Efficient Multi-Source Prompt Adaptation for Cross-Domain Open-Vocabulary Learning

Xiaofan Que[1] **Dingrong Wang**[1] **Daniel Krutz**[1] **Xumin Liu**[1] **Qi Yu**[1,2*]
[1]Rochester Institute of Technology [2]Amazon Inc.
{xq5054, dw7445, dskvse, xmlics, qi.yu }@rit.edu

Abstract

Cross-domain open-vocabulary learning poses a unique and underexplored challenge, requiring models to generalize across both domain shifts and category shifts. To tackle this, we propose MAP: a parameter-efficient Multi-source domain-Adaptive Prompt tuning framework that leverages multiple labeled source domains to improve learning in novel, unlabeled target domains with unseen categories. MAP consists of two key components: Multi-Source Prompt Learning (MSPL) and Unsupervised Target Prompt Learning (UTPL). MSPL disentangles domain-invariant category semantics from domain-specific visual patterns by jointly learning shared and domain-aware prompts. UTPL enhances generalization in the unlabeled target domain by enforcing prediction consistency under text-guided style augmentations, introducing a novel entropy-minimization objective without relying on pseudo-labels. Together, these components enable effective alignment of visual and textual representations across both domains and categories. In addition, we present a theoretical analysis of the proposed prompts, examining their behavior through the lenses of fidelity and distinction. Extensive experiments on challenging CDOV benchmarks demonstrate that MAP achieves state-of-the-art performance with significantly fewer additional parameters.

1 Introduction

Cross-Domain Open-Vocabulary Learning (CDOV) aims to enable models to recognize and generalize to novel categories under distribution shifts, a setting that naturally arises in real-world applications. As a representative instantiation of this problem, object detection has witnessed significant progress with the emergence of large-scale vision-language models (VLMs), which enable open-vocabulary detection (OVD) by leveraging textual supervision to recognize a wide range of categories, including those unseen during training [17, 28, 13, 51, 34, 52, 57, 1, 8]. However, most existing OVD methods assume that training and test data share the same underlying domain, which is often unrealistic in practical scenarios where data distribution shifts are common. This motivates the task of Cross-domain Open-Vocabulary Learning (CDOV), where models must generalize to unseen categories in a novel domain.

Prompt learning has recently emerged as a powerful and flexible strategy for adapting VLMs to downstream tasks. Instead of fine-tuning the entire model, prompt learning optimizes a small set of learnable tokens that guide the pre-trained model to focus on task-relevant features. This parameter-efficient approach has shown strong performance in domain adaptation [46, 42, 41, 27, 14]. In the context of CDOV where models must handle both domain and category shifts, prompt learning offers a natural and scalable solution, as it enables the model to flexibly align textual and visual representations across diverse distributions. In this work, we introduce Multi-Source Prompt Learning (MSPL) for the first time in CDOV, leveraging multiple labeled source domains to jointly capture class-specific and domain-specific knowledge. By learning both domain-adaptive and class-aware prompts,

*Work not related to the position at Amazon.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

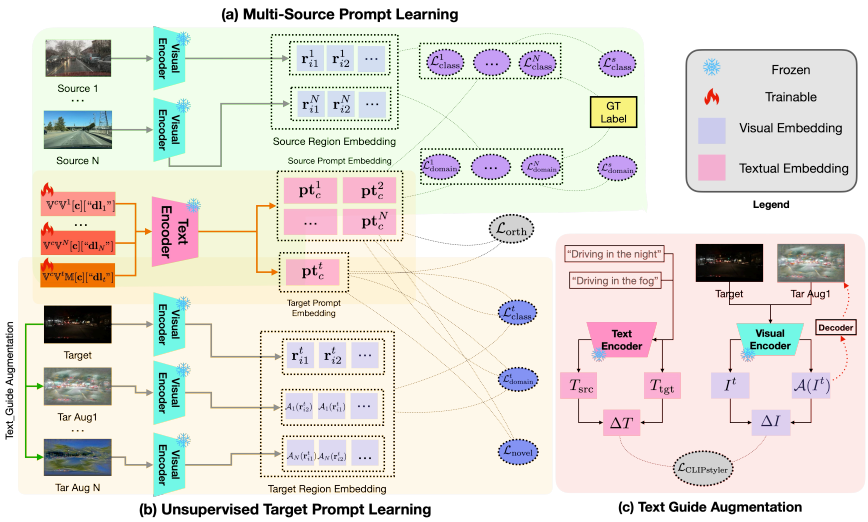


Figure 1: Framework overview: (a) Multi-Source Prompt Learning (MSPL) disentangles domain-invariant category semantics from domain-specific visual patterns by jointly learning shared and domain-aware prompts. (b) Unsupervised Target Prompt Learning (UTPL) enhances generalization in the unlabeled target domain by enforcing prediction consistency under text-guided style augmentations (c), introducing a novel entropy-minimization objective without relying on pseudo-labels. See appendix C for detailed loss illustration.

