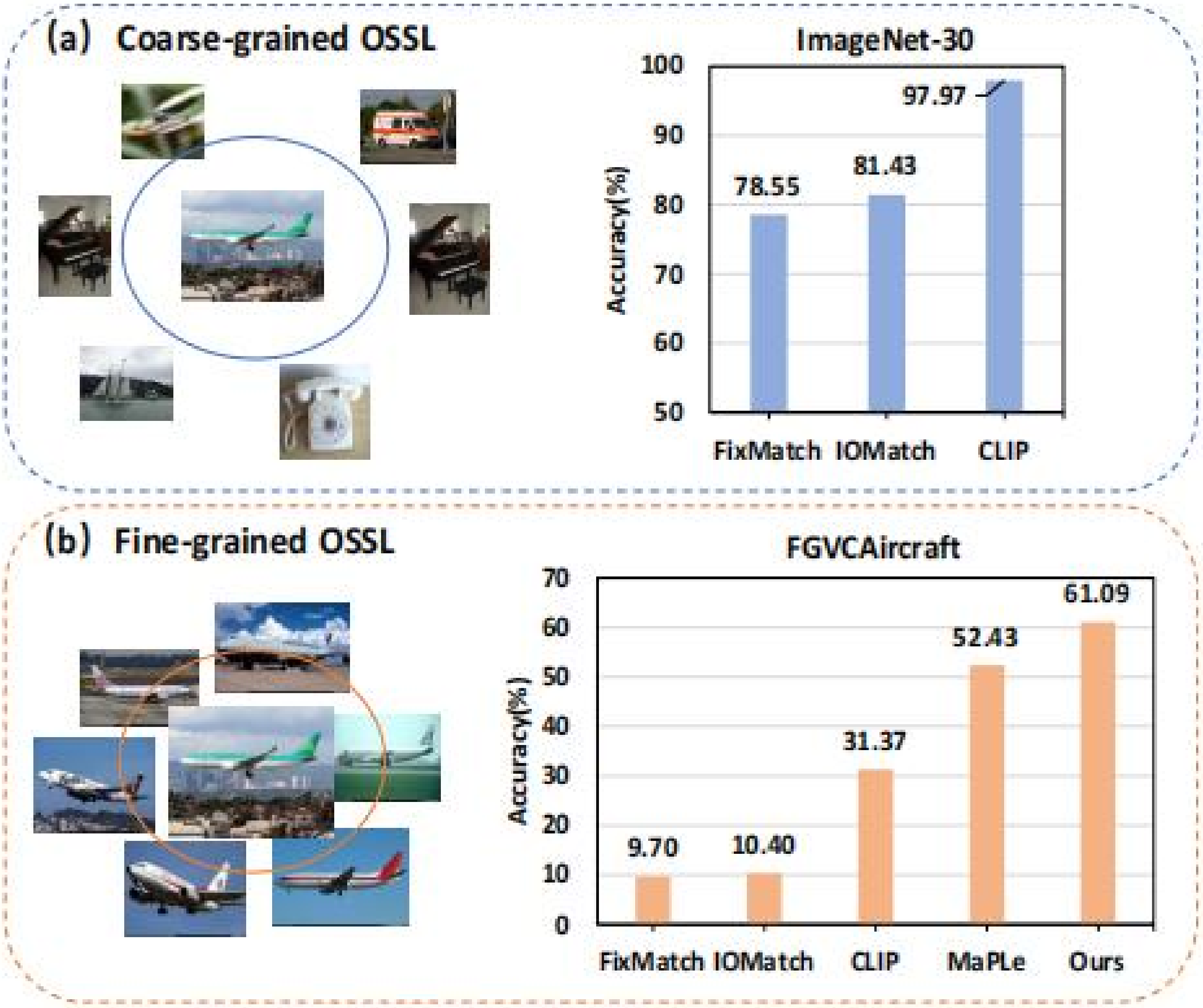
CLIP-driven Coarse-to-fine Semantic Guidance for Fine-grained Open-set Semi-supervised Learning

Xiaokun Li, Yaping Huang*, Qingji Guan*
Beijing Key Laboratory of Traffic Data Mining and Embodied Intelligence, Beijing Jiaotong University
22110102@bjtu.edu.cn, yphuang@bjtu.edu.cn, qjguan@bjtu.edu.cn

CVPR 2025

Background



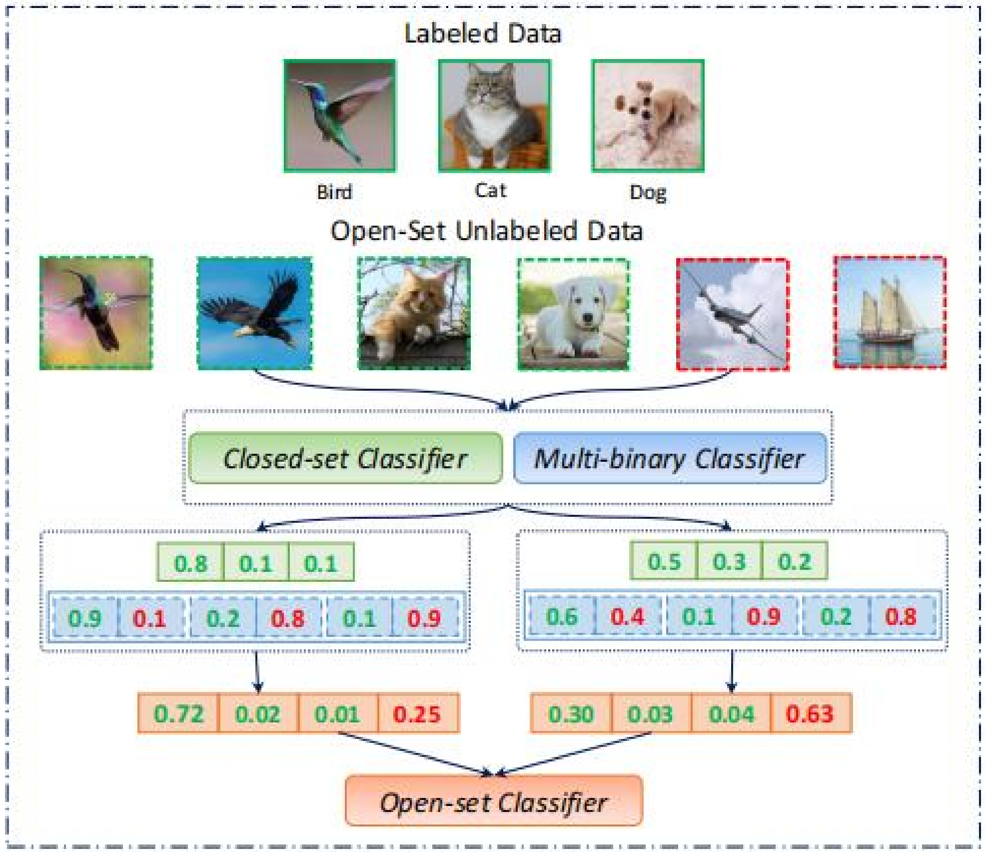
Existing OSSL methods:
have not sufficiently explored more practical fine-grained OSSL tasks.

