南京航空航天大學
Nanjing University of Aeronautics and Astronautics

模式分析与机器智能
工业和信息化部重点实验室
MIIT Key Laboratory of
Pattern Analysis & Machine Intelligence

Texts as Images in Prompt Tuning for Multi-Label Image Recognition

Zixian Guo[1*] Bowen Dong[2] Zhilong Ji[1] Jinfeng Bai[1] Yiwen Guo[4] Wangmeng Zuo[2,3✉]

[1]Tomorrow Advancing Life [2]Harbin Institute of Technology [3]Pazhou Lab, Guangzhou [4]Independent Researcher

zixian_guo@foxmail.com cndongsky@gmail.com zhilongji@hotmail.com

jfbai.bit@gmail.com guoyiwen89@gmail.com wmzuo@hit.edu.cn

CVPR 2023

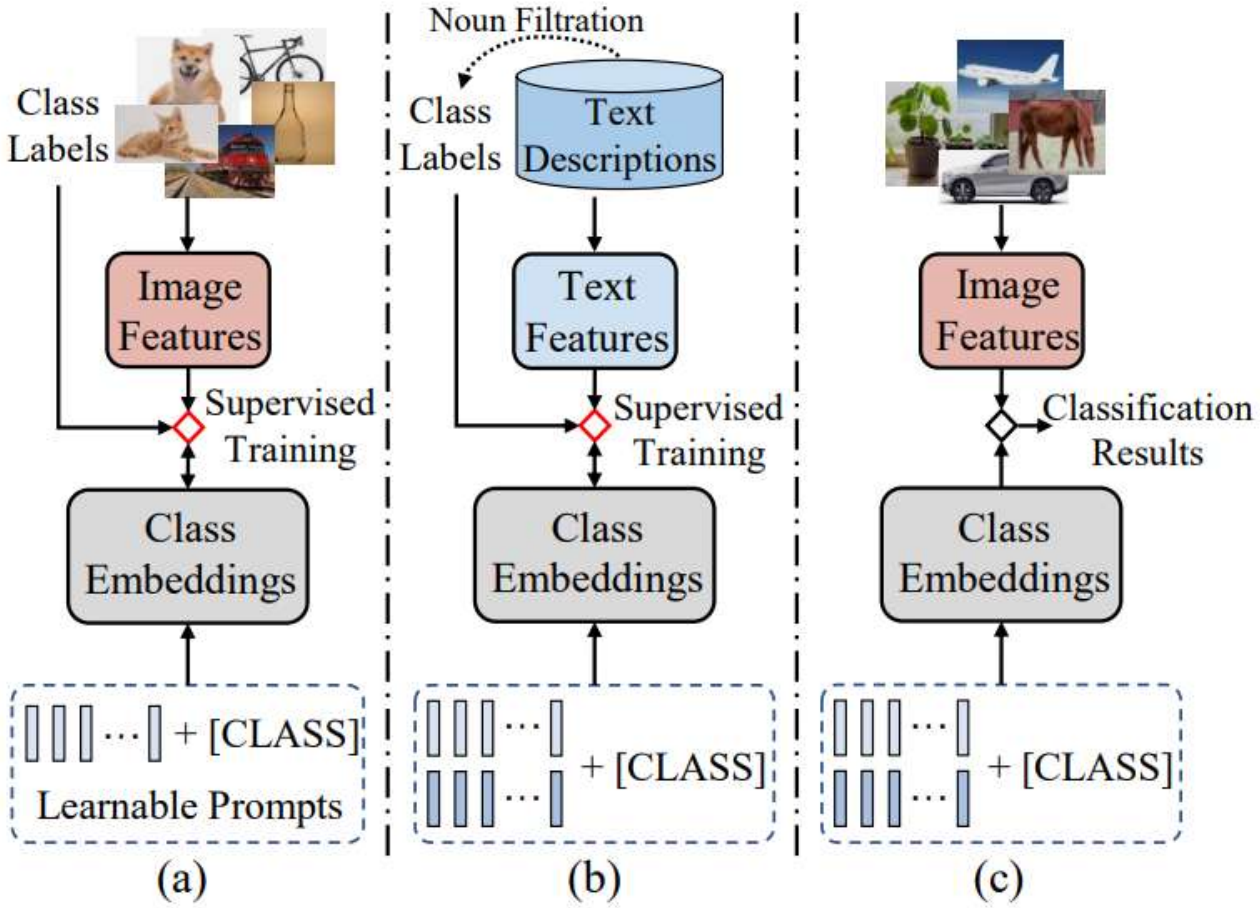
1 Background & Motivation

- Multi-label image recognition is important yet difficult in data-scarce settings.
- Existing prompt tuning methods (e.g., CoOp) rely heavily on labeled images.
- Text data, however, is easier to collect and naturally labelable.