MSPL enables effective visual-textual alignment across domains and categories, and supports robust generalization to unseen target domains with novel classes.

For target adaptation, previous methods rely heavily on pseudo labels generated from source-trained models [7, 49, 50, 30, 45]. Such pseudo labels are often unreliable due to domain and category shifts, especially in open-vocabulary settings where unseen classes lack meaningful supervision. This limitation hinders robust generalization and leads to error propagation. To address these challenges, we propose Unsupervised Target Prompt Learning (UTPL), which eliminates the need for pseudo labels by enforcing prediction consistency under style-guided augmentations through a novel entropy-based objective. Together with MSPL, UTPL forms the core of our framework, Multi-source domain-Adaptive Prompt tuning (MAP), which enables robust adaptation to novel domains and categories in the CDOV setting.

Beyond the algorithmic design, we provide a theoretical analysis of the learnable prompts in MAP by introducing two key properties: fidelity and distinction, to characterize their effectiveness. We rigorously define these properties through an information-theoretic lens, connecting high-fidelity prompts to low predictive entropy and high mutual information with latent labels, and high-distinction prompts to mutual orthogonality across domains or classes. Furthermore, we propose a novel definition of unsupervised fidelity, enabling the evaluation of prompt quality even in the absence of labeled target data. This theoretical grounding helps to clarify why MAP is effective and provides a principled framework for future developments in prompt-based domain adaptation.

Our main contributions are as follows: (1) We propose MAP, a novel framework for multi-source cross-domain open-vocabulary learning that integrates MSPL and UTPL to align visual–textual representations across domains and categories. (2) We provide a theoretical analysis of prompt fidelity and distinction, introducing a new formulation of unsupervised fidelity based on entropy and mutual information. (3) We achieve state-of-the-art performance on standard CDOV benchmarks with minimal additional parameters.

2 Related Works

Open vocabulary object detection. Open-vocabulary detection was first proposed by Bansal et al. [3], introducing the visual-semantic embeddings to replace the classification layer of a closed vocabulary detector. This design has now become a common practice for many subsequent open-vocabulary object detection works [17, 28, 13, 51, 34, 52, 57, 1, 8]. To name a few, ViLD [17]

uses a teacher-student framework to distill knowledge from a pre-trained open-vocabulary image classification model (i.e., teacher model) and detect region boxes with the student model by aligning the text and image embeddings. PromptDet [13] scales up the OV detector with noisy uncurated web images using the proposed regional prompt learning technique which aligns text embedding with visual features. DetClip [51] designs a concept dictionary with a paralleled concept formulation to facilitate learning from heterogeneous datasets including detection, grounding, and image-text pairs. YOLO-World [8] is built upon the YOLO backbone and highly generalizable vision-language models; a vision-language path aggregation network is further utilized to enhance the interaction between visual and semantic information. In this work, we use several pre-trained open-vocabulary object detectors as backbones to evaluate their ability over cross domain adaptation.

Domain adaptive prompt learning. Prompting origins in NLP tasks which prepend instructions to the input to help train the language model [25, 39]. With the emergence of vision-language models such as CLIP [36], ALIGN [22], which embed a shared visual-semantic space by learning from a large scale of image alt-text pairs using a contrastive loss, domain-adaptive prompt learning has been used to address domain shifts between the pre-trained VLMs and the downstream tasks [14, 27**?** , 2, 53, 46]. One line of the work constructs domain adaptive prompts by directly embedding domain information [14, 27, 2]. The other line of work aims to align source and target domains with a prompt-based framework [**?** 6, 2]. To facilitate multi-source adaptive prompt learning (MSPL), MPA [7] first trains individual prompt for a source and target domain pair with a contrastive loss then utilizes auto-encoding to denoise the learned prompt and align all the reconstructed prompts. SAP-CLIP [49] designs learnable prompts with class context vectors and domain representation vectors based on CLIP to leverage textual information of both class semantics and domain representations. APNE-CLIP utilizes negative textual semantics and energy-based uncertainty to transfer knowledge from multiple source domains [50]. VAMP proposes a vision-aware multimodal prompt tuning framework to address multi-source domain few-shot adaptation [30]. However, these multi-source prompt learning methods heavily rely on pseudo labels generated by source-trained detectors, which are often noisy and unreliable—especially under significant domain and category shifts. In this work, we propose MAP (Multi-source domain-Adaptive Prompt tuning) that circumvents the reliance on pseudo labels by introducing a novel unsupervised target prompt learning (UTPL) strategy.

