Advancing Textual Prompt Learning with Anchored Attributes

ICCV 2025

Zheng Li, Yibing Song, Ming-Ming Cheng, Xiang Li, Jian Yang.

Nankai University & DAMO Academy, Alibaba Group

Email: zhengli97@qq.com

Outlines

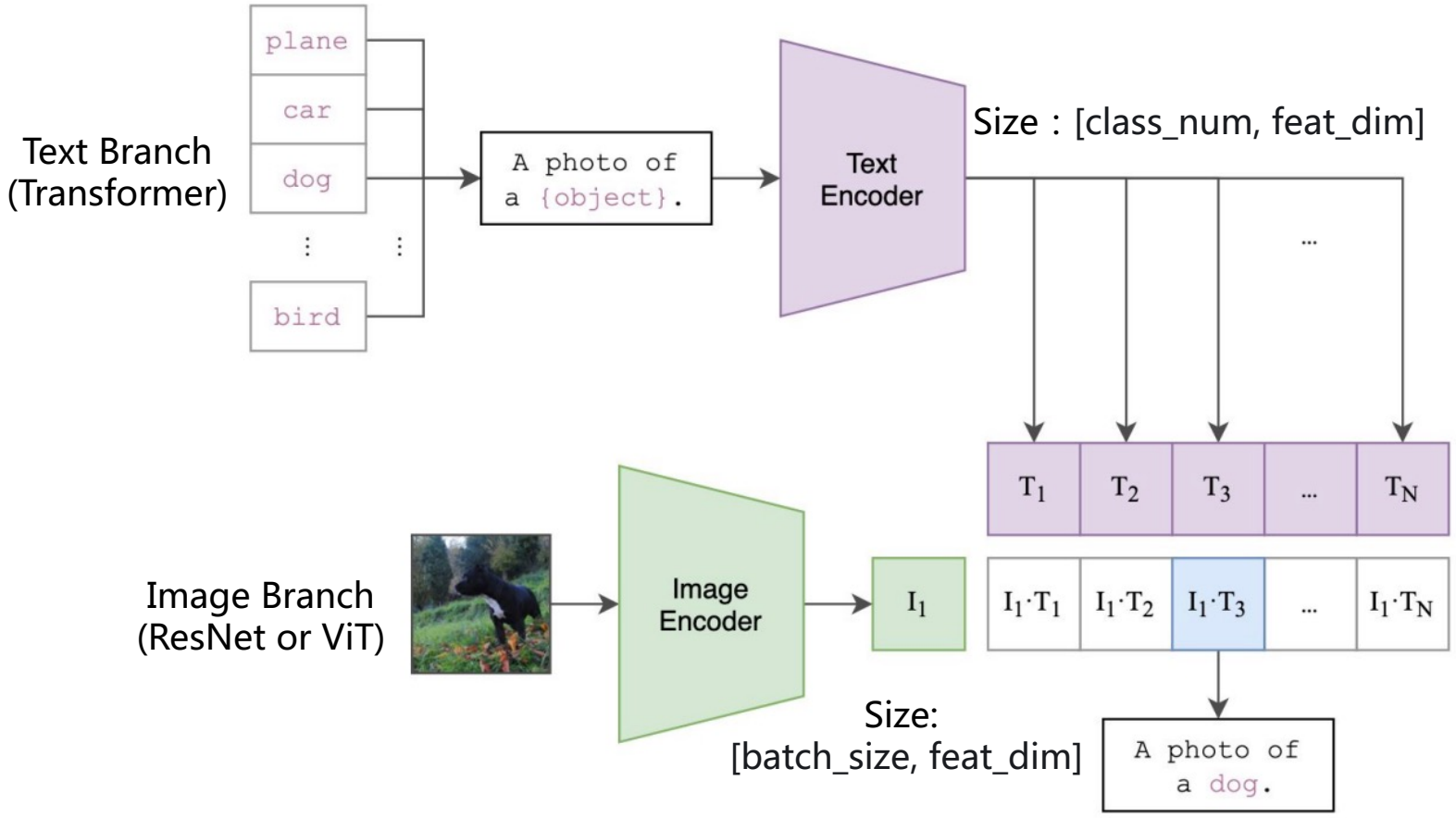
- Background
  - Vision-Language Models (VLMs)
  - Prompt Learning
  - Evaluation Metric
- Questions in Existing Methods
- [ICCV 25] Advancing Textual Prompt Learning with Anchored Attributes
  - Framework
  - Differentiable Attribute Search
  - Experiments
- Conclusion

Outlines

- Background
  - Vision-Language Models (VLMs)
  - Prompt Learning
  - Evaluation Metric
- Questions in Existing Methods
- [ICCV 25] Advancing Textual Prompt Learning with Anchored Attributes
  - Framework
  - Differentiable Attribute Search
  - Experiments
- Conclusion

What is Vision-Language Models (VLMs)?

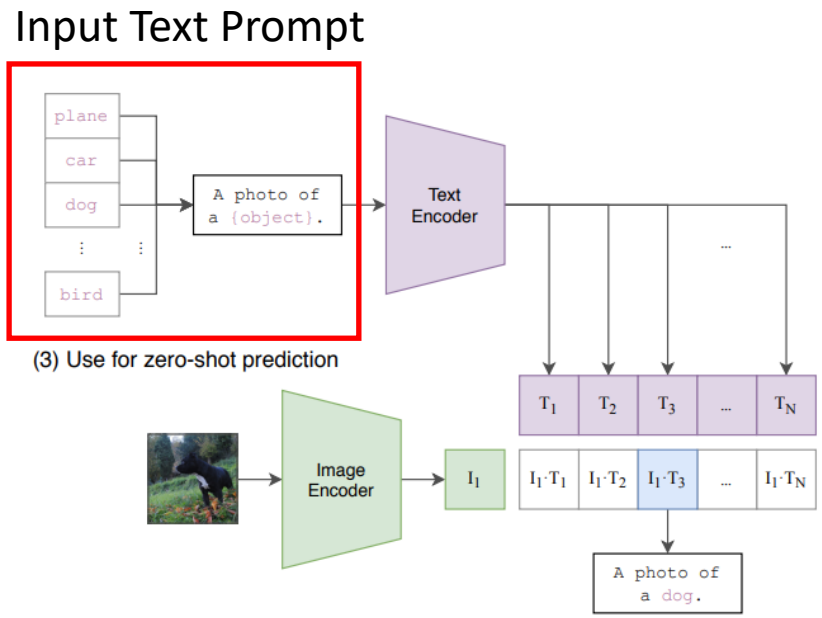
VLMs usually consists of two parts: text branch and image branch which process the information of the corresponding modality respectively.

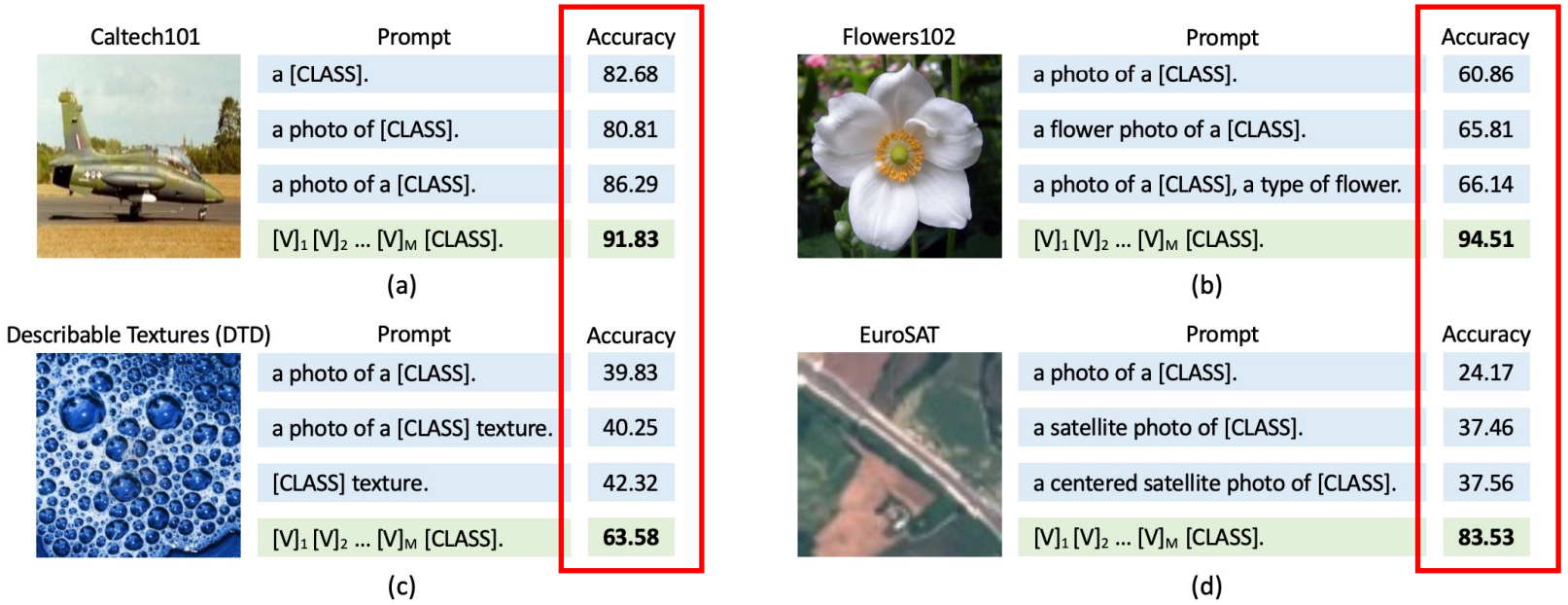


Radford et al. Learning transferable visual models from natural language supervision. ICML 2021

What is prompt learning?

