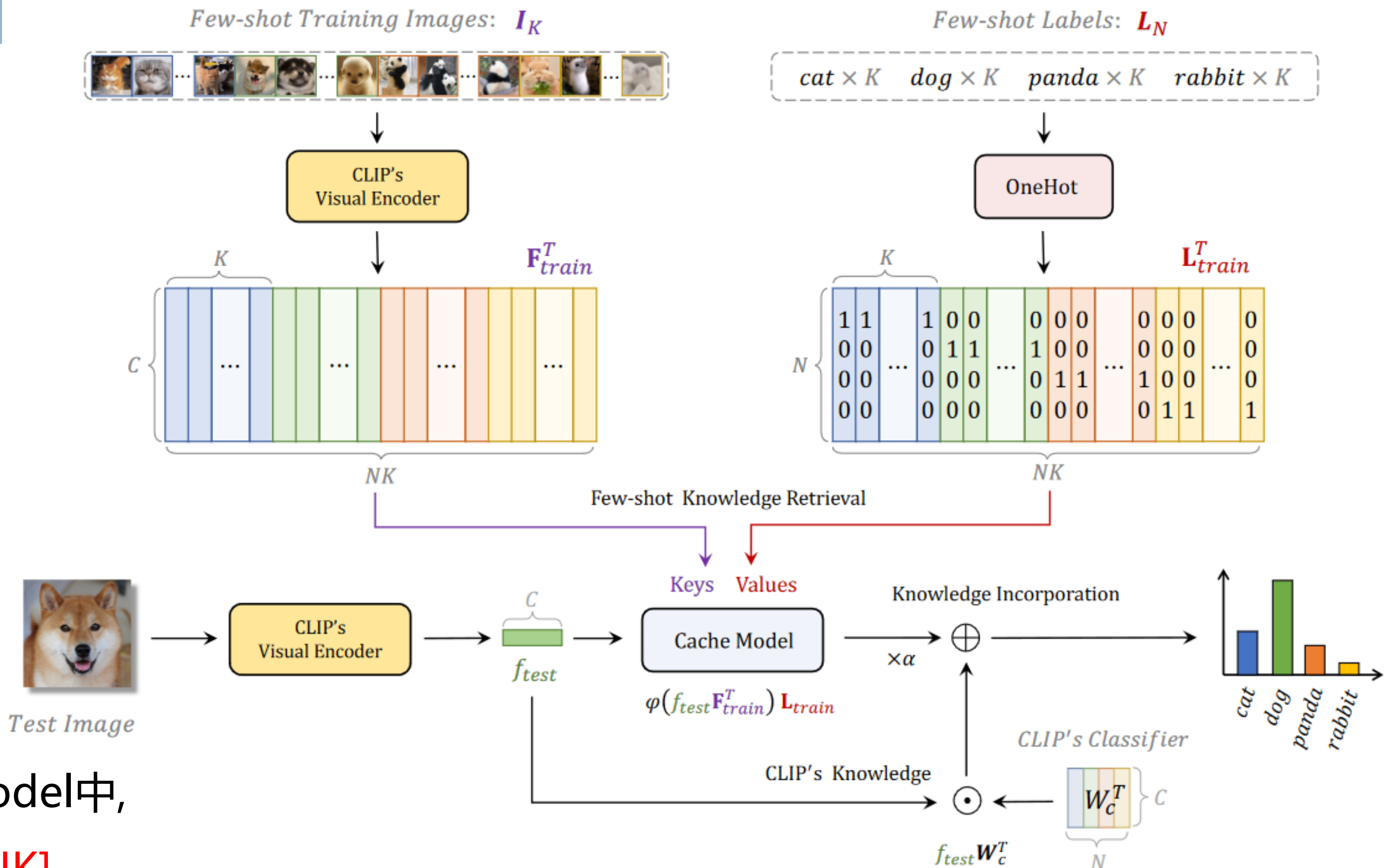
- 同时结合CLIP-Adapter和CoOp的优点，并且无需训练

Tip-Adapter

$\mathbf{F}_{\text{train}} \in \mathbb{R}^{NK \times C}$

$\mathbf{L}_{\text{train}} \in \mathbb{R}^{NK \times N}$

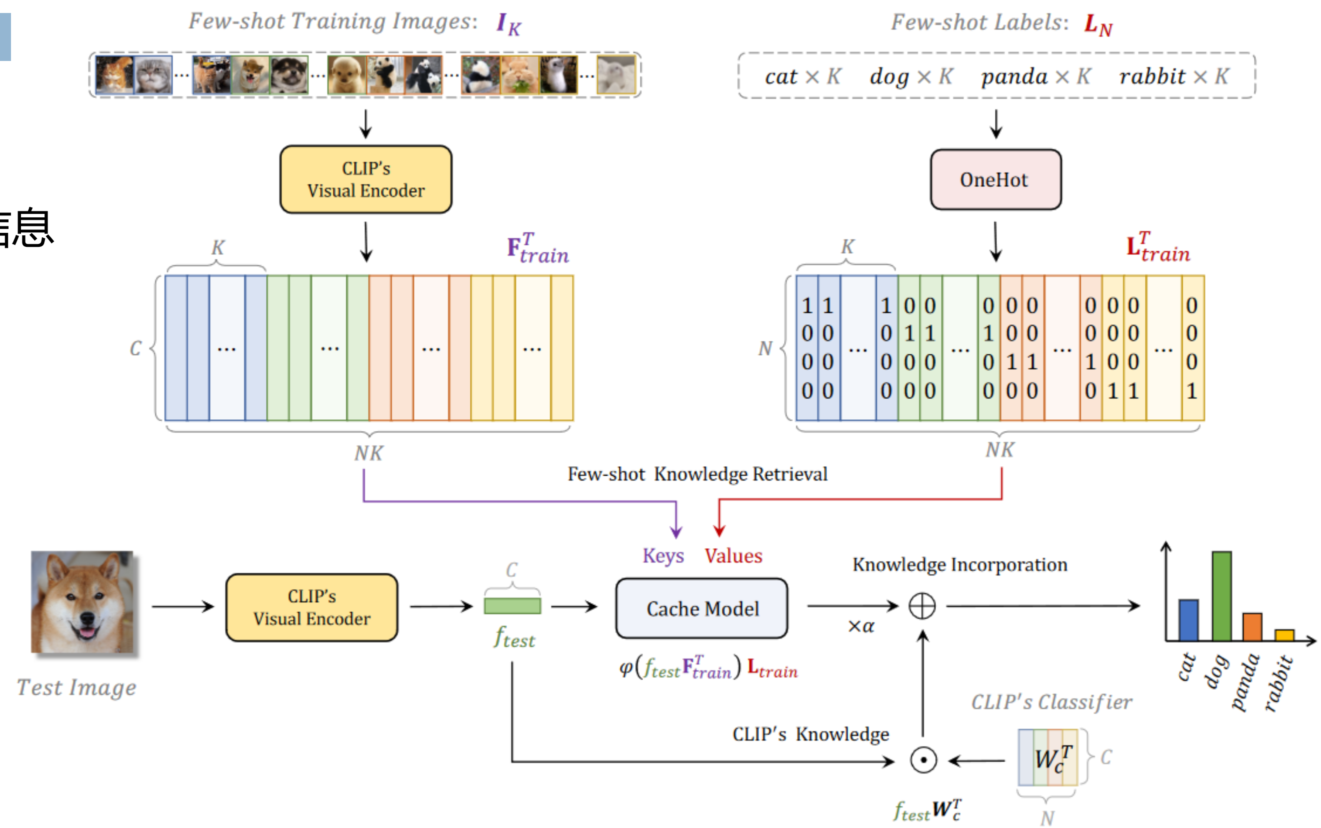
$$\mathbf{F}_{\text{train}} = \text{VisualEncoder}(I_K),$$
$$\mathbf{L}_{\text{train}} = \text{OneHot}(L_N).$$