3 Methodology

CLIP. The CLIP model [36] consists of a visual encoder f and a text encoder g that map images and text into a shared embedding space. To enable zero-shot classification, class labels are converted into textual prompts, e.g., "a photo of a [CLASS]". After tokenization, the prompt $\mathbf{pt}$ is encoded into a text embedding $\mathbf{z} = g(\mathbf{pt})$. Given an image $\mathbf{x}$ with label y, the visual encoder produces an embedding $f(\mathbf{x})$. The probability of class k is then computed using cosine similarity between the visual and text embeddings:

$$p(\hat{y} = k|\mathbf{x}) = \frac{\exp(\texttt{cos}(f(\mathbf{x}), g(\mathbf{pt}_k))/T)}{\sum c = 1^K \exp(\texttt{cos}(f(\mathbf{x}), g(\mathbf{pt}_c))/T)}, \tag{1}$$

where K is the number of categories, $\texttt{cos}(\cdot, \cdot)$ denotes cosine similarity, and T is a temperature parameter.

Cross-domain open-vocabulary learning. We consider multiple labeled source datasets $\{\mathcal{D}^s = \{I_i^s, y_i^s\}_{i=1}^{N_s}\}_{s=1}^{N}$ from domains $\{\mathbb{D}^s\}_{s=1}^{N}$, and an unlabeled target dataset $\mathcal{D}^t = \{I_i^t\}_{i=1}^{N_t}$ from domain $\mathbb{D}^t$, where N_s and N_t denote the number of samples in source and target domains, respectively. For each image $I_i^s \in \mathcal{D}^s$, its annotation y_i^s corresponds to a class label in $\mathbb{C}^s$. Here, $\mathbb{C}^s$ denotes the label space of domain $\mathbb{D}^s$ containing K_s pre-defined categories. We define the base label space as the union of all source-domain label sets, i.e., $\mathbb{C}^{\text{base}} = \bigcup_{s=1}^{N} \mathbb{C}^s$. The target label space is denoted by $\mathbb{C}^t$, which includes both base and novel categories: $\mathbb{C}^t = \mathbb{C}^{\text{base}} \cup \mathbb{C}^{\text{novel}}$, with $\mathbb{C}^{\text{novel}} \cap \mathbb{C}^{\text{base}} = \varnothing$. Here, $\mathbb{C}^{\text{novel}}$ represents categories not seen in any source domain, however, the category names can be utilized for domain adaptation. For simplicity, we denote the total number of categories across domains by K.

3.1 Multi-Source Prompt Learning

We present Multi-Source Prompt Learning (MSPL) for MAP, which learns class- and domain-specific tokens using a standard domain-adaptive prompt structure that separates domain-invariant semantics

from domain-specific patterns for better cross-domain generalization. Prompts are optimized with a class-specific loss $\mathcal{L}^s_{\text{class}}$ to align with category semantics and a domain-specific loss $\mathcal{L}^s_{\text{domain}}$ to capture domain variations. As shown in Section 4, these objectives are theoretically grounded and provably contribute to effective prompt learning.

Source prompt design. Following the domain-adaptive prompt formulation, each source-specific prompt is defined as:

$$\mathbf{pt}^s_c = \mathbb{V}^c\mathbb{V}^s[\mathbf{c}][\mathbf{dl}_s], \tag{2}$$

where $\mathbb{V}^c = [\mathbf{v}^c_1][\mathbf{v}^c_2]...[\mathbf{v}^c_{M_1}]$ contains M_1 learnable tokens and $\mathbf{c} = e(\text{“CLASS-}c\text{”})$ is the tokenized class label. The domain-specific component $\mathbb{V}^s = [\mathbf{v}^s_1][\mathbf{v}^s_2]...[\mathbf{v}^s_{M_2}]$ is combined with a tokenized domain label $\mathbf{dl}_s = e(\text{“Domain-}s\text{”})$. While $\mathbb{V}^c$ and $[\mathbf{c}]$ capture category semantics shared across domains, $\mathbb{V}^s$ and $[\mathbf{dl}_s]$ encode domain-specific variations independent of categories. This structure allows each prompt to jointly model general category knowledge and domain-specific patterns, providing a strong foundation for cross-domain open-vocabulary learning.