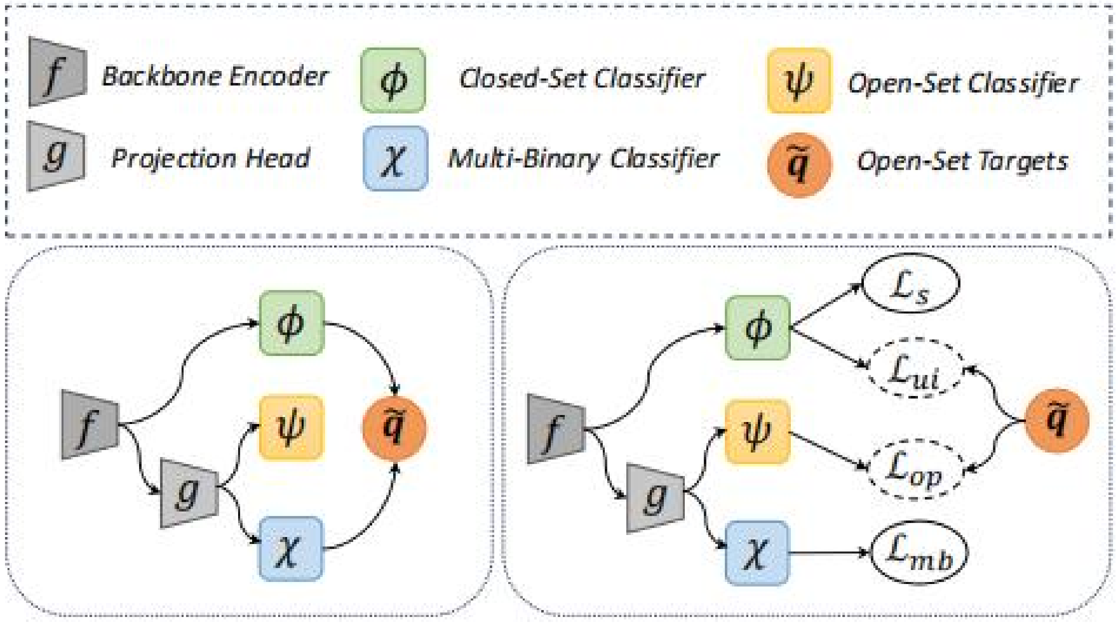
Visual language models(CLIP):
Focus on capturing global, general attributes.
Difficult to concentrate on fine-grained features.

Core Challenges:
Visual differences between ID and OOD samples are not significant.
Existing models struggle to capture fine-grained distinguishing features.

Related Work

OSSL methods: 对OOD样本进行丢弃/加权/视为负样本
IOMatch/SCOMatch: 将 OOD 样本归类为新类别,优化一个类别为K+1的开集分类器

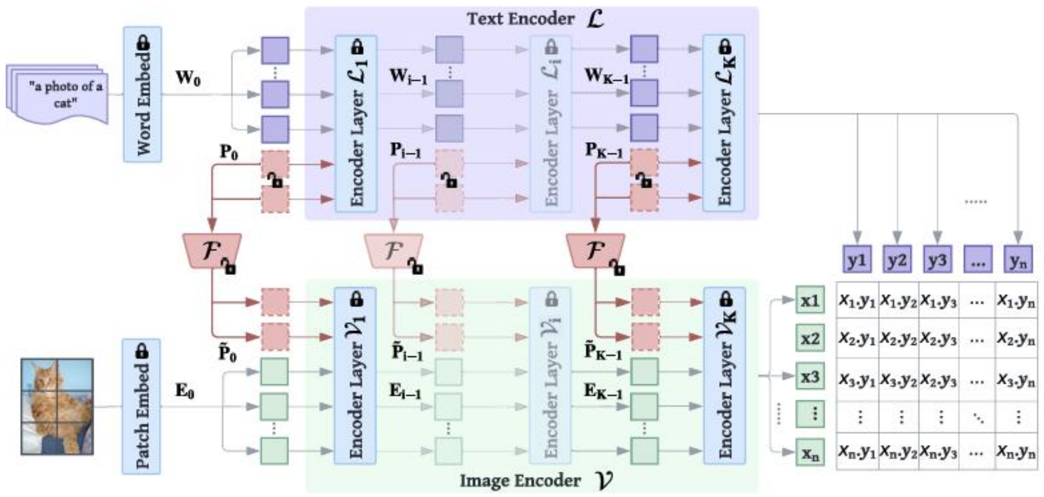




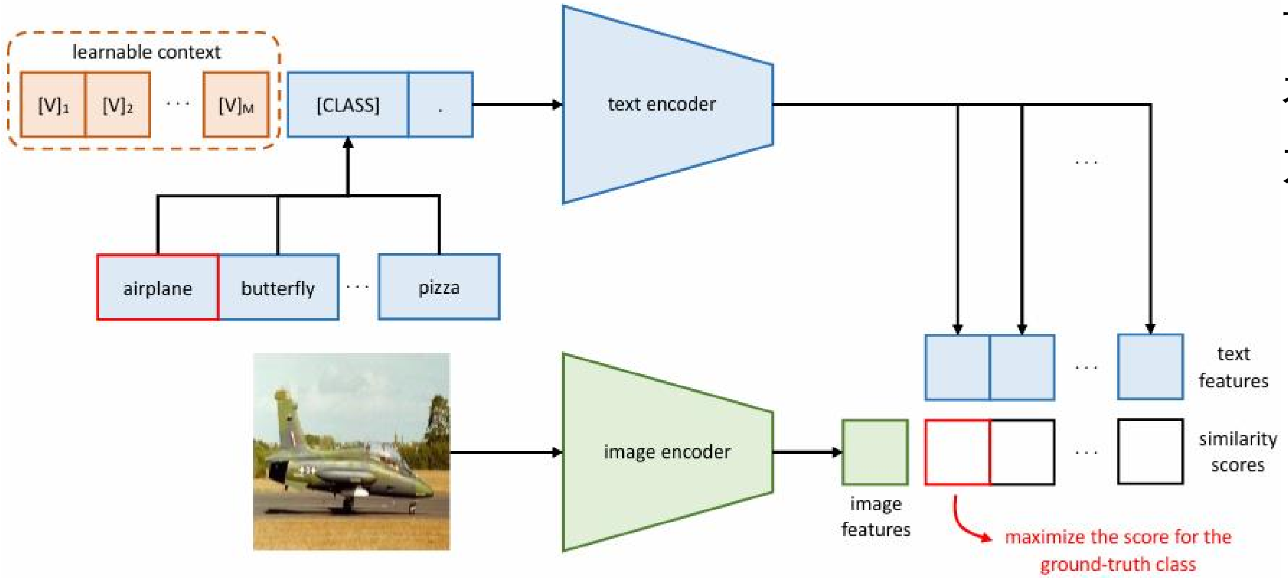
(a) Produce Open-Set Targets (b) Optimize Whole Framework

不足：仅能处理粗粒度OOD样本带来的影响，忽略了ID/OOD样本之间的细粒度差异。面对紧密特征分布边界时性能下降。

Related Work



MaPLe



CoOp

CLIP based method:

