Prompt Learning Meets Visual Context

Guangyi Chen 陈广义

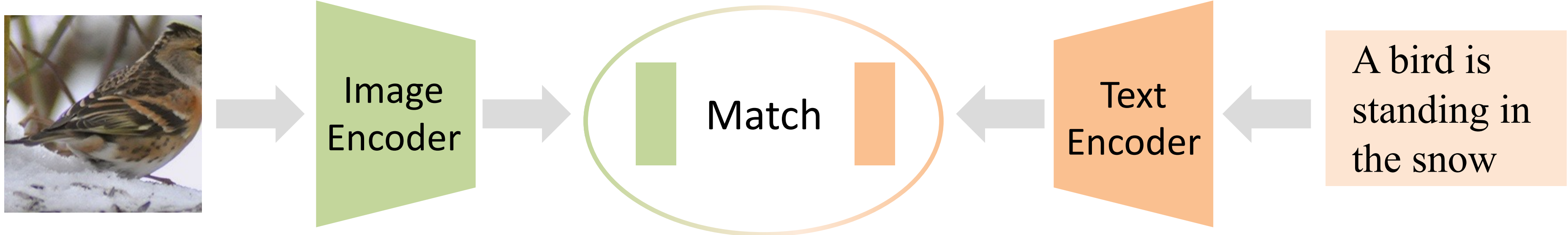
https://chengy12.github.io/

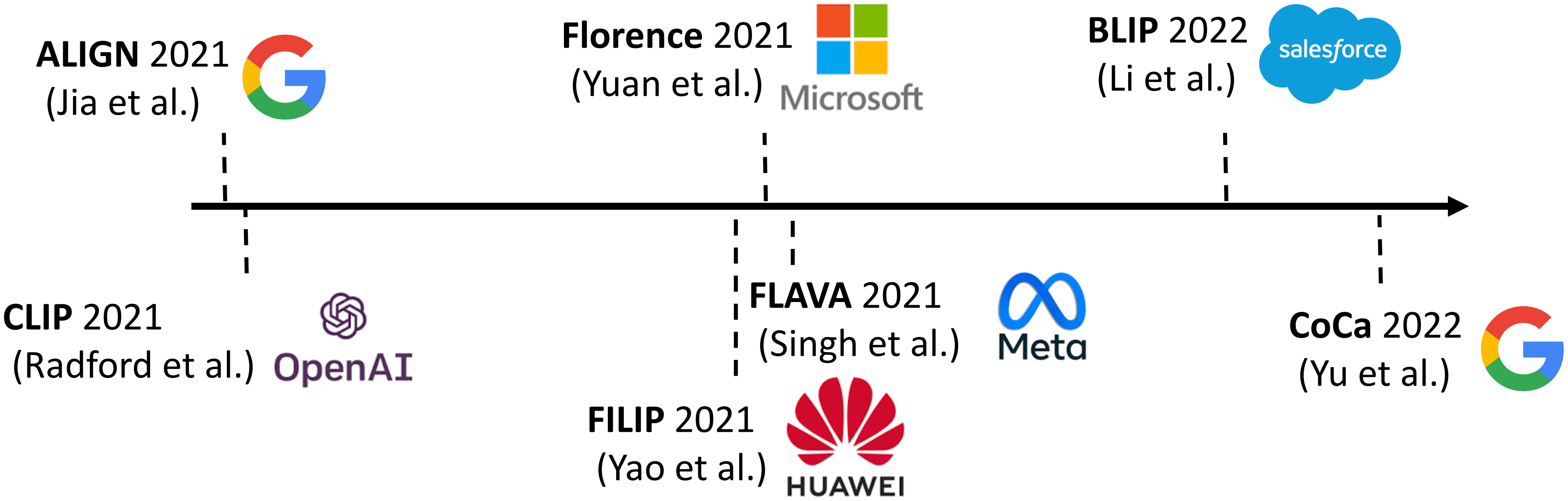
Carnegie Mellon University, Pittsburgh PA, USA

Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE

Contrastive Vision Language Model

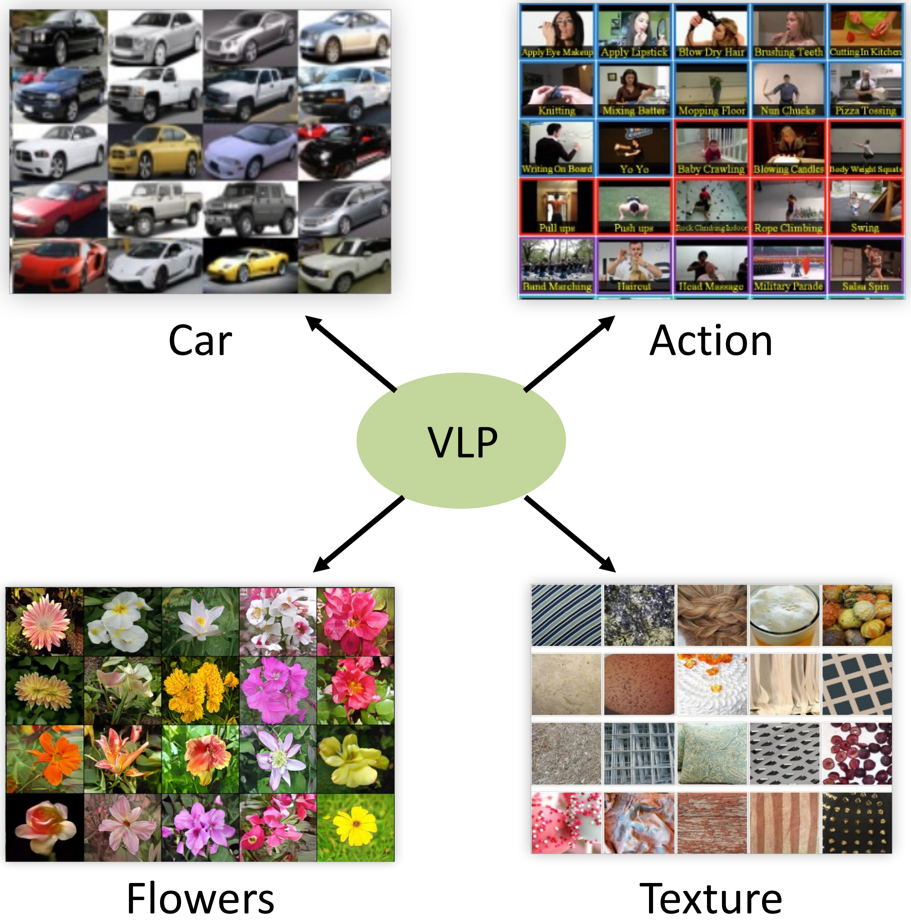
- Recently, contrastive vision-language pre-trained models, which learn visual representation with natural language supervision, have achieved significant success.





Contrastive Vision Language Model

- ➢ These vision language pretrained (VLP) models show promising generalization ability by explicitly leveraging the neural languages.



Dataset Examples	ImageNet ResNet101	Zero-Shot CLIP	Δ Score
ImageNet	76.2	76.2	0%
ImageNetV2	64.3	70.1	+5.8%
ImageNet-R	37.7	88.9	+51.2%
ObjectNet	32.6	72.3	+39.7%
ImageNet Sketch	25.2	60.2	+35.0%
ImageNet-A	2.7	77.1	+74.4%

Illustration from CLIP (Radford et al., 2021)

Contrastive Vision Language Model

- How to efficiently adapt the knowledge from pretraining to the downstream tasks is an important question, given these models are typical of massive sizes.

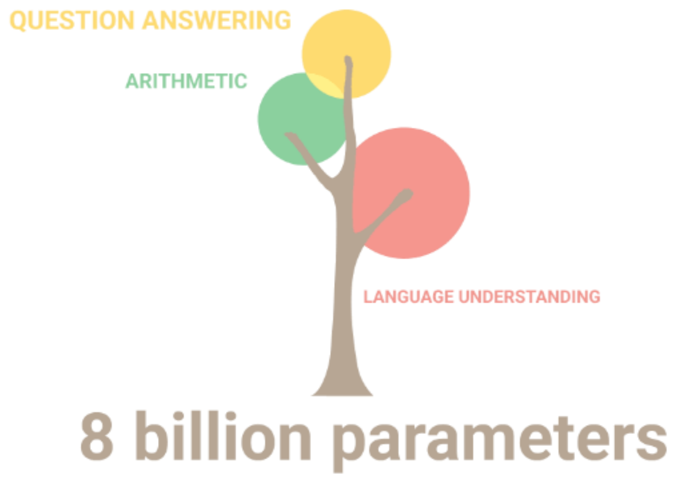
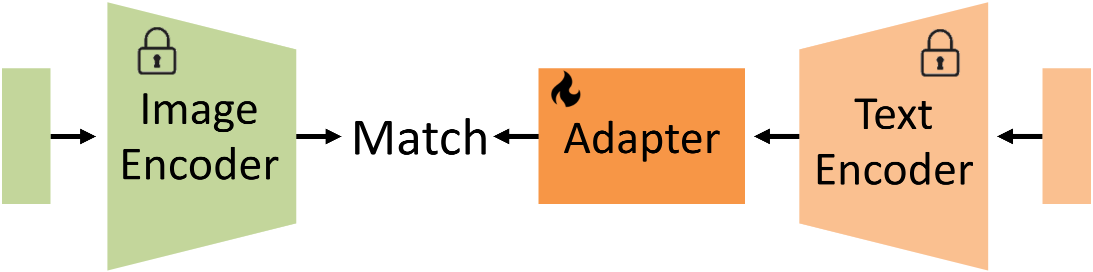


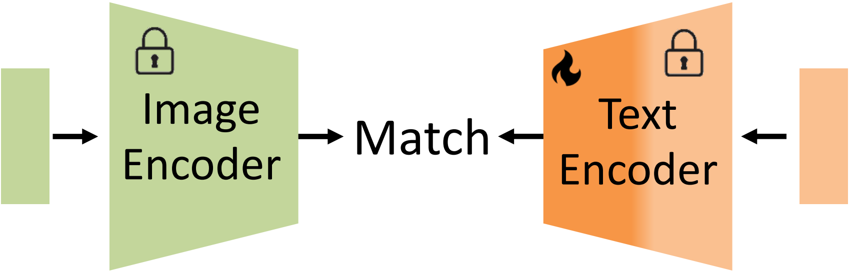
Illustration from Google AI Blog

Parameter-Efficient Fine-Tuning

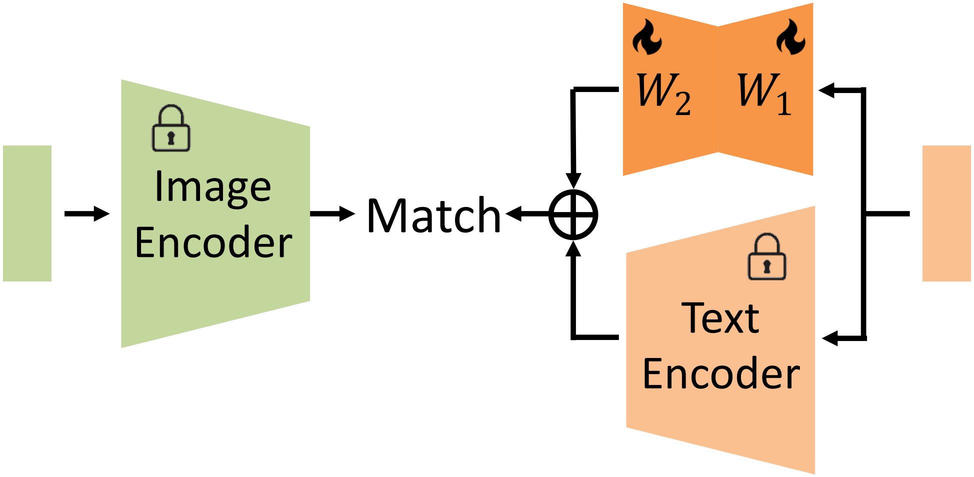
- Recently, many parameter-efficient finetuning methods have been proposed, including Adapter, Partial Finetuning, LoRA, and Prompt Learning.