Learning class-specific tokens. To learn class-specific domain invariant tokens $\mathbb{V}^c$ and $[\mathbf{c}]$, we propose to optimize the standard cross-entropy loss over each source domain for input $\mathbf{r}^s$ of category k:

$$\mathcal{L}^s_{\text{class}} = \mathbb{E}_{\mathcal{D}^s}[-\log p^s_{\text{class}}(\hat{y} = k|\mathbf{r}^s)], \tag{3}$$

where $\mathbf{r}^s$ is an sample in the s-th source domain. The classification probability p^s_{class} is obtained by aligning the visual features $f(\mathbf{r}^s)$ and corresponding *class* prompt $g(\mathbf{pt}^s_k)$:

$$p^s_{\text{class}}(\hat{y} = k|\mathbf{r}^s) = \frac{\exp(\texttt{cos}(f(\mathbf{r}^s), g(\mathbf{pt}^s_k))/T)}{\sum_{c=1}^{K}\exp(\texttt{cos}(f(\mathbf{r}^s), g(\mathbf{pt}^s_c))/T)}, \tag{4}$$

By minimizing eq. (3) across all N source domains, class-level semantic knowledge ($\mathbb{V}^c$ and $[\mathbf{c}]$) is distilled into each source prompt $\mathbf{pt}^s_c$, enabling accurate category-level discrimination within its domain and promoting domain-invariant understanding.

Learning domain-specific tokens. To learn domain-specific tokens $\mathbb{V}^s$ and $[\mathbf{dl}_s]$ for each source domain prompt, we minimize the following domain-specific loss function:

$$\mathcal{L}^s_{\text{domain}} = \mathbb{E}_{\mathcal{D}^s}[-\log p^s_{\text{domain}}(\hat{y} = k|\mathbf{r}^s)], \tag{5}$$

where the probability formulation p^s_{domain} defined in eq. (6) is a softmax over all class-domain prompt combinations, promoting the *domain* and *class* alignment with the visual feature:

$$p^s_{\text{domain}}(\hat{y} = k|\mathbf{r}^s) = \frac{\exp(\texttt{cos}(f(\mathbf{r}^s), g(\mathbf{pt}^s_k))/T)}{\sum_{d=1}^{N}\sum_{c=1}^{K}\exp(\texttt{cos}(f(\mathbf{r}^s), g(\mathbf{pt}^d_c))/T)}. \tag{6}$$

By minimizing eq. (5), we encourage the source image $\mathbf{r}^s$ to be close with domain-specific prompt $\mathbf{pt}^s_k$ corresponding to its true class label k in the vision language space, while simultaneously suppressing similarity with prompts from other domains ($d \neq s$). This contrastive design in eq. (6) enforces discriminative associations between domain-specific prompts and their corresponding visual features, facilitating the learning of fine-grained, domain-aware representations. A formal theoretical analysis is presented in Section 4, which rigorously justifies these objectives and establishes a principled, theoretically grounded framework for effective prompt learning.

3.2 Unsupervised Target Prompt Learning

We introduce Unsupervised Target Prompt Learning (UTPL) in MAP, which designs prompts for the target domain in CDOV using a novel class mask to handle unseen categories. UTPL learns prompts without supervision via three objectives guiding class-specific, domain-specific, and novel-mask tokens. As detailed in Section 4, these objectives are theoretically justified, providing a principled framework for effective target prompt learning.

Target domain prompt design To enable open-vocabulary learning in the target domain, we introduce a learnable novel mask $\mathbb{M} = [m_1][m_2], ...[m_{M_3}]$, which serves to specialize the prompt representation for novel classes. The target prompt is defined as:

$$\mathbf{pt}^t = \mathbb{V}^c\mathbb{V}^t\mathbb{M}[\mathbf{c}][\text{“}\mathbf{dl}_t\text{”}]. \tag{7}$$

Similar to the source prompt, $\mathbb{V}^c$ and $[\mathbf{c}]$ encodes class-specific context, while $\mathbb{V}^t$ and $[\text{“}\mathbf{dl}_t\text{”}]$ captures target-domain-specific context. The mask $\mathbb{M}$ is designed to enhance attention toward novel categories in the target domain by providing an explicit cue to the model, helping it distinguish between seen and unseen categories and thus improving open-vocabulary learning performance.