We generally use the template ``a photo of a [CLASS]" as the input of the text branch.





Caltech101

Prompt	Accuracy
a [CLASS].	82.68
a photo of [CLASS].	80.81
a photo of a [CLASS].	86.29
$[V]_1\,[V]_2 \ldots [V]_M$ [CLASS].	**91.83**

(a)

Flowers102

Prompt	Accuracy
a photo of a [CLASS].	60.86
a flower photo of a [CLASS].	65.81
a photo of a [CLASS], a type of flower.	66.14
$[V]_1\,[V]_2 \ldots [V]_M$ [CLASS].	**94.51**

(b)

Describable Textures (DTD)

Prompt	Accuracy
a photo of a [CLASS].	39.83
a photo of a [CLASS] texture.	40.25
[CLASS] texture.	42.32
$[V]_1\,[V]_2 \ldots [V]_M$ [CLASS].	**63.58**

(c)

EuroSAT

Prompt	Accuracy
a photo of a [CLASS].	24.17
a satellite photo of [CLASS].	37.46
a centered satellite photo of [CLASS].	37.56
$[V]_1\,[V]_2 \ldots [V]_M$ [CLASS].	**83.53**

(d)

Different text prompts can lead to noticeable performance change.

Zhou et al. Learning to prompt for vision-language models. IJCV 2022

What is prompt learning?

However, for different datasets, this type of text template is too common and does not perform well.

Caltech101

Prompt	Accuracy
a [CLASS].	82.68
a photo of [CLASS].	80.81
a photo of a [CLASS].	86.29
$[V]_1$ $[V]_2$ ... $[V]_M$ [CLASS].	**91.83**

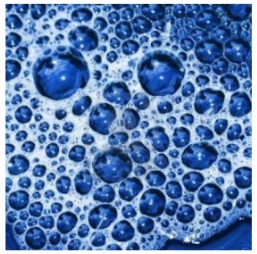
(a)

Flowers102

Prompt	Accuracy
a photo of a [CLASS].	60.86
a flower photo of a [CLASS].	65.81
a photo of a [CLASS], a type of flower.	66.14
$[V]_1$ $[V]_2$ ... $[V]_M$ [CLASS].	**94.51**

(b)

Describable Textures (DTD)



Prompt	Accuracy
a photo of a [CLASS].	39.83
a photo of a [CLASS] texture.	40.25
[CLASS] texture.	42.32
$[V]_1$ $[V]_2$ ... $[V]_M$ [CLASS].	**63.58**

(c)

EuroSAT

Prompt	Accuracy
a photo of a [CLASS].	24.17
a satellite photo of [CLASS].	37.46
a centered satellite photo of [CLASS].	37.56
$[V]_1$ $[V]_2$ ... $[V]_M$ [CLASS].	**83.53**

(d)

Zhou et al. Learning to prompt for vision-language models. IJCV 2022

What is prompt learning?

If we manually design text prompts for each dataset, it will be time-consuming and laborious.

Flowers102

Prompt	Accuracy
a photo of a [CLASS].	60.86
a flower photo of a [CLASS].	65.81
a photo of a [CLASS], a type of flower.	66.14
$[V]_1$ $[V]_2$ ... $[V]_M$ [CLASS].	**94.51**

(b)

EuroSAT

Prompt	Accuracy
a photo of a [CLASS].	24.17
a satellite photo of [CLASS].	37.46
a centered satellite photo of [CLASS].	37.56
$[V]_1$ $[V]_2$ ... $[V]_M$ [CLASS].	**83.53**

(d)

Problems:

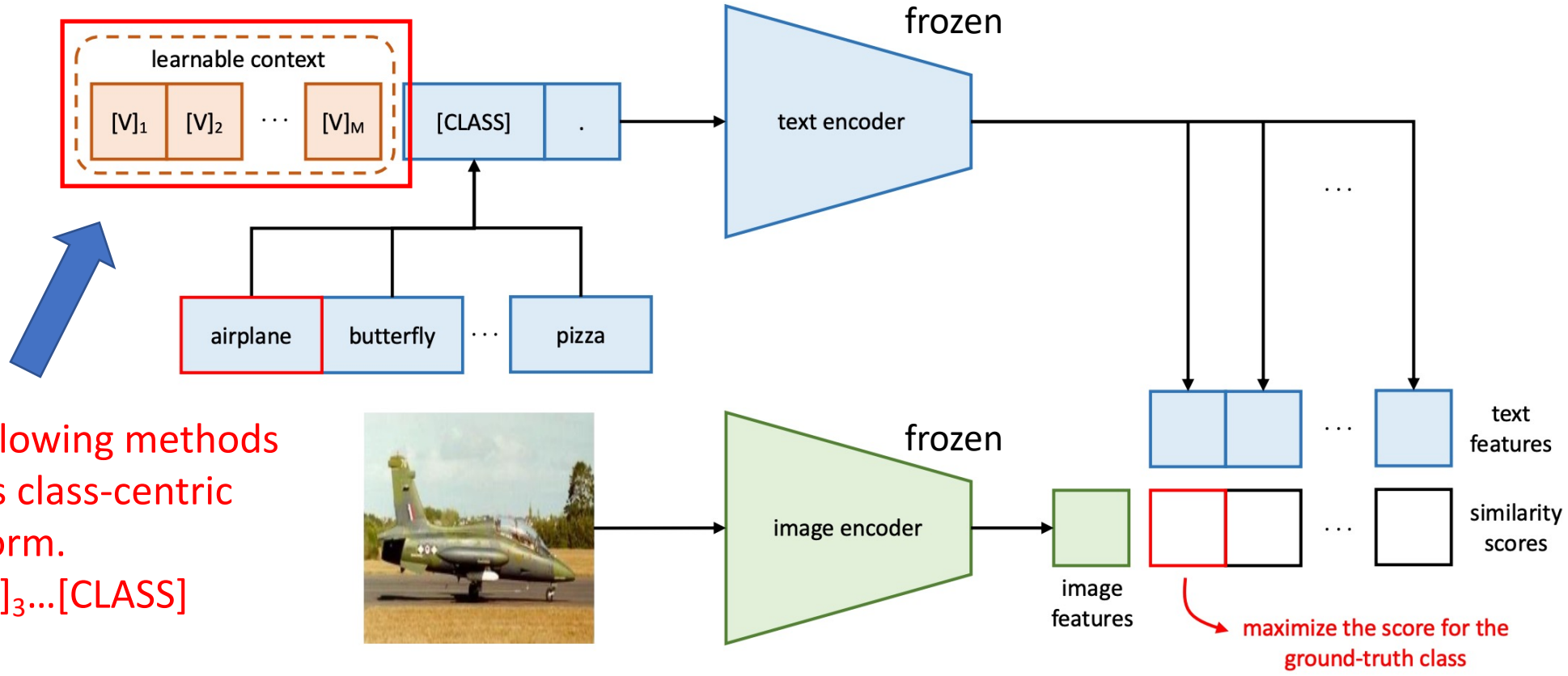
(1). Fixed template is not optimal for each dataset.

(2). Manually designed prompts are time-consuming and difficult to generalize.

Zhou et al. Learning to prompt for vision-language models. IJCV 2022

Can we let the model learn appropriate text prompts on its own?

The first prompt learning method for VLMs: <CoOp>



All the following methods adopt this class-centric prompt form.
$[V]_1[V]_2[V]_3$…[CLASS]

Zhou et al. Learning to prompt for vision-language models. IJCV 2022

Prompt Learning

Through training, soft tokens can learn general representations on downstream datasets and achieve better performance.

Caltech101

Prompt	Accuracy
a [CLASS].	82.68
a photo of [CLASS].	80.81
a photo of a [CLASS].	86.29
$[V]_1\,[V]_2\,\ldots\,[V]_M$ [CLASS].	**91.83**

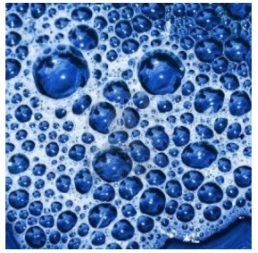
(a)

Flowers102

Prompt	Accuracy
a photo of a [CLASS].	60.86
a flower photo of a [CLASS].	65.81
a photo of a [CLASS], a type of flower.	66.14
$[V]_1\,[V]_2\,\ldots\,[V]_M$ [CLASS].	**94.51**

(b)

Describable Textures (DTD)



Prompt	Accuracy
a photo of a [CLASS].	39.83
a photo of a [CLASS] texture.	40.25
[CLASS] texture.	42.32
$[V]_1\,[V]_2\,\ldots\,[V]_M$ [CLASS].	**63.58**

(c)