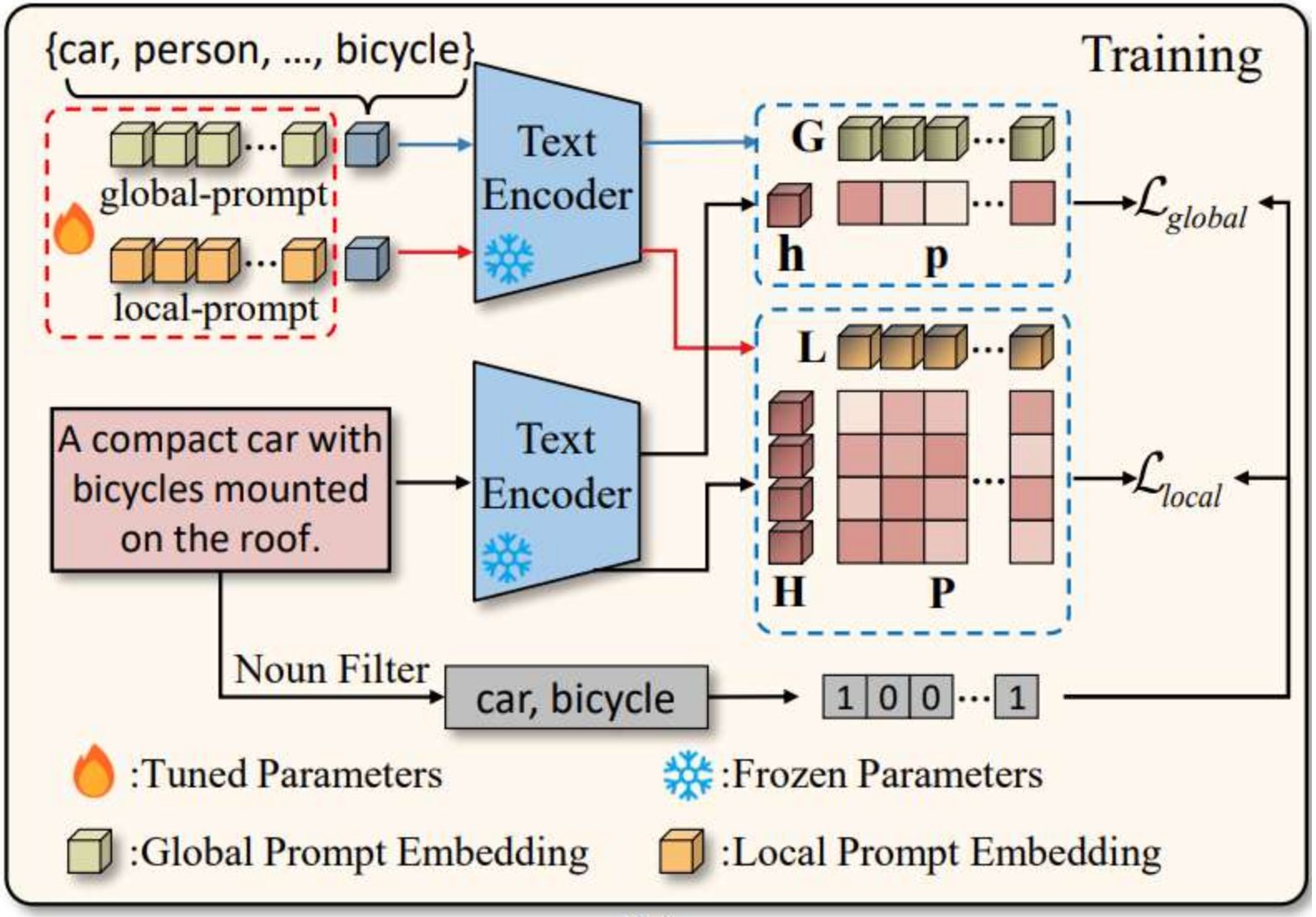
- Idea: Can we learn prompts using only texts instead of images?