To learn the target prompt without supervision, instead of relying on potentially unreliable pseudo labels, we propose an unsupervised target prompt optimization objective that encourages consistency in the model's predictions across different augmentations of the same image.

Discussion of target image augmentation. Various image augmentation techniques—such as random cropping, translation, rotation, and Gaussian noise—are commonly used in vision tasks. However, for CDOV, it is crucial to preserve the semantic integrity of the image and minimize noise to maintain accurate predictions. In the multi-source domain adaptation setting, we aim to fully leverage pre-trained source models. To this end, we propose narrowing the domain gap by augmenting target images with source domain styles. Traditional methods like AdaIN [19] can transfer the style from a source image to a target image. More recently, text-guided augmentation techniques [11, 24, 43] have enabled style transfer based on textual descriptions, offering greater flexibility without requiring explicit source images—thus preserving source domain privacy. In our work, we adopt CLIPstyler [24], a text-guided augmentation method, for target image style transfer. Formally, given an image $I \in \mathcal{D}^t$, we generate a stylized image $\mathcal{A}_s(I)$ using CLIPstyler with a style text prompt T_s . Additional augmentation details and visual examples are provided in the Appendix.

After obtaining stylized target augmentations, the learning of target prompt is formulated in an unsupervised manner with the following three parts:

Learning class-specific tokens. To learn class-specific tokens for the unsupervised target domain, we propose minimizing the entropy of the averaged prediction probability distribution across different prompts and augmentations. The objective is formulated as:

$$\mathcal{L}^t_{\text{class}} = \mathbb{E}_{\mathcal{D}^t}[-\sum_{k=1}^{K} \tilde{p}_{\text{class}}(\hat{y}=k|\mathbf{r}^t)\log\tilde{p}(\hat{y}=k|\mathbf{r}^t)], \tag{8}$$

where $\tilde{p}_{\text{class}}(\hat{y}=k|\mathbf{r}^t) = \frac{1}{N+1}\sum_{s=0}^{N} p^t_{\text{class}}(\hat{y}=k|\mathcal{A}_s(\mathbf{r}^t))$ denotes the average prediction probability across N augmented views and the original view. Each prediction is computed by aligning (stylized) visual feature and its corresponding *class* prompt:

$$p^t_{\text{class}}(\hat{y}=k|\mathcal{A}_s(\mathbf{r}^t)) = \frac{\exp\left(\texttt{cos}(f(\mathcal{A}_s(\mathbf{r}^t)), g(\mathbf{pt}^s_k))/T\right)}{\sum_{c=1}^{K}\exp\left(\texttt{cos}(f(\mathcal{A}_s(\mathbf{r}^t)), g(\mathbf{pt}^s_c))/T\right)},$$

where $\mathcal{A}_s(I)$ is an augmented view of I with s source's style, and $\mathcal{A}_0(\mathbf{r}^t)$ denotes the original view. Minimizing $\mathcal{L}^t_{\text{class}}$ encourages the model to produce consistent and confident (i.e., low-entropy) predictions across different domain-style augmentations of the same target image, each paired with its corresponding source-specific prompts. By enforcing agreement across these views, the model is incentivized to extract features that are stable under stylistic shifts. This effectively forces the class-specific tokens to capture semantic information that is invariant across diverse domain styles.

Learning domain-specific tokens. To facilitate the learning of target domain-specific tokens, we minimize the following entropy loss over all prompt–image pairs:

$$\mathcal{L}^t_{\text{domain}} = \mathbb{E}_{\mathcal{D}^t}[-\sum_{k=1}^{K} \tilde{p}_{\text{domain}}(\hat{y}=k|\mathbf{r}^t)\log\tilde{p}_{\text{domain}}(\hat{y}=k|\mathbf{r}^t)], \tag{9}$$

where $\tilde{p}_{\text{domain}}(\hat{y}=k|\mathbf{r}^t) = \frac{1}{N+1}\sum_{s=0}^{N} p^t_{\text{domain}}(\hat{y}=k|\mathcal{A}_s(\mathbf{r}^t))$ and the prediction probability promotes the alignment of *domain* and *class* with the (stylized) visual feature:

$$p^t_{\text{domain}}(\hat{y}=k|\mathcal{A}_s(\mathbf{r}^t)) = \frac{\exp\left(\texttt{cos}(f(\mathcal{A}_s(\mathbf{r}^t)), g(\mathbf{pt}^s_k))/T\right)}{\sum_{d=0}^{N}\sum_{c=1}^{K}\exp\left(\texttt{cos}(f(\mathcal{A}_d(\mathbf{r}^t)), g(\mathbf{pt}^d_c))/T\right)}. \tag{10}$$