- N为类别数量
- K为每类的shot数
- C为emb dim
- 将img对应到cache model中,
  [1, C] * [C, NK] = [1, NK]
- 利用cache将特征加权到类别,
  [1, NK] * [NK, N] = [1, N]

Tip-Adapter

- 利用残差连接将cache的信息融入到CLIP的logit中
- 利用相似度进行加权



$$
\begin{aligned}
\text{logits} &= \alpha A \mathbf{L}_{\text{train}} + f_{\text{test}} W_c^T \\
&= \alpha \varphi(f_{\text{test}} \mathbf{F}_{\text{train}}^T) \mathbf{L}_{\text{train}} + f_{\text{test}} W_c^T, \qquad A = \exp\left(-\beta(1 - f_{\text{test}} \mathbf{F}_{\text{train}}^T)\right),
\end{aligned}
$$

Tip-Adapter

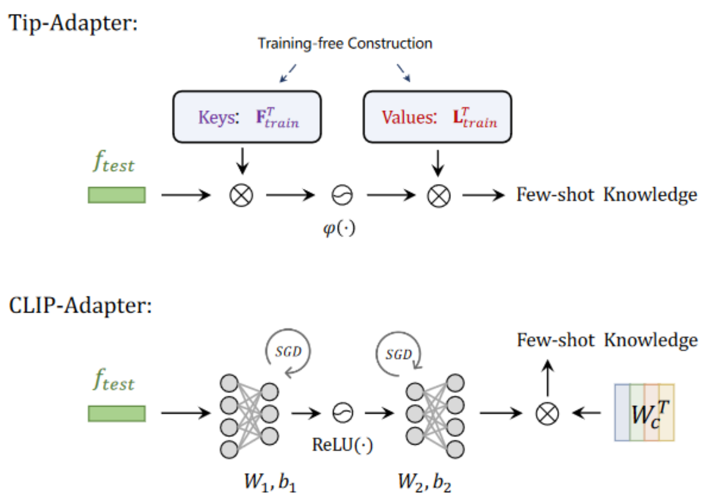


Fig. 2. Comparison of Tip-Adapter and CLIP-Adapter [16] to acquire few-shot knowledge. Tip-Adapter retrieves from the constructed cache model, but CLIP-Adapter encodes the knowledge by the learnbale adapter and obtains it aided by CLIP's classifier W_c.

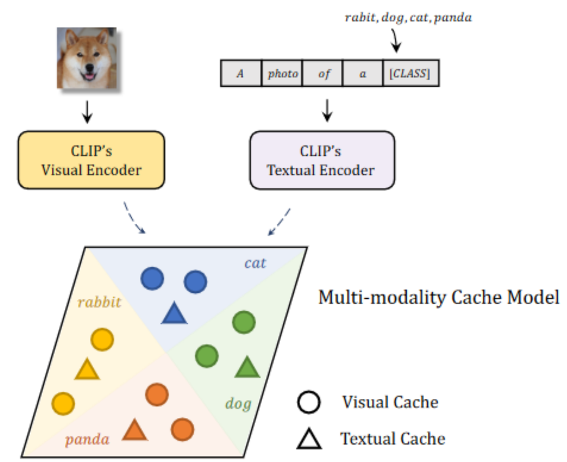


Fig. 3. The multi-modality cache model of Tip-Adapter. Different from previous networks only with visual cache, Tip-Adapter caches both visual and textual knowledge by CLIP's encoders.