CLIP-Adapter/CoOp/PLOT/MaPLe

通常通过计算每种模态全局特征之间的相似性来进行跨模态交互,忽略了CLIP中局部特征包含大量与类别语义无关的信息干扰,无法有效筛选出细粒度任务所需局部关键特征

LoCoOP

面向少样本分类、 OOD 检测等任务设计,在细粒度开集数据上无法匹配同类别方差较大，跨类别方差较小的精细区分需求

Method

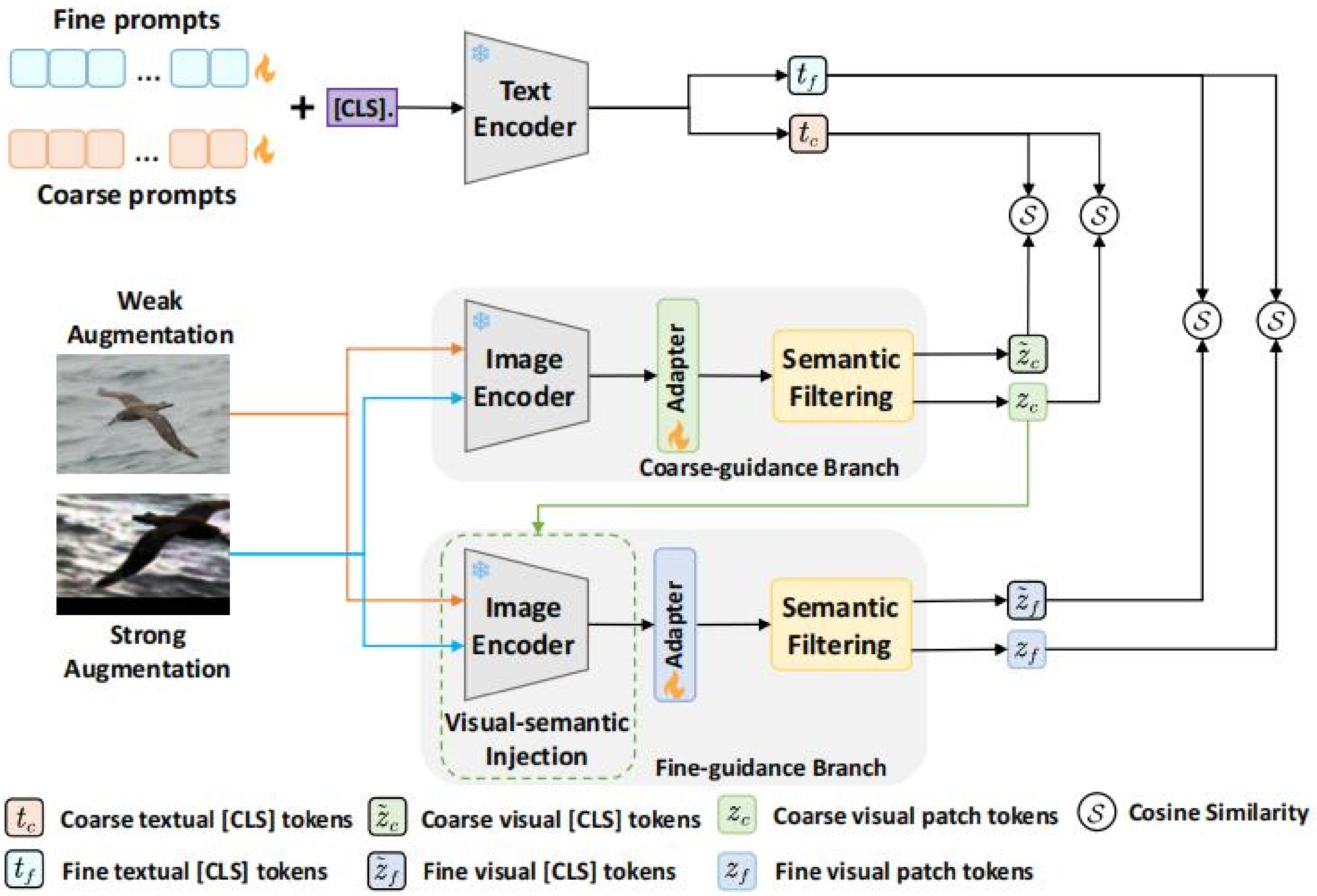


Figure 2. Overview of the proposed CFSG-CLIP framework. CFSG-CLIP is composed of a coarse-guidance branch and a fine-guidance branch based on the pre-trained CLIP model. In the coarse-guidance branch, we design a semantic filtering module to initially capture global and local visual features. In the fine-guidance branch, we design a visual-semantic injection strategy to embed category-related visual cues into the visual encoder for further refining the local visual features. For brevity, we omit the SSL training process.

Method

Problem Setting

$$\mathcal{D}^l = \{(x_1^l, y_1^l), (x_2^l, y_2^l), ..., (x_N^l, y_N^l)\}$$

$$\mathcal{D}^u = \{(x_1^u), (x_2^u), ..., (x_U^u)\}$$ (ID/OOD samples)

$$y^l \in \{1, \ldots, M\}$$

test phase: ID samples

Semantic Filtering Module

$$\mathcal{Z} = \{\tilde{z}_c, z_{c_1}, z_{c_2}, ..., z_{c_P}\} \in \mathbb{R}^{(P+1)\times d}$$

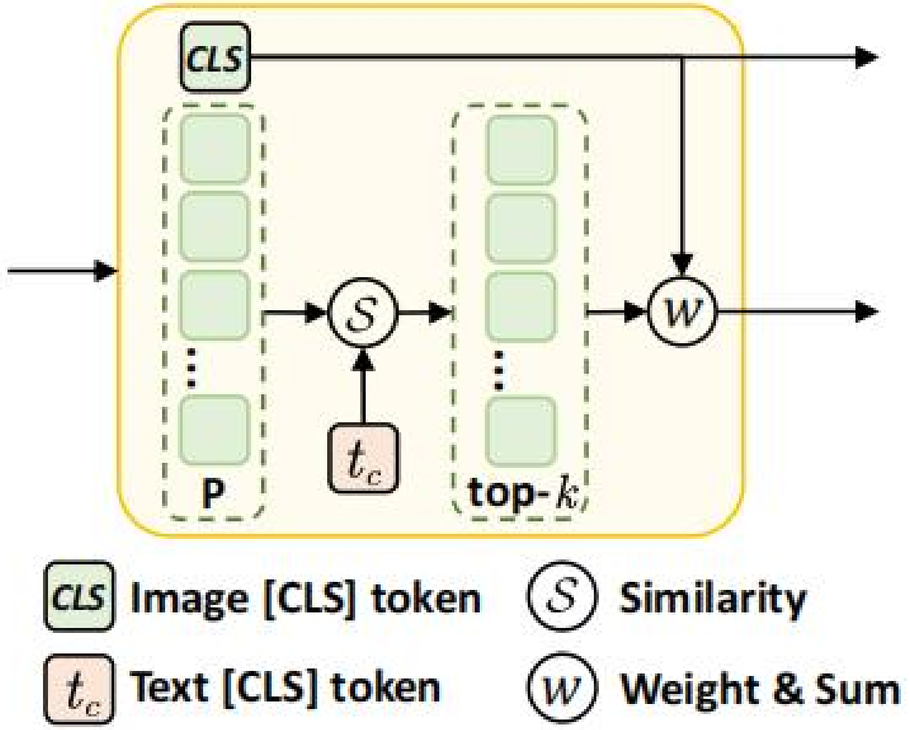
coarse prompt $p_c^m = \{v_1, v_2, ..., v_n, \mathcal{C}^m\}$

$$s_{c_i} = \mathrm{sim}(z_{c_i}, t_c^m),$$

$$\mathcal{K} = \{i \in P : \mathrm{rank}(s_{c_i}) \leq k\},$$

$$w_{c_i} = \frac{\exp(\mathrm{sim}(z_{c_i}, \tilde{z}_c))}{\sum \exp(\mathrm{sim}(z_{c_i}, \tilde{z}_c))}, i \in \mathcal{K}.$$

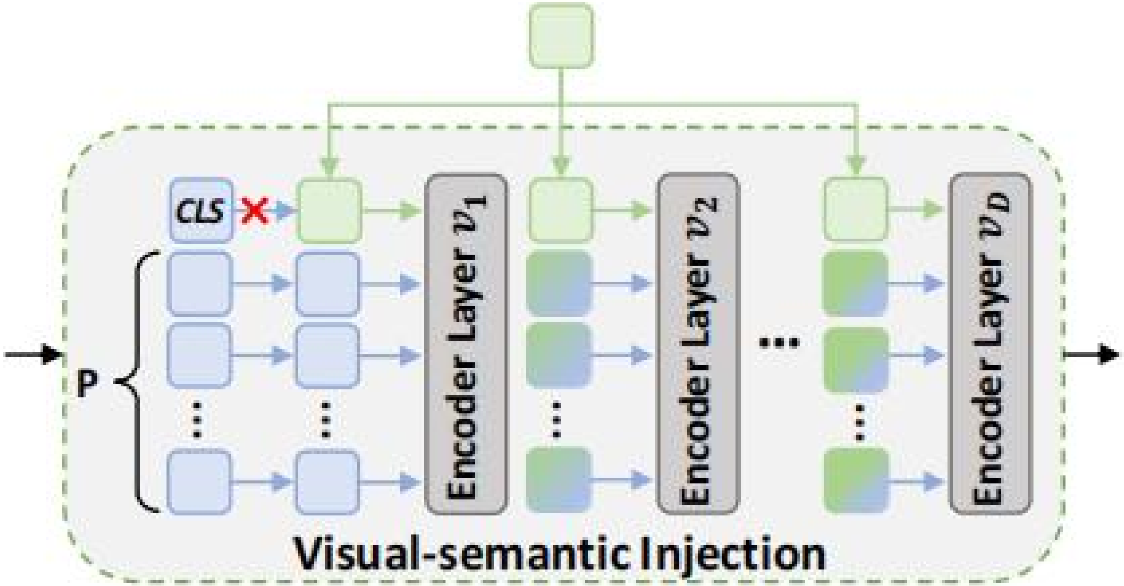
$$z_c = \sum_{i \in \mathcal{K}} w_{c_i} z_{c_i}.$$



