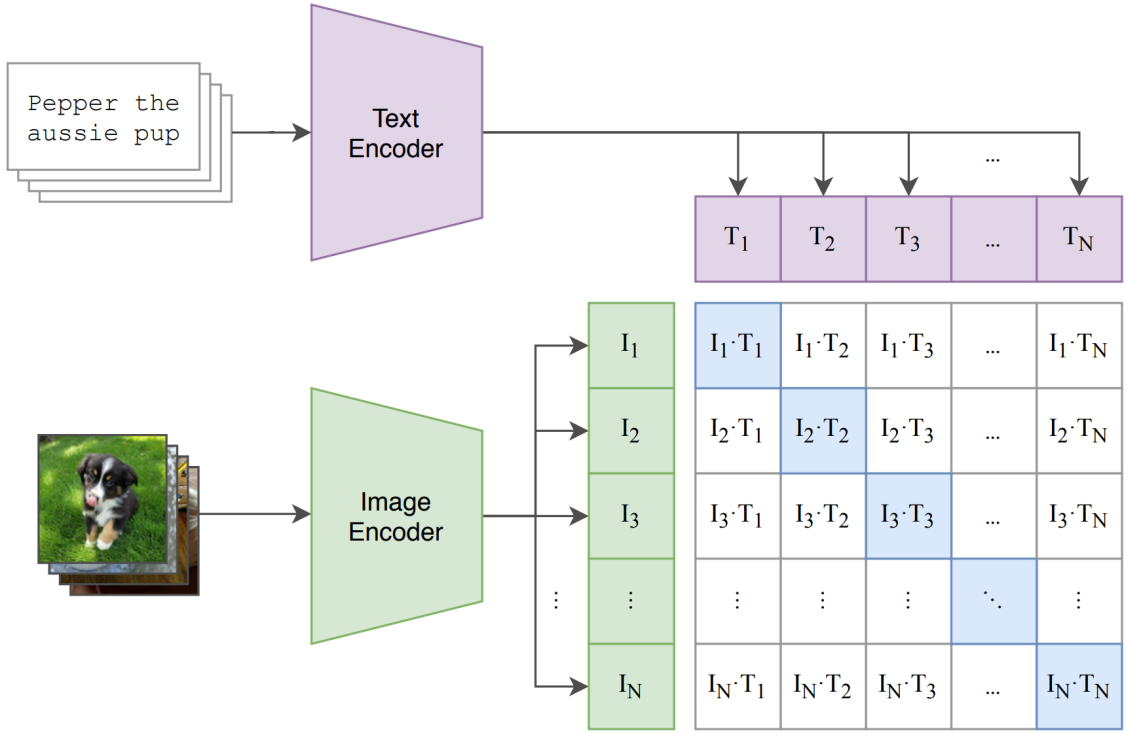
Beyond Sole Strength: Customized Ensembles for Generalized Vision-Language Models

Zhihe Lu[1] Jiawang Bai[12] Xin Li[13] Zeyu Xiao[13] Xinchao Wang[1]

[1]National University of Singapore
[2]Tsinghua University
[3]University of Science and Technology of China

Background: CLIP

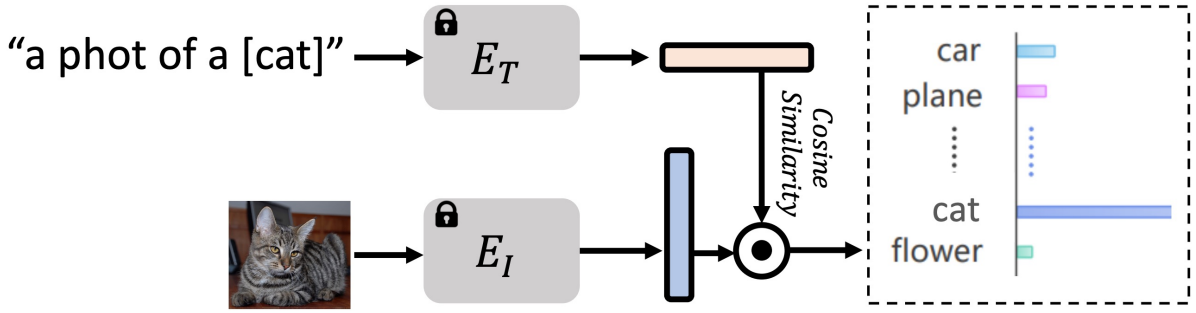
❑ Contrastive Language-Image Pretraining (CLIP [1]) Model



[1] Radford, Alec, et al. "Learning transferable visual models from natural language supervision." ICML 2021.