Note that the denominator sum over all source and target domain, where $d=0$ indicates the target domain. Similar to the formulation in eq. (6), the probability expression in eq. (10) pushes the model to select the correct class–domain pair under entropy minimization, while discouraging the alignment of mismatched *domain* and *class* prompts with the (stylized) visual image, thereby facilitating the unsupervised learning of the target domain-specific tokens.

Learning novel class mask. To learn the mask for novel classes, we propose encouraging the novel class prompts to diverge from the base class features by minimizing the orthogonality loss:

$$\mathcal{L}_{\text{novel}} = \sum_{c \in \mathbb{C}^{\text{base}}} \sum_{k \in \mathbb{C}^{\text{novel}}} (g(\mathbf{pt}_c)^\intercal g(\mathbf{pt}_k))^2 \tag{11}$$

This minimizes dot product between base and novel prompt embeddings, making them directionally independent. Similarly, for all source and target prompts, to promote their distinctions, we also impose the mutual orthogonality on them:

$$\mathcal{L}_{\text{orth}} = \sum_{i \in \mathbb{P}^{\text{source}}} \sum_{i \in \mathbb{P}^{\text{target}}} 1_{\text{diff}}(i,j)(g(\mathbf{pt}^i)^\intercal g(\mathbf{pt}^j))^2, \tag{12}$$

where we use $\mathbb{P}^{\text{source}}$ and $\mathbb{P}^{\text{target}}$ to denote the set of source and target prompts, respectively. $1_{\text{diff}}(i,j) = 1$ when i and j are not the same class label (to avoid penalizing matched semantics). Minimizing $\mathcal{L}_{\text{orth}}$ encourages each prompt to be dissimilar to the rest, effectively promoting mutual orthogonality. This enforces that prompts encode distinct semantic information, thereby reducing redundancy and enhancing domain diversity.

3.3 Training and Inference

During adaptation, the pre-trained visual backbones and the text encoder are kept frozen. In each iteration, only the learnable prompts are updated via backward gradients, which is performed by minimizing the overall learning objective $\mathcal{L}_{\text{total}}$, and the gradients can be back-propagated all the way through the text encoder $g(\cdot)$, making use of the rich knowledge encoded in the parameters to optimize the context. The the overall learning objective is defined as:

$$\mathcal{L}_{\text{total}} = (1-\eta)(\mathcal{L}^{\text{src}} + \lambda \mathcal{L}^{\text{tgt}}) + \eta \mathcal{L}_{\text{orth}}, \tag{13}$$

where $\mathcal{L}^{\text{src}} = \frac{1}{N}\sum_{s=1}^{N} \mathcal{L}^s_{\text{class}} + \mathcal{L}^s_{\text{domain}}$, and $\mathcal{L}^{\text{tgt}} = \mathcal{L}^t_{\text{class}} + \mathcal{L}^t_{\text{domain}} + \mathcal{L}_{\text{novel}}$. The source and target prompt loss $\mathcal{L}^{\text{src}}$ and $\mathcal{L}^{\text{tgt}}$ play equally important roles during cross-domain adaptation. The additional orthogonal regularization term $\mathcal{L}_{\text{orth}}$, balanced by a hyperparameter λ, brings mutual orthogonality among different domain prompts, enforcing distinction for each domain prompt. After training, the learned prompts are saved and used for inference.

4 Theoretical Analysis

In this section, we provide a theoretical analysis of the proposed prompts through the lens of their associated loss functions. Inspired by [45], we analyze the behavior of the prompts from two perspectives: fidelity and distinction. Fidelity describes a prompt's ability to faithfully encode either domain-specific and class-specific information by maintaining strong alignment with the label semantics of its corresponding domain or class. That is, a prompt with high fidelity captures the underlying semantics of the class it represents or preserving the characteristics of the domain it originates from. Distinction, on the other hand, requires minimal overlap between prompts from different domains or classes, ensuring that each prompt captures unique domain or class-specific characteristics without redundancy. Prompts with high distinction are mutually dissimilar, which facilitates better disentanglement of semantic and domain-specific information and reduces ambiguity during alignment.