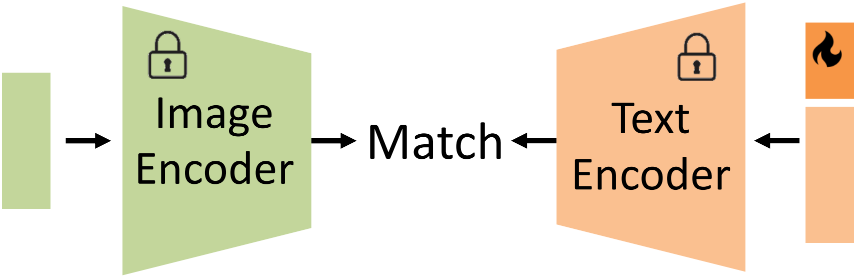
(a) Adapter



(b) Partial Finetuning



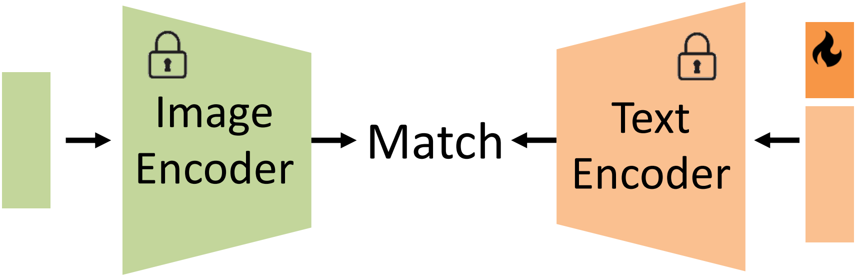
(c) LoRA



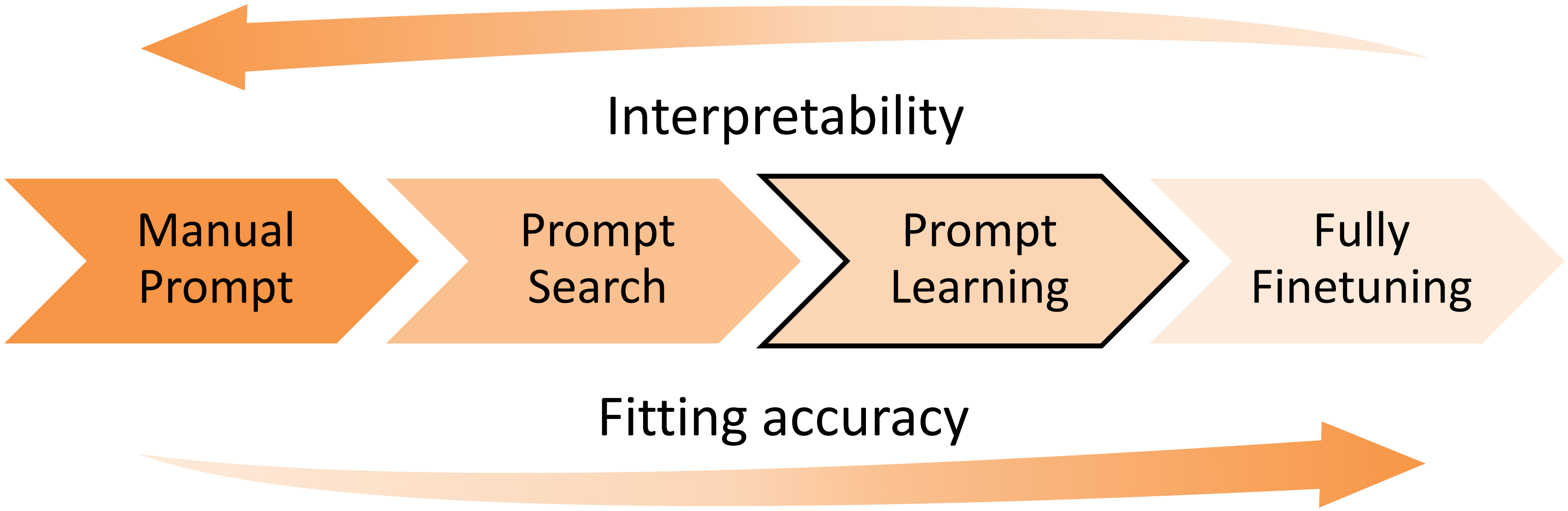
(d) Prompt Learning

Prompt Learning

➢ Prompt learning offers better interpretability because humans more easily understand textual tokens than model parameters.

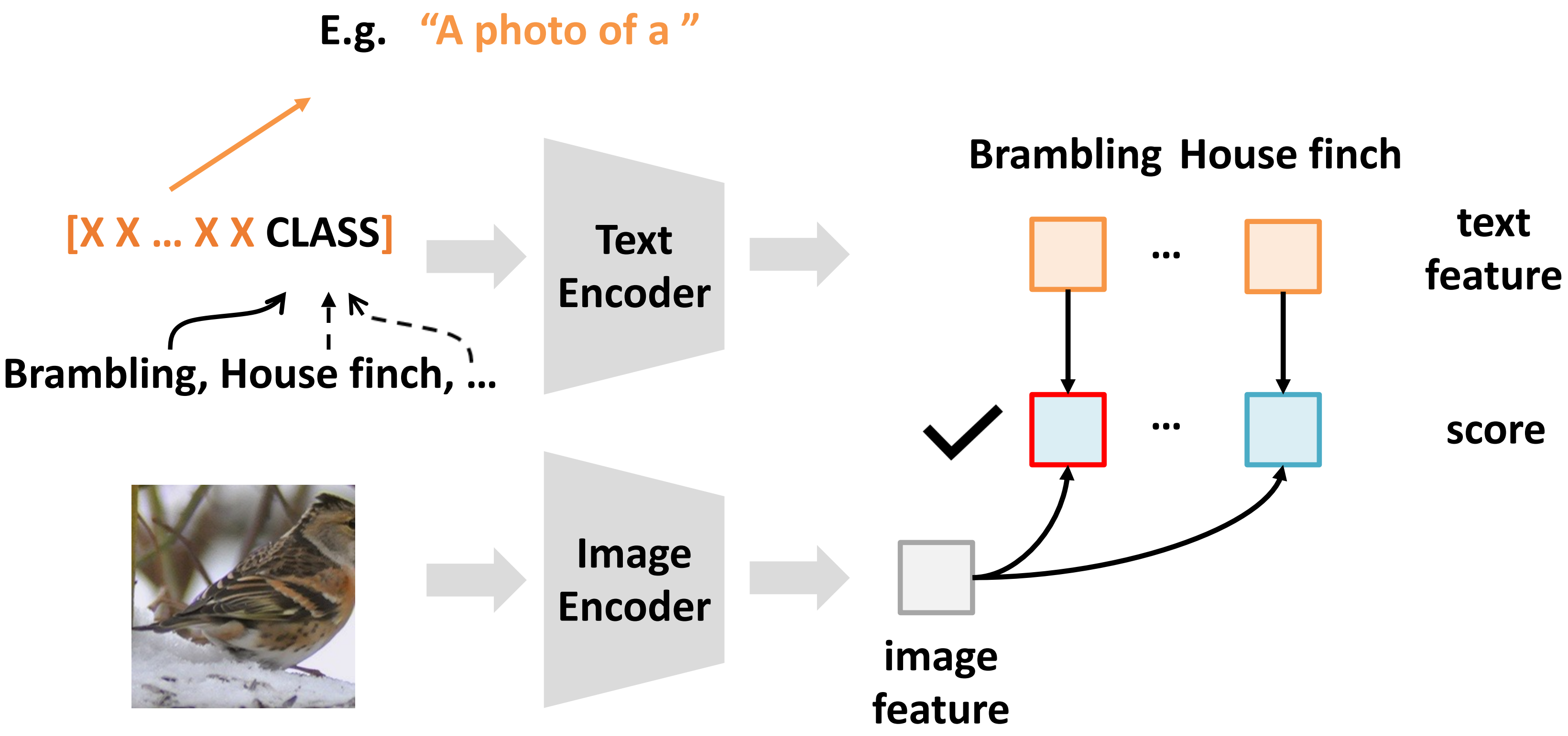


Prompt learning is a method that prepends a sequence of tokens to inputs for optimization, while keeping model parameters frozen.



Prompt Learning for VLP Models

- Build the template as: learnable prompts + classname
- Fix the model parameters and learn the prompts with few-shot annotation



The Gap Across Modalities

- The granularity of information in images and captions for contrastive pretraining is mismatched. Images contain more detailed visual contexts than the high-level overviews provided in captions.

The image contains detailed, fine-grained context.

The textural prompts are always coarse, high-level overviews.

How to bridge this gap and leverage visual context

Local Language Compatibility

- **Key findings**: The local features extracted from the CLIP image encoder are language-compatible, enabling fine-grained local text-patch matching.

DenseCLIP
Learning prompts for dense prediction

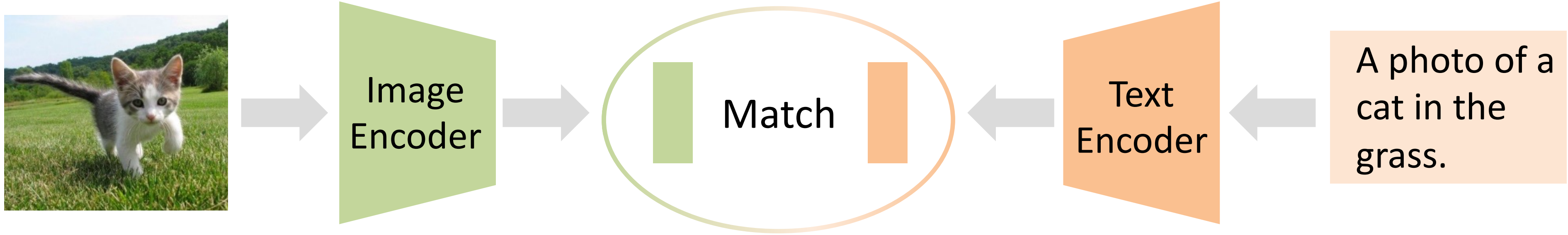
PLOT
Learning multiple comprehensive prompts

DenseCLIP: Language-Guided Dense Prediction with Context-Aware Prompting. Yongming Rao, Wenliang Zhao, Guangyi Chen, Yansong Tang, Zheng Zhu, Guan Huang, Jie Zhou, Jiwen Lu. CVPR, 2022.

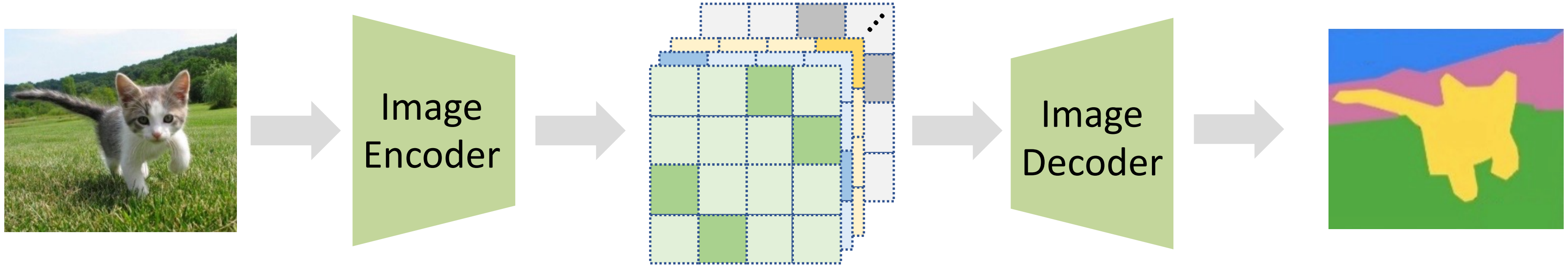
PLOT: Prompt Learning with Optimal Transport for Vision-Language Models. Guangyi Chen, Weiran Yao, Xiangchen Song, Xinyue Li, Yongming Rao, Kun Zhang. ICLR, 2023. (Spotlight)

Prompt Learning for Dense Prediction

- How to transfer the image-text matching pre-trained model to more complex dense prediction tasks, such as segmentation, and detection.



