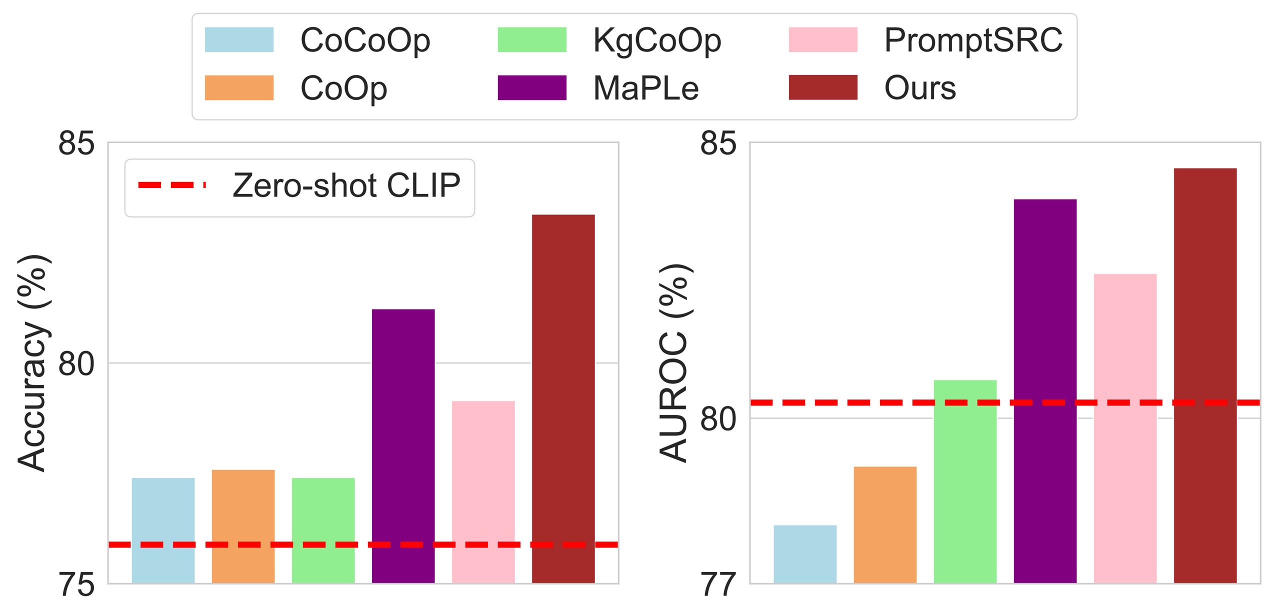
Catching the Unknown with Limited Data: Bi-Directional Prompt Tuning in CLIP for Few-Shot Open-Set Adaptation

Moloud Abdar[1], Md Mehedi Hasan[2], Biplab Banerjee[3], Abbas Khosravi[2], Pietro Lio[4]
[1]The University of Queensland, [2]Deakin University, [3]IIT Bombay, [4]University of Cambridge

BMVC 2025

Problem Definition and Contribution

Challenges: Domain Adaptive Few-Shot Open-Set Learning (DAFSOS) combines two core challenges: (1) adjusting to domain shifts when labeled data from the source and target domains may have fundamentally different distributions, and (2) recognizing open-set unknowns, not seen during training.



Key Contributions: We introduce a multi-modal prompt learning with CLIP-based models that 1) utilizes bidirectional prompt coupling. 2) It incorporates class discriminative attributes in the text prompt. 3) Finally, it generalizes well on the challenging DAFSOS problem.

Formulation

Text Encoder Equation:

$$[\_, E_i] = S_i([\tilde{W}_{i-1}, E_{i-1}]), i = 1, 2, ..J. \quad (1)$$

$$[\tilde{W}_i, E_i] = S_i([\tilde{W}_{i-1}, E_{i-1}]), i = J+1, ..K. \quad (2)$$

$$\tilde{W}_i = [W_i, \mathcal{F}_i^{vl}(Q_i)], i = 0, 1, ..J-1, \quad (3)$$

Image Encoder Equation:

$$[c_i, P_i, \_] = T_i([c_{i-1}, P_{i-1}, \tilde{Q}_{i-1}]), i = 1, 2, .., J+1. \quad (4)$$