EuroSAT

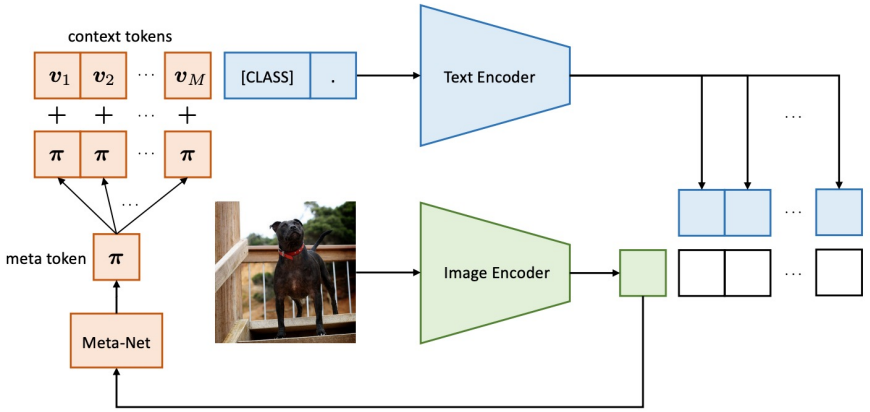
Prompt	Accuracy
a photo of a [CLASS].	24.17
a satellite photo of [CLASS].	37.46
a centered satellite photo of [CLASS].	37.56
$[V]_1\,[V]_2\,\ldots\,[V]_M$ [CLASS].	**83.53**

(d)

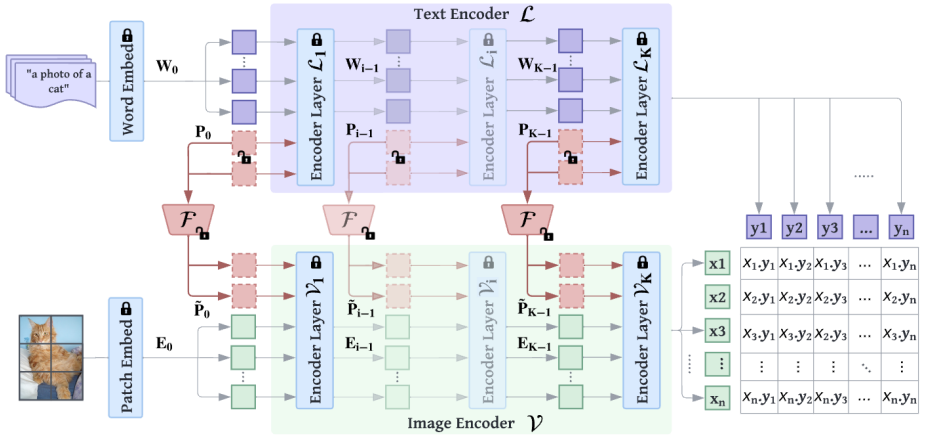
Zhou et al. Learning to prompt for vision-language models. IJCV 2022

Prompt Learning

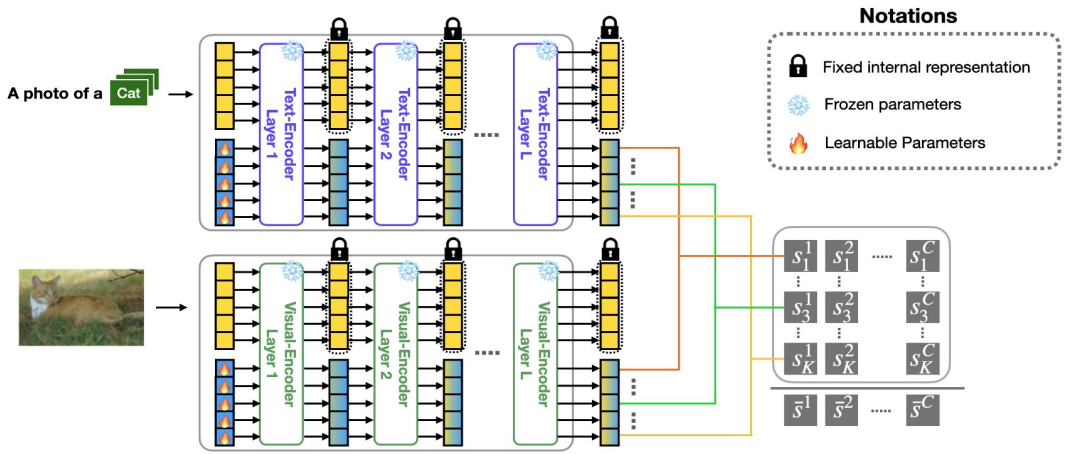
Nowadays, researchers have developed various methods to enhance the learning ability of soft prompts.



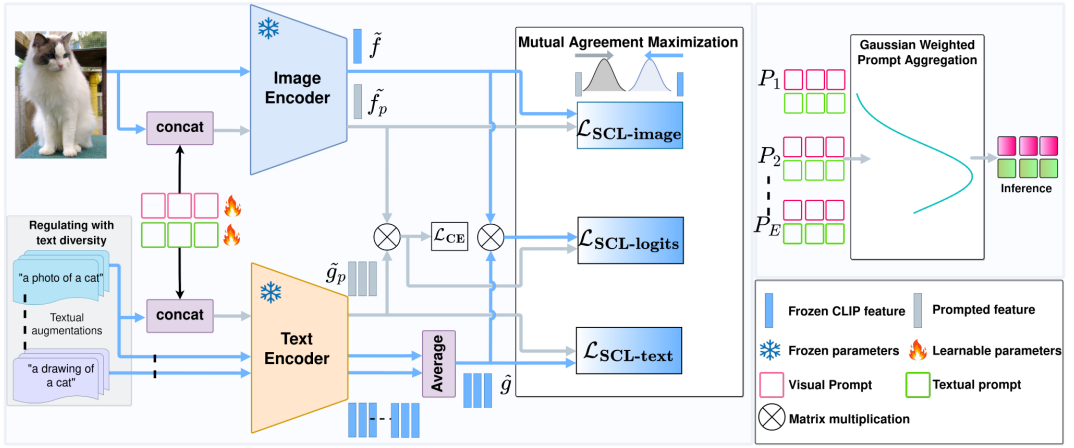
CoCoOp (CVPR 22)



MaPLe (CVPR 23)



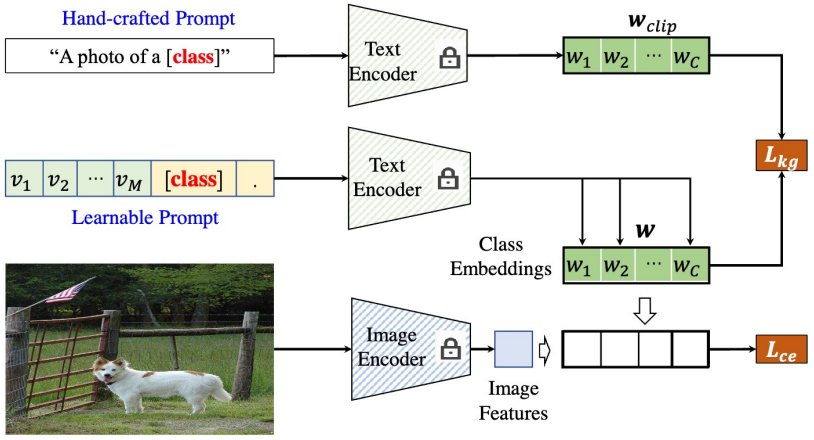
RPO (ICCV 23)



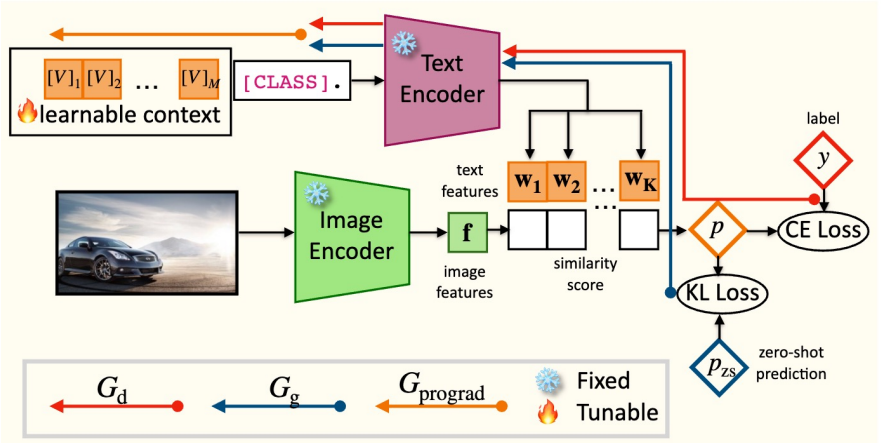
PromptSRC (ICCV 23)

Prompt Learning

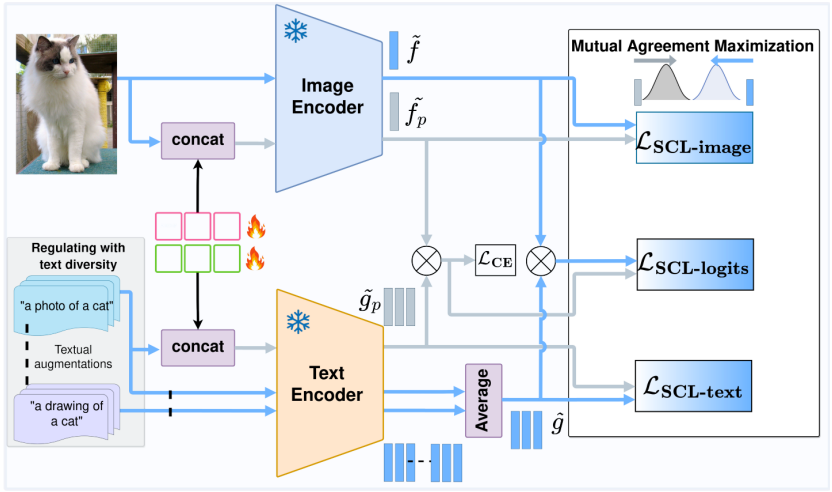
Among all the methods, the regularization-based forms are the most popular and effective.