Background: CLIP

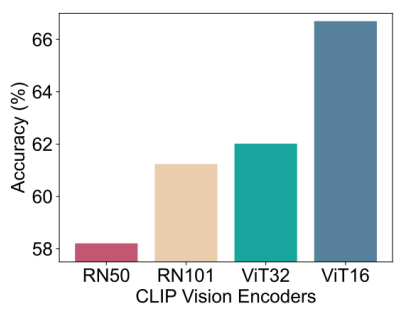
❑ Zero-shot CLIP



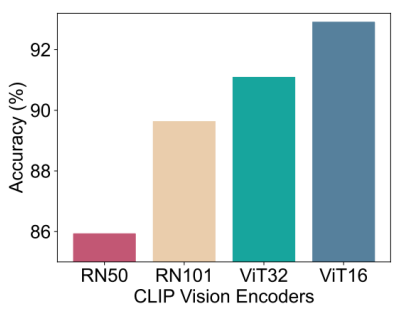
$$p(y=i|\mathbf{z}) = \frac{\exp(\text{sim}(\mathbf{z}, \mathbf{t}_i)/\tau)}{\sum_{j=1}^{K} \exp(\text{sim}(\mathbf{z}, \mathbf{t}_j)/\tau)}$$

Motivation

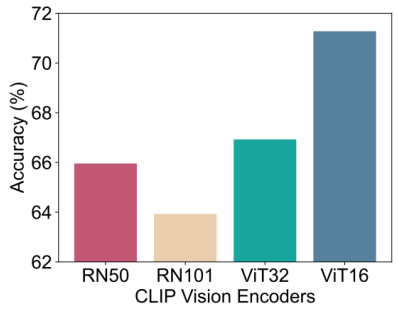
- Ensemble learning of pre-trained foundation models is timely research, but no study has been specifically explored this for VLMs.