2 Comparison of Prompting Strategies

- (a) Traditional prompting: learns prompts using labeled images.
- (b) **TaI prompting**: uses free-form text descriptions & noun filtering to guide training.
- (c) Inference: learned prompts are used to classify new images.



3 Method Overview – Tal Prompting Pipeline



(a)

$$t_i^G = [v_1, v_2, v_3, ..., v_M, s_i]$$

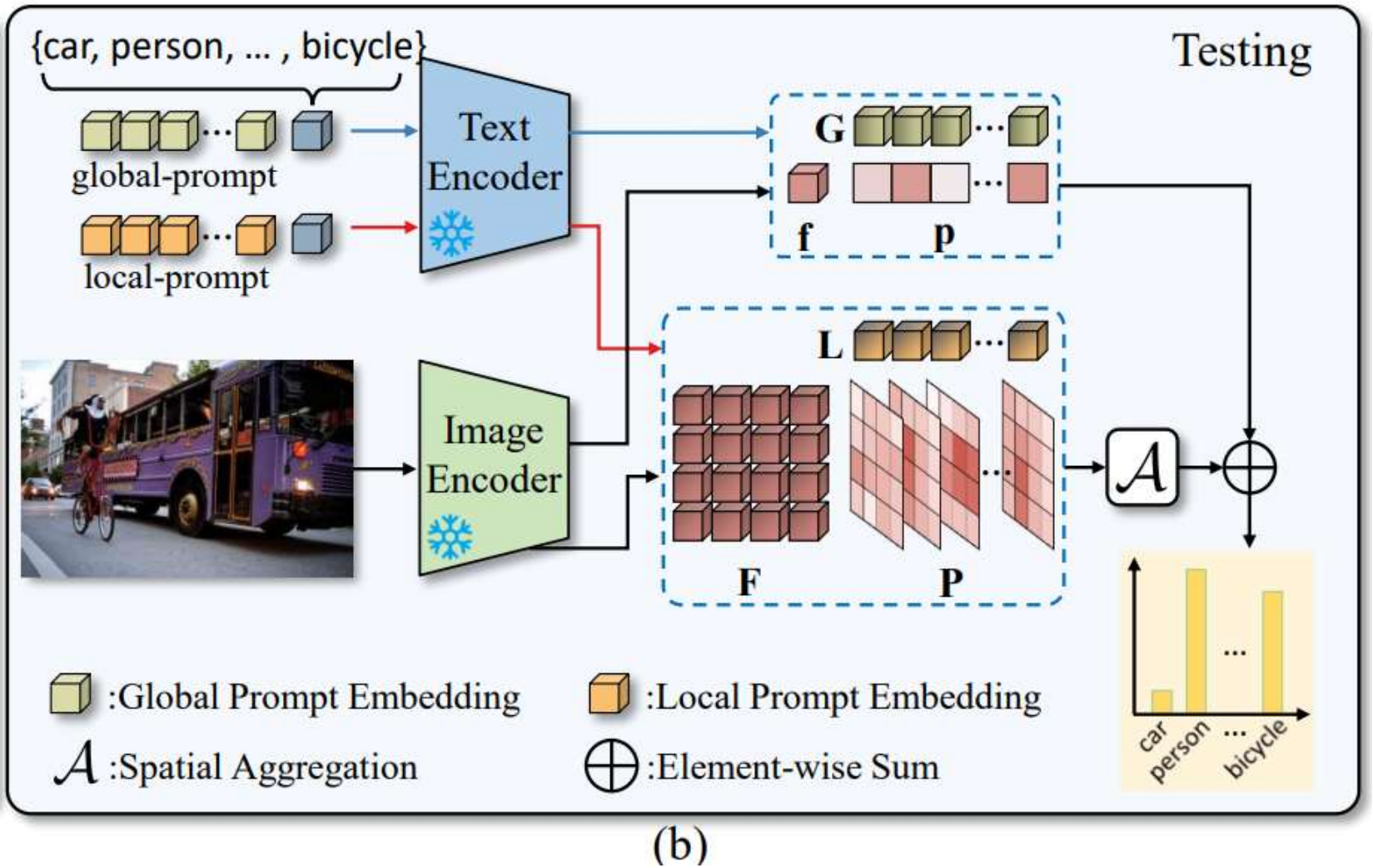
$$t_i^L = [v'_1, v'_2, v'_3, ..., v'_M, s_i]$$

$$\mathcal{L}_{global} = \sum_{i\in\{c^+\}} \sum_{j\in\{c^-\}} \max(0, m - p_i + p_j)$$

$$\mathcal{L}_{local} = \sum_{i\in\{c^+\}} \sum_{j\in\{c^-\}} \max(0, m - p'_i + p'_j)$$