To formalize these concepts, we first introduce propositions for both fidelity (including both supervised and unsupervised formulations for class and domain prompts) and distinction (including both inter-class and inter-domain variants). Based on these formulations, we analyze our proposed prompts from these two perspectives.

Definition 4.1 (**Fidelity**)**.** A prompt $\mathbf{pt}$ is said to have high *fidelity* if it retains task-relevant semantic information that aligns closely with the target label y. Formally, fidelity can be measured by the mutual information $\text{MI}(\mathbf{pt}, y)$ between the prompt and the label. A prompt with high fidelity satisfies a high $\text{MI}(\mathbf{pt}, y)$, indicating strong semantic alignment with class-specific and domain-specific cues.

Definition 4.2 (**Distinction**)**.** A prompt $\mathbf{pt}^i$ is said to exhibit *distinction* if it captures information that is discriminative for its associated class and domain, while remaining minimally redundant with prompts from other classes or domains. Formally, distinction can be quantified by the (negative) mutual information with other prompts $\mathbf{pt}^j$ ($j \neq i$): $-\text{MI}(\mathbf{pt}^i, \mathbf{pt}^j)$, where lower mutual information indicates higher distinction. Specifically, for class-specific prompts, this is referred to as *inter-class distinction*. For domain-specific prompts, this is referred to as *inter-domain distinction*.

Definition 4.3 (**Unsupervised Fidelity**)**.** A prompt $\mathbf{pt}$ is said to exhibit *unsupervised fidelity* if it induces confident and semantically meaningful predictions in the absence of ground-truth labels. Given an unlabeled input image $\mathbf{x} \in \mathcal{X}$, the predicted class distribution is denoted by $p(\hat{y}|\mathbf{pt}, \mathbf{x})$. Assuming a non-degenerate label distribution, unsupervised fidelity is measured by the negative Shannon entropy of the prediction distribution:

$$\mathrm{UnFI}(\mathbf{pt}) = -H(p(\hat{y}|\mathbf{pt}, \mathbf{x})), \tag{14}$$

where $H(p) = -\sum_k p_k \log p_k$. Lower entropy reflects higher confidence and alignment with semantic content, indicating greater fidelity.

Based on the definitions introduced in theorem 4.1 through theorem 4.3, we establish the following propositions to analyze the behavior of the proposed prompt tokens with respect to fidelity, unsupervised fidelity, and distinction, including both inter-class and inter-domain distinctions:

Proposition 4.4 (Source Class-Specific Prompt)**.** *Minimizing the source class-specific loss $\mathcal{L}^s_{\mathrm{class}}$ maximizes a lower bound of the mutual information between the source class-specific prompt and its corresponding class label, thereby promoting* high class fidelity. *Jointly minimizing the orthogonality loss $\mathcal{L}_{\mathrm{orth}}$ further reduces the mutual information among different class prompts, explicitly enforcing* high inter-class distinction.

Proposition 4.5 (Source Domain-Specific Prompt)**.** *Minimizing the source domain-specific loss $\mathcal{L}^s_{\mathrm{domain}}$ maximizes a lower bound of the mutual information between the source domain-specific prompt and its corresponding domain label, thereby promoting* high domain fidelity. *Jointly minimizing $\mathcal{L}_{\mathrm{orth}}$ reduces the mutual information among prompts from different domains, explicitly enforcing* high inter-domain distinction.

Proposition 4.6 (Target Class-Specific Prompt)**.** *Minimizing the target entropy objective $\mathcal{L}^t_{\mathrm{class}}$ maximizes a lower bound of the mutual information between the target image and its predicted class, thereby promoting* high unsupervised fidelity. *Since the entropy is minimized over the averaged predictions of multiple domain-style augmentations, the learned class prompt is encouraged to preserve semantic information that is invariant across domain styles. Jointly minimizing $\mathcal{L}_{\mathrm{orth}}$ further enforces* high inter-class distinction.

Proposition 4.7 (Target Domain-Specific Prompt)**.** *Minimizing the target domain entropy objective $\mathcal{L}^t_{\mathrm{domain}}$ maximizes a lower bound of the mutual information between the target image and its domain-aware prediction, thereby promoting* high unsupervised fidelity. *In addition, minimizing $\mathcal{L}_{\mathrm{orth}}$ reduces the mutual information among different domain prompts, explicitly enforcing* high inter-domain distinction.