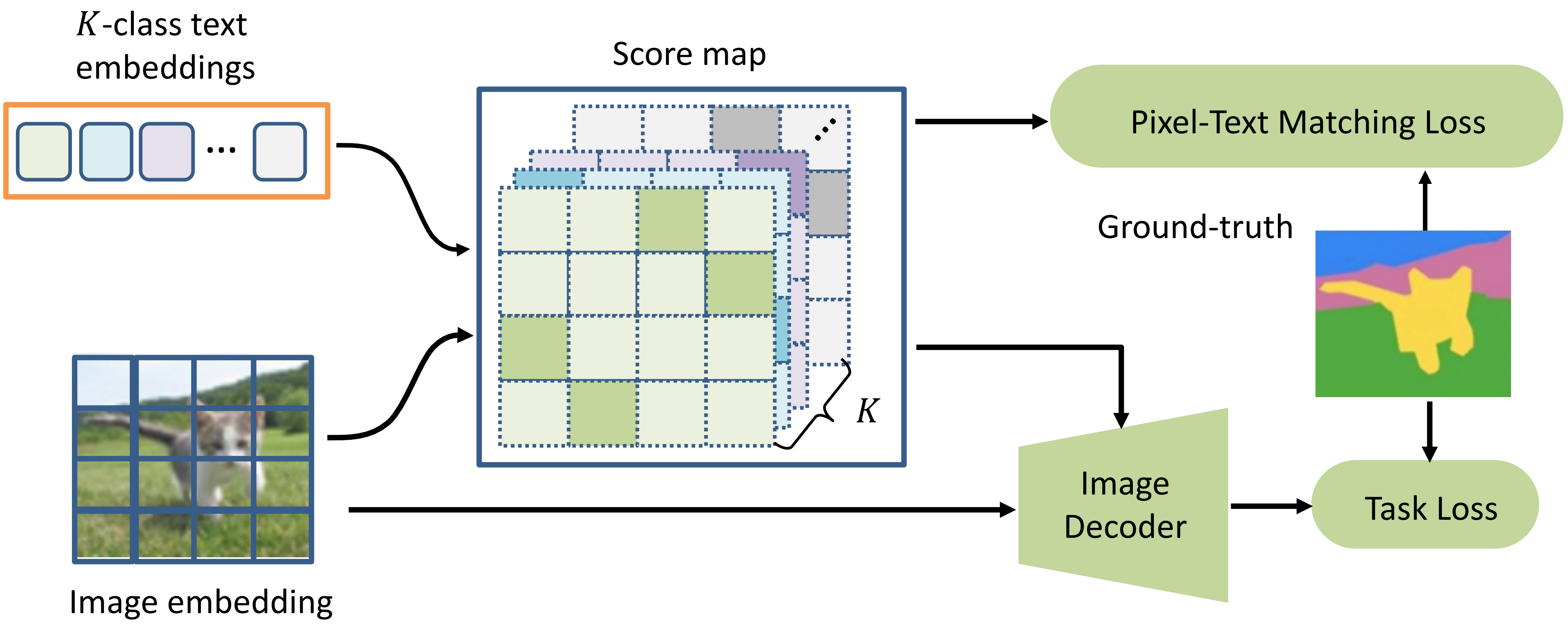
Pretraining with image-text matching task



Downstream dense prediction task

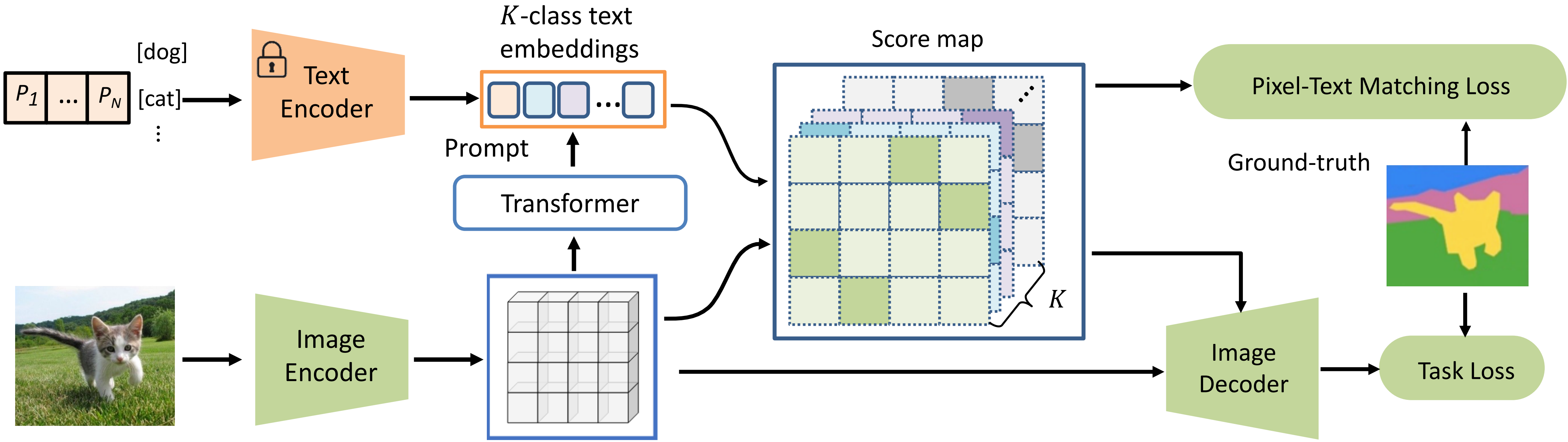
How to Use Local Matches

- Compute the pixel-text score maps using local visual and textual embeddings
- Apply pixel-text matching task as an auxiliary loss to refine features
- Concatenate the score maps as features to incorporate language priors



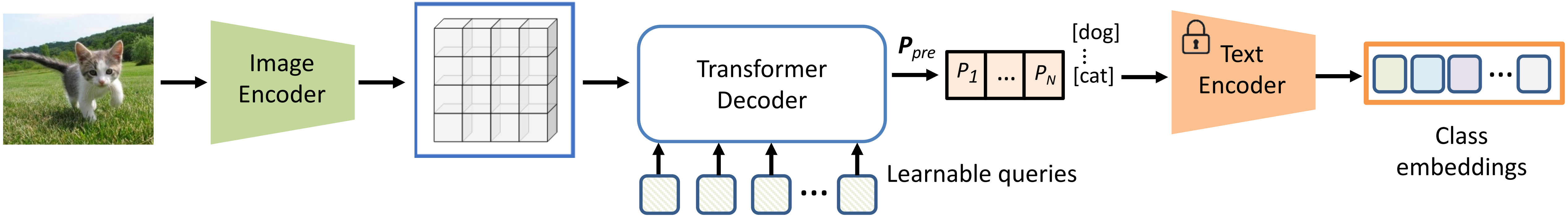
How to Learn Local Matches

- Finetune the image encoder for visual embeddings and learn prompts to refine textual embeddings
- Involve visual contexts to refine text features with the cross-attention mechanism

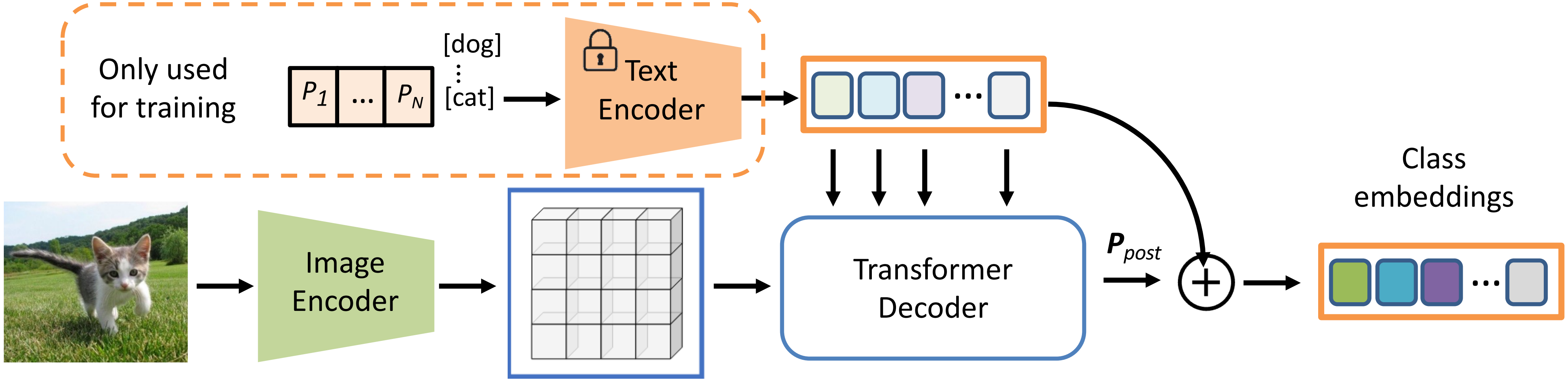


Pre/Post Model Prompting

- ➢ Two different strategies of context-aware prompting, pre/post text encoder
- ➢ Post-model prompting is more effective and efficient



Pre Model Context-aware prompting



Post Model Context-aware prompting

Results on Semantic Segmentation

- CLIP pre-train shows better performance than the ImageNet pre-train
- DenseCLIP better utilizes CLIP's knowledge than direct finetuning

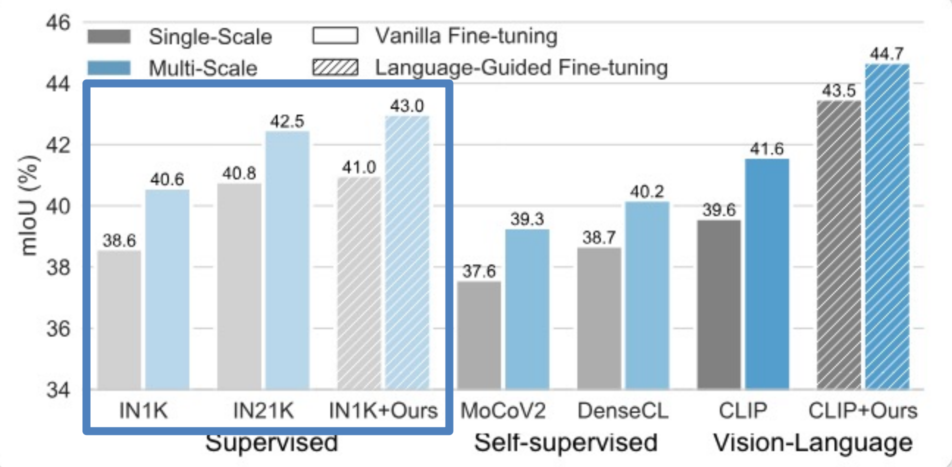
Backbone	Method	Pre-train	mIoU (SS)	mIoU (MS)	GFLOPs	Params (M)
ResNet-50	FCN [30]	ImageNet	36.1	38.1	793.3	49.6
	EncNet [55]	ImageNet	40.1	41.7	565.6	36.1
	PSPNet [57]	ImageNet	41.1	41.9	716.2	49.1
	CCNet [20]	ImageNet	42.1	43.1	804.0	49.9
	DeeplabV3+ [7]	ImageNet	42.7	43.8	711.5	43.7
	UperNet [45]	ImageNet	42.1	42.8	953.2	66.5
	DNL [52]	ImageNet	41.9	43.0	939.3	50.1
	Semantic FPN [21]	ImageNet	38.6	40.6	227.1	31.0
	CLIP + Semantic FPN	CLIP	39.6	41.6	248.8	31.0
	DenseCLIP + Semantic FPN	CLIP	**43.5**	**44.7**	269.2	50.3
ResNet-101	FCN [30]	ImageNet	39.9	41.4	1104.4	68.6
	EncNet [55]	ImageNet	42.6	44.7	876.8	55.1
	PSPNet [57]	ImageNet	43.6	44.4	1027.4	68.1
	CCNet [20]	ImageNet	44.0	45.2	1115.2	68.9
	DeeplabV3+ [7]	ImageNet	44.6	46.1	1022.7	62.7
	UperNet [45]	ImageNet	43.8	44.8	1031.0	85.5
	OCRNet [54]	ImageNet	45.3	-	923.9	55.5
	DNL [52]	ImageNet	44.3	45.8	1250.5	69.1
	Semantic FPN [21]	ImageNet	40.4	42.3	304.9	50.0
	CLIP + Semantic FPN	CLIP	42.7	44.3	326.6	50.0
	DenseCLIP + Semantic FPN	CLIP	**45.1**	**46.5**	346.3	67.8
ViT-B	SETR-MLA-DeiT [58]	ImageNet	46.2	47.7	-	-
	Semantic FPN [21]	ImageNet	48.3	50.9	1037.4	100.8
	Semantic FPN [21]	ImageNet-21K	49.1	50.4	1037.4	100.8
	CLIP + Semantic FPN	CLIP	49.4	50.3	1037.4	100.8
	DenseCLIP + Semantic FPN	CLIP	**50.6**	**51.3**	1043.1	105.3

More Experimental Analysis

- DenseCLIP can be applied for other dense prediction tasks, such as detection and instance segmentation, other base backbones, and other pre-trained datasets.