Training Phase:
Two frozen CLIP text encoders process:
- - Text descriptions → features (h, H)
- - Learnable prompts → class emb (G, L)
Supervision from pseudo-labels generated via noun filtering.
Use ranking loss to optimize similarity between:
- - Global (G vs h)
- - Local (L vs H)

3 Method Overview – Tal Prompting Pipeline



$$p_i' = \sum_{j=1}^{N} \frac{\exp(P_{ij}/\tau_s)}{\sum_{j=1}^{N} \exp(P_{ij}/\tau_s)} \cdot P_{ij}$$

Testing Phase:
Replace input with images.
Use CLIP image encoder to extract (f, F), the
- Compare with G and L
- Fuse global & local scores for classificatio

Visualization

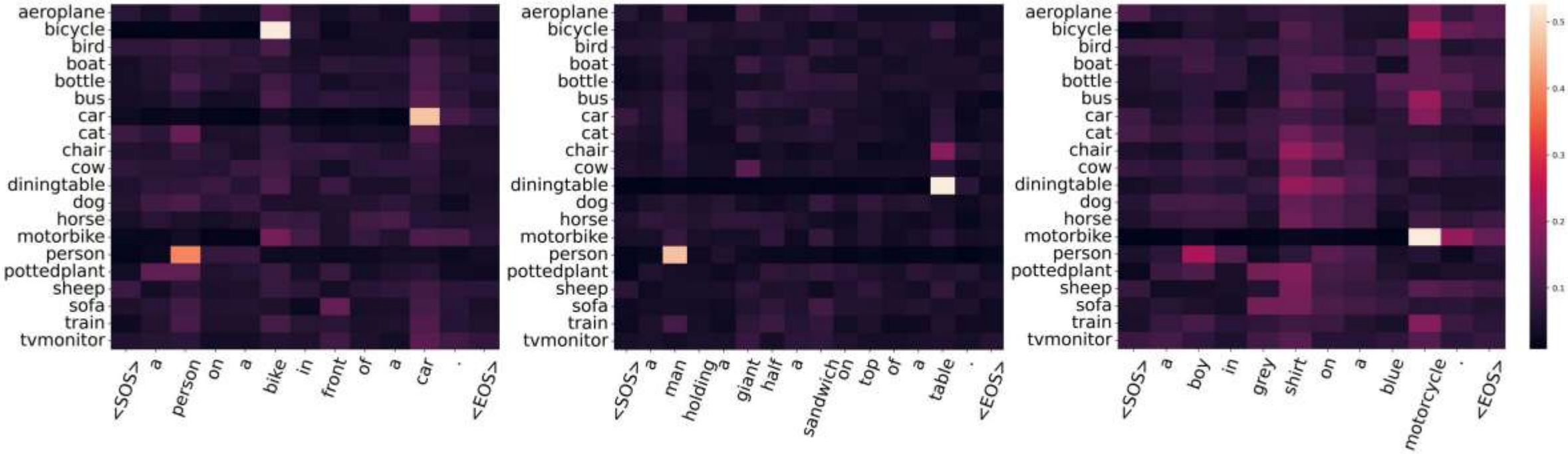


Figure 3. Visualization of correlations between the local class embedding $\boldsymbol{L}$ and sequential token feature from texts. Each class embedding clearly correlates to words that describe the corresponding class (shown in highlight regions) rather than the global <EOS> token.

- Correlation heatmaps between class embedding \(L \) and tokens.
- High responses appear at meaningful class tokens (e.g., "car", "table") instead of generic tokens like `<EOS>`.