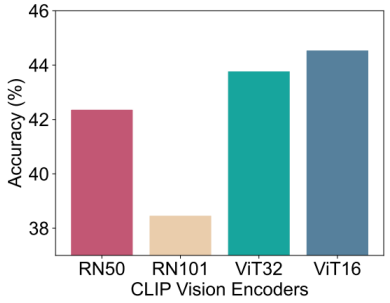
(a) ImageNet



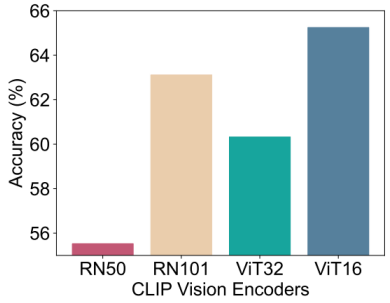
(b) Caltech101



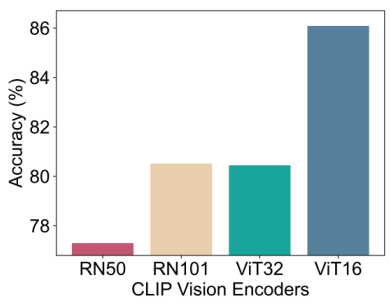
(c) Flowers102



(d) DTD



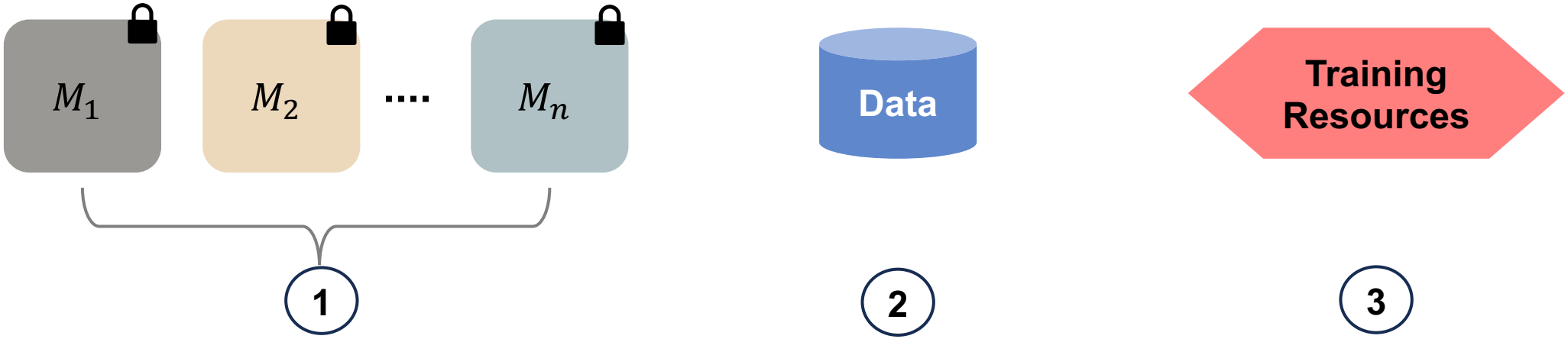
(e) Stanford Cars



(f) Food101

Motivation

❑ How to design ensemble learning given specific conditions.



Set-up 1: given solely pretrained CLIP models (1)

Set-up 2: given pretrained CLIP models and an extra few-shot dataset, but *no training*! (1) + (2)

Set-up 3: given pretrained CLIP models and an extra few-shot dataset with *training* (1) + (2) + (3)

Methodology

- Set-up 1: given solely pretrained CLIP models

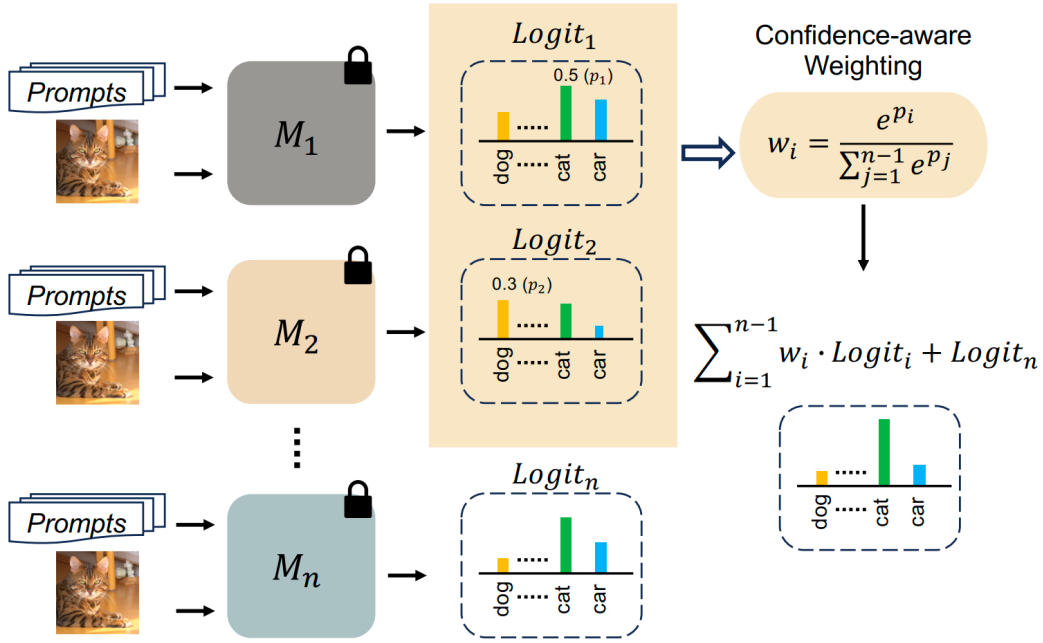


Figure 3. **Illustration of our zero-shot ensemble ($\mathbf{ZS}_{En}$).** We assign a weight 1.0 to the best performing model, *i.e.*, CLIP-ViT-B/16, and use the confidence-aware weights for other VLMs.

Methodology

- Set-up 2: given pretrained CLIP models and an extra few-shot dataset, but *no training*!