Method

南京航空航天大學
NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

Visual-semantic Injection



$$[\_, E_j] = \mathcal{V}_j([\mathrm{proj}(z_c), E_{j-1}]) \quad j = 1, 2, \ldots, D,$$

$$[\tilde{z}_{f_j}, E_j] = \mathcal{V}_j([\mathrm{proj}(z_c), E_{j-1}]) \quad j = D+1, \ldots, T,$$

Dual-branch Training

$$\tilde{p}_c^l = \frac{\exp(\mathrm{sim}(\tilde{z}_c^l, t_c^m)/\tau)}{\sum_{m'} \exp(\mathrm{sim}(\tilde{z}_c^l, t_c^{m'})/\tau)},$$

$$p_c^l = \frac{\exp(\mathrm{sim}(z_c^l, t_c^m)/\tau)}{\sum_{m'} \exp(\mathrm{sim}(z_c^l, t_c^{m'})/\tau)},$$

$$L_c = H(y^l, \tilde{p}_c^l) + H(y^l, p_c^l) + \lambda_c(\mathcal{F}(x^u) H(\tilde{p}_c^{u_w}, \tilde{p}_c^{u_s}) + \mathcal{F}(x^u) H(p_c^{u_w}, p_c^{u_s})),$$

$$L_f = H(y^l, \tilde{p}_f^l) + H(y^l, p_f^l) + \lambda_f(\mathcal{F}(\tilde{u}_f) H(\tilde{p}_f^{u_w}, \tilde{p}_f^{u_s}) + \mathcal{F}(u_f) H(p_f^{u_w}, p_f^{u_s})),$$

Experiments

Method	Stanford Dogs		Stanford Cars		CUB-200-2011		FGVCAircraft	
	5	20	5	20	5	20	5	20
CLIP [27]	79.25±0.00	79.25±0.00	75.97±0.00	75.97±0.00	66.00±0.00	66.00±0.00	31.37±0.00	31.37±0.00
CLIP-LORA [41]	83.81±0.37	84.31±0.27	82.71±0.56	82.45±1.30	70.95±0.85	73.40±0.75	40.89±1.73	42.37±0.71
CLIP-Adapter [7]	82.91±0.25	86.02±0.27	84.31±0.02	87.13±0.28	80.03±0.29	84.77±1.02	47.77±0.90	55.79±0.73
CoOp [45]	83.01±0.26	85.68±0.37	85.45±0.31	87.64±0.46	80.10±0.29	85.40±0.37	45.39±0.96	55.43±0.30
LoCoOp [24]	83.08±0.25	86.26±0.11	84.10±0.72	87.83±0.66	79.27±0.45	85.63±0.54	45.53±1.36	54.67±1.59
PLOT [3]	84.46±0.07	87.11±0.09	86.28±0.30	88.59±0.45	81.43±0.66	87.20±0.14	49.59±0.37	58.25±0.93
MaPLe [14]	**85.64**±0.15	87.64±0.20	88.16±0.25	90.34±0.25	83.30±0.33	88.77±0.21	52.43±0.47	64.33±1.21
Ours	85.48±0.21	**89.42**±0.16	**90.38**±0.09	**93.08**±0.08	**84.73**±0.17	**91.75**±0.24	**61.09**±0.27	**73.56**±0.58

Method	Stanford Dogs		Stanford Cars		CUB-200-2011		FGVCAircraft	
	5	20	5	20	5	20	5	20
CLIP [27]	77.17±0.00	77.17±0.00	75.70±0.00	75.70±0.00	64.10±0.00	64.10±0.00	31.08±0.00	31.08±0.00
CLIP-LORA [41]	82.34±0.57	82.67±0.36	82.10±0.56	81.03±1.24	70.52±1.34	72.20±0.45	40.10±1.69	41.57±0.69
CLIP-Adapter [7]	81.63±0.14	84.36±0.25	83.65±0.04	86.50±0.25	81.36±0.60	85.20±0.95	46.87±0.88	53.52±1.35
CoOp [45]	81.65±0.20	84.25±0.41	84.63±0.36	86.84±0.42	81.04±0.06	84.92±0.51	44.55±0.94	54.35±0.30
LoCoOp [24]	81.78±0.23	84.68±0.20	83.25±0.76	87.17±0.64	80.45±0.86	85.46±0.32	44.67±1.35	53.60±1.57
PLOT [3]	82.95±0.07	85.54±0.11	85.58±0.32	87.83±0.34	83.32±0.31	87.16±0.17	48.68±0.37	57.13±0.92
MaPLe [14]	**84.09**±0.19	86.02±0.25	87.43±0.27	89.48±0.25	84.72±0.53	88.66±0.60	51.79±0.43	63.12±1.19
Ours	84.02±0.15	**87.77**±0.19	**89.65**±0.05	**92.34**±0.10	**86.46**±0.25	**90.92**±0.24	**59.92**±0.26	**72.13**±0.57

Open-set classification balanced accuracy $BA = \frac{1}{K+1}\sum_{k=1}^{K+1} Recall_k$

Experiments

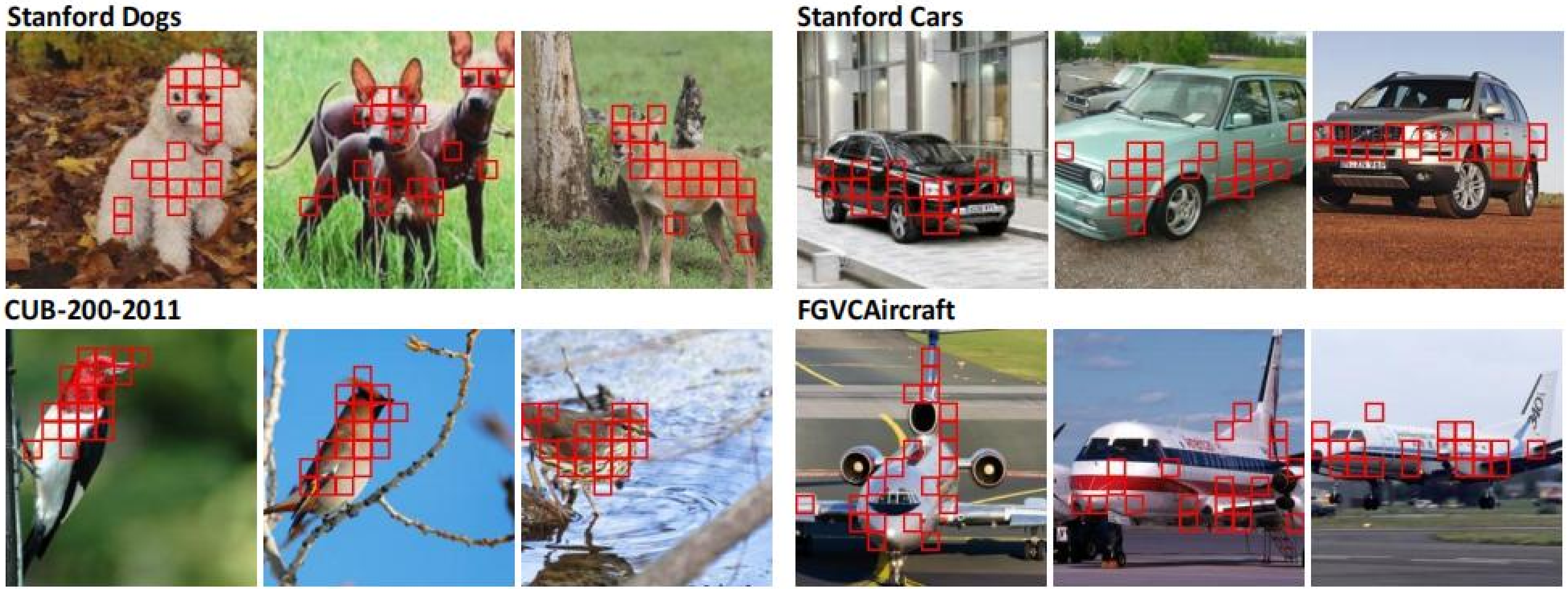


Figure 5. Visualization of patch-tokens extracted by semantic filtering module. We find that the semantic filtering module can correctly extract local visual regions on different fine-grained datasets.