Visualization

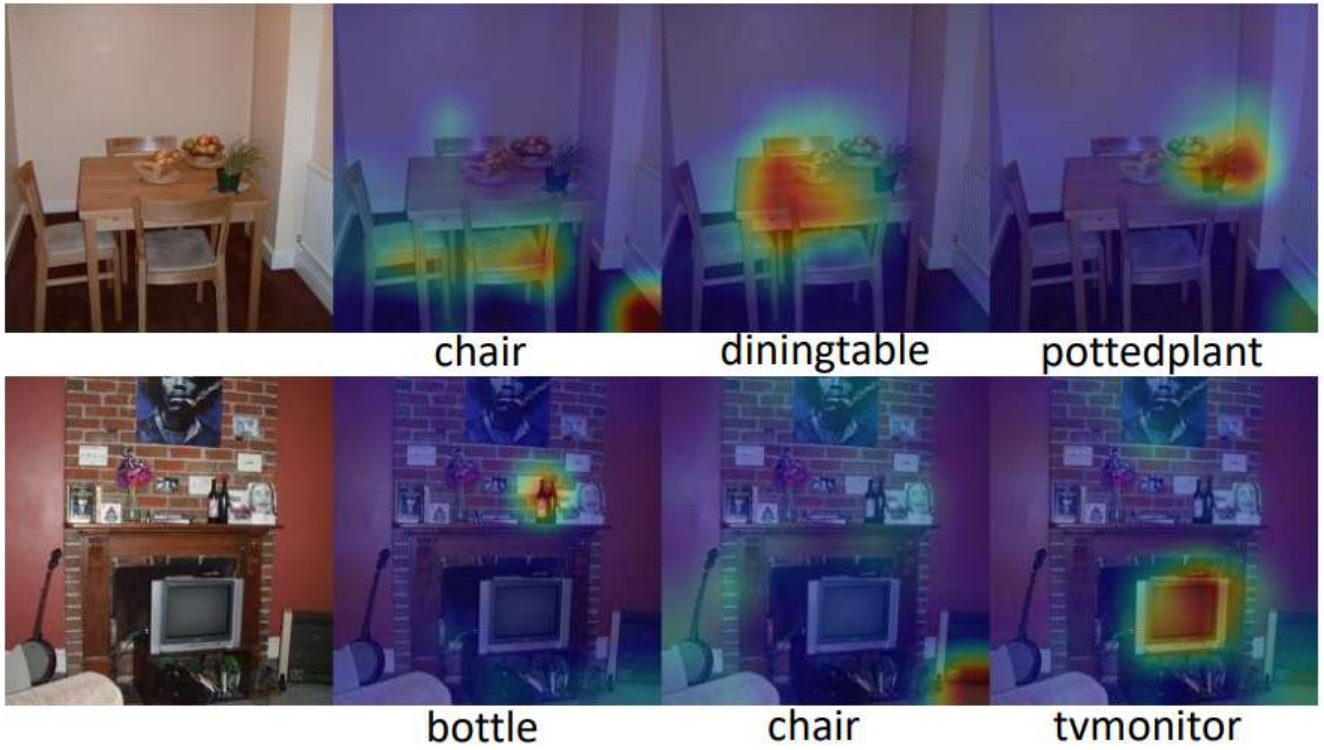


Figure 4. Visualization of correlations between the local class embedding $\boldsymbol{L}$ and dense image feature. The learned class embeddings can focus on the location of the object effectively.