Model	FLOPs (G)	Params (M)	AP^b	AP^b_{50}	AP^b_{75}	AP^b_S	AP^b_M	AP^b_L	AP^m	AP^m_{50}	AP^m_{75}	AP^m_S	AP^m_M	AP^m_L
RN50-IN1K [18]	275	44	38.2	58.8	41.4	21.9	40.9	49.5	34.7	55.7	37.2	18.3	37.4	47.2
RN50-CLIP [33]	301	44	39.3	61.3	42.7	24.6	42.6	50.1	36.8	58.5	39.2	18.6	39.9	51.8
RN50-DenseCLIP	327	67	**40.2**	**63.2**	**43.9**	**26.3**	**44.2**	**51.0**	**37.6**	**60.2**	**39.8**	**20.8**	**40.7**	**53.7**
RN101-IN1K [18]	351	63	40.0	60.5	44.0	22.6	44.0	52.6	36.1	57.5	38.6	18.8	39.7	49.5
RN101-CLIP [33]	377	63	42.2	64.2	**46.5**	26.4	46.1	54.0	38.9	61.4	41.8	20.5	42.3	55.1
RN101-DenseCLIP	399	84	**42.6**	**65.1**	**46.5**	**27.7**	**46.5**	**54.2**	**39.6**	**62.4**	**42.4**	**21.4**	**43.0**	**56.2**

Decoder	Method	mIoU (SS) (%)	mIoU (MS) (%)
Semantic FPN [21]	RN50 [18]	38.6	40.6
	RN50 + DenseCLIP	**41.0** (+2.4)	**43.0** (+2.4)
	RN101 [18]	40.4	42.3
	RN101 + DenseCLIP	**43.0** (+2.6)	**45.2** (+2.9)
UperNet [45]	Swin-T [29]	44.5	45.8
	Swin-T + DenseCLIP	**45.4** (+0.9)	**46.5** (+0.7)
	Swin-S [29]	47.6	49.5
	Swin-S + DenseCLIP	**48.3** (+0.7)	**49.7** (+0.2)



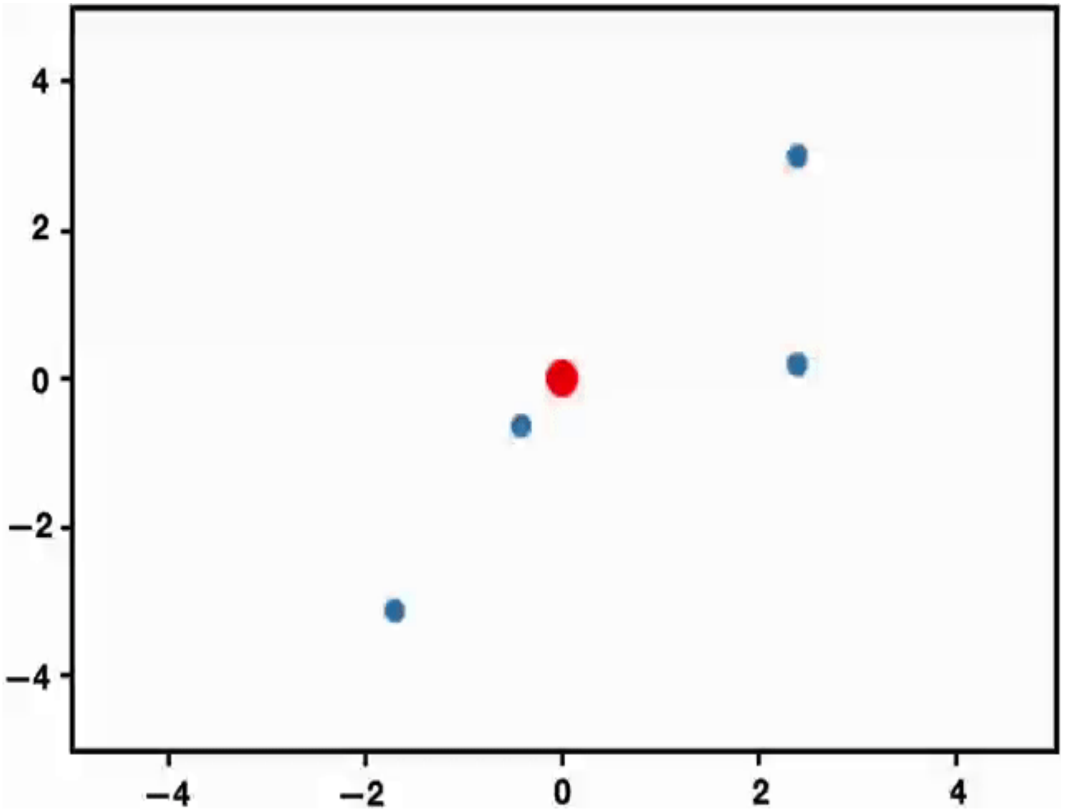
PLOT: Learning Multiple Prompts

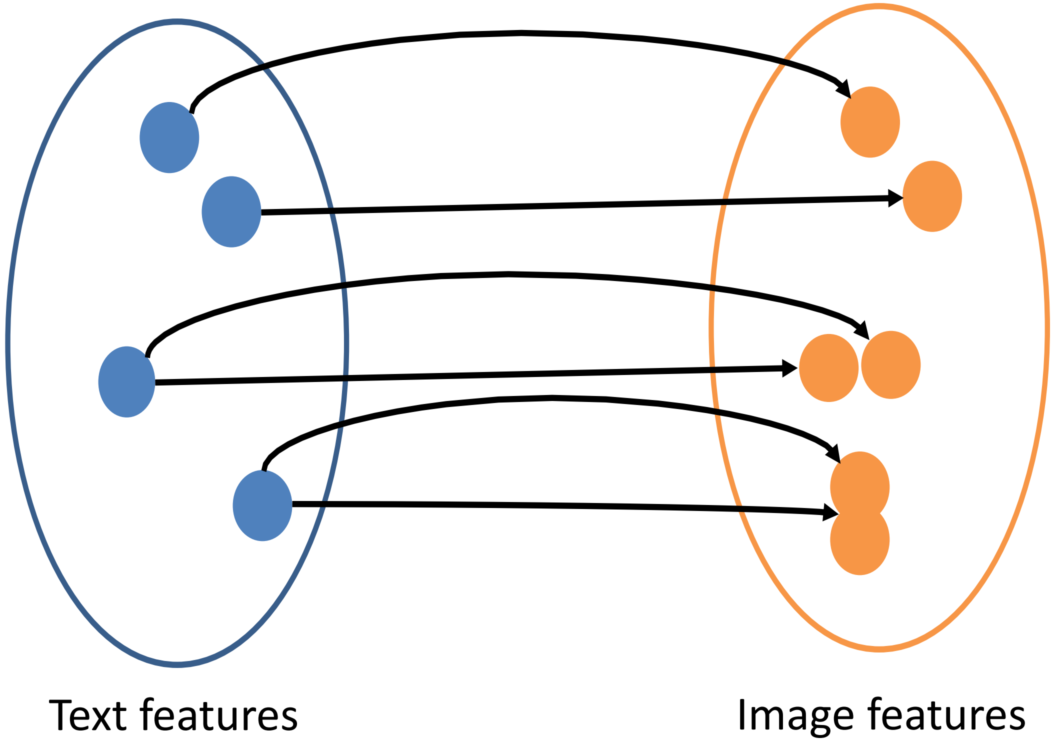
- Learning one sentence is intuitively insufficient to describe a class.
- One class can be described by many intrinsic characteristics and even extrinsic context relations.

Brambling

Key Idea

- ➢ How about directly learning multiple prompts? (even adding some constraints to push away the prompt from each other)
- ➢ The key idea is to use different local features to guide different prompts
- ➢ Formulate it as a set-to-set matching problem and use Optimal Transport to solve

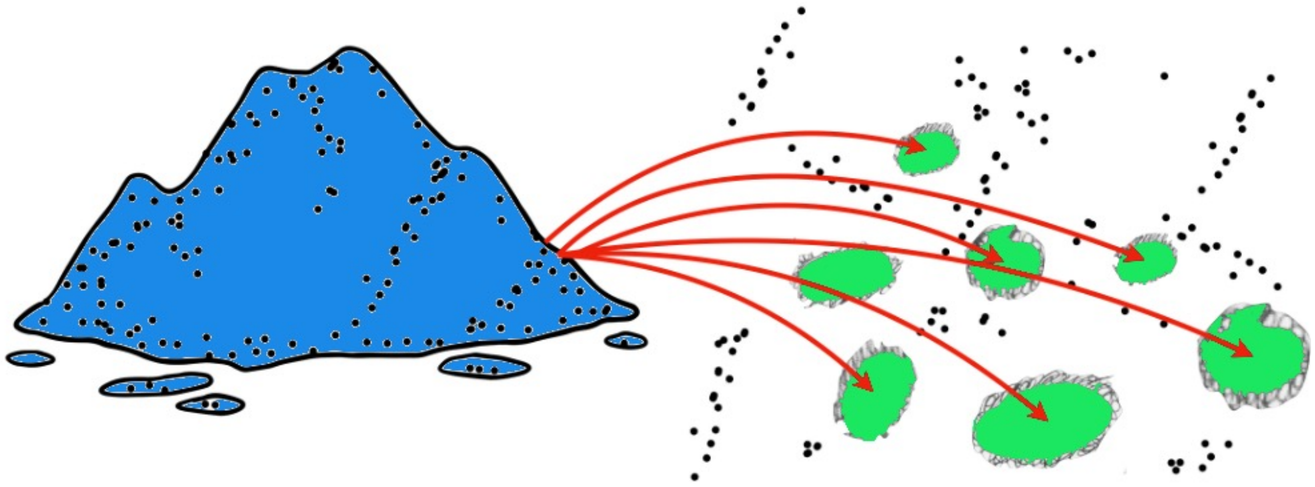




Optimal Transport (OT)

- The role of OT: minimize the cost when moving several items simultaneously.
- Extended role: compare two distributions given the cost function.

$$D_{OT}(U,V) = \inf_{\Gamma} \int_{\mathcal{X}\times\mathcal{Y}} \boldsymbol{c}(\boldsymbol{x},\boldsymbol{y}) d_{\gamma}(\boldsymbol{x},\boldsymbol{y})$$



Optimal Transport (OT)

- Here, we assume U, V are two discrete distributions (feature sets)

$$U = \sum_{m=1}^{M} \mu_m \delta_{f_m} \quad , \qquad V = \sum_{n=1}^{N} v_n \delta_{g_n}$$

- The cost function is defined by the distance between visual and textual features.

$$\boldsymbol{C}_{m,n} = 1 - \mathrm{sim}\,(\boldsymbol{f}_m, g_n)$$

- Optimal transport plan $\boldsymbol{T}$ is to minimize the total distance of two distributions (U, V)

$$d_{OT}(U, V|\boldsymbol{C}) = \min_{\boldsymbol{T}} \sum_{m=1}^{M} \sum_{n=1}^{N} \boldsymbol{T}_{m,n} \boldsymbol{C}_{m,n}$$