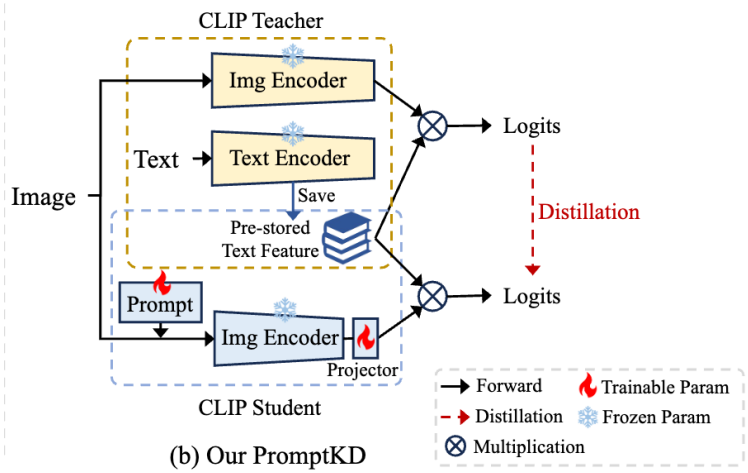
KgCoOp (CVPR 23)



ProGrad (ICCV 23)



PromptSRC (ICCV 23)



(b) Our PromptKD

PromptKD (CVPR 24)

Evaluation Metric

Base-to-Novel Generalization

Three metrics:
(1). Base Acc: Base class classification accuracy.
(2). Novel Acc: Novel class classification accuracy.
(3). HM: Harmonic Mean.

Evaluation Procedure:
Train the model only on the base class training set,
then evaluate it on the base and novel classes' test set.

For example:
For imagenet-1k, we usually divide the one thousand classes equally, the first 500 classes (base) are used for training, and the last 500 classes (novel) are used for testing.

Methods	Pub	Base	Novel	HM (main)
CLIP	ICML 21	69.34	74.22	71.70
CoOp	IJCV 22	82.69	63.22	71.66
CoCoOp	CVPR 22	80.47	71.69	75.83
ProDA	CVPR 22	81.56	72.30	76.65
KgCoOp	CVPR 23	80.73	73.60	77.00
RPO	ICCV 23	81.13	75.00	77.78
MaPLe	CVPR 23	82.28	75.14	78.55
DePT	CVPR 24	83.62	75.04	79.10
TCP	CVPR 24	84.13	75.36	79.51
MMA	CVPR 24	83.20	76.80	79.87
PromptSRC	ICCV 23	84.26	76.10	79.97
HPT	AAAI 24	84.32	76.86	80.23
CoPrompt	ICLR 24	84.00	77.23	80.48
CasPL	ECCV 24	86.11	79.54	82.69
PromptKD	CVPR 24	86.96	80.73	83.73

Regularization-based

Outlines

- Background
  - Vision-Language Models (VLMs)
  - Prompt Learning
  - Evaluation Metric
- Questions in Existing Methods
- [ICCV 25] Advancing Textual Prompt Learning with Anchored Attributes
  - Framework
  - Differentiable Attribute Search
  - Experiments
- Conclusion

Questions in existing methods

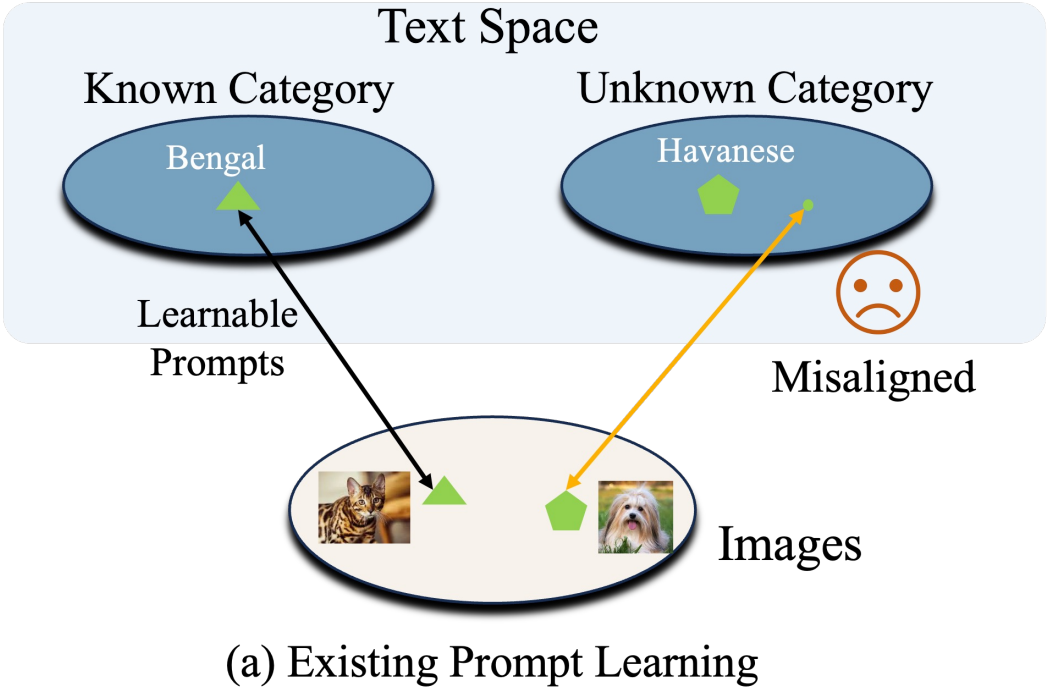
Why we need regularization in prompt learning?

Existing class-centric prompt form ($[V]_1[V]_2...[V]_M[CLASS]$) has shortcomings:

1. Soft prompts can only learn representations related to known categories. It cannot establish accurate association with unknown categories.

2. The prompt form is simple and it is easy to overfit to the training data during training. This further weakens the generalization ability of the model.

Lead to

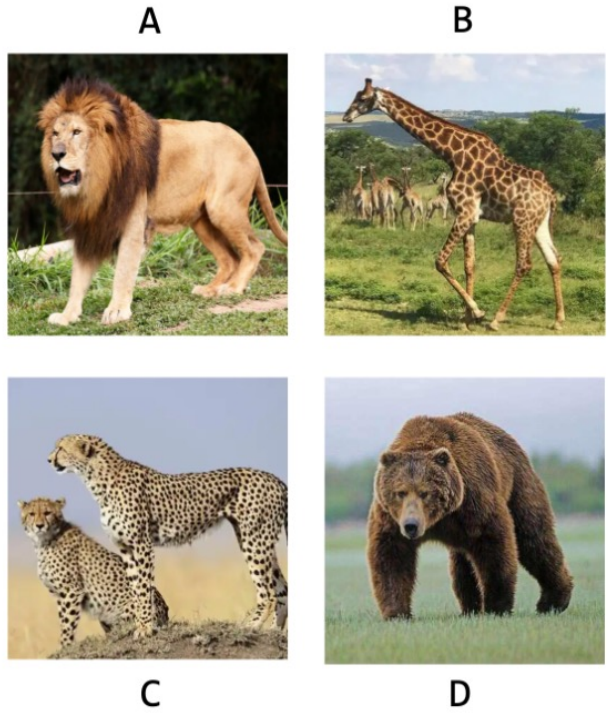


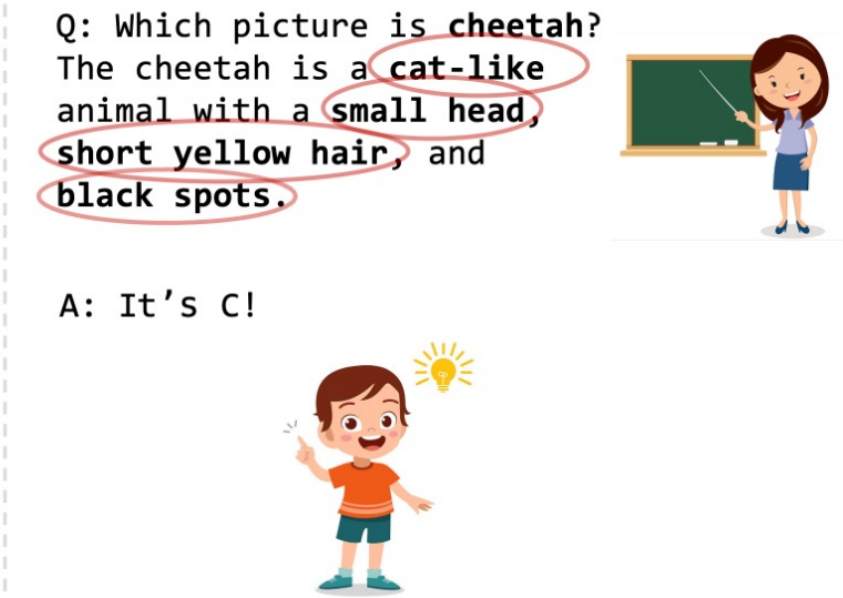
(a) Existing Prompt Learning

We need to build accurate associations between images and unknown categories.

Inspirations from real life

In our real life, people usually combine relevant attribute information (e.g. shape, color, texture) to facilitate the recognition of unknown categories.