Tip-Adapter

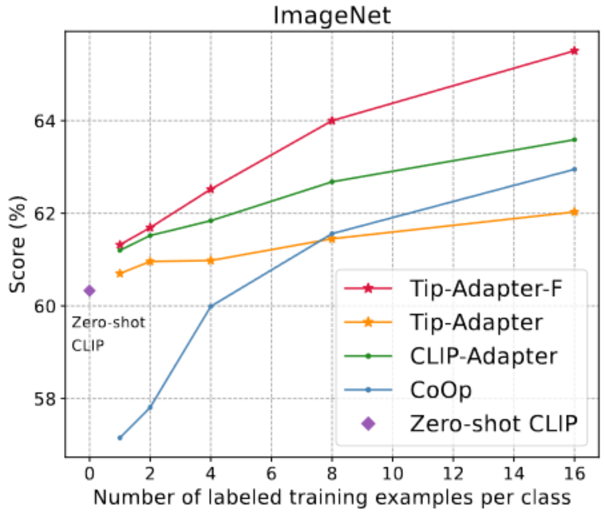
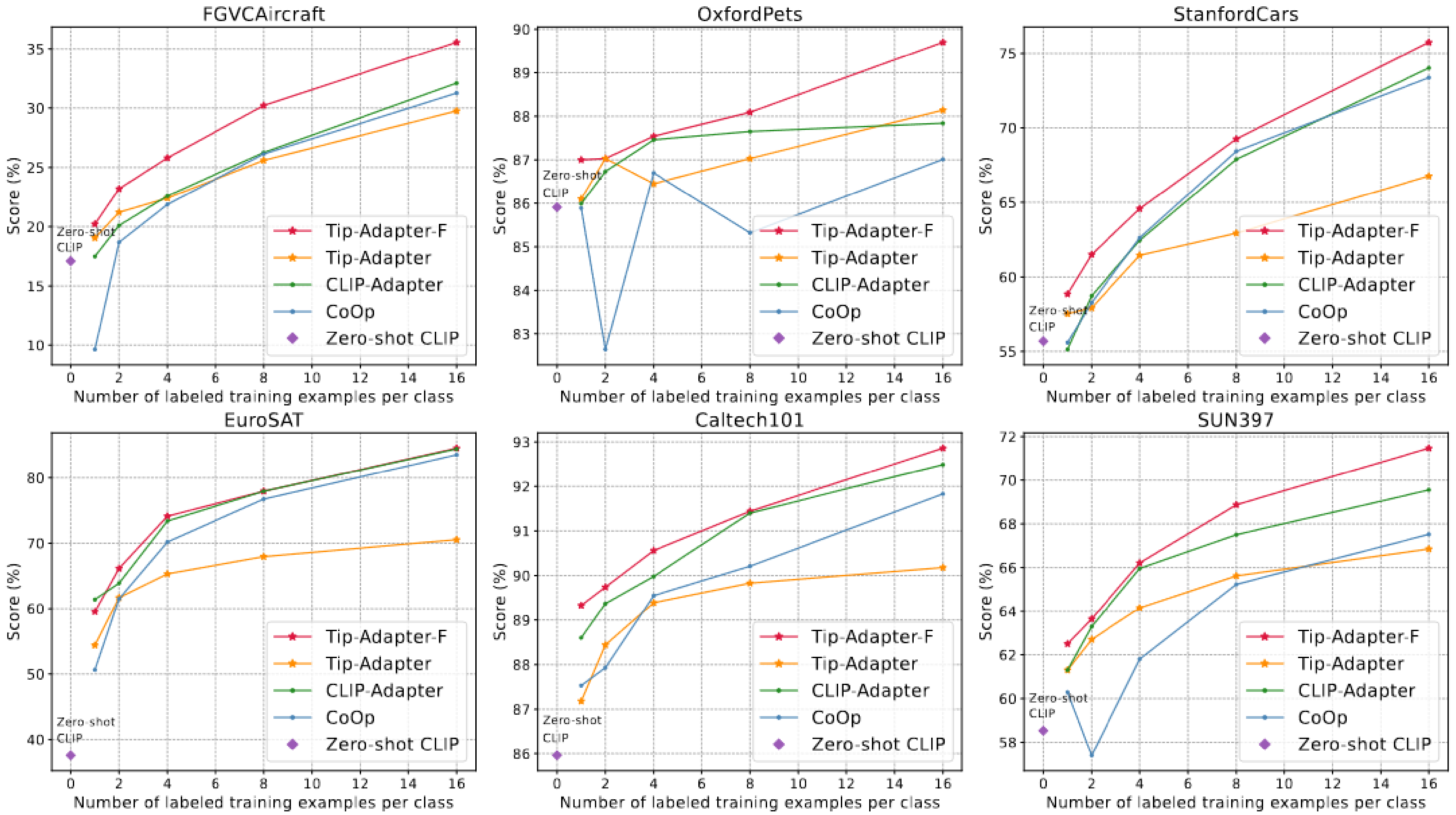


Fig. 4. Few-shot classification accuracy of different models on ImageNet [10].

Few-shot Setup	1	2	4	8	16
Zero-shot CLIP [48]: 60.33					
Linear-probe CLIP [48]	22.17	31.90	41.20	49.52	56.13
CoOp [73]	57.15	57.81	59.99	61.56	62.95
CLIP-Adapter [16]	61.20	61.52	61.84	62.68	63.59
Tip-Adapter	60.70	60.96	60.98	61.45	62.03
Tip-Adapter-F	**61.32**	**61.69**	**62.52**	**64.00**	**65.51**
	+0.62	+0.73	+1.54	+2.55	+3.48

Table 2. Classification accuracy (%) on ImageNet [10] of different models with quantitative values. The last row in blue records the performance gain of Tip-Adapter-F brought by further fine-tuning over Tip-Adapter.

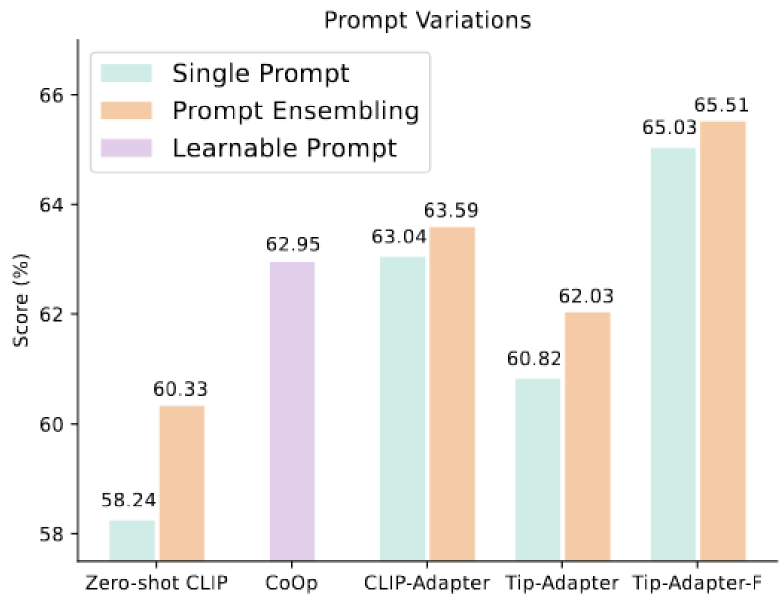
Tip-Adapter


FGVCAircraft
OxfordPets
StanfordCars
EuroSAT
Caltech101
SUN397
Score (%)
Number of labeled training examples per class
Zero-shot CLIP
Tip-Adapter-F
Tip-Adapter
CLIP-Adapter
CoOp
Zero-shot CLIP