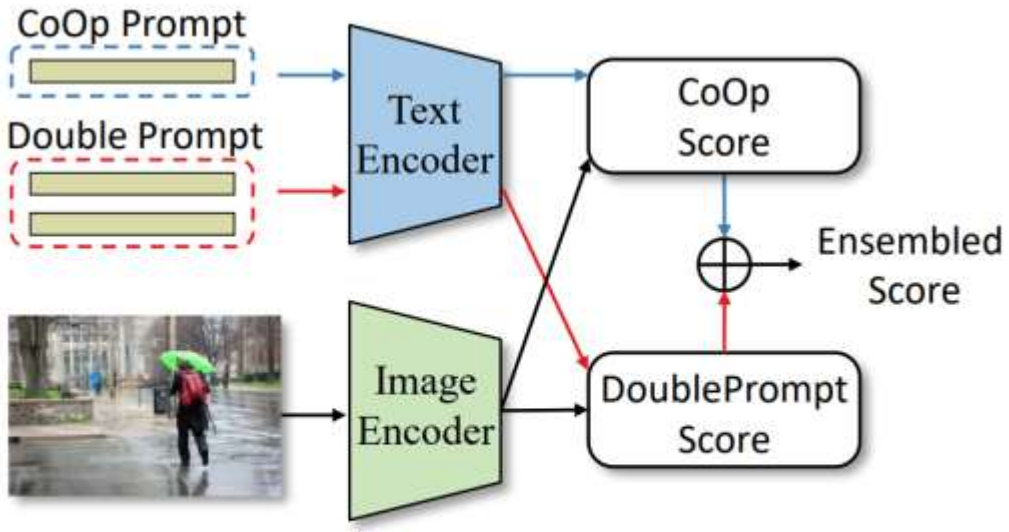


Figure 5. Our learned double-grained prompt tuning is easy to combine with existing prompt tuning methods with ensemble.

4 Experiments

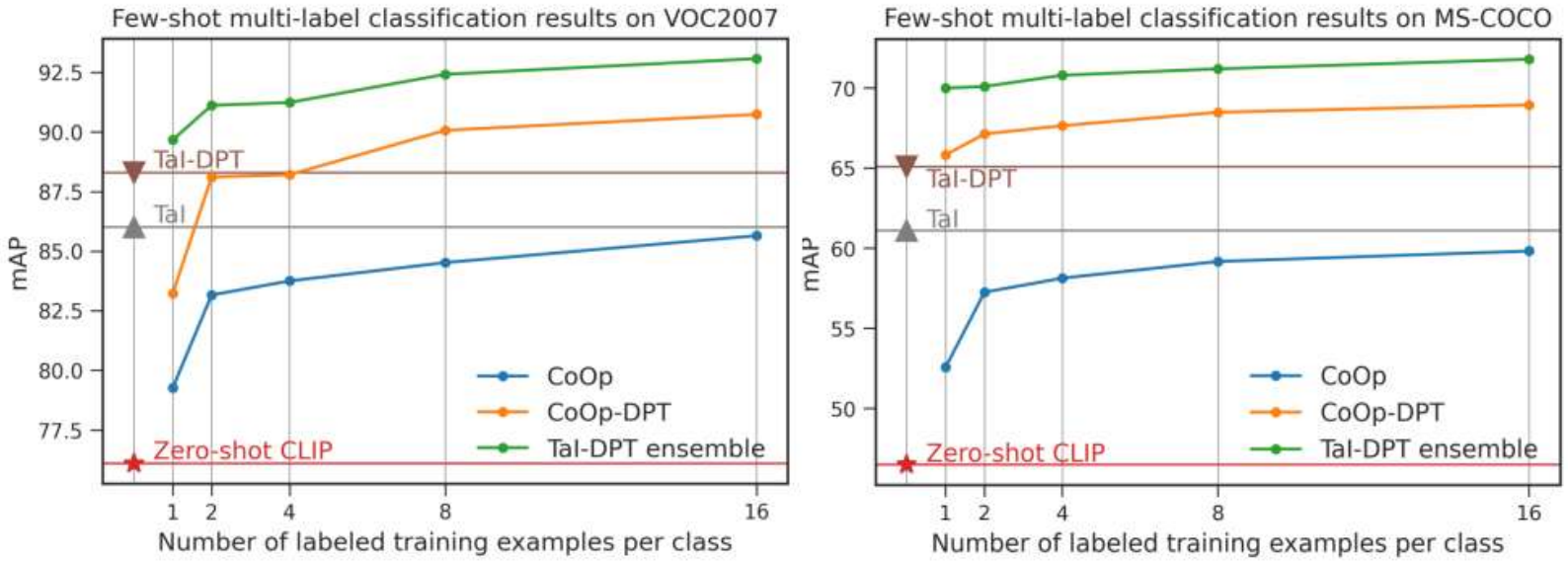


Figure 6. Comparison of different methods in few-shot multi-label recognition on VOC2007 and MS-COCO. Our zero-shot TaI-DPT can achieve comparable results with methods trained by 16-shot labeled image samples. And learned prompt ensemble proofs the complementarity between images and texts.