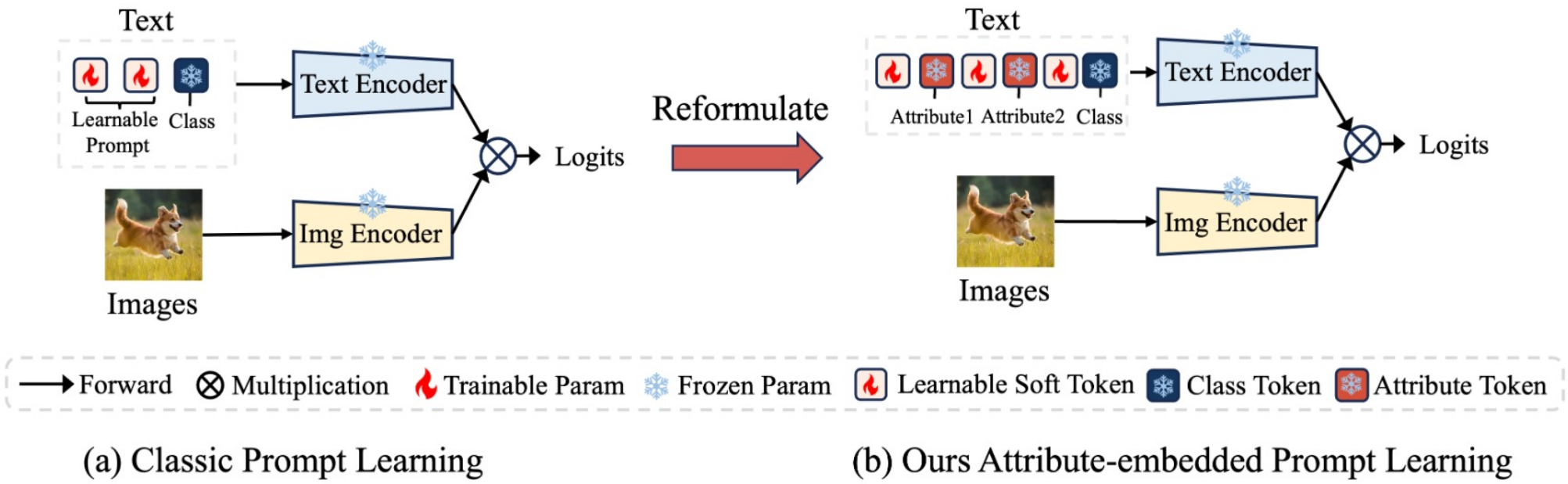
Attributes can provide additional information to facilitate recognition.

Outlines

- Background
  - Vision-Language Models (VLMs)
  - Prompt Learning
  - Evaluation Metric
- Questions in Existing Methods
- [ICCV 25] Advancing Textual Prompt Learning with Anchored Attributes
  - Framework
  - Differentiable Attribute Search
  - Experiments
- Conclusion

Can we let soft prompts learn attribute-related representations?

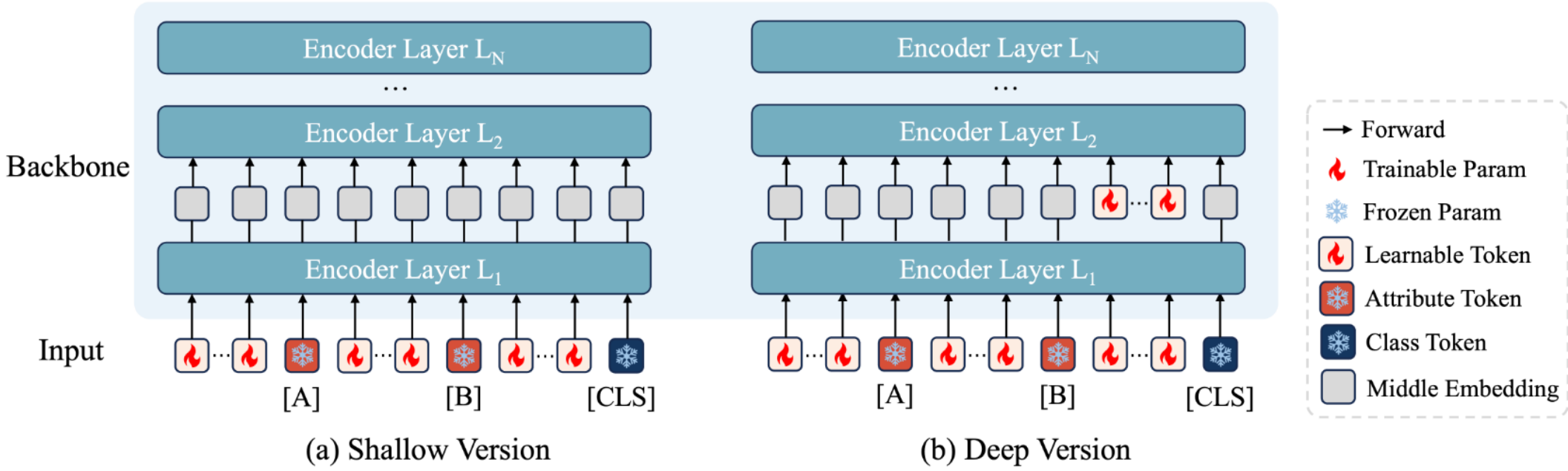
Of Course! Here we present **ATPrompt**, an attribute-embedded prompt learning method for VLMs.



In this work, we propose to embeds multiple fixed attribute tokens into the set of soft tokens, transforming the original form (a) into an attribute-class mixed form (b) for prompt learning.

ATPrompt

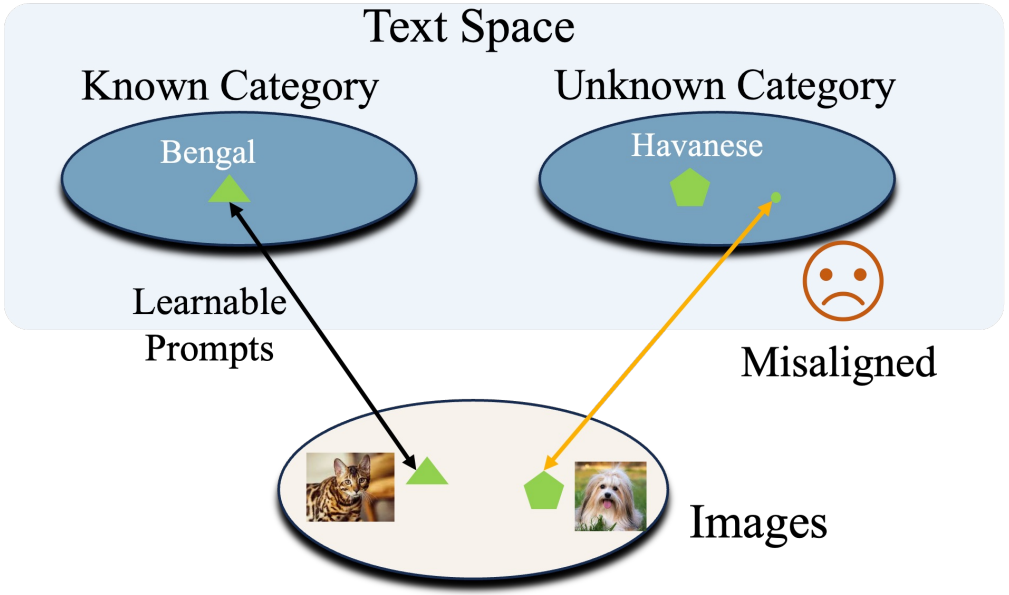
Furthermore, we introduce two versions: shallow and deep version.



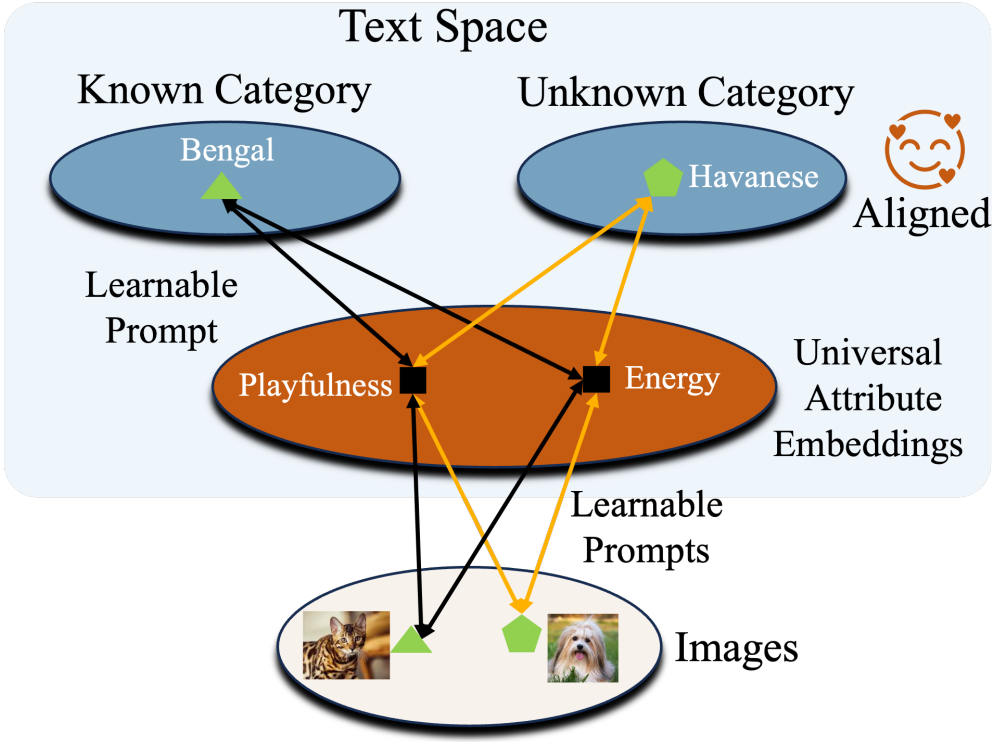
The deep version uses the same input but discards the class-related soft prompt tokens after calculating the self-attention and introduces them again before the next layer.