Tip-Adapter

Ablation Studies on Tip-Adapter

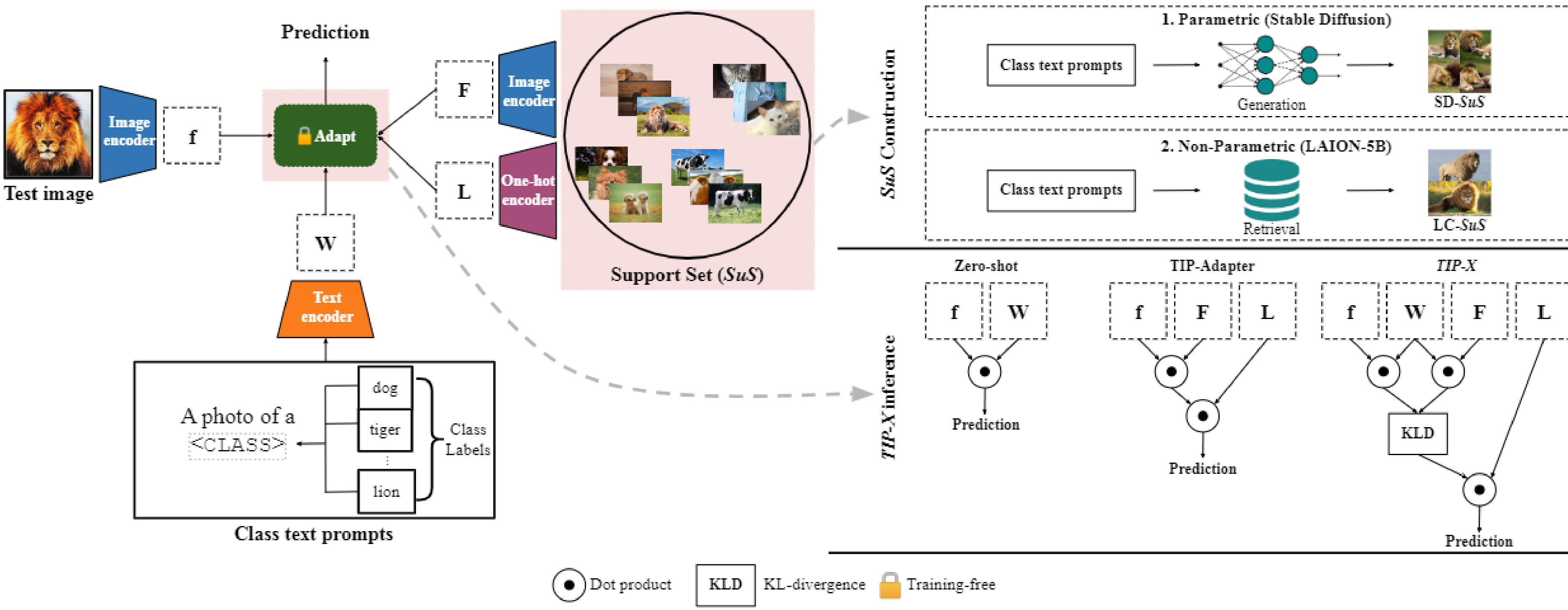
Residual Ratio α	0.0	0.5	**1.0**	2.0	3.0	4.0
	60.33	61.44	**62.03**	61.41	60.36	59.14
Sharpness Ratio β	1.5	3.5	**5.5**	7.5	9.5	11.5
	61.82	61.91	**62.03**	61.76	61.62	61.40
Cache Size	0	1	2	4	8	**16**
	60.33	61.45	61.71	61.79	61.83	**62.03**
More Shots than 16	Shot Setup		16	32	64	128
	Tip-Adapter		62.03	62.51	62.88	63.15
	Tip-Adapter-F		65.47	66.58	67.96	69.74



Models	ResNet-50	ResNet-101	ViT-B/32	ViT-B/16	RN50×16
Zero-shot CLIP [48]	60.33	62.53	63.80	68.73	70.94
CoOp [73]	62.95	66.60	66.85	71.92	-
CLIP-Adapter [16]	63.59	65.39	66.19	71.13	-
Tip-Adapter	62.03	64.78	65.61	70.75	72.95
Tip-Adapter-F	**65.51**	**68.56**	**68.65**	**73.69**	**75.81**

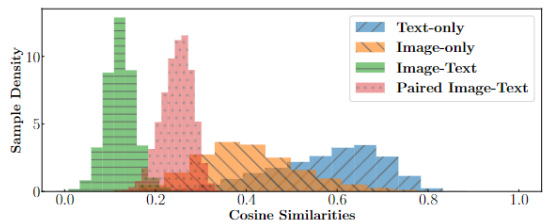
- 作者介绍
- 研究背景
- Tip-Adapter
- SuS-X
- 总结