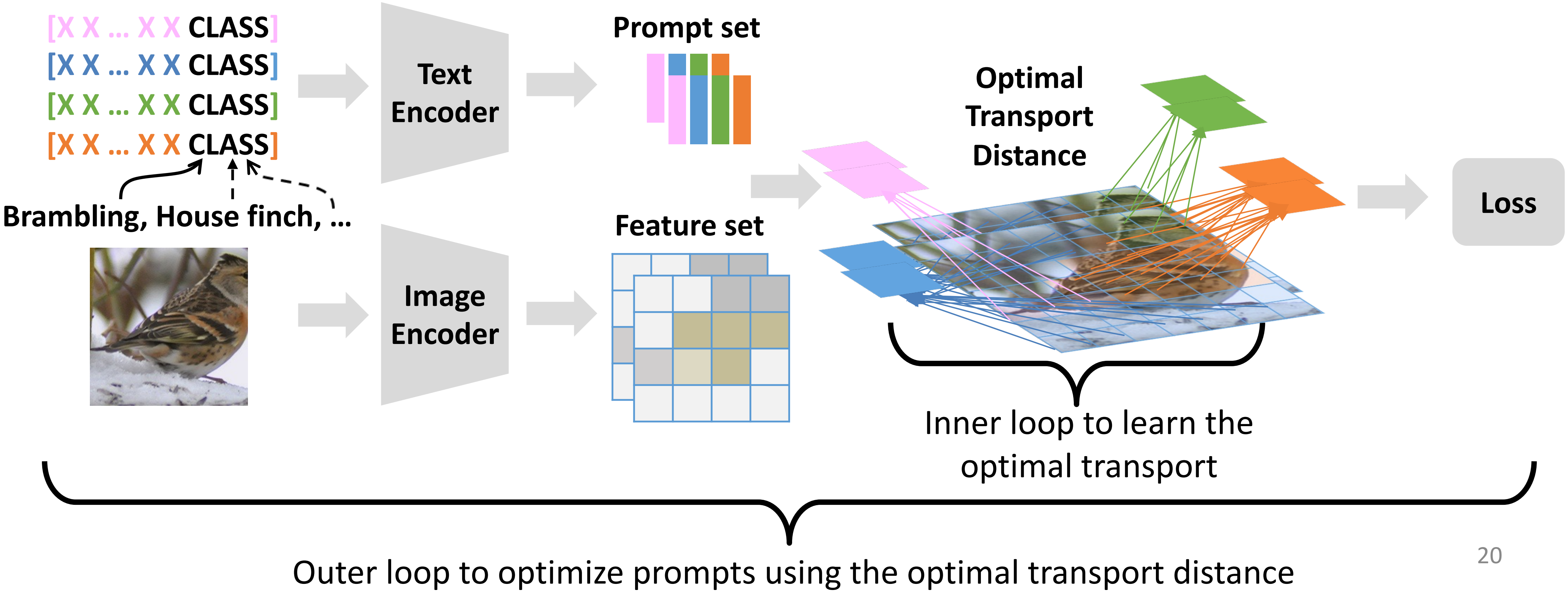
$$D_{OT}(U, V) = \inf_{\Gamma} \int_{\mathcal{X} \times \mathcal{Y}} \boldsymbol{c}(\boldsymbol{x}, \boldsymbol{y}) d_{\gamma}(\boldsymbol{x}, \boldsymbol{y})$$

$$\text{s.t.} \quad \boldsymbol{T}\mathbf{1}_{\mathrm{N}} = \boldsymbol{\mu}, \boldsymbol{T}^{T}\mathbf{1}_{M} = \boldsymbol{v}, \boldsymbol{T} \in \mathbb{R}_{+}^{M \times N}$$

- We apply the Sinkhorn algorithm (Curuti, 2013) for fast optimization

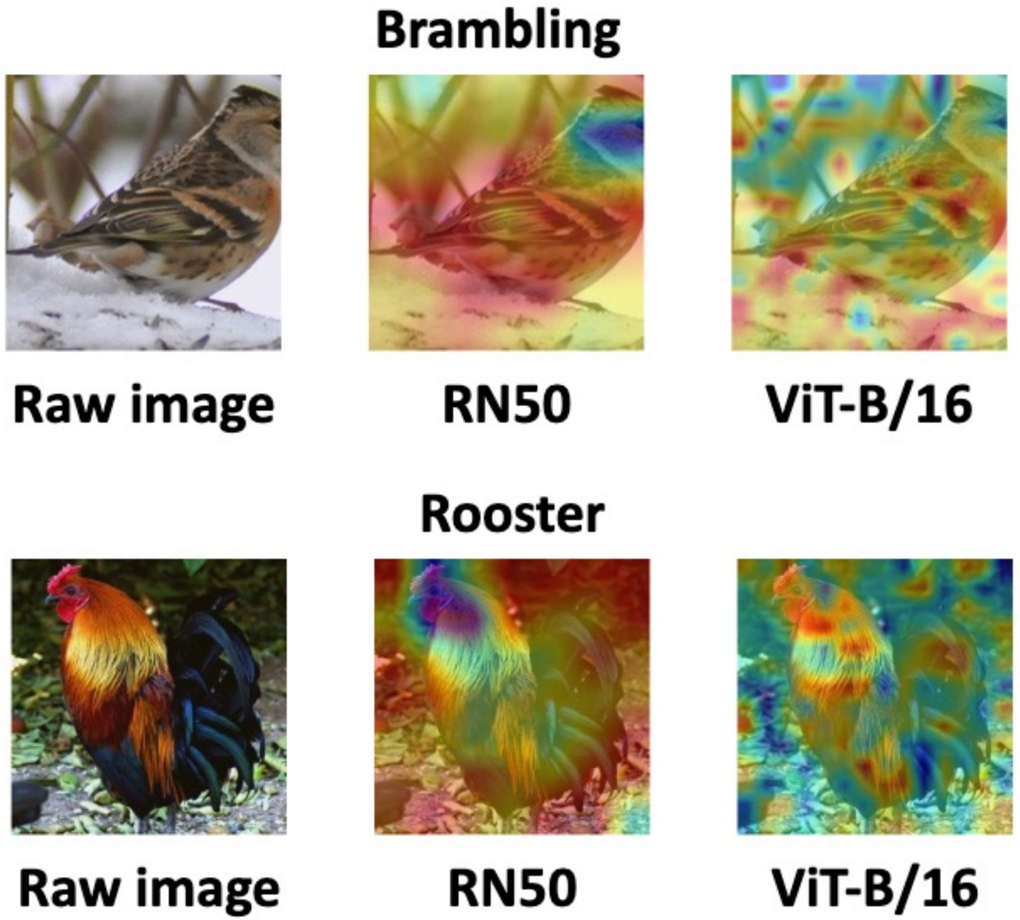
Overall Framework

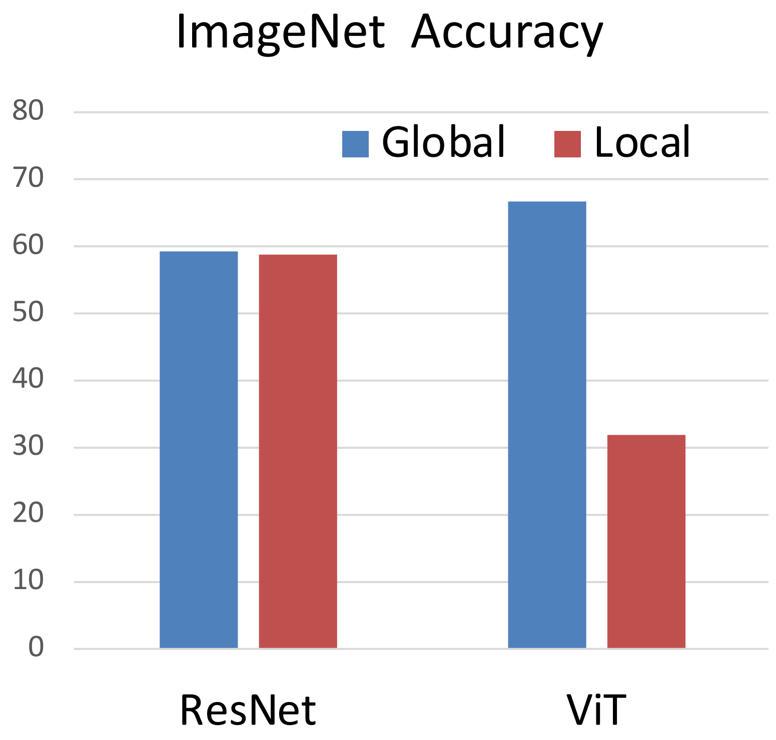
- Initialize multiple prompts and obtain the local feature sets
- Calculate the set-to-set distance between feature sets of prompts and visual patches using Optimal Transport.
- Two-stage optimization for learning optimal transport and prompts



When Local Matching Fails

- ➢ The success relies on effective language-compatible local features.
- ➢ What would happen if local matching fails, such as in ViT?

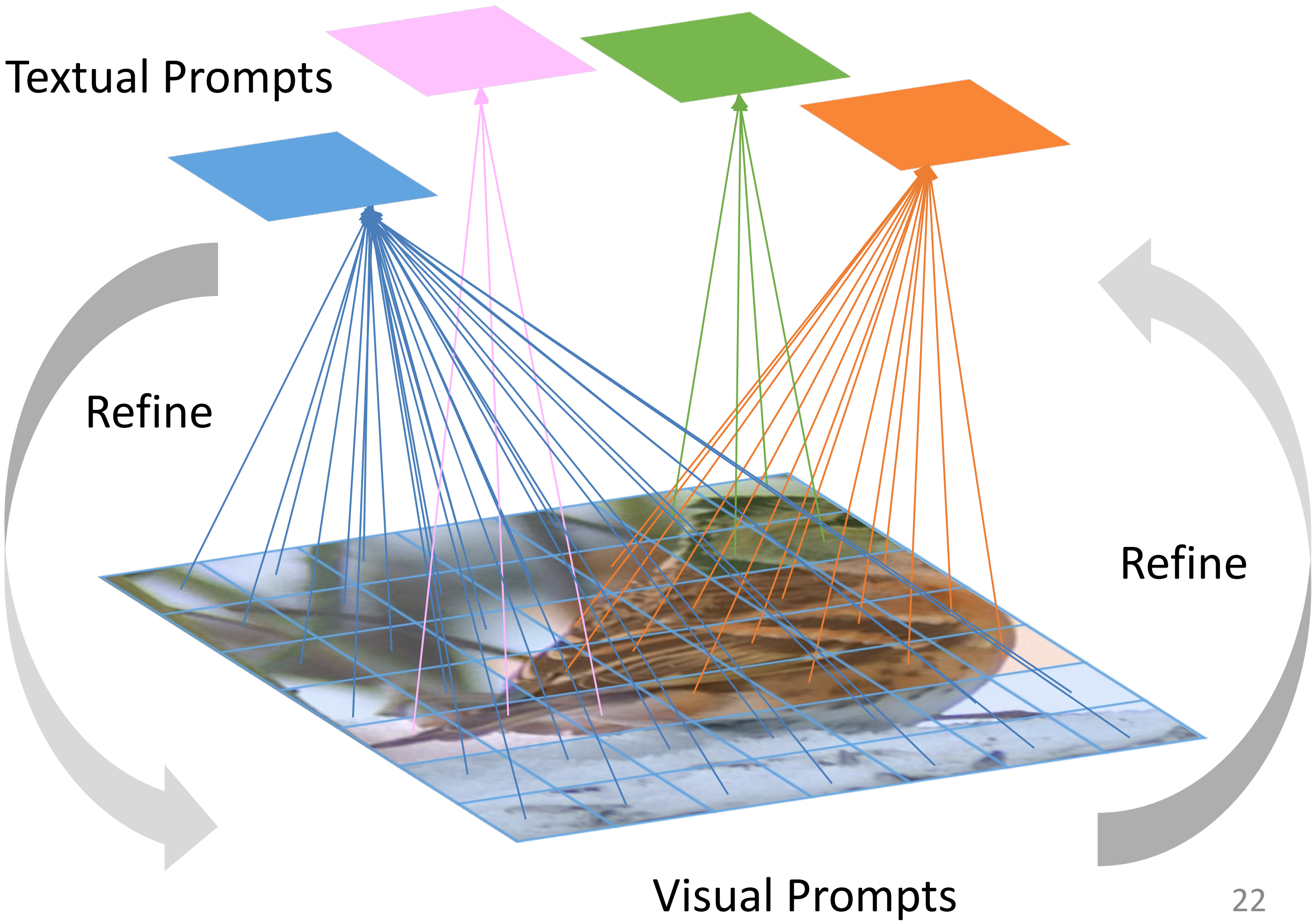




When Local Matching Fails (PLOT++)

- Jointly learning multiple visual and textual prompts, by guidance from each other.
- How do we get the starting point of this refinement?