南京航空航天大學
NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

OSLOPROMPT: Bridging Low-Supervision Challenges and Open-Set Domain Generalization in CLIP

Mohamad Hassan N C[1] Divyam Gupta[1] Mainak Singha[1] Sai Bhargav Rongali[1] Ankit Jha[2]
Muhammad Haris Khan[3] Biplab Banerjee[1]

[1]Indian Institute of Technology Bombay [2]The LNM Institute of Information Technology (LNMIIT)
[3]Mohamed Bin Zayed University of Artificial Intelligence

CVPR 2025

Background

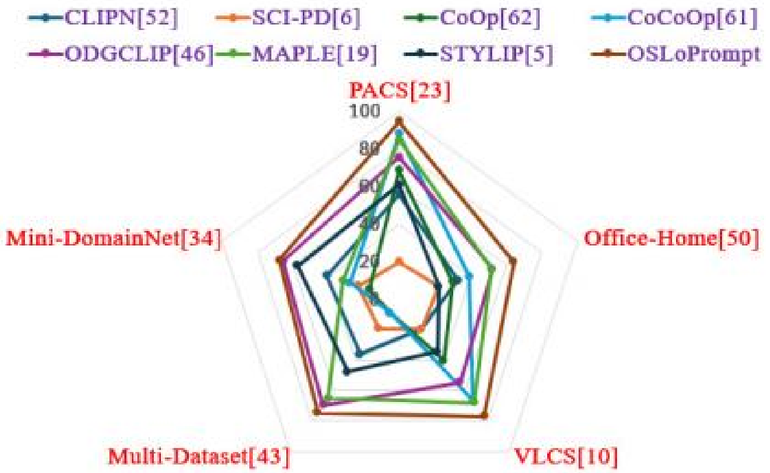


Figure 1. **Harmonic score (H-score) (between known and novel class performances) comparisons** of various CLIP-based DG/ODG/open-set recognition techniques versus our approach in LSOSDG setting with one-training example per known class, demonstrating the improved performances of OSLoPrompt.

- 传统DG/ODG方法依赖充足训练数据，无法应对稀缺训练样本+动态未知类的情况
- CLIP-based方法缺乏细粒度开放样本区分能力，无法应对细粒度开集检测
- Prompt learning的方法对于域无关提示缺乏结构化知识，低数据下易受误导
- 现有的ODG方法(ODG-CLIP)生成的伪开放样本语义相关性差，影响清晰的闭集/开集边界构建。

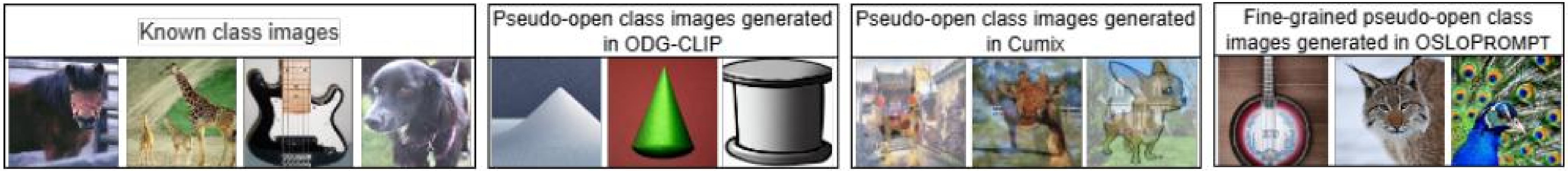


Figure 6. **Pseudo-open images** generated by ODG-CLIP [46] are highly **coarse-grained** in relation to the known classes. While CuMix [28] provides improved fine-grained details compared to ODG-CLIP, it still lacks proper semantic coherence. Our pseudo-open image generation achieves a **fine-grained** level of detail, maintaining both semantic relevance and class-specific granularity (for PACS).

Problem Setting

$\mathcal{D}_s = \{(x_i^s, y_i^s)\}_{i=1}^{n_s} \quad \mathcal{D} = \{\mathcal{D}_s\}_{s=1}^{\mathcal{N}}$

$(\mathcal{P}(\mathcal{D}_s) \neq \mathcal{P}(\mathcal{D}_{s'})$ for $s \neq s')$

$\mathcal{C} = \bigcup_{s=1}^{\mathcal{N}} \mathcal{Y}_s$

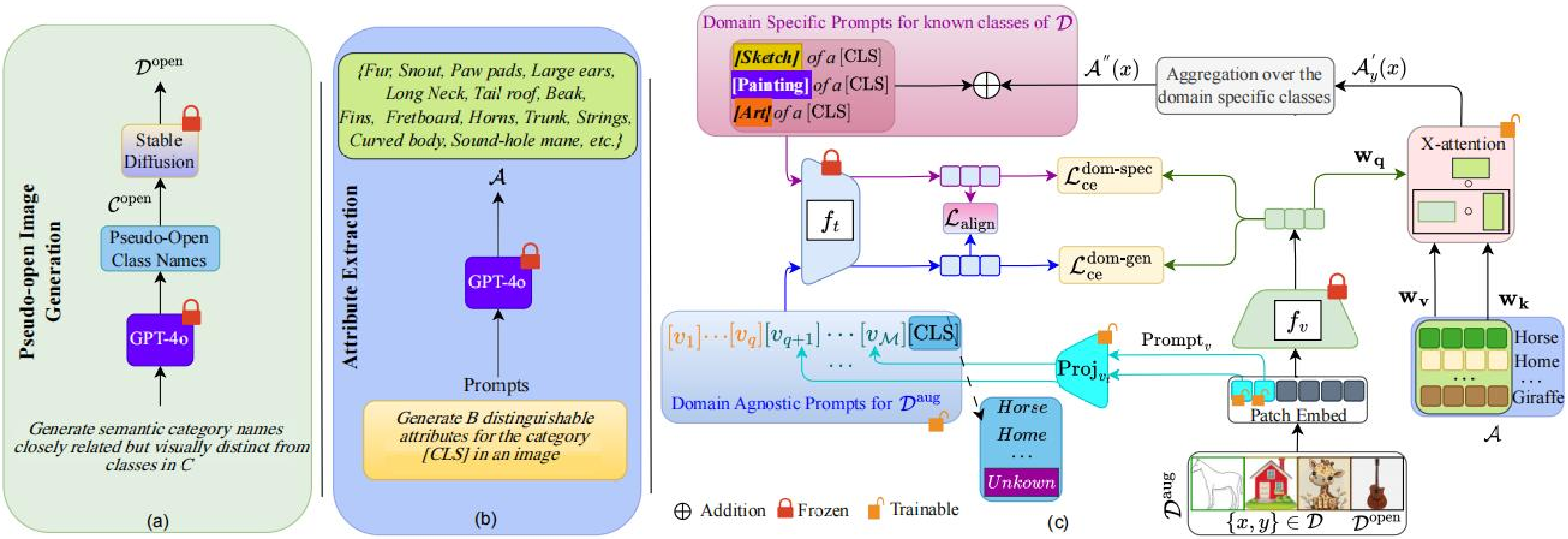
each class typically having limited samples
（1-shot / 5-shot）

test phase:

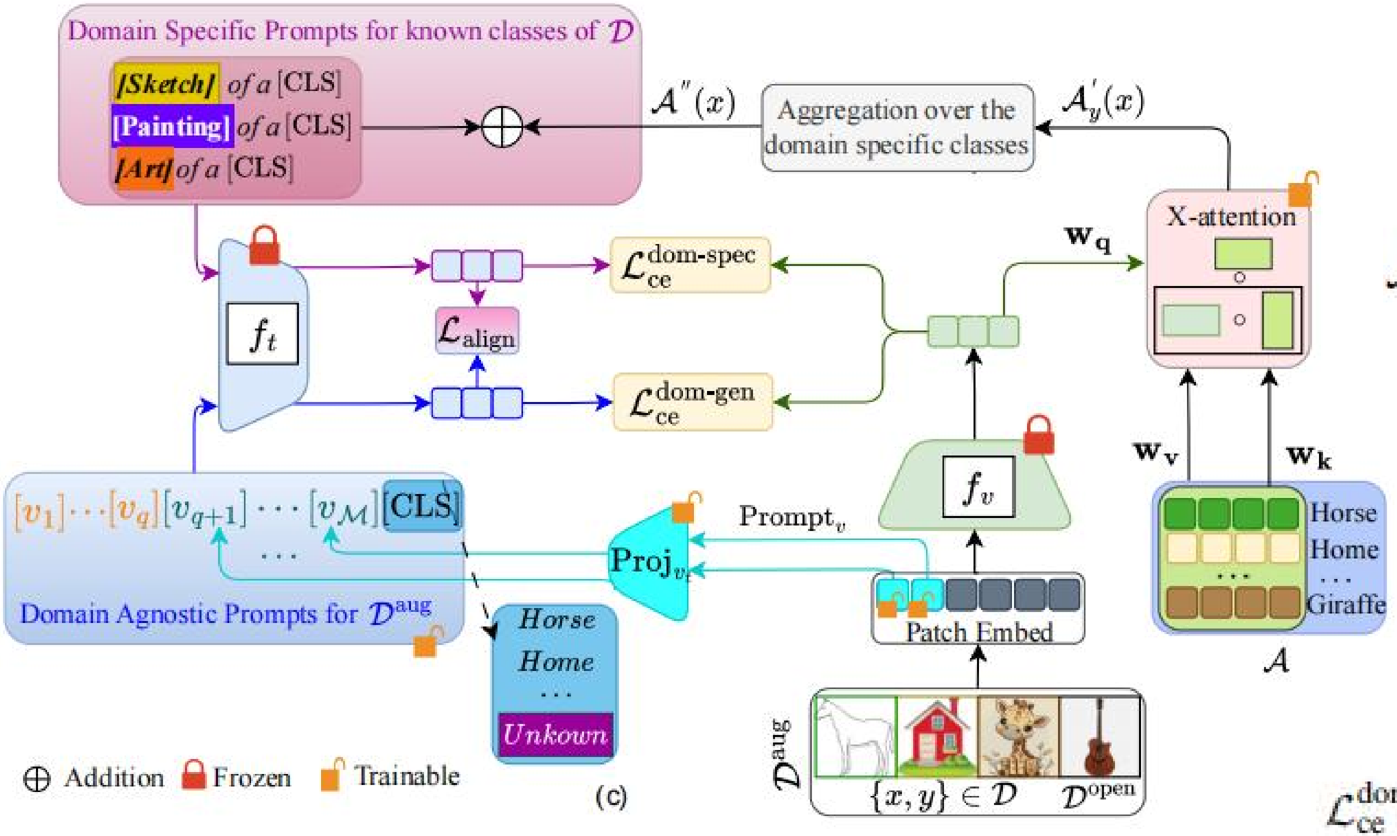
$\mathcal{D}_t = \{x_j^t\}_{j=1}^{n_t}$

target domain’s label set $\begin{cases} \mathcal{Y}_t^{\text{known}} = \mathcal{C} \\ \mathcal{Y}_t^{\text{novel}} = \mathcal{Y}_t \setminus \mathcal{Y}_t^{\text{known}} \end{cases}$

Method


Pseudo-open Image Generation
$\mathcal{D}^{open}$
Stable Diffusion
$\mathcal{C}^{open}$
Pseudo-Open Class Names
GPT-4o
Generate semantic category names closely related but visually distinct from classes in C
(a)
Attribute Extraction
{Fur, Snout, Paw pads, Large ears, Long Neck, Tail roof, Beak, Fins, Fretboard, Horns, Trunk, Strings, Curved body, Sound-hole mane, etc.}
$\mathcal{A}$
GPT-4o
Prompts
Generate B distinguishable attributes for the category [CLS] in an image
(b)
Domain Specific Prompts for known classes of $\mathcal{D}$
[Sketch] of a [CLS]
[Painting] of a [CLS]
[Art] of a [CLS]
$\mathcal{A}''(x)$
Aggregation over the domain specific classes
$\mathcal{A}'_y(x)$
X-attention
$\mathbf{w_q}$
f_t
$\mathcal{L}_{ce}^{dom\text{-}spec}$
$\mathcal{L}_{align}$
$\mathcal{L}_{ce}^{dom\text{-}gen}$
$\mathbf{w_v}$
$\mathbf{w_k}$
f_v
$[v_1]\cdots[v_q][v_{q+1}]\cdots[v_{\mathcal{M}}]$[CLS]
Domain Agnostic Prompts for $\mathcal{D}^{aug}$
Prompt$_v$
Proj$_v$
Horse
Home
Giraffe
$\mathcal{A}$
Horse
Home
Unkown
Patch Embed
$\mathcal{D}^{aug}$
$\{x, y\} \in \mathcal{D}$
$\mathcal{D}^{open}$
⊕ Addition
Frozen
Trainable
(c)

prompt learning strategy



$$\mathcal{A}_{y^s} = [a^1_{y^s}, a^2_{y^s}, \ldots, a^{\mathcal{B}}_{y^s}]$$

"Generate $\mathcal{B}$ distinguishable attributes for the category [CLS] in an image."

Domain-specific prompts

$$\mathbf{Prompt}_{\mathbf{s}}^{\mathbf{y}^{\mathbf{s}}} = \text{“}[Domain_s] \text{ of a } [CLS_{y^s}]\text{”}$$

$$\mathcal{A}'_{y^s}(x^s) = \text{Softmax}\left[\frac{\mathcal{F}^q_v(x^s)\mathcal{F}^k_t(\mathcal{A}_{y^s})^T}{\sqrt{d}}\right]\mathcal{F}^v_t(\mathcal{A}_{y^s})$$

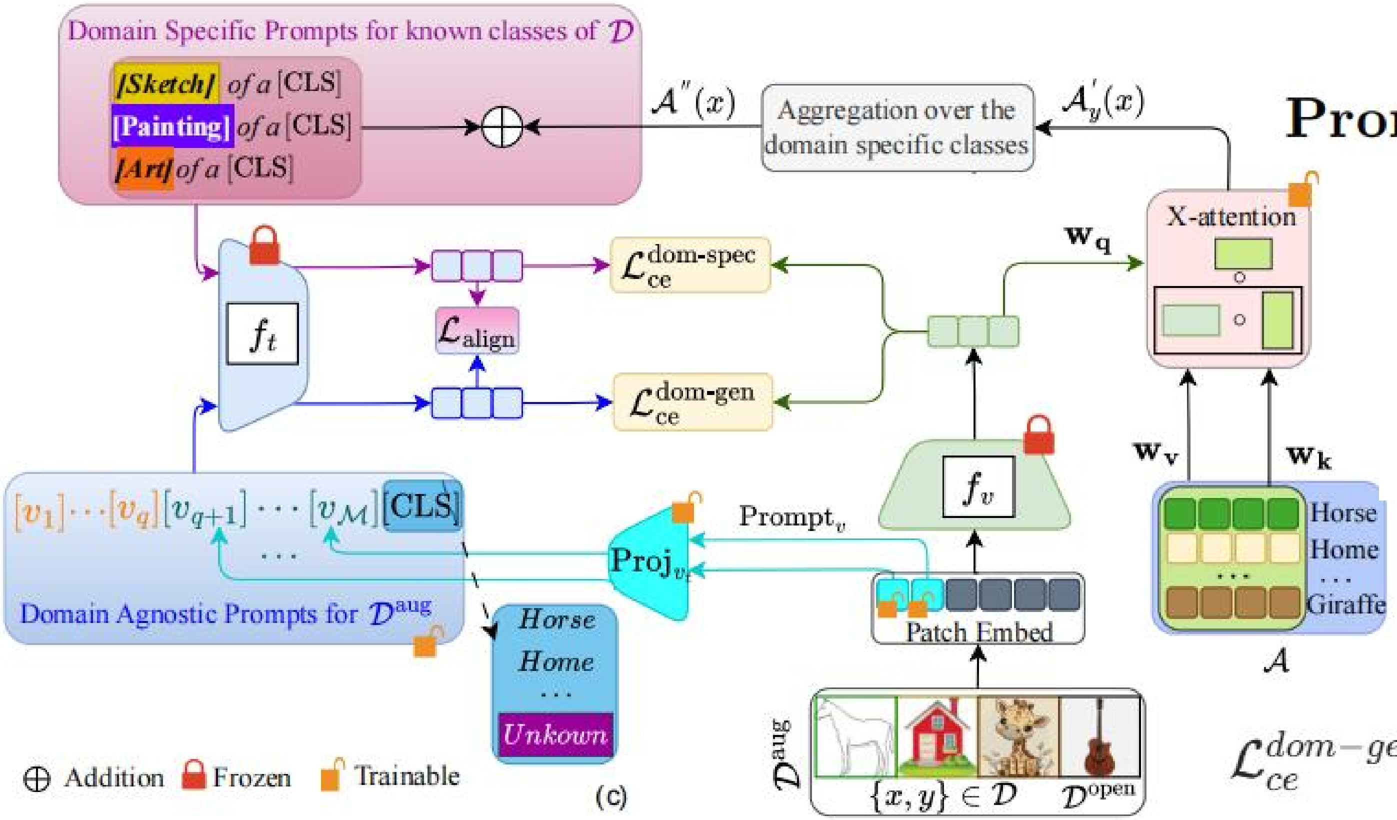
$$\mathcal{A}''(x^s) = \frac{1}{|\mathcal{Y}_s|}\sum_{y^{s'} \in \mathcal{Y}_s} \mathcal{A}'_{y^{s'}}(x^s)$$