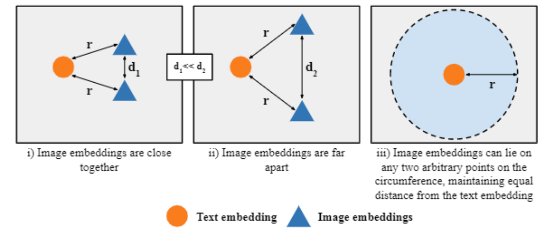
SuS-X

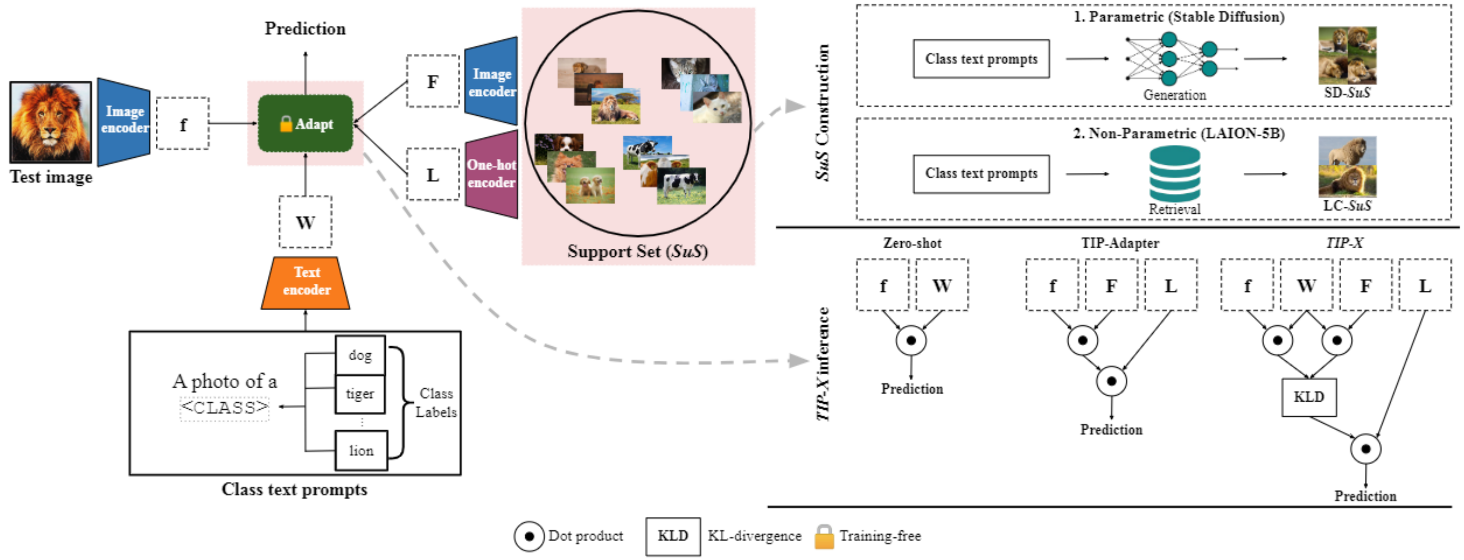


- 在Tip-Adapter基础上，引入了Diffusion图像生成或LAION-5B检索构建cache
- 利用image feature和cache feature分别与text feature对齐

SuS-X







$$W \in \mathbb{R}^{C\times d}$$

$$f_i = \texttt{CLIPImageEncoder}(y_i), i \in [1,t], f_i \in \mathbb{R}^d$$
$$f = \texttt{Concat}([f_1, f_2, \dots, f_t]), f \in \mathbb{R}^{t\times d}$$

$$F_i = \texttt{CLIPImageEncoder}(x_i), i \in [1,CK], F_i \in \mathbb{R}^d$$
$$F = \texttt{Concat}([F_1, F_2, \dots, F_{CK}]), F \in \mathbb{R}^{CK\times d}$$

$$\texttt{TL} = \alpha AL + fW^T$$

$$\texttt{KL}(P||Q) = \textstyle\sum_i P_i \log \frac{P_i}{Q_i}.$$

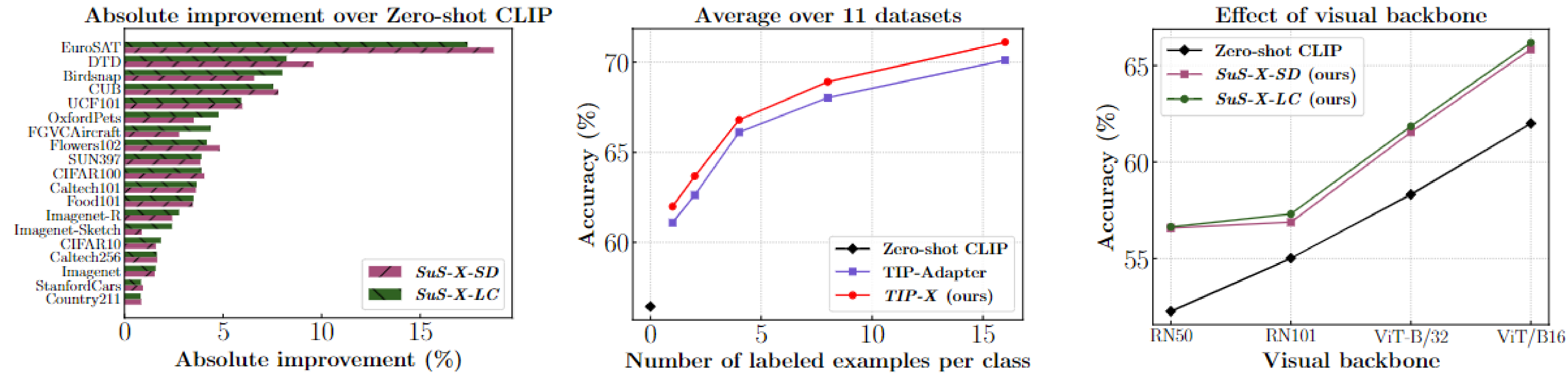
$$S = \texttt{softmax}(FW^T), S \in \mathbb{R}^{CK\times C}$$
$$s = \texttt{softmax}(fW^T), s \in \mathbb{R}^{t\times C}$$

$$M_{i,j} = \texttt{KL}(s_i||S_j), i \in [1,t], j \in [1,CK]$$

$$\texttt{TXL} = fW^T + \alpha AL + \gamma\psi(-M)L$$

SuS-X

	Method	Average*	ImageNet [18]	ImageNet-R [38]	ImageNet-Sketch [73]	EuroSAT [37]	DTD [14]	Birdsnap [5]
Zero-shot	Zero-shot CLIP [61]	52.27	60.31	59.34	35.42	26.83	41.01	30.56
	CALIP [34]	–	60.57	–	–	38.90	42.39	–
	CALIP [34]†	52.37	60.31	59.33	36.10	26.96	41.02	30.68
	CLIP+DN [89]	53.02	60.16	60.37	35.95	28.31	41.21	31.23
Name-only	CuPL [60]	55.50	61.45	61.02	35.13	38.38	48.64	35.65
	CuPL+e	55.76	61.64	61.17	35.85	37.06	47.46	35.80
	VisDesc [53]	53.76	59.68	57.16	33.78	37.60	41.96	35.65
	SuS-X-SD (ours)	56.73	61.84	61.76	36.30	**45.57**	**50.59**	37.14
	SuS-X-LC (ours)	**56.87**	**61.89**	**62.10**	**37.83**	44.23	49.23	**38.50**



(a) (b) (c)

SuS-X

Table 3: ***SuS-X* generalises to different VLMs.** *Average reported across 19 datasets.