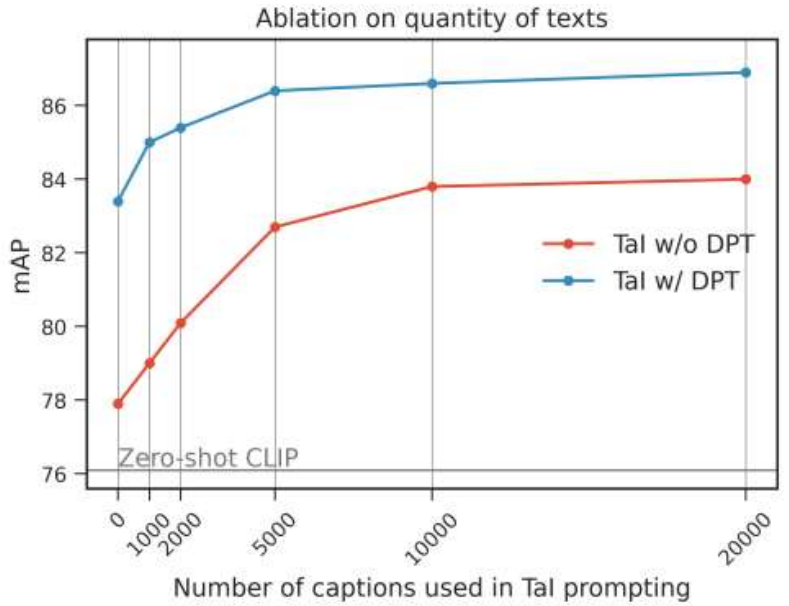


Figure 7. Ablation experiment on number of texts and performance of TaI prompting on VOC2007.

4 Experiments

Table 1. Comparison with zero-shot methods on VOC2007, MS-COCO, and NUS-WIDE. Our proposed TaI-DPT outperforms CLIP [33] by a large margin on all datasets.

Method	DPT	VOC2007	MS-COCO	NUSWIDE
ZSCLIP	✗	76.2	47.3	36.4
	✓	77.3	49.7	37.4
TaI	✗	86.0	61.1	44.9
	✓	**88.3**	**65.1**	**46.5**

Table 3. Comparison with existing multi-label few-shot learning methods on MS-COCO. The evaluation is based on mAP for zero-shot, 1-shot and 5-shot with 16 novel classes.

Method	**0-shot**	1-shot	5-shot
LaSO [2]	-	45.3	58.1
ML-FSL [36]	-	54.4	63.6
TaI-DPT	59.2	-	-

4 Experiments

Table 2. Results of integrating our TaI-DPT with partial-label multi-label recognition method based on pre-trained CLIP. Our approach further improves the frontier performance of DualCoOp [37]. $*$ indicates the results of our reproduction.

Datasets	Method	10%	20%	30%	40%	50%	60%	70%	80%	90%	Avg.
MS-COCO	SARB [32]	71.2	75.0	77.1	78.3	78.9	79.6	79.8	80.5	80.5	77.9
	DualCoOp [37]	78.7	80.9	81.7	82.0	82.5	82.7	82.8	83.0	83.1	81.9
	DualCoOp*	81.0	82.3	82.9	83.4	83.5	83.9	84.0	84.1	84.3	83.3
	+TaI-DPT	**81.5**	**82.6**	**83.3**	**83.7**	**83.9**	**84.0**	**84.2**	**84.4**	**84.5**	**83.6**
PascalVOC 2007	SARB [32]	83.5	88.6	90.7	91.4	91.9	92.2	92.6	92.8	92.9	90.7
	DualCoOp [37]	90.3	92.2	92.8	93.3	93.6	93.9	94.0	94.1	94.2	93.2
	DualCoOp*	91.4	93.8	93.8	94.3	94.6	94.7	94.8	94.9	94.9	94.1
	+TaI-DPT	**93.3**	**94.6**	**94.8**	**94.9**	**95.1**	**95.0**	**95.1**	**95.3**	**95.5**	**94.8**
NUS-WIDE	DualCoOp*	54.0	56.2	56.9	57.4	57.9	57.9	57.6	58.2	58.8	57.2
	+TaI-DPT	**56.4**	**57.9**	**57.8**	**58.1**	**58.5**	**58.8**	**58.6**	**59.1**	**59.4**	**58.3**

5 Conclusion

- Propose **Text-as-Image (Tal) prompting** : no image labels needed.
- Introduce **Double-Grained Prompt Tuning (DPT)** for global & local understanding.
- Demonstrated effectiveness in: Zero-shot 、Few-shot 、Partial-label settings
- Compatible with existing image-prompt methods via ensemble.

THANKS