[] exhibits a range of
[] is in the image
…
[] is a real-world object
[] exists in real-word scenario

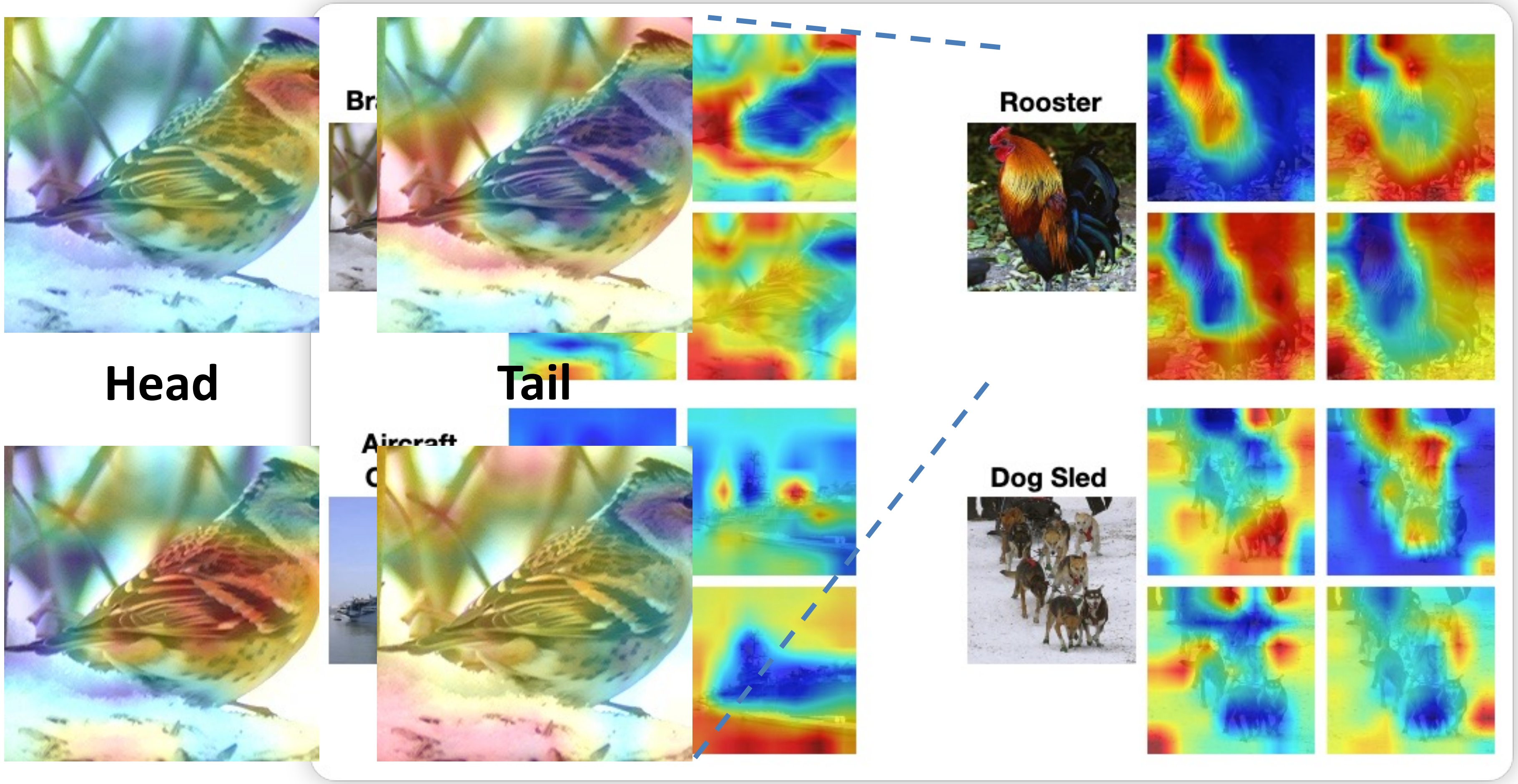


Generate the shared templates without knowing all class knowledge.

- Commonality
- Obviousness
- Diversity
- Brevity

What Can PLOT Learn

➢ We provide the heatmaps of transport plan related to each prompt on 4 classes in ImageNet. Different transport plans focus on different attributes.



Wings Environment

What Can PLOT Learn

MOHAMED BIN ZAYED UNIVERSITY OF ARTIFICIAL INTELLIGENCE

VALSE

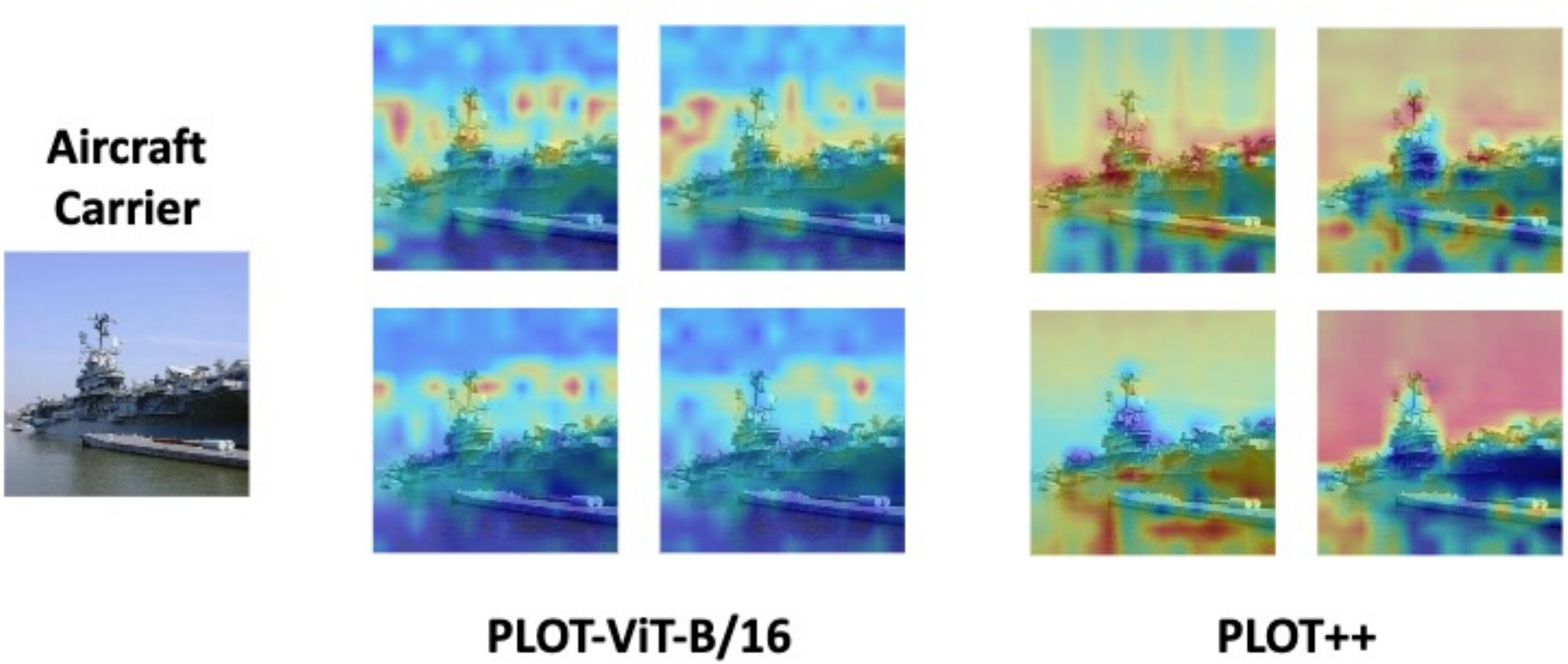
Carnegie Mellon University

➢ We show the nearest words for 16 context vectors of all 4 prompts learned by PLOT.

Number	Prompt 1	Prompt 2	Prompt 3	Prompt 4
1	ag	pa	trying	gaz
2	flint	as	field	**white**
3	leaving	wit	N/A	t
4	sot	1	icons	ario
5	tint	N/A	eclub	safe
6	tar	yl	indiffe	class
7	attn	N/A	ts	represented
8	2	job	cold	attend
9	rollingstones	built	yeah	vie
10	N/A	brought	band	recognized
11	N/A	or	love	old
12	bel	j	late	stel
13	**head**	ag	industry	awhile
14	artifact	bad	N/A	ded
15	an	chie	across	these
16	5	in	actual	visiting

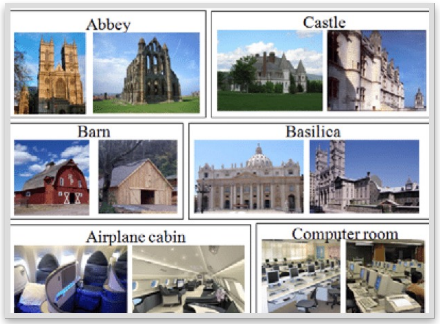
What Can PLOT++ Learn

➢ We provide the heatmaps of transport plans related to each prompt. The refined local visual features are more complementary and meaningful.