Why ATPrompt can work?

Principle:



(a) Existing Prompt Learning



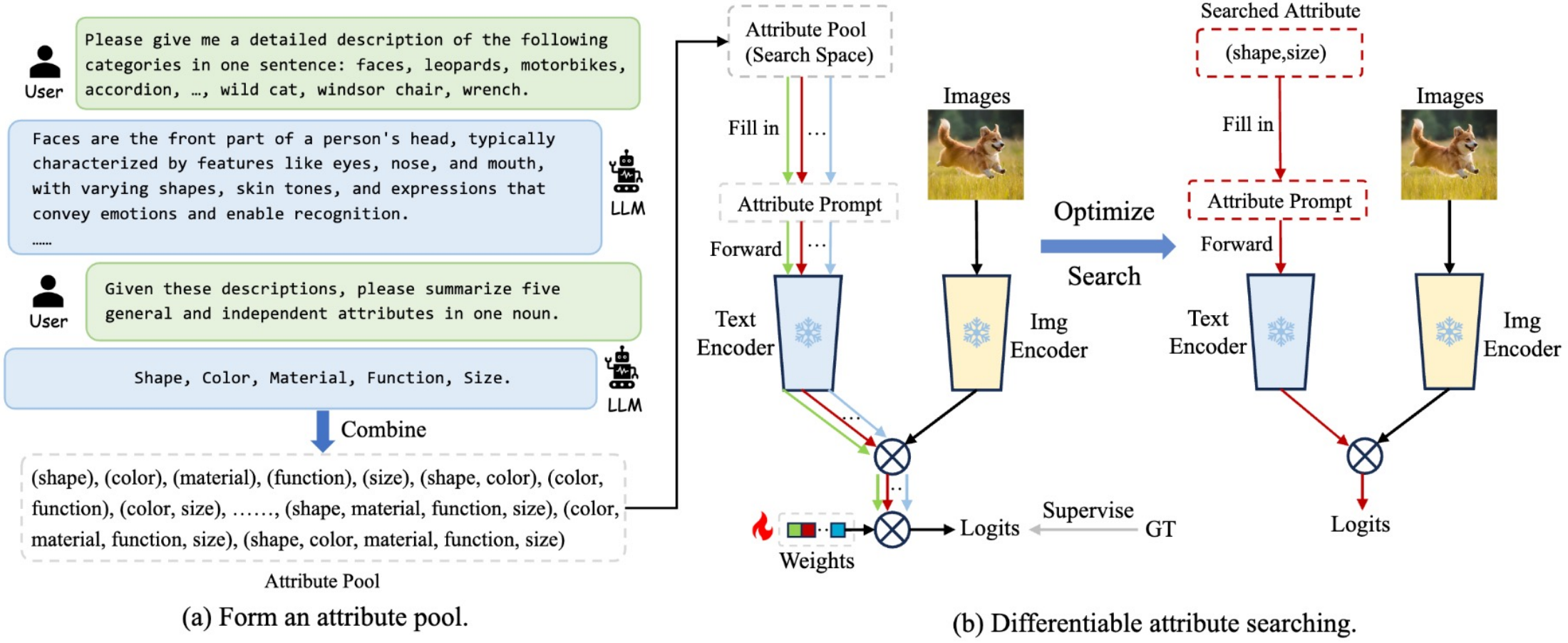
(b) Ours Attribute-embedded Prompt Learning

(a) Current prompt learning methods align images with predefined categories but fail to establish accurate associations with unknown categories.

(b) ATPrompt leverages universal attributes as an intermediary to create more accurate alignments between images and unknown categories

How to determine attributes?

We introduce a differentiable attribute search method that learns to select appropriate attributes.

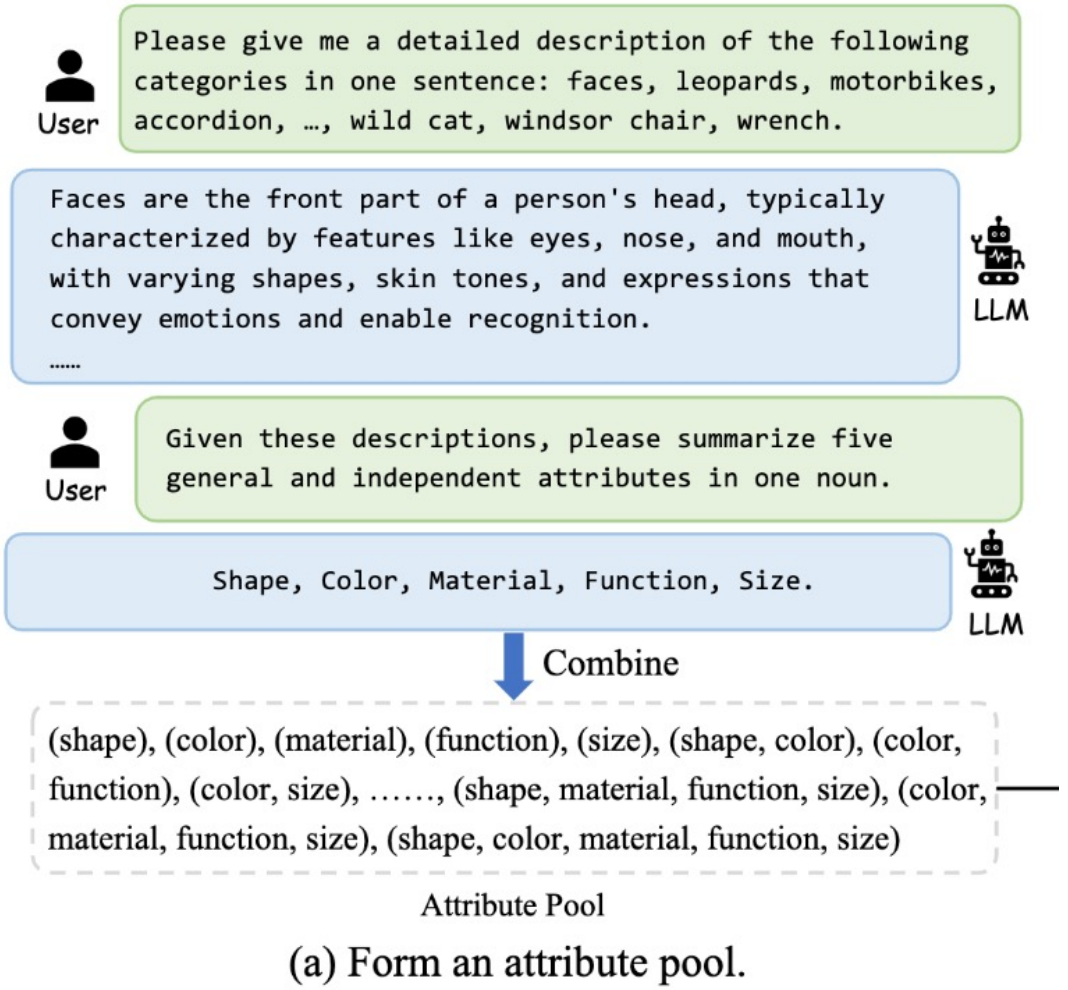


(a) Form an attribute pool.

(b) Differentiable attribute searching.

Overview of differentiable attribute search framework

Differentiable Attribute Search



(a) Form an attribute pool.

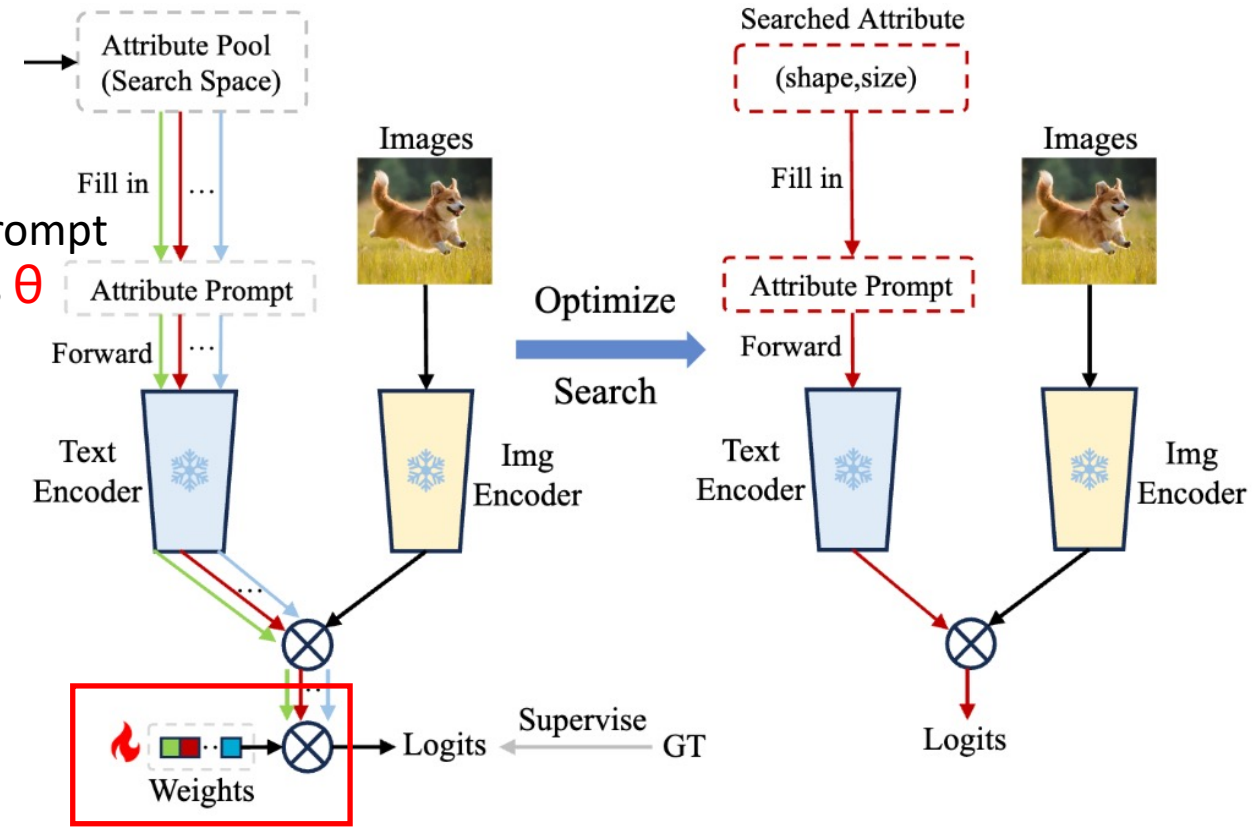
Inspired by Chain-of-Thought, we divide the entire process into multiple steps to enhance the reasoning ability of LLMs.