$$\overline{\mathbf{Prompt}_{\mathbf{s}}^{\mathbf{y}^{\mathbf{s}}}}(x^s) = \mathbf{Prompt}_{\mathbf{s}}^{\mathbf{y}^{\mathbf{s}}} + \mathcal{A}''(x^s)$$

$$\mathcal{L}_{\text{ce}}^{\text{dom-spec}} = \min_{\substack{\mathbf{w^q}, \mathbf{w^k} \\ \mathbf{w^v}, \mathbf{Prompt}_v}} \sum_{s=1}^{N} \mathbb{E}_{(x^s, y^s) \sim \mathcal{P}(\mathcal{D}_s)} [-\log p(y^s | x^s)]$$

$$p(y^s|x^s) = \frac{\exp\left(\delta\left(\mathcal{F}_t\left(\overline{\mathbf{Prompt}_s^{y^s}}(x^s)\right), \mathcal{F}_v(x^s)\right)/\tau\right)}{\sum_{y^{s'} \in \mathcal{Y}_s} \exp\left(\delta\left(\mathcal{F}_t\left(\overline{\mathbf{Prompt}_s^{y^{s'}}}(x^s)\right), \mathcal{F}_v(x^s)\right)/\tau\right)}$$

prompt learning strategy



$$\mathcal{A}_{y^s} = [a^1_{y^s}, a^2_{y^s}, \ldots, a^{\mathcal{B}}_{y^s}]$$

"Generate $\mathcal{B}$ distinguishable attributes for the category [CLS] in an image."

Domain-generic prompts

$$\mathbf{Prompt}^{\mathbf{y}}_{\text{gen}} = [\nu_{1:q}][\text{Proj}_{vt}(\mathbf{Prompt_v})]_{q+1:\mathcal{M}}[\text{CLS}_y]$$

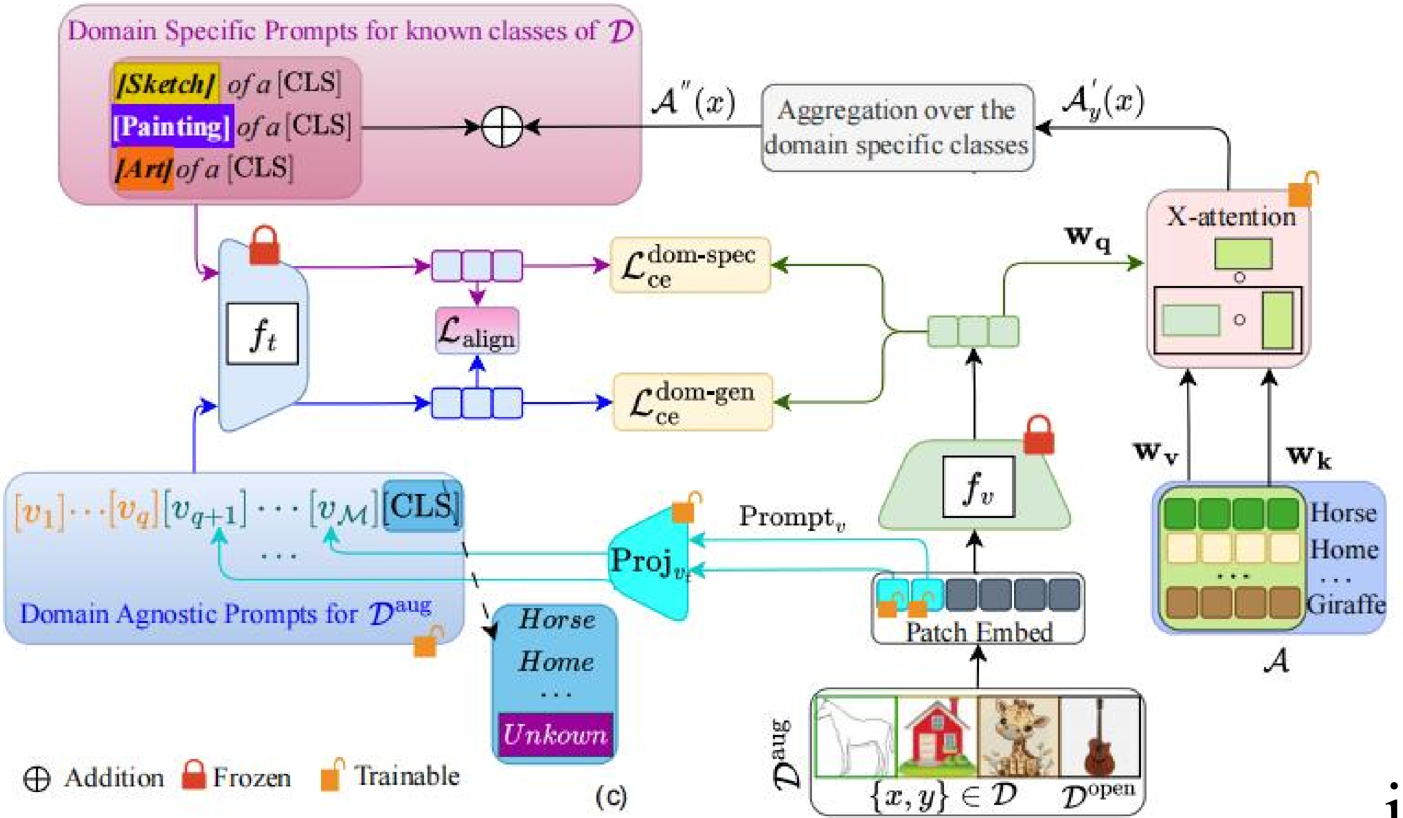
$$[c_1, \_, E_1] = \mathcal{F}^1_v([c_0, \mathbf{Prompt_v}, E_0])$$

$$[c_l, E_l] = \mathcal{F}^l_v([c_{l-1}, E_{l-1}]), \quad l = 2, 3, \ldots$$

$$\mathcal{L}^{dom-gen}_{ce} = \frac{1}{|D|+|D^{open}|}\Big[\sum_{x^{known}\in D}\mathcal{L}^{single}_{ce}(x^{known}, y^{known}) + \sum_{x^{open}\in D^{open}}\mathcal{L}^{single}_{ce}(x^{open}, Unknown)\Big]$$

$$\mathcal{L}_{\text{align}} = \min_{\substack{\{\nu_{1:q}\},\mathbf{w^q},\mathbf{w^k},\mathbf{w^v},\\ \text{Proj}_{vt},\mathbf{Prompt_v}}} \sum_{s=1}^{\mathcal{N}} \mathop{\mathbb{E}}_{(x^s,y^s)\in\mathcal{P}(\mathcal{D}_s)} \Big[1 - \text{cosine}\left(\mathcal{F}_t(\mathbf{Prompt}^{\mathbf{y^s}}_{\text{gen}}), \mathcal{F}_t(\overline{\mathbf{Prompt}^{\mathbf{y^s}}_{\mathbf{s}}}(x^s))\right)\Big]$$

prompt learning strategy



inference:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{ce}}^{\text{dom-gen}} + \mathcal{L}_{\text{ce}}^{\text{dom-spec}} + \mathcal{L}_{\text{align}}$$

$$\bar{y}^t = \underset{y^t \in \mathcal{C} \cup \text{Unknown}}{\text{argmax}} \; p(y^t | x^t, \mathcal{F}_v, \mathcal{F}_t, \mathbf{Prompt}_{\text{gen}}^{\mathbf{y}^t})$$