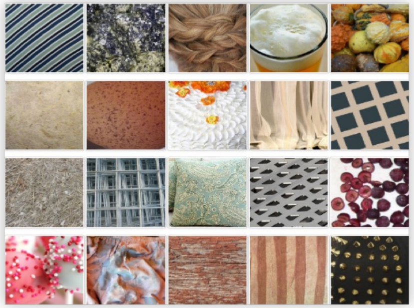


Experiments on 11 Datasets


StanfordCars

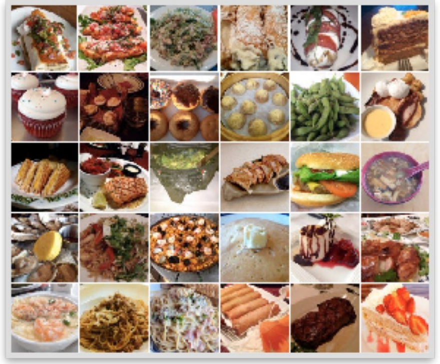


SUN397


DTD

UCF101


FGVCAircraft


Food101


EuroSAT

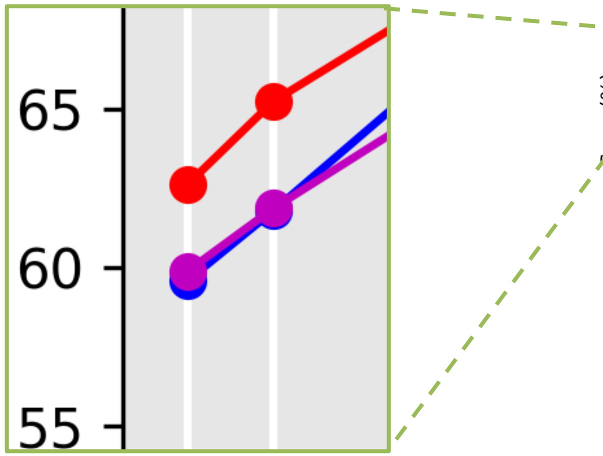
ImageNet

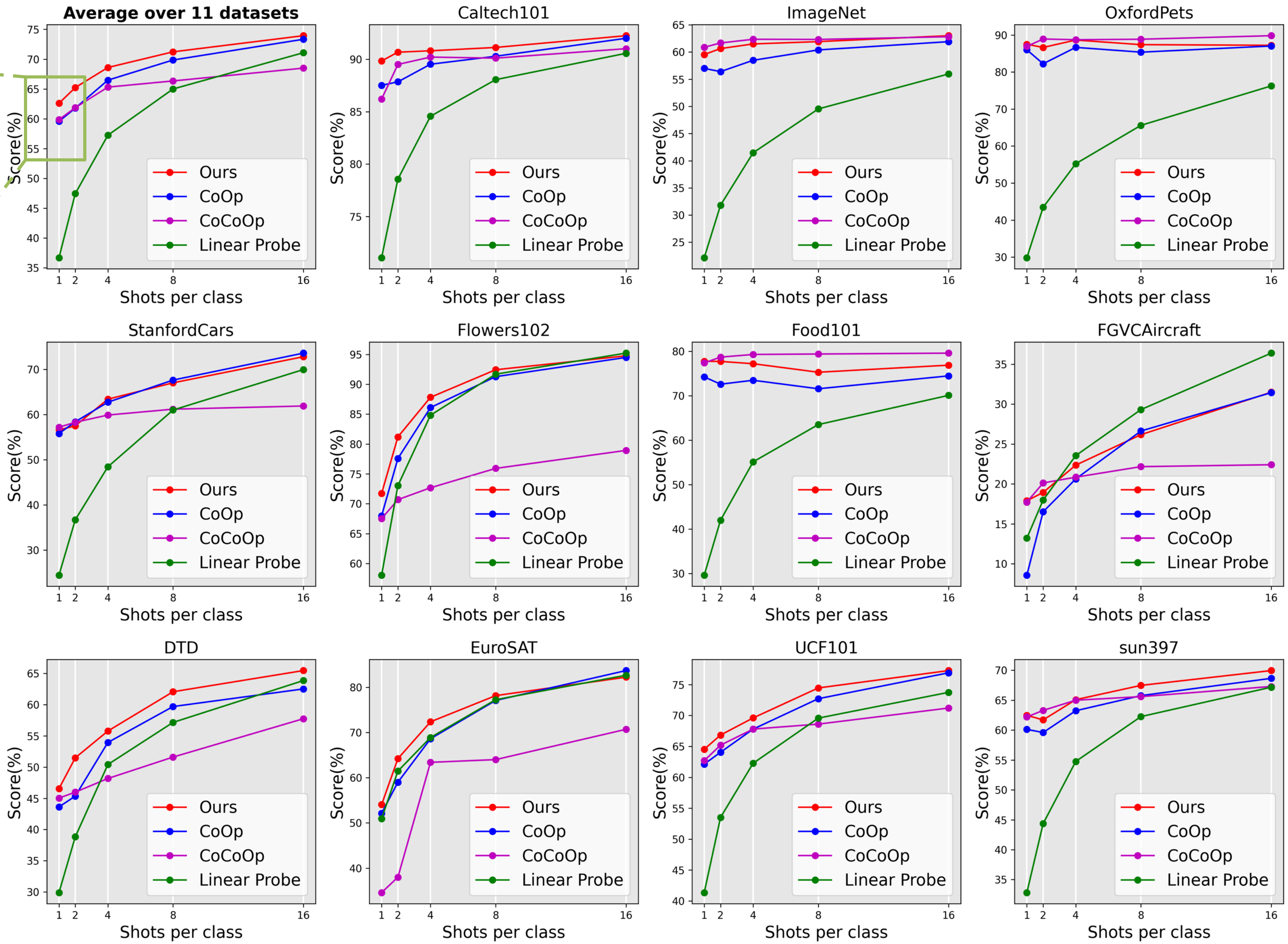

Caltech101


OxfordPets


Flowers102

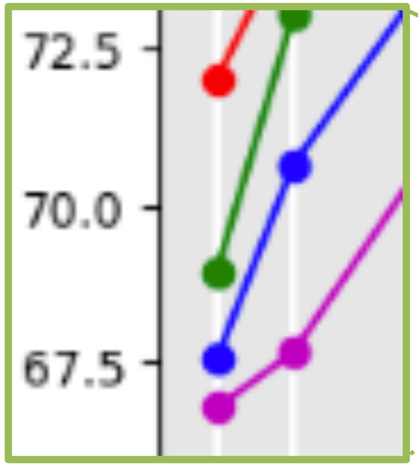
Few-shot Learning Results

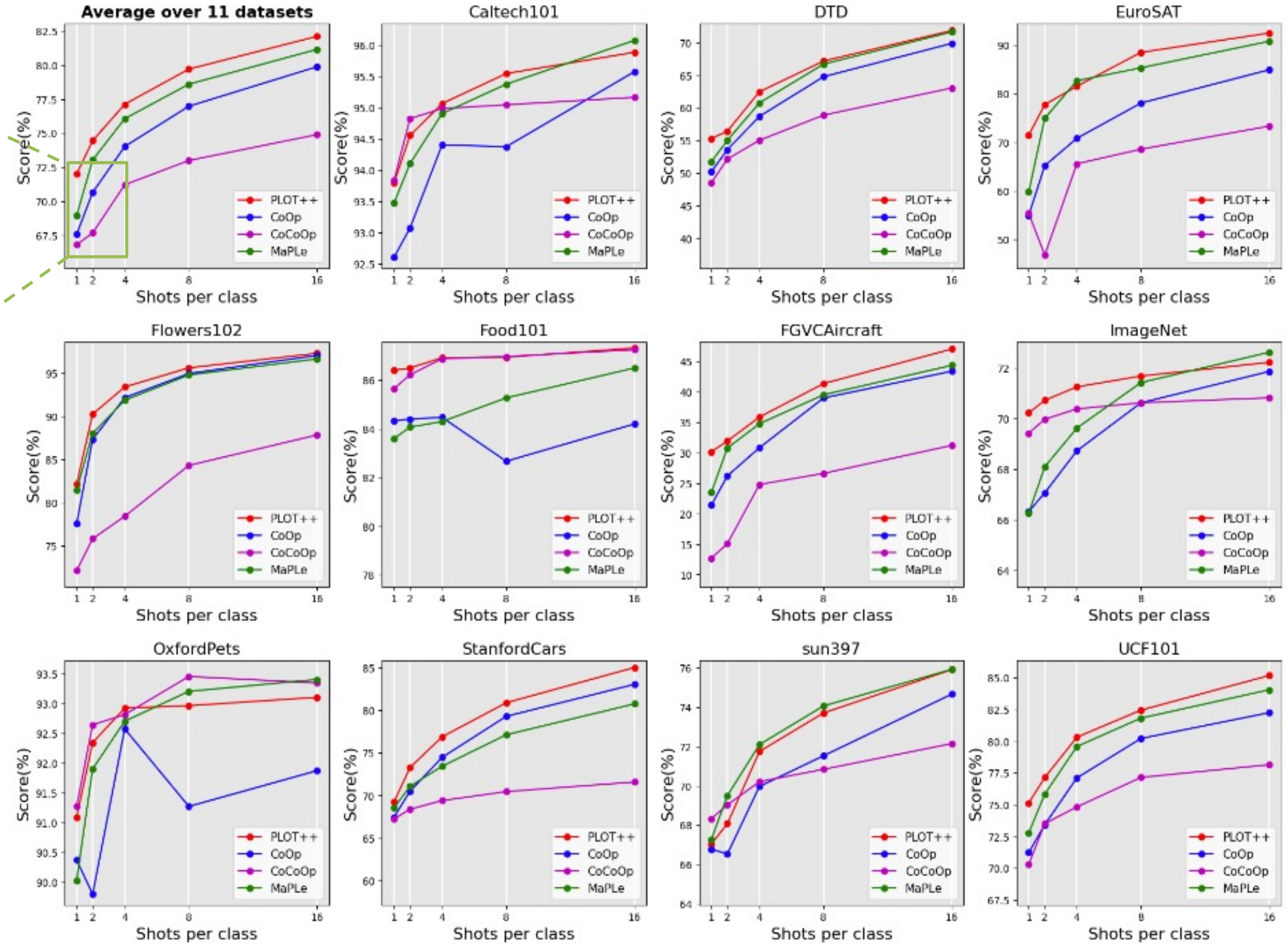




> 3% 1-shot performance improvement over CoOp (Zhou et al. 2021) and CoCoOp (Zhou et al. 2022)

Few-shot Learning Results of PLOT++





> 4.5% 1-shot performance improvement over CoOp (Zhou et al. 2021) and CoCoOp (Zhou et al. 2022)

Ablation Studies

- We conducted the ablation studies on three datasets.

Dataset	Settings	1 shot	2 shots	4 shots	8 shots	16 shots
Caltech101	PLOT	**89.83 ± 0.33**	**90.67 ± 0.21**	**90.80 ± 0.20**	**91.54 ± 0.33**	**92.24 ± 0.38**
	CoOp	87.51 ± 1.02	87.84 ± 1.10	89.52 ± 0.80	90.28 ± 0.42	91.99 ± 0.31
	G	88.13 ± 0.36	86.98 ± 1.25	88.45 ± 0.79	90.16 ± 0.22	90.72 ± 0.18
	G+V	88.28 ± 0.43	87.72 ± 1.25	88.45 ± 0.30	89.82 ± 0.20	92.00 ± 0.13
	M	69.78 ± 1.75	71.57 ± 1.59	77.18 ± 2.16	81.77 ± 0.47	86.21 ± 0.20
	M+V	66.11 ± 8.29	71.45 ± 3.98	79.30 ± 3.96	86.96 ± 0.78	89.80 ± 0.17
DTD	PLOT	**46.55 ± 2.62**	**51.24 ± 1.95**	**56.03 ± 0.43**	**61.70 ± 0.35**	**65.60 ± 0.82**
	CoOp	43.62 ± 1.96	45.35 ± 0.31	53.94 ± 1.37	59.69 ± 0.13	62.51 ± 0.25
	G	45.12 ± 1.69	48.39 ± 2.08	54.75 ± 0.48	60.15 ± 0.70	63.59 ± 0.76
	G+V	45.90 ± 2.00	48.50 ± 0.99	53.96 ± 0.48	59.69 ± 1.01	63.51 ± 0.66
	M	13.18 ± 4.57	12.25 ± 3.86	13.00 ± 4.73	20.76 ± 5.42	26.99 ± 1.98
	M+V	12.61 ± 5.93	15.11 ± 1.81	20.35 ± 1.33	44.13 ± 2.39	56.85 ± 0.54
FOOD101	PLOT	**77.74 ± 0.47**	**77.70 ± 0.02**	**77.21 ± 0.43**	**75.31 ± 0.30**	**77.09 ± 0.18**
	CoOp	74.25 ± 1.52	72.61 ± 1.33	73.49 ± 2.03	71.58 ± 0.79	74.48 ± 0.15
	G	74.63 ± 0.11	70.15 ± 0.49	70.41 ± 0.46	70.72 ± 0.98	73.68 ± 0.46
	G+V	74.83 ± 0.31	70.09 ± 0.85	70.86 ± 0.22	70.80 ± 0.68	73.93 ± 0.35
	M	52.02 ± 4.86	46.12 ± 1.46	46.86 ± 1.39	53.43 ± 0.88	61.28 ± 0.23
	M+V	46.52 ± 1.15	45.95 ± 2.66	53.57 ± 0.83	62.95 ± 0.37	67.63 ± 1.11

MOHAMED BIN ZAYED UNIVERSITY OF ARTIFICIAL INTELLIGENCE

Ablation Studies

VALSE Carnegie Mellon University

- Q: Can we directly learn multiple prompts by matching it with the global visual feature? A: **No**.