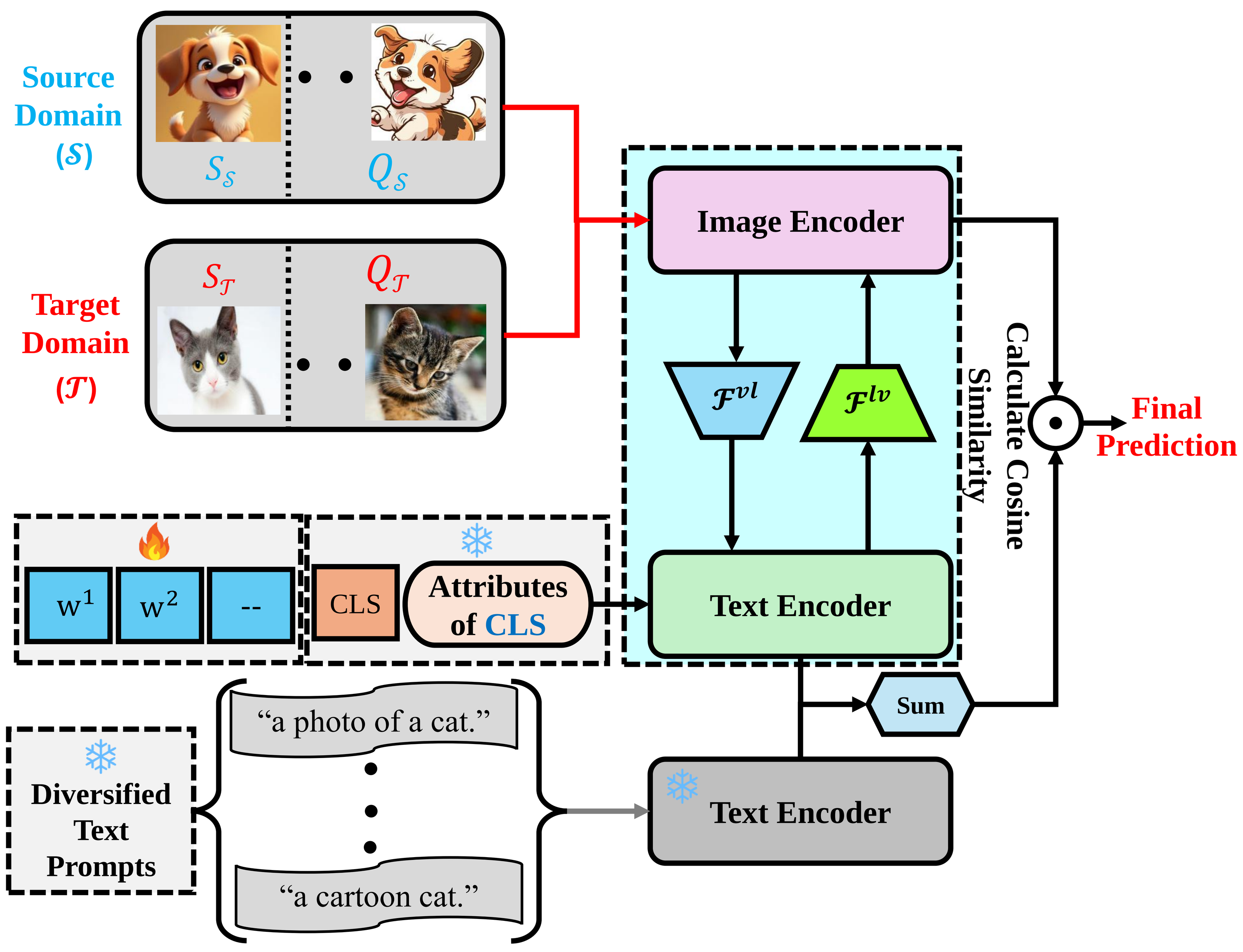
$$[c_i, P_i, \tilde{Q}_i] = T_i([c_{i-1}, P_{i-1}, \tilde{Q}_{i-1}]), i = J+2, .., K. \quad (5)$$

$$\tilde{Q}_i = \begin{cases} Q_i & for \quad i = 0 \\ [Q_i, \mathcal{F}_i^{lv}(W_{i-1})] & for \quad 1 \le i \le J-1 \\ \mathcal{F}_i^{lv}(W_{i-1}) & for \quad i = J \end{cases} \quad (6)$$