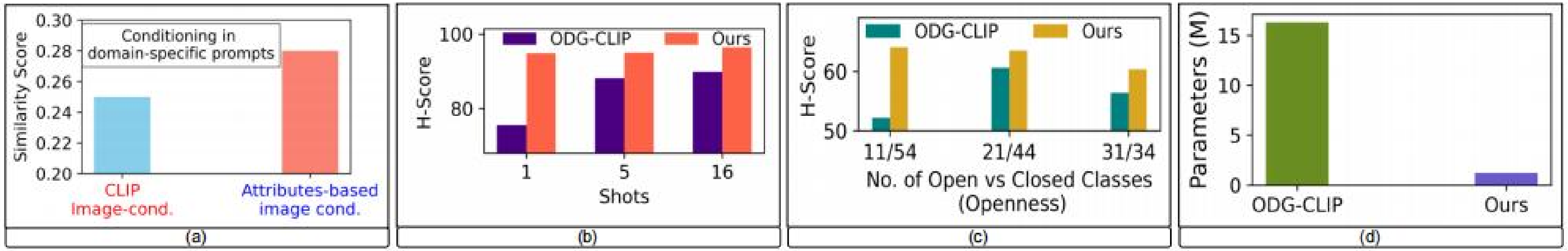
Experiments

南京航空航天大學
NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

Methods	CLIP-based	Venue	PACS		VLCS		OfficeHome		Multi-Dataset		Mini-DomainNet		Average	
			Acc	H-score	Acc	H-score	Acc	H-score	Acc	H-score	Acc	H-score	Acc	H-score
CLIP + OpenMax (OSR) [2]	✓	CVPR'16	20.24	31.97	20.59	31.83	20.00	32.64	11.74	20.87	16.92	28.05	17.90	29.07
CLIPN (OSR) [52]	✓	ICCV'23	64.03	55.79	25.34	19.49	44.18	32.83	39.84	36.28	47.63	40.91	44.20	37.06
MORGAN (FS-OSR) [30]	×	WACV'23	37.40	19.06	31.35	27.22	19.21	18.51	30.00	37.26	22.40	15.70	28.07	23.55
StyLIP (DG + OSR) [5]	✓	WACV'24	74.89	60.99	27.94	34.61	52.34	21.95	51.50	47.64	59.44	57.46	53.22	44.53
PromptSRC (DG + OSR) [20]	✓	ICCV'23	35.72	27.09	24.98	20.04	22.02	14.85	30.16	31.18	25.20	20.44	27.62	22.72
2LM (FSDG + OSR) [36]	×	CVPR'23	35.22	21.42	31.61	28.76	21.30	13.60	29.73	34.80	24.50	17.75	28.47	23.27
ODG-Net (ODG) [4]	×	TMLR'23	34.82	21.67	32.33	29.17	20.47	11.45	29.16	29.40	22.05	19.08	27.77	22.15
MEDIC (ODG) [53]	×	ICCV'23	33.91	21.40	32.94	26.28	21.31	11.75	30.35	33.11	23.73	19.05	28.45	22.32
SCI-PD (ODG) [6]	✓	CVPR'24	23.40	25.84	19.88	19.60	35.27	44.31	16.95	19.18	16.25	23.33	22.35	26.45
ODG-CLIP (ODG) [46]	✓	CVPR'24	68.89	75.56	52.43	54.70	48.69	52.93	63.74	69.53	61.05	65.50	58.96	63.64
OSLoPrompt (Ours)	✓	-	**92.71**	**94.86**	**78.89**	**76.89**	**69.73**	**64.04**	**76.30**	**74.49**	**69.00**	**67.57**	**77.32**	**75.57**

Methods	CLIP-based	Venue	PACS		VLCS		OfficeHome		Multi-Dataset		Mini-DomainNet		Average	
			Acc	H-score	Acc	H-score	Acc	H-score	Acc	H-score	Acc	H-score	Acc	H-score
CLIP + OpenMax (OSR) [2]	✓	CVPR'16	68.75	80.98	66.25	74.74	35.59	49.28	56.59	68.84	32.46	48.20	51.93	64.41
CLIPN (OSR) [52]	✓	ICCV'23	78.04	71.14	32.92	27.95	47.94	40.33	46.50	39.23	55.78	48.53	52.24	45.44
MORGAN (FS-OSR) [30]	×	WACV'23	46.27	24.06	42.16	38.70	36.20	18.63	35.47	42.80	37.81	27.06	39.58	30.25
StyLIP (DG + OSR)[5]	✓	WACV'24	80.10	70.01	45.78	48.93	61.87	42.46	54.58	49.76	64.03	60.68	61.27	54.37
PromptSRC (DG + OSR) [20]	✓	ICCV'23	46.86	30.23	36.16	32.36	31.10	20.35	35.68	38.12	36.37	31.32	37.24	30.28
2LM (FSDG + OSR) [36]	×	CVPR'23	46.70	24.06	41.67	37.36	29.38	18.95	35.04	35.38	38.43	28.70	38.24	28.89
ODG-Net (ODG) [4]	×	TMLR'23	46.66	25.92	43.05	37.71	34.52	15.96	34.20	36.93	39.95	23.72	39.68	28.05
MEDIC (ODG) [53]	×	ICCV'23	44.88	25.05	40.53	35.56	30.40	18.45	35.42	36.26	36.95	30.60	37.64	29.18
SCI-PD (ODG) [6]	✓	CVPR'24	35.16	34.53	30.11	30.48	32.98	42.50	32.20	28.89	21.25	30.57	30.34	33.39
ODG-CLIP (ODG) [46]	✓	CVPR'24	83.65	88.16	62.93	56.89	55.32	49.31	74.40	76.14	74.38	65.49	70.14	67.20
OSLoPrompt (Ours)	✓	-	**93.72**	**95.01**	**79.04**	**77.34**	**75.33**	**62.08**	**79.75**	**80.05**	**74.52**	**66.58**	**80.47**	**76.21**

Experiments



Methods	O.H.	M.DNet
Analysis of domain-specific prompts		
✓ Manual prompting: `Domain of a CLS`	59.33	65.88
✓ Manual prompting with image conditioning	62.96	67.27
✓ Manual prompting expanded with ad-hoc attributes from $\mathcal{A}$ [29]	60.69	63.82
✓ Manual prompting with ad-hoc attributes and image conditioning	62.16	65.11
✓ Visual attributes learning [21]	58.35	60.57
✓ **Proposed cross-attention approach**	**64.04**	**67.57**
Analysis of domain-agnostic prompts		
✓ Full context learning [62]	60.81	53.43
✓ Image-cond. context learning [61]	63.10	59.61
✓ **Proposed multi-modal prompting**	**64.04**	**67.57**
Sensitivity to the number of attributes per class in $\mathcal{A}$		
✓ **4**	**64.04**	**67.57**
✓ 8	63.97	66.36
✓ 12	63.79	65.14
Importance of the loss terms		
✓ $\mathcal{L}_{ce}^{\text{dom-gen}}$ (no domain-specific guidance)	62.51	63.86
✓ $\mathcal{L}_{ce}^{\text{dom-gen}} + \mathcal{L}_{ce}^{\text{dom-spec}}$ (partial domain-specific guidance)	62.52	65.56
✓ $\mathcal{L}_{ce}^{\text{dom-gen}} + \mathcal{L}_{ce}^{\text{dom-spec}} + \mathcal{L}_{\text{align}}$	**64.04**	**67.57**
Pseudo-open image synthesis		
✓ Generic sample generation of [46]	41.09	49.07
✓ Mixup-based [28] pseudo-open images	57.26	64.85
✓ **Our fine-grained sample generation**	**64.04**	**67.57**

Thanks