First, we employ the LLM to generate descriptive sentences for each known category, thereby enriching category-related information.

Subsequently, based on these sentences, we prompt the LLM to summarize multiple independent attribute basesacross these categories.

Finally, an attribute pool is formed by combining different bases.

Differentiable Attribute Search



Attribute weight α (b) Differentiable attribute searching.

We apply the alternating algorithm to solve this problem, fixing one set of variables and solving for the other set. Formally, we can alternate between solving these two subproblems:

$$\hat{\alpha} = \arg\min_{\alpha} L_{val}(f(x, v; \alpha, \hat{\theta}), c),$$

$$\hat{\theta} = \arg\min_{\theta} L_{train}(f(x, v; \hat{\alpha}, \theta), c).$$

where denote attribute weight, is the L_{train} and L_{val} uses cross-entropy as the loss function.

Differentiable Attribute Search

Output results of our attribute search method.

Attributes	shape, color, material, function, size
Combinations & Weights	(shape), score: 0.298 (color), score: 0.004 (material), score: 0.002 (function), score: 0.002 (size), score: 0.003 (shape, color), score: 0.003 (shape, material), score: 0.006 (shape, function), score: 0.000 **(shape, size), score: 0.565** (color, material), score: 0.000 (color, function), score: 0.001 (color, size), score: 0.005 (material, function), score: 0.000 (material, size), score: 0.002 (function, size), score: 0.002 (shape, color, material), score: 0.002 (shape, color, function), score: 0.002 (shape, color, size), score: 0.000 (shape, material, function), score: 0.001 (shape, material, size), score: 0.085 (shape, function, size), score: 0.001 (color, material, function), score: 0.001 (color, material, size), score: 0.000 (color, function, size), score: 0.002 (material, function, size), score: 0.001 (shape, color, material, function), score: 0.001 (shape, color, material, size), score: 0.001 (shape, color, function, size), score: 0.001 (shape, material, function, size), score: 0.005 (color, material, function, size), score: 0.001 (shape, color, material, function, size), score: 0.001

Highest weights (confidence)

Experiments

Base-to-Novel Generalization

Our ATPrompt achieves consistent improvement over 11 datasets.

Method	Average			ImageNet			Caltech101			OxfordPets		
	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM
CoOp (IJCV 22)	82.69	63.22	71.66	76.47	67.88	71.92	98.00	89.81	93.73	93.67	95.29	94.47
CoCoOp (CVPR 22)	80.47	71.69	75.83	75.98	70.43	73.10	97.96	93.81	95.84	95.20	97.69	96.43
MaPLe (CVPR 23)	82.28	75.14	78.55	76.66	70.54	73.47	97.74	94.36	96.02	95.43	97.76	96.58
PromptSRC (ICCV 23)	84.26	76.10	79.97	77.60	70.73	74.01	98.10	94.03	96.02	95.33	97.30	96.30
ArGue (CVPR 24)	83.69	78.07	80.78	76.92	72.06	74.41	98.43	95.20	96.79	95.36	97.95	96.64
DePT (CVPR 24)	83.66	71.82	77.29	77.13	70.10	73.45	98.33	94.33	96.29	94.70	97.63	96.14
CoPrompt (ICLR 24)	84.00	77.23	80.48	77.67	71.27	74.33	98.27	94.90	96.55	95.67	98.10	96.87
PromptKD (CVPR 24)	86.96	80.73	83.73	80.83	74.66	77.62	98.91	96.65	97.77	96.30	98.01	97.15
CoOp + ATPrompt	82.68	68.04	**74.65** (+2.99)	76.27	70.60	**73.33**	97.95	93.63	**95.74**	94.77	96.59	**95.67**
CoCoOp + ATPrompt	81.69	74.54	**77.95** (+2.12)	76.43	70.50	**73.35**	97.96	95.27	**96.60**	95.46	97.89	**96.66**
MaPLe + ATPrompt	82.98	75.76	**79.21** (+0.66)	76.94	70.72	**73.70**	98.32	95.09	**96.68**	95.62	97.63	**96.61**
DePT + ATPrompt	83.80	73.75	**78.45** (+1.16)	77.32	70.65	**73.83**	98.48	94.60	**96.50**	94.65	97.99	**96.29**
PromptKD + ATPrompt	87.05	81.82	**84.35** (+0.62)	80.90	74.83	**77.75**	98.90	96.52	97.70	96.92	98.27	**97.59**

Method	StanfordCars			Flowers102			Food101			FGVCAircraft		
	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM
CoOp (IJCV 22)	78.12	60.40	68.13	97.60	59.67	74.06	88.33	82.26	85.19	40.44	22.30	28.75
CoCoOp (CVPR 22)	70.49	73.59	72.01	94.87	71.75	81.71	90.70	91.29	90.99	33.41	23.71	27.74
MaPLe (CVPR 23)	72.94	74.00	73.47	95.92	72.46	82.56	90.71	92.05	91.38	37.44	35.61	36.50
PromptSRC (ICCV 23)	78.27	74.97	76.58	98.07	76.50	85.95	90.67	91.53	91.10	42.73	37.87	40.15
ArGue (CVPR 24)	75.64	73.38	74.49	98.34	75.41	85.36	92.33	91.96	92.14	40.46	38.03	39.21
DePT (CVPR 24)	79.67	72.40	75.86	98.20	72.00	83.08	90.43	91.33	90.88	42.53	22.53	29.46
CoPrompt (ICLR 24)	76.97	74.40	75.66	97.27	76.60	85.71	90.73	92.07	91.40	40.20	39.33	39.76
PromptKD (CVPR 24)	82.80	83.37	83.13	99.42	82.62	90.24	92.43	93.68	93.05	49.12	41.81	45.17
CoOp + ATPrompt	77.43	66.55	**71.58**	97.44	67.52	**79.77**	88.74	87.44	**88.09**	40.38	27.22	**32.52**
CoCoOp + ATPrompt	74.50	73.47	**73.98**	96.52	73.59	**83.51**	90.59	91.74	**91.16**	37.30	33.15	**35.10**
MaPLe + ATPrompt	75.39	73.84	**74.61**	97.82	75.07	**84.95**	90.65	92.00	91.32	37.61	36.15	**36.87**
DePT + ATPrompt	79.29	73.47	**76.27**	98.20	73.69	**84.20**	90.42	91.69	**91.05**	43.19	33.23	**37.56**
PromptKD + ATPrompt	82.51	84.03	**83.26**	99.15	82.03	89.78	92.48	93.86	**93.22**	49.63	42.35	**45.70**