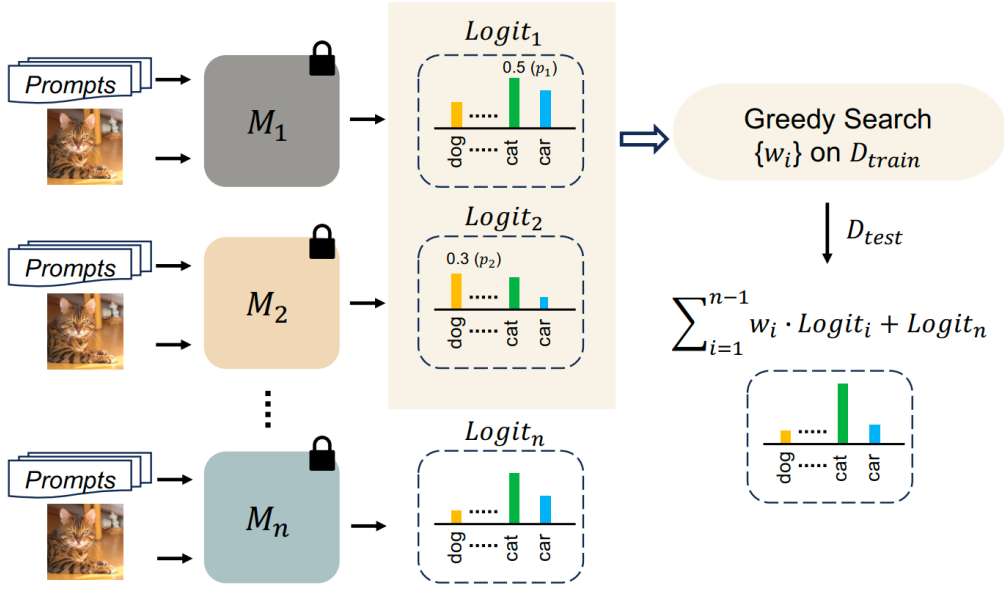


Figure 4. **Illustration of our training-free ensemble (TF$_{En}$).** We assign a weight 1.0 to the best performing model, *i.e.*, CLIP-ViT-B/16, and determine the weights of other VLMs by greedy searching on a given "training" set without training.

Methodology

- Set-up 3: given pretrained CLIP models and an extra few-shot dataset with *training*.

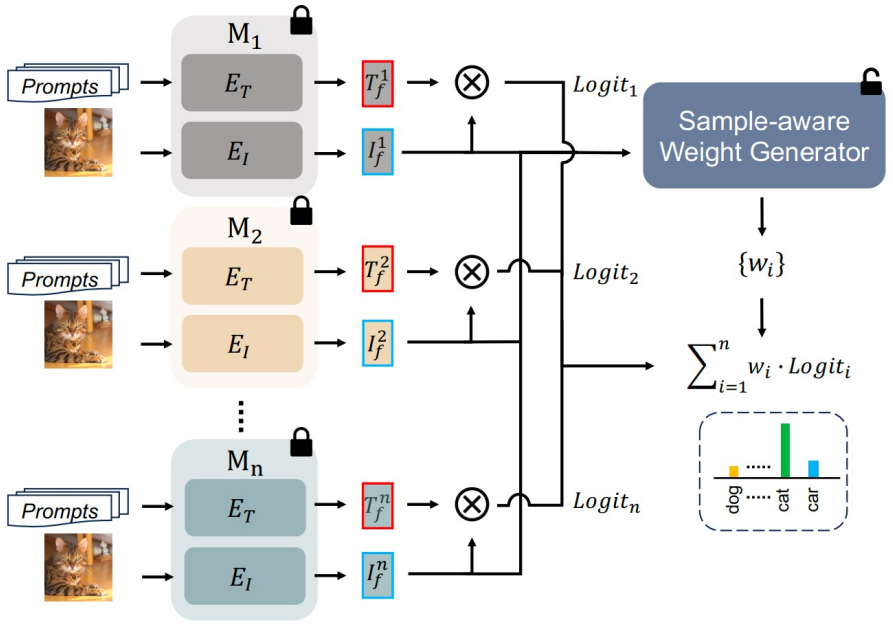


Figure 5. **Illustration of our tuning ensemble ($\mathbf{T}_{En}$).** The proposed sample-aware weight generator (SWIG) takes sample features as input to generate sample-aware weights, which are then used for weighted prediction.