VLM	Method	Average*	ImageNet	EuroSAT	DTD	Birdsnap
TCL	Zero-shot	31.38	35.55	20.80	28.55	4.51
	CuPL	34.79	41.60	26.30	42.84	6.83
	CuPL+e	32.79	41.36	25.88	41.96	6.60
	VisDesc	33.94	40.40	21.27	34.28	5.69
	SuS-X-SD	41.49	52.29	28.75	**48.17**	13.60
	SuS-X-LC	**42.75**	**52.77**	**36.90**	46.63	**17.93**
BLIP	Zero-shot	48.73	50.59	44.10	44.68	10.21
	CuPL	51.11	52.96	39.37	52.95	12.24
	CuPL+e	51.36	53.07	41.48	53.30	12.18
	VisDesc	49.91	50.94	42.25	47.45	11.69
	SuS-X-SD	53.20	55.93	45.36	**56.15**	16.95
	SuS-X-LC	**54.64**	**56.75**	**51.62**	55.91	**23.78**

Table 4: **Component Analysis of *SuS-X*.**

Text Prompts	Method	*SuS*	*TIP-X*	Average Accuracy
Default	Zero-shot CLIP	✗	✗	52.27
	SuS-TIP-SD	✓	✗	53.49 (+1.22%)
	SuS-X-SD	✓	✓	53.69 (+1.42%)
	SuS-TIP-LC	✓	✗	53.83 (+1.56%)
	SuS-X-LC	✓	✓	54.20 (+1.93%)
CuPL+e	CuPL+e	✗	✗	55.76 (+3.49%)
	SuS-TIP-SD	✓	✗	56.63 (+4.36%)
	SuS-X-SD	✓	✓	56.73 (+4.46%)
	SuS-TIP-LC	✓	✗	56.72 (+4.45%)
	SuS-X-LC	✓	✓	**56.87** (+4.60%)

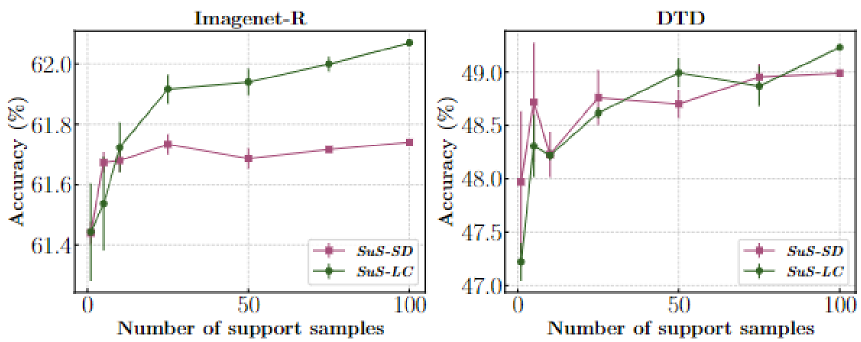
Table 5: **Prompting strategies for *SuS* construction.**

SuS method	Average Acc.		ImageNet Acc.		Diversity	
	Photo	*CuPL*	*Photo*	*CuPL*	*Photo*	*CuPL*
LC	**56.87**	56.20	**61.89**	61.79	0.28	**0.32**
SD	56.32	56.73	61.79	61.84	0.17	**0.20**

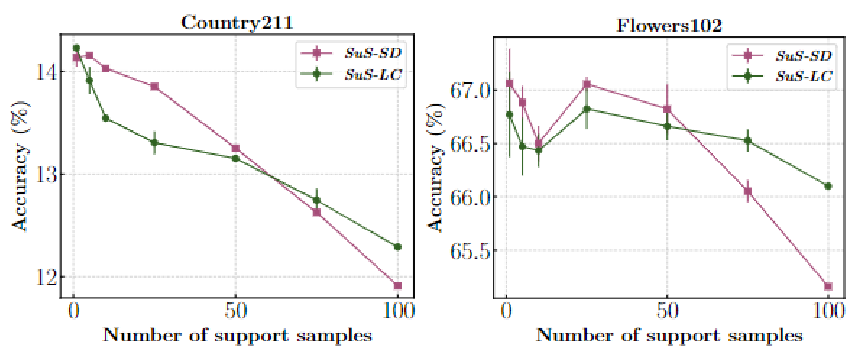
Table 6: **Hyperparameter sensitivity for γ**

Dataset	γ value						
	0	0.1	0.2	0.3	0.5	0.75	1
ImageNet-R	60.87	60.98	61.03	**61.05**	61.00	60.89	60.65
OxfordPets	76.76	77.17	**77.58**	77.44	77.17	77.17	76.90
DTD	47.16	47.16	47.51	47.69	47.87	**47.96**	47.60