Experiments

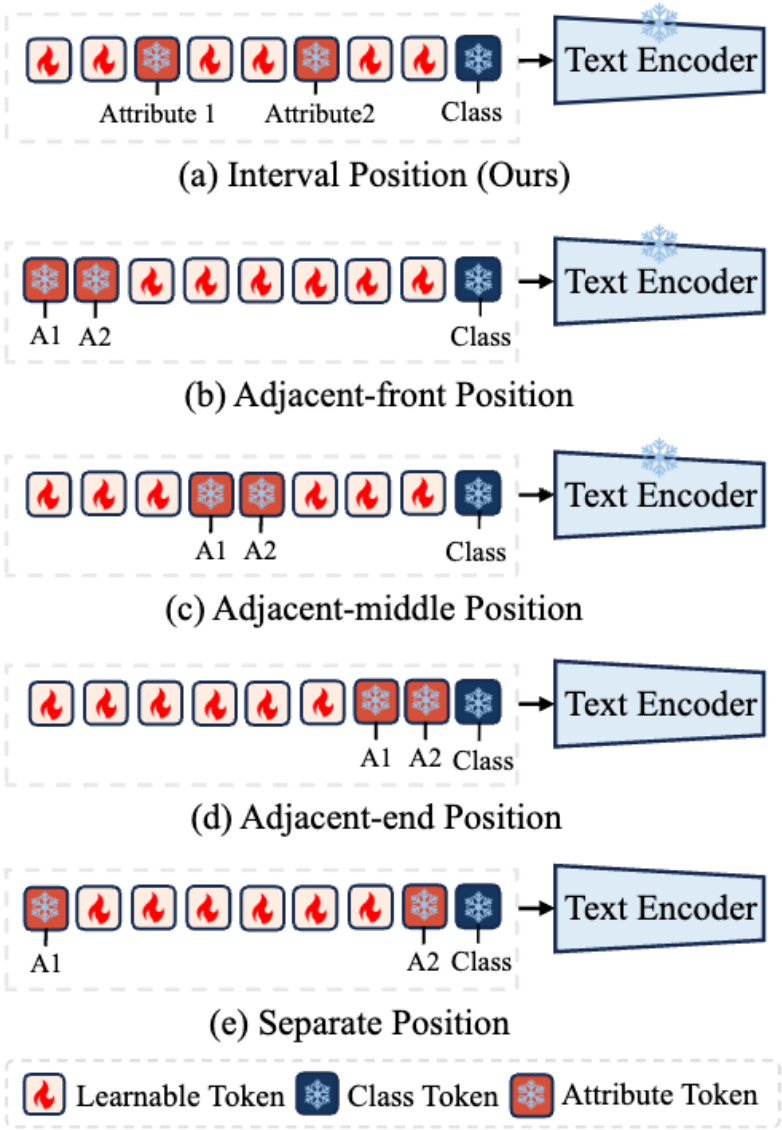
Cross-dataset Experiments

Method	Source	Target Dataset										
	Image Net	Caltech 101	Oxford Pets	Stanford Cars	Flowers 102	Food 101	FGVC Aircraft	SUN 397	DTD	Euro SAT	UCF 101	Avg.	Δ
CoOp	71.51	93.70	89.14	64.51	68.71	85.30	18.47	64.15	41.92	46.39	66.55	63.88	–
+ATPrompt	71.67	93.96	90.65	65.01	70.40	85.86	20.97	65.77	43.44	46.59	69.92	65.26	(+1.38)
CoCoOp	71.02	94.43	90.14	65.32	71.88	86.06	22.94	67.36	45.73	45.37	68.21	65.74	
+ATPrompt	71.27	93.79	90.62	65.90	71.17	86.03	23.22	66.63	44.44	48.70	70.71	66.59	(+0.85)
MaPLe	70.72	93.53	90.49	65.57	72.23	86.20	24.74	67.01	46.49	48.06	68.69	66.30	–
+ATPrompt	70.69	94.04	91.03	66.06	71.99	86.33	24.42	67.05	45.21	48.63	69.15	66.75	(+0.45)

Domain Generalization

Version	Method	Source	Target Dataset					
		ImageNet	-V2	-S	-A	-R	Avg.	Δ
Shallow	CoOp	71.51	64.20	47.99	49.71	75.21	59.28	
	+ATPrompt	71.67	64.43	49.13	50.91	76.24	60.18	(+0.90)
Shallow	CoCoOp	71.02	64.07	48.75	50.63	76.18	59.91	
	+ATPrompt	71.27	64.66	49.15	51.44	76.33	60.40	(+0.49)
Deep	MaPLe	70.72	64.07	49.15	50.90	76.98	60.27	
	+ATPrompt	70.69	64.40	49.10	51.77	77.11	60.60	(+0.33)

Ablation Studies

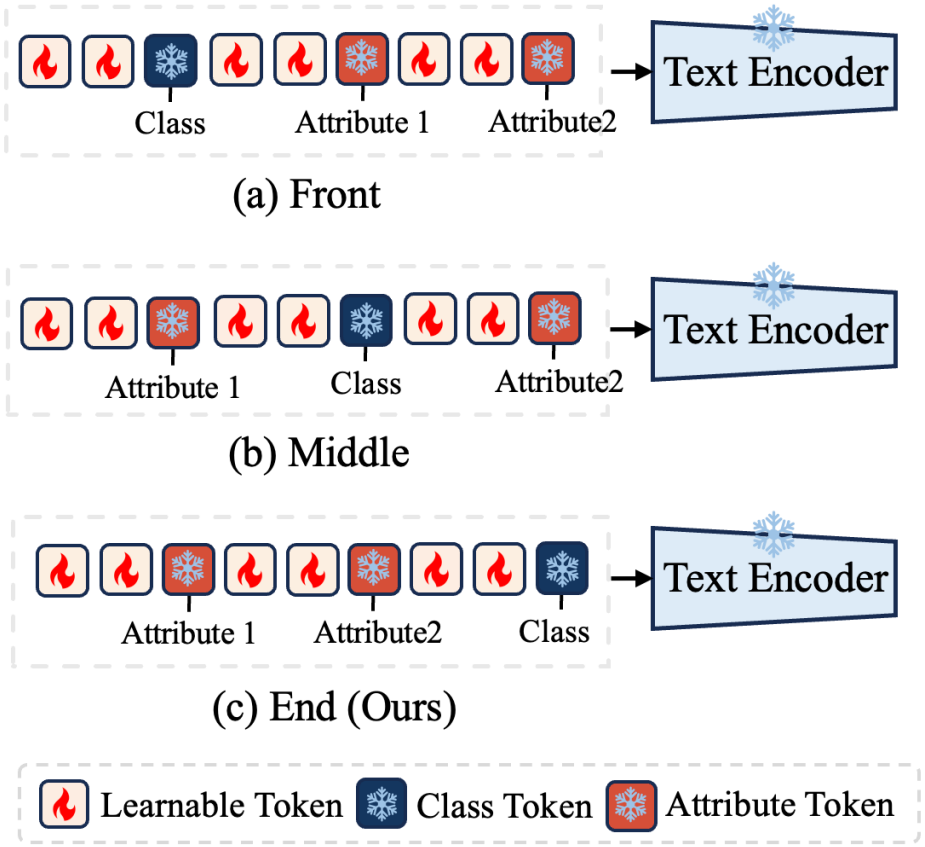


Attribute Position.

Version	Base	Novel	HM
Baseline (CoOp)	76.47	67.88	71.92
(a) Interval (Ours)	76.27	70.60	73.33
(b) Adjacent-front	76.39	70.22	73.18
(c) Adjacent-middle	76.46	70.11	73.15
(d) Adjacent-end	76.34	70.31	73.20
(e) Separate	76.48	70.08	73.14

The results indicate that the interval version achieves the best performance among all variations.

Ablation Studies



Class Token Position.

Position	Base	Novel	HM
Front	76.12	70.50	73.20
Middle	76.13	70.29	73.09
End	**76.27**	**70.60**	**73.33**

The results demonstrate that optimal performance is achieved when the class token is placed at the end, which aligns with the conclusions of CoOp.

Ablation Studies

Attribute Order

Attributes	Base	Novel	HM
(shape, color)	76.32	70.39	73.24
(color, shape)	76.27	70.60	**73.33**
(size, habitat)	76.44	70.23	73.20
(habitat, size)	76.46	70.16	73.14
(material, function)	76.40	70.13	73.13
(function, material)	76.28	70.00	73.01
(growth, season)	76.46	70.18	73.19
(season, growth)	76.40	70.21	73.17
(color, size, shape)	76.27	69.95	72.97
(shape, size, color)	76.32	70.19	73.13
(habitat, size, shape)	76.50	70.21	73.22
(habitat, shape, size)	76.46	70.08	73.13
Searched Attributes (color, shape)	76.27	70.60	**73.33**

From this table, we observe that despite variations in order, similar results are consistently produced, and the performance fluctuations across different orders remain within a reasonable range.

Comparison to Other Attributes

Type	Attributes	Base	Novel	HM
Common	(shape, size)	88.48	86.87	87.67
	(color, texture)	88.50	87.17	87.83
Irrelevant	(plane, engines)	88.64	86.31	87.46
	(football, sport)	88.60	86.18	87.37
	(leather, silk)	88.56	86.82	87.68
Searched	(flavor, preparation)	88.74	**87.44**	**88.09**

This suggests that incorrect attribute tokens cause the soft tokens to develop biased representations, thereby diminishing their zero-shot generalization ability.

Visualization

Attribute Bases and Searched Results

Attribute Pool

Dataset	Attribute Bases	Searched Attributes
ImageNet-1K	color, size, shape, habitat, behavior	(color, shape)
Caltech-101	shape, color, material, function, size	(shape,size)
Oxford Pets	loyalty, affection, playfulness, energy, intelligence	(playfulness, energy)
Stanford Cars	design, engine, performance, luxury, color	(luxury)
Flowers-102	color, flower, habitat, growth, season	(color, habitat, growth)
Food-101	flavor, texture, origin, ingredients, preparation	(flavor, preparation)
FGVC Aircraft	design, capacity, range, engines, liveries	(design, range)
SUN-397	architecture, environment, structure, design, function	(function)
DTD	pattern, texture, color, design, structure	(pattern, color, design)
EuroSAT	habitat, foliage, infrastructure, terrain, watercourse	(habitat)
UCF-101	precision, coordination, technique, strength, control	(precision)

Conclusion

(1) We introduce an attribute-templated prompt learning method for VLMs that utilizes universal attributes to guide the learning of soft prompts.

(2) We introduce a differentiable attribute search method that learns to determine the appropriate attribute content and quantity.

(3) Both shallow and deep versions of ATPrompt are introduced to achieve compatibility with existing methods.

(4) ATPrompt can be seamlessly integrated into existing textual-based methods and brings general improvement at a negligible computational cost.

Thanks!

Paper: https://arxiv.org/abs/2412.09442

Code: https://github.com/zhengli97/ATPrompt

中文解读：https://zhuanlan.zhihu.com/p/11787739769