Experiment

- Zero-shot generalization and ablation study

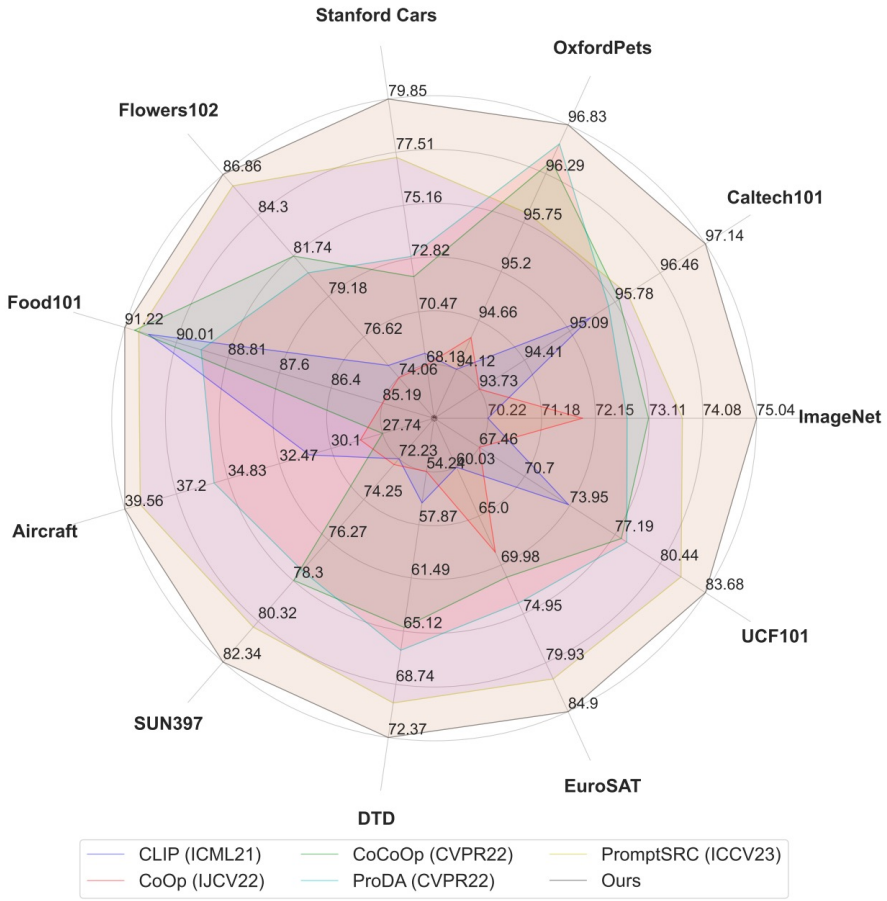
Dataset	CLIP RN50	CLIP RN101	CLIP ViT-B/32	CLIP ViT-B/16	ZS_{En} Ours	Δ
Average	58.71	59.52	61.87	65.27	67.88	+2.61
ImageNet	58.23	61.26	62.04	66.72	70.66	+3.94
Caltech101	85.96	89.66	91.12	92.94	93.79	+0.85
OxfordPets	85.80	86.89	87.49	89.07	90.57	+1.50
Stanford Cars	55.57	63.16	60.37	65.29	70.76	+5.47
Flowers102	65.98	63.95	66.95	71.30	73.16	+1.86
Food101	77.32	80.54	80.47	86.11	86.78	+0.67
FGVC Aircraft	17.16	18.18	19.20	24.87	25.68	+0.81
SUN397	58.53	58.96	62.00	62.62	65.91	+3.29
DTD	42.38	38.48	43.79	44.56	49.35	+4.79
EuroSAT	37.42	32.62	45.12	47.69	50.20	+2.51
UCF101	61.43	61.04	62.07	66.77	69.84	+3.07

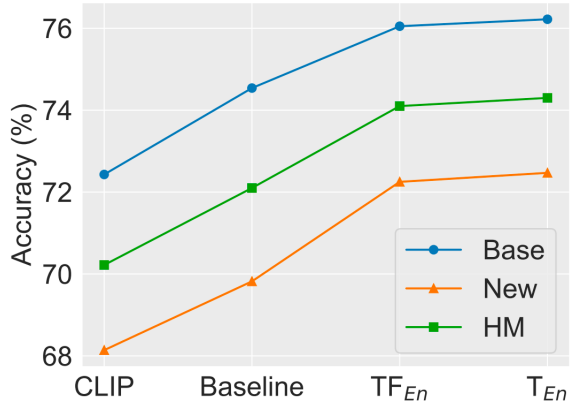
CLIP ViT-B/16	CLIP RN50	CLIP RN101	CLIP ViT-B/32	Acc	Δ
✓				66.72	+0.00
✓	✓			67.75	+1.03
✓	✓	✓		68.34	+1.62
✓	✓	✓	✓	68.76	+2.04