SuS-X



(a) Tasks where larger support sets are beneficial



(b) Tasks where larger support sets are harmful

Figure 6: **Effect of support size.**

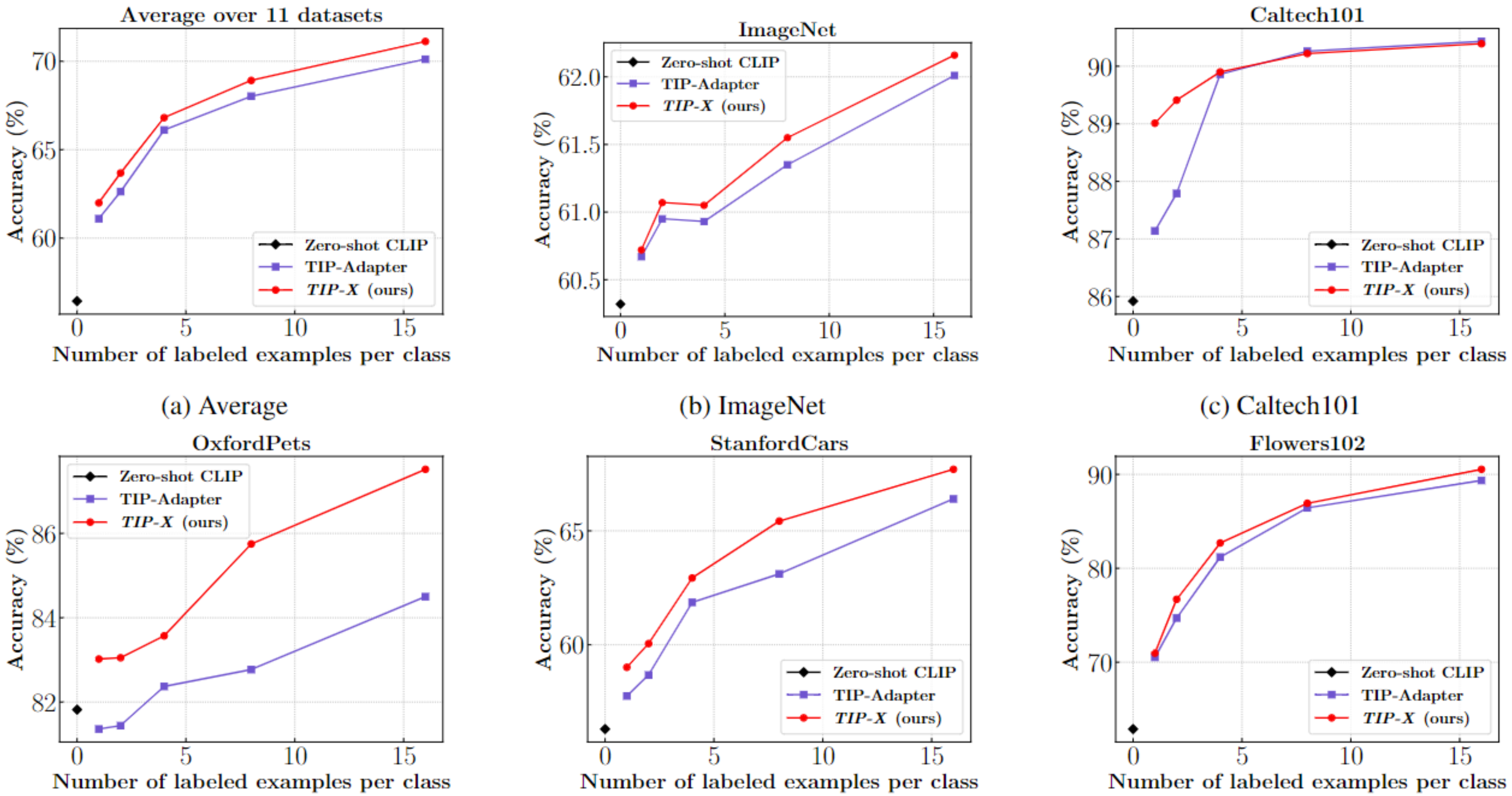
(a) Dishwasher

(c) Australian Kelpie

(b) Split Rail Fence

(d) Bulbul

SuS-X


Average over 11 datasets
Accuracy (%)
Number of labeled examples per class
Zero-shot CLIP
TIP-Adapter
TIP-X (ours)
(a) Average
ImageNet
(b) ImageNet
Caltech101
(c) Caltech101
OxfordPets
StanfordCars
Flowers102

SuS-X

Dataset	Classes	Val	Test
UCF-101	101	1898	3783
CIFAR-10	10	10000	10000
CIFAR-100	100	10000	10000
Caltech101	100	1649	2465
Caltech256	257	6027	9076
ImageNet	1000	50000	50000
SUN397	397	3970	19850
FGVCAircraft	100	3333	3333
Birdsnap	500	7774	11747
StanfordCars	196	1635	8041
CUB	200	1194	5794
Flowers102	102	1633	2463
Food101	101	20200	30300
OxfordPets	37	736	3669
DTD	47	1128	1692
EuroSAT	10	5400	8100
ImageNet-Sketch	1000	50889	50889
ImageNet-R	200	30000	30000
Country211	211	10550	21100

Dataset	Support Set Size
UCF-101	5858
CIFAR-10	50
CIFAR-100	4700
Caltech101	101
Caltech256	3084
ImageNet	36000
SUN397	397
FGVCAircraft	7900
Birdsnap	39000
StanfordCars	980
CUB	400
Flowers102	3162
Food101	3434
OxfordPets	2627
DTD	188
EuroSAT	150
ImageNet-Sketch	42000
ImageNet-R	10200
Country211	844

Dataset	α	β	γ
UCF-101	0.10	8.59	0.10
CIFAR-10	5.09	5.41	0.10
CIFAR-100	0.10	1.49	0.10
Caltech101	0.10	1.27	0.10
Caltech256	0.10	12.76	0.10
ImageNet	10.08	39.46	0.10
SUN397	2.60	8.35	0.10
FGVCAircraft	2.60	24.52	0.69
Birdsnap	48.53	22.55	0.69
StanfordCars	0.10	1.58	0.10
CUB	0.10	8.84	0.10
Flowers102	0.10	2.72	0.10
Food101	17.56	49.02	0.10
OxfordPets	10.08	41.91	1.29
DTD	5.09	23.79	0.70
EuroSAT	2.60	1.00	0.10
ImageNet-Sketch	30.04	38.48	0.69
ImageNet-R	2.60	30.65	0.70
Country211	12.57	22.31	0.10

SuS-X

$$\text{TXL} = \underbrace{fW^T}_{\text{1. zero-shot component}} + \underbrace{\alpha AL}_{\text{2. intra-modal distance component}} + \underbrace{\gamma\psi(-M)L}_{\text{3. inter-modal distance component}}$$

Table 15: **Contribution of intra-modal and inter-modal distances.**

Dist. terms used	**1** (Zero-shot)	**1+3** (Inter-modal)	**1+2** (Intra-modal)	**1+2+3** (Both)
Average Acc.	52.27	56.30	56.56	56.87
Gain	0	+4.03	+4.29	+4.60

Table 21: ***SuS-X-SD*** **Results with additional T2I models.**

T2I Model	**ImageNet**	**EuroSAT**	**DTD**	**OxfordPets**	**Average**
ZS-CLIP (baseline)	60.31	26.83	41.01	81.82	52.49
StableDiffusion-1.4 (from main paper)	**61.84**	45.57	50.59	85.34	60.84 (+8.35%)
Kandinsky2.1	61.83	44.96	49.17	**85.47**	60.36 (+7.87%)
OpenJourney-4	61.81	45.00	**50.71**	85.17	60.67 (+8.18%)
Protogen-2.2	61.82	**48.67**	50.35	85.26	**61.52** (+9.03%)