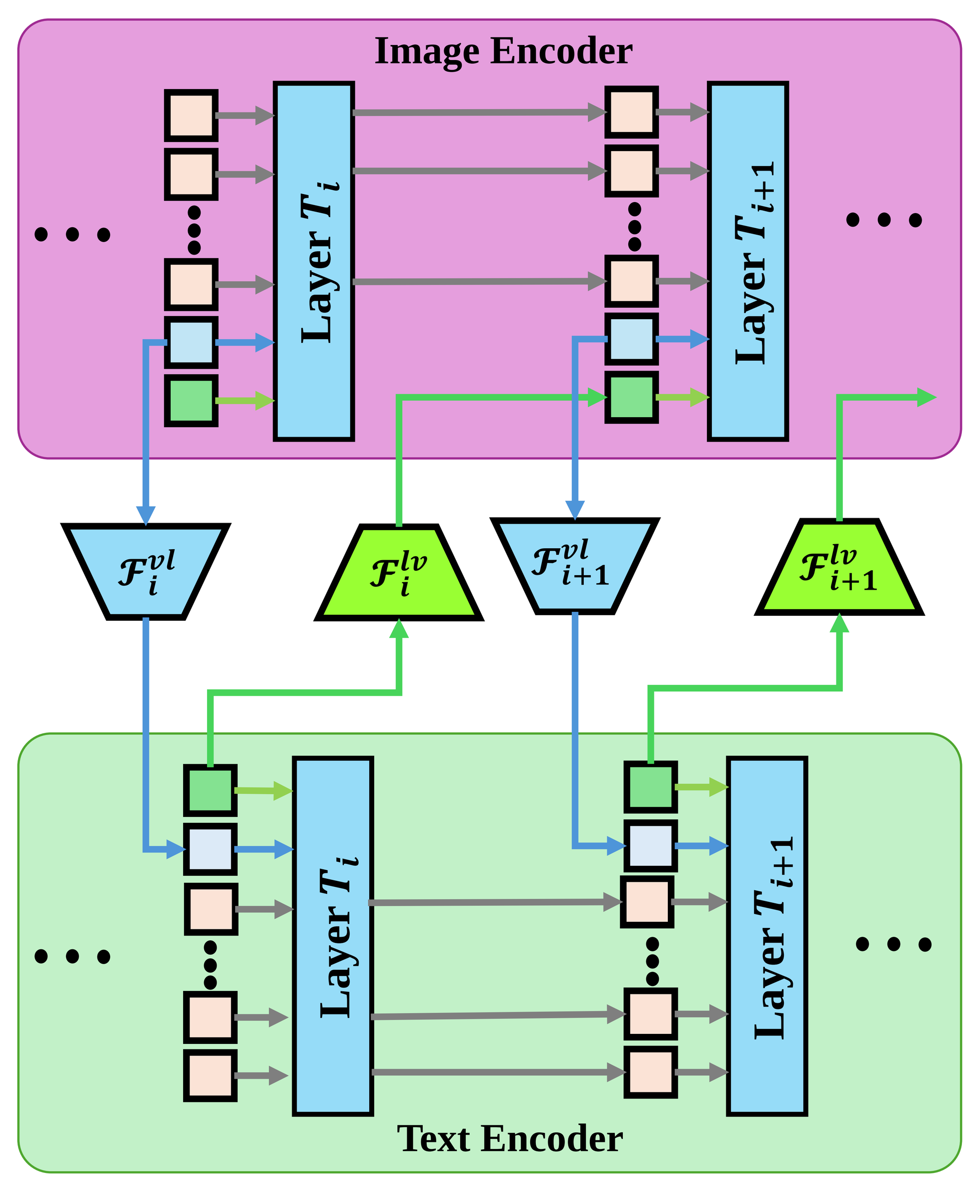
Proposed Method

Overall Architecture:

In this study, we propose Bi-directional Multi-modal Attribute-enriched Prompting for CLIP (BIMAP-CLIP).



Bidirectional Coupling:



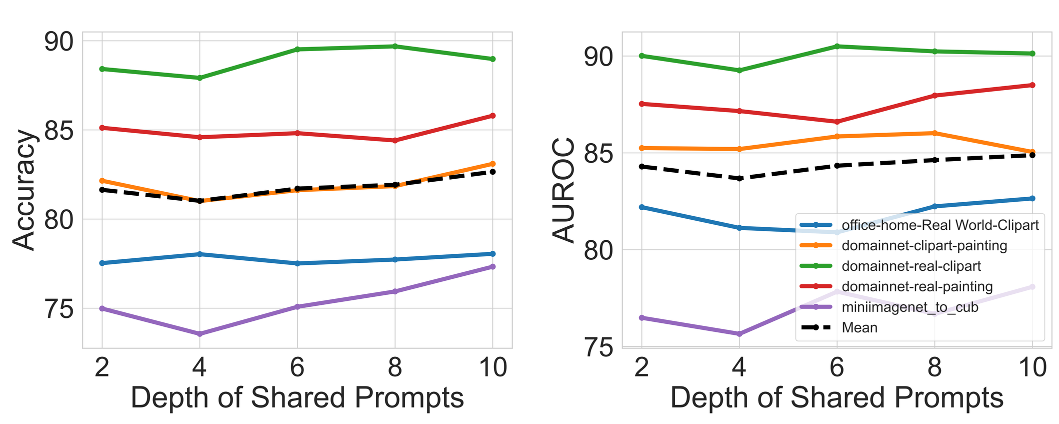
Experiments & Results

Quantitative Results on 5 domain adaptive setups:

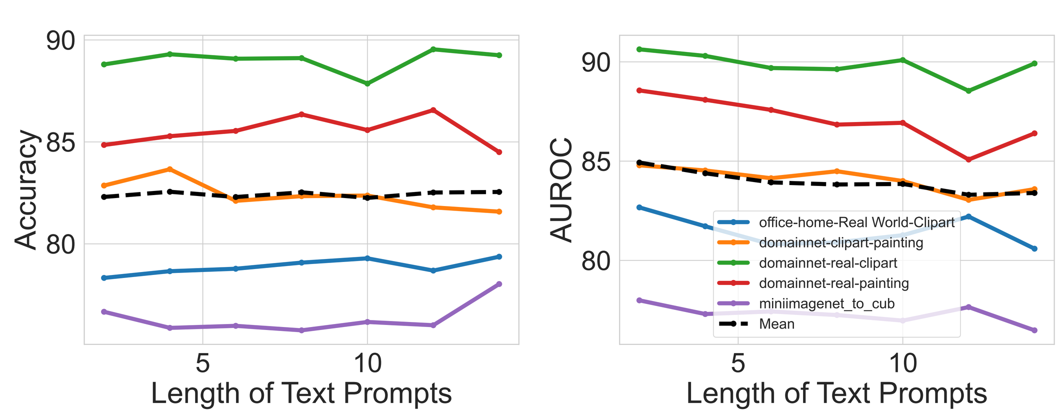
Architecture	Method	Office-home Real World-Clipart ACC	Office-home Real World-Clipart AUROC	Clipart-Painting ACC	Clipart-Painting AUROC	Domainnet Real-Clipart ACC	Domainnet Real-Clipart AUROC	Real-Painting ACC	Real-Painting AUROC	Mini-imagenet to CUB Mini-imagenet-CUB ACC	Mini-imagenet to CUB Mini-imagenet-CUB AUROC
ViT-B/32	DAFOS-NET	51.79±0.67	50.54±0.54	56.21±0.21	57.26±0.49	52.42±0.37	54.84±0.36	55.06±0.34	54.16±0.33	49.25±0.62	50.17±0.28
	CLIP	75.29±0.2	78.63±0.1	77.15±0.13	81.44±0.09	84.92±0.06	85.54±0.04	81.45±0.13	83.8±0.07	60.61±0.54	72.0±0.21
	CoCoOp	73.5±0.28	77.0±0.08	80.27±0.28	78.82±0.19	86.94±0.04	85.21±0.12	83.66±0.15	80.1±0.15	62.68±0.12	69.22±0.27
	CoOp	70.34±0.41	74.29±0.08	79.1±0.14	80.74±0.09	86.27±0.06	85.78±0.09	83.56±0.15	83.97±0.06	68.7±0.24	70.88±0.18
	KgCoOp	71.57±0.36	78.84±0.1	78.85±0.29	81.15±0.08	85.91±0.09	87.87±0.06	82.99±0.29	83.97±0.04	67.72±0.39	71.67±0.21
	MaPLe	77.71±0.23	**81.45±0.07**	81.24±0.18	84.3±0.09	88.88±0.06	89.36±0.13	85.31±0.1	**87.53±0.1**	73.02±0.3	77.25±0.39
	PromptSRC	72.95±0.37	80.17±0.03	81.46±0.15	84.19±0.1	88.2±0.07	88.8±0.06	85.37±0.07	86.22±0.04	67.77±0.44	73.74±0.28
	Ours	**79.26±0.23**	81.34±0.08	**82.22±0.08**	**84.67±0.13**	**90.65±0.04**	**90.77±0.05**	**86.55±0.11**	87.38±0.1	**78.19±0.33**	**78.54±0.53**

Ablation Study

Effect of the depth of the shared prompts:



Effect of the length of the learnable text prompts:



Conclusion

we introduce:

- A bi-directional multi-modal prompt-learning mechanism, named BIMAP-CLIP, that strengthens visual–semantic alignment by sharing learnable tokens.
- A hybrid prompt initialization strategy that combines manual and learnable prompts, enhancing their resilience across domains.