Method	Acc	Δ
CLIP ViT-B/16	66.72	+0.00
Baseline	68.76	+2.04
CAW of 3 models (w/o RN50)	70.01	+3.29
CAW of 4 models	70.19	+3.47
CLIP ViT-B/16 + CAW of 3 other models	70.66	+3.94

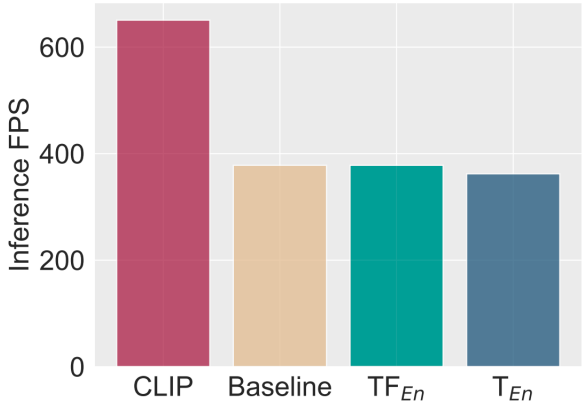
Experiment

❑ Base-to-new generalization





(a) Base-to-new performance



(b) Inference FPS

Experiment

- Cross-dataset evaluation

	Source	Target										
	ImageNet	Caltech101	OxfordPets	StanfordCars	Flowers102	Food101	Aircraft	SUN397	DTD	EuroSAT	UCF101	*Average*
CLIP (Radford et al., 2021)	66.72	92.94	89.07	65.29	71.30	86.11	24.87	62.62	44.56	47.69	66.77	65.12
CoOp (Zhou et al., 2022b)	**71.51**	93.70	89.14	64.51	68.71	85.30	18.47	64.15	41.92	46.39	66.55	63.88
CoCoOp (Zhou et al., 2022a)	71.02	**94.43**	90.14	65.32	71.88	86.06	22.94	**67.36**	45.73	45.37	68.21	65.74
MaPLe (Khattak et al., 2023a)	70.72	93.53	90.49	65.57	72.23	86.20	24.74	67.01	46.49	48.06	68.69	66.30
PromptSRC (Khattak et al., 2023b)	71.27	93.60	90.25	65.70	70.25	86.15	23.90	67.10	46.87	45.50	68.75	65.81
T_{En}	70.88	93.91	**90.72**	**71.94**	**72.59**	**86.68**	**26.03**	66.07	**49.31**	**48.18**	**69.14**	**67.46**

Experiment

❑ Domain generalization

	Source	Target				
	ImageNet	-V2	-S	-A	-R	Avg.
CLIP (Radford et al., 2021)	66.73	60.83	46.15	47.77	73.96	57.18
CoOp (Zhou et al., 2022b)	71.51	64.20	47.99	49.71	75.21	59.28
CoCoOp (Zhou et al., 2022a)	71.02	64.07	48.75	50.63	76.18	59.91
MaPLe (Khattak et al., 2023a)	70.72	64.07	49.15	50.90	76.98	60.28
PromptSRC (Khattak et al., 2023b)	71.27	64.35	49.55	50.90	77.80	60.65
CLIP + T_{En}	70.88	62.87	48.97	49.97	75.98	59.45
CoCoOp + T_{En}	**73.25**	**65.73**	**50.70**	**52.11**	**78.11**	**61.66**

Summary

- ❑ The first comprehensive study for ensemble learning of VLMs.
- ❑ We propose three customized ensemble learning strategies tailored for three practical scenarios.
- ❑ Our method has shown its effectiveness on diverse datasets and tasks.

Thanks!

Contact to Zhihe Lu: zhihelu.academic@gmail.com
Code: https://github.com/zhiheLu/Ensemble_VLM