Table 22: **Fine-tuning methods vs** ***SuS-X*****.**

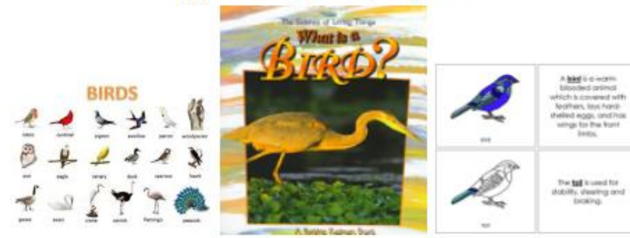
Method	ZS-CLIP (No adaptation)	FT-CLIP (Full fine-tuning)	CoOp [88] (PromptTuning)	CLIP-Adapter [28] (Adapters)	***SuS-X*** (Ours)	***SuS-X-F*** (Ours)
ImageNet	60.31	60.35	60.96	61.61	<u>61.89</u>	**63.22**
EuroSAT	26.83	55.37	52.12	<u>57.00</u>	44.23	**59.22**
DTD	41.01	<u>50.35</u>	45.66	49.29	49.23	**52.30**
OxfordPets	81.82	84.51	85.99	85.06	<u>86.59</u>	**87.77**

SuS-X

(a) *SuS-LC*, *Photo*, Airplane

(b) *SuS-LC*, *CuPL*, Airplane

(c) *SuS-LC*, *Photo*, Bird



(d) *SuS-LC*, *CuPL*, Bird

(e) *SuS-SD*, *Photo*, Airplane

(f) *SuS-SD*, *CuPL*, Airplane

(g) *SuS-SD*, *Photo*, Bird

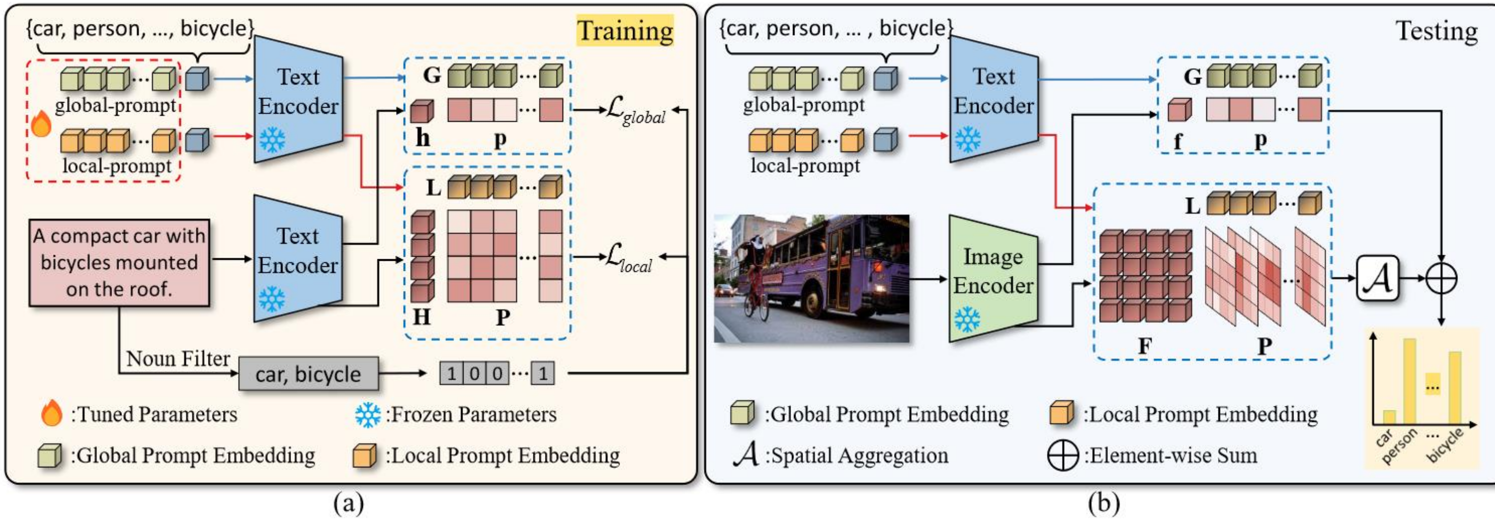
(h) *SuS-SD*, *CuPL*, Bird

- 作者介绍
- 研究背景
- 方法
- 实验效果
- 总结

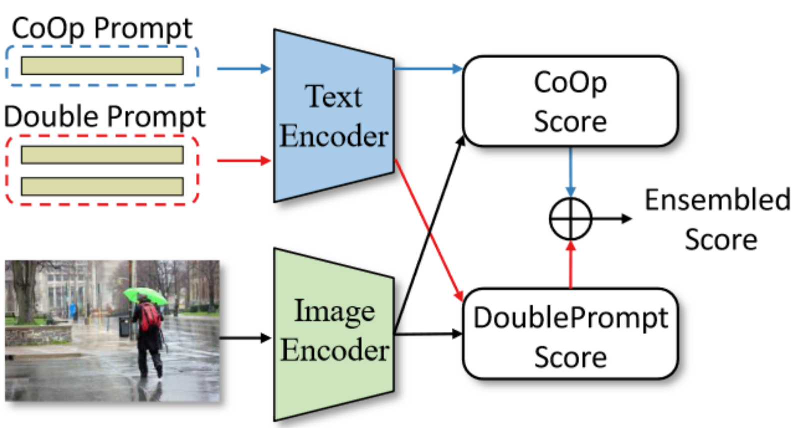
总结反思

- 多模态大模型的能力已经足够实现无需训练即可适配到下游任务中
- 适配任务时需要合理利用多模态模型中嵌入的知识，结合prompt，adapter系列方法合理利用这些特征

总结反思



{car, person, ..., bicycle}
Training
global-prompt
local-prompt
Text Encoder
G
h
p
$\mathcal{L}_{global}$
L
A compact car with bicycles mounted on the roof.
Text Encoder
H
P
$\mathcal{L}_{local}$
Noun Filter
car, bicycle
1 0 0 ··· 1
:Tuned Parameters
:Frozen Parameters
:Global Prompt Embedding
:Local Prompt Embedding
(a)
{car, person, ... , bicycle}
Testing
global-prompt
local-prompt
Text Encoder
G
f
p
L
Image Encoder
F
P
$\mathcal{A}$
car
person
...
bicycle
:Global Prompt Embedding
:Local Prompt Embedding
:Spatial Aggregation
:Element-wise Sum
(b)



CoOp Prompt
Double Prompt
Text Encoder
Image Encoder
CoOp Score
Ensembled Score
DoublePrompt Score

谢谢！