Dataset	Settings	1 shot	2 shots	4 shots	8 shots	16 shots
Caltech101	PLOT	**89.83 ± 0.33**	**90.67 ± 0.21**	**90.80 ± 0.20**	**91.54 ± 0.33**	**92.24 ± 0.38**
	CoOp	87.51 ± 1.02	87.84 ± 1.10	89.52 ± 0.80	90.28 ± 0.42	91.99 ± 0.31
	G	88.13 ± 0.36	86.98 ± 1.25	88.45 ± 0.79	90.16 ± 0.22	90.72 ± 0.18
	G+V	88.28 ± 0.43	87.72 ± 1.25	88.45 ± 0.30	89.82 ± 0.20	92.00 ± 0.13
	M	69.78 ± 1.75	71.57 ± 1.59	77.18 ± 2.16	81.77 ± 0.47	86.21 ± 0.20
	M+V	66.11 ± 8.29	71.45 ± 3.98	79.30 ± 3.96	86.96 ± 0.78	89.80 ± 0.17
DTD	PLOT	**46.55 ± 2.62**	**51.24 ± 1.95**	**56.03 ± 0.43**	**61.70 ± 0.35**	**65.60 ± 0.82**
	CoOp	43.62 ± 1.96	45.35 ± 0.31	53.94 ± 1.37	59.69 ± 0.13	62.51 ± 0.25
	G	45.12 ± 1.69	48.39 ± 2.08	54.75 ± 0.48	60.15 ± 0.70	63.59 ± 0.76
	G+V	45.90 ± 2.00	48.50 ± 0.99	53.96 ± 0.48	59.69 ± 1.01	63.51 ± 0.66
	M	13.18 ± 4.57	12.25 ± 3.86	13.00 ± 4.73	20.76 ± 5.42	26.99 ± 1.98
	M+V	12.61 ± 5.93	15.11 ± 1.81	20.35 ± 1.33	44.13 ± 2.39	56.85 ± 0.54
FOOD101	PLOT	**77.74 ± 0.47**	**77.70 ± 0.02**	**77.21 ± 0.43**	**75.31 ± 0.30**	**77.09 ± 0.18**
	CoOp	74.25 ± 1.52	72.61 ± 1.33	73.49 ± 2.03	71.58 ± 0.79	74.48 ± 0.15
	G	74.63 ± 0.11	70.15 ± 0.49	70.41 ± 0.46	70.72 ± 0.98	73.68 ± 0.46
	G+V	74.83 ± 0.31	70.09 ± 0.85	70.86 ± 0.22	70.80 ± 0.68	73.93 ± 0.35
	M	52.02 ± 4.86	46.12 ± 1.46	46.86 ± 1.39	53.43 ± 0.88	61.28 ± 0.23
	M+V	46.52 ± 1.15	45.95 ± 2.66	53.57 ± 0.83	62.95 ± 0.37	67.63 ± 1.11

Ablation Studies

➢ Q: Can methods that encourage the variety of prompts work well? A: **Not** really.

Dataset	Settings	1 shot	2 shots	4 shots	8 shots	16 shots
Caltech101	PLOT	**89.83 ± 0.33**	**90.67 ± 0.21**	**90.80 ± 0.20**	**91.54 ± 0.33**	**92.24 ± 0.38**
	CoOp	87.51 ± 1.02	87.84 ± 1.10	89.52 ± 0.80	90.28 ± 0.42	91.99 ± 0.31
	G	88.13 ± 0.36	86.98 ± 1.25	88.45 ± 0.79	90.16 ± 0.22	90.72 ± 0.18
	G+V	88.28 ± 0.43	87.72 ± 1.25	88.45 ± 0.30	89.82 ± 0.20	92.00 ± 0.13
	M	69.78 ± 1.75	71.57 ± 1.59	77.18 ± 2.16	81.77 ± 0.47	86.21 ± 0.20
	M+V	66.11 ± 8.29	71.45 ± 3.98	79.30 ± 3.96	86.96 ± 0.78	89.80 ± 0.17
DTD	PLOT	**46.55 ± 2.62**	**51.24 ± 1.95**	**56.03 ± 0.43**	**61.70 ± 0.35**	**65.60 ± 0.82**
	CoOp	43.62 ± 1.96	45.35 ± 0.31	53.94 ± 1.37	59.69 ± 0.13	62.51 ± 0.25
	G	45.12 ± 1.69	48.39 ± 2.08	54.75 ± 0.48	60.15 ± 0.70	63.59 ± 0.76
	G+V	45.90 ± 2.00	48.50 ± 0.99	53.96 ± 0.48	59.69 ± 1.01	63.51 ± 0.66
	M	13.18 ± 4.57	12.25 ± 3.86	13.00 ± 4.73	20.76 ± 5.42	26.99 ± 1.98
	M+V	12.61 ± 5.93	15.11 ± 1.81	20.35 ± 1.33	44.13 ± 2.39	56.85 ± 0.54
FOOD101	PLOT	**77.74 ± 0.47**	**77.70 ± 0.02**	**77.21 ± 0.43**	**75.31 ± 0.30**	**77.09 ± 0.18**
	CoOp	74.25 ± 1.52	72.61 ± 1.33	73.49 ± 2.03	71.58 ± 0.79	74.48 ± 0.15
	G	74.63 ± 0.11	70.15 ± 0.49	70.41 ± 0.46	70.72 ± 0.98	73.68 ± 0.46
	G+V	74.83 ± 0.31	70.09 ± 0.85	70.86 ± 0.22	70.80 ± 0.68	73.93 ± 0.35
	M	52.02 ± 4.86	46.12 ± 1.46	46.86 ± 1.39	53.43 ± 0.88	61.28 ± 0.23
	M+V	46.52 ± 1.15	45.95 ± 2.66	53.57 ± 0.83	62.95 ± 0.37	67.63 ± 1.11

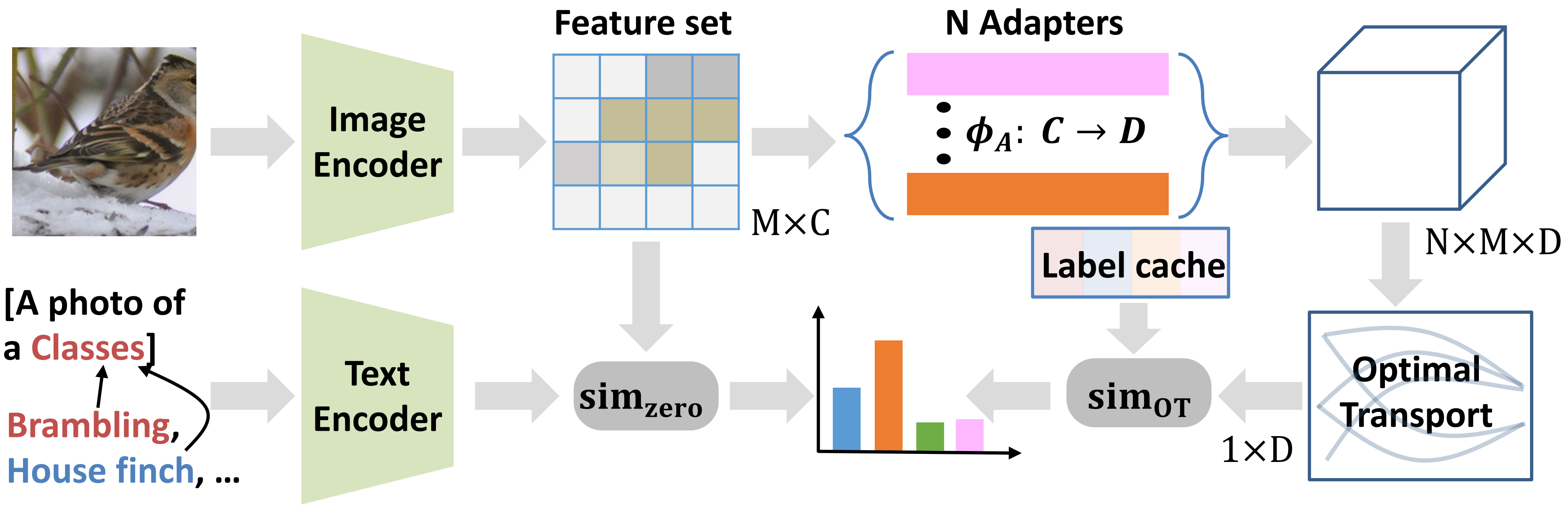
Ablation Studies

➢ Q: Does the improvement mainly come from using all local feature maps? A: **No**.

Dataset	Settings	1 shot	2 shots	4 shots	8 shots	16 shots
Caltech101	PLOT	**89.83 ± 0.33**	**90.67 ± 0.21**	**90.80 ± 0.20**	**91.54 ± 0.33**	**92.24 ± 0.38**
	CoOp	87.51 ± 1.02	87.84 ± 1.10	89.52 ± 0.80	90.28 ± 0.42	91.99 ± 0.31
	G	88.13 ± 0.36	86.98 ± 1.25	88.45 ± 0.79	90.16 ± 0.22	90.72 ± 0.18
	G+V	88.28 ± 0.43	87.72 ± 1.25	88.45 ± 0.30	89.82 ± 0.20	92.00 ± 0.13
	M	69.78 ± 1.75	71.57 ± 1.59	77.18 ± 2.16	81.77 ± 0.47	86.21 ± 0.20
	M+V	66.11 ± 8.29	71.45 ± 3.98	79.30 ± 3.96	86.96 ± 0.78	89.80 ± 0.17
DTD	PLOT	**46.55 ± 2.62**	**51.24 ± 1.95**	**56.03 ± 0.43**	**61.70 ± 0.35**	**65.60 ± 0.82**
	CoOp	43.62 ± 1.96	45.35 ± 0.31	53.94 ± 1.37	59.69 ± 0.13	62.51 ± 0.25
	G	45.12 ± 1.69	48.39 ± 2.08	54.75 ± 0.48	60.15 ± 0.70	63.59 ± 0.76
	G+V	45.90 ± 2.00	48.50 ± 0.99	53.96 ± 0.48	59.69 ± 1.01	63.51 ± 0.66
	M	13.18 ± 4.57	12.25 ± 3.86	13.00 ± 4.73	20.76 ± 5.42	26.99 ± 1.98
	M+V	12.61 ± 5.93	15.11 ± 1.81	20.35 ± 1.33	44.13 ± 2.39	56.85 ± 0.54
FOOD101	PLOT	**77.74 ± 0.47**	**77.70 ± 0.02**	**77.21 ± 0.43**	**75.31 ± 0.30**	**77.09 ± 0.18**
	CoOp	74.25 ± 1.52	72.61 ± 1.33	73.49 ± 2.03	71.58 ± 0.79	74.48 ± 0.15
	G	74.63 ± 0.11	70.15 ± 0.49	70.41 ± 0.46	70.72 ± 0.98	73.68 ± 0.46
	G+V	74.83 ± 0.31	70.09 ± 0.85	70.86 ± 0.22	70.80 ± 0.68	73.93 ± 0.35
	M	52.02 ± 4.86	46.12 ± 1.46	46.86 ± 1.39	53.43 ± 0.88	61.28 ± 0.23
	M+V	46.52 ± 1.15	45.95 ± 2.66	53.57 ± 0.83	62.95 ± 0.37	67.63 ± 1.11

Application on Adapter

➢ Q: Can PLOT benefit Adapter-based methods? A: **Yes**.



Average	PLOT-A	65.45	68.63	71.23	73.49	76.20
	Tip	64.62	66.65	69.67	72.45	75.83

Consistent performance improvement over Tip-Adapter (Zhang et al. 2022)

Computation Cost

- Q: What is the extra computation time cost of PLOT over CoOp baseline? A: Around **10%** inference speed and **5%** training time.

Settings	CoOp	PLOT (N=1)	PLOT (N=2)	PLOT (N=4)	PLOT (N=8)
Training Time (s)	1.127	1.135	1.148	1.182	1.267
Inference Time (images/s)	719.1	714.4	690.7	653.0	519.8

	CoOp	CoCoOp	MaPLe	PLOT	PLOT++
Model Size	8.2k	41.5k	3,555.1k	32.8k	14.3k

Takeaway Points

- 1) There are gaps in information granularity between image contexts and text captions in current contrastive vision-language pre-trained models.
- 2) Good finding: For CLIP, the local visual features are language-compatible.
- 3) This property can help prompt learning, such as learning prompts for dense prediction, and learning multiple comprehensive prompts.
- 4) However, in ViT-based CLIP models, local visual features are not sufficiently language-compatible. In such cases, it becomes beneficial to jointly refine both prompts and visual features.

Thanks for your listening