Proposition 4.8 (Novel Mask Prompt)**.** *Minimizing the novel mask loss $\mathcal{L}_{\mathrm{novel}}$ reduces the correlation between novel-class prompts and base-class prompts. Consequently, the mutual information between base and novel prompts is reduced, encouraging* high distinction *between seen and unseen categories.*

Theoretical Insights. Our analysis provides a unified information-theoretic framework for understanding prompt learning in cross-domain open-vocabulary detection. Rather than analyzing each optimization objective independently, we show that all components of MAP optimize two fundamental principles: *fidelity* and *distinction*. Specifically, we establish that (1) both cross-entropy and entropy minimization maximize lower bounds of mutual information, providing a unified interpretation of supervised and unsupervised prompt learning under the common notion of fidelity; and (2) orthogonality regularization explicitly reduces the mutual information between prompts, thereby promoting semantic distinction. These results reveal that the source and target objectives, although optimized with different supervision signals, are theoretically aligned through the same information-theoretic principles, providing insights beyond empirical validation and offering a principled understanding of the proposed prompt learning framework.

Through the above propositions, we establish a theoretical foundation for designing prompts that simultaneously preserve semantic fidelity and promote prompt distinction, enabling robust generalization across both domains and categories. Proofs are deferred to Appendix E.

5 Empirical Evaluation

Datasets. We perform extensive experiments on datasets, including (1) the Art Image dataset with different artistic styles including Clipart1k, Comic2k, and Watercolor2k [20]; (2) Diverse

Weather Dataset (DWD) [47], and (3) Cityscapes [10], FoggyCityscapes [37], and KITTI [15]. Clipart1k contains 1000 clipart images across 20 classes, Watercolor2k and Comic2k contains 2000 watercolor/comic images across 6 classes. DWD covers five domains: Day Clear (DC), Night Clear (NC), Dusk Rainy (DR), Night Rainy (NR), and Day Foggy (DF). Each domain collects images from different weather and lighting conditions. During adaptation, we detect novel classes beyond the six classes, such as traffic light, and train, which do not appear in the source domains, as open-vocabulary evaluation. Cityscapes and FoggyCityscapes contains 7 classes, KITTI contains synthetic images of cars.

Evaluation metrics. For object detection task, Mean Average Precision (mAP) is used as the evaluation metric in all our experiments. We report mAP@0.5, which considers a prediction as a true positive if it matches the ground-truth label and has an intersection over union (IOU) score of more than 0.5 with the ground-truth bounding box. To evaluate performance across different class groups, we report AP_{novel} and AP_{base}, which correspond to the average precision computed over novel and base classes, respectively.

Table 1: Domain adaptation results (APs) on Clipart1k. For multi-source domain adaptation methods (highlighted with *), Watercolor2k and Comic2k are used jointly as source domains. For single-source domain adaptation methods, either Watercolor2k or Comic2k is used as the source domain, and only the best performance is reported. LB denotes Lower Bound, where the backbone detector is directly applied to the target domain without adaptation.

Backbone	RegionCLIP [54]			OVMR [33]			YOLO-World [8]			GDINO [32]		
Methods	AP	AP_{base}	AP_{novel}	AP	AP_{base}	AP_{novel}	AP	AP_{base}	AP_{novel}	AP	AP_{base}	AP_{novel}
LB	41.14	42.85	39.03	42.86	43.33	40.57	41.96	42.87	39.83	43.06	43.94	40.79
CoOP [56]	41.26	42.98	39.18	43.05	43.55	40.72	42.12	43.02	39.97	43.19	44.05	41.03
DAPL [14]	41.57	43.25	39.47	43.36	43.86	41.05	42.54	43.41	40.42	43.48	44.49	41.48
DAPro [27]	42.42	44.13	40.36	44.29	44.75	41.93	43.47	44.33	41.35	44.38	45.40	42.37
MPA* [7]	45.38	47.05	43.34	47.08	47.65	44.84	46.42	47.30	44.28	47.35	48.26	45.19
SAP* [49]	46.57	48.31	44.42	48.14	48.72	45.96	47.51	48.42	45.41	48.44	49.35	46.28
APNE* [50]	46.88	48.65	44.85	48.47	49.15	46.34	47.92	48.85	45.83	48.89	49.78	46.72
POND* [45]	49.54	51.36	47.45	51.12	51.89	49.02	50.58	51.49	48.49	51.50	52.47	49.45
MAP*(Ours)	**52.54**	**53.25**	**52.18**	**55.35**	**56.42**	**55.14**	**54.86**	**55.98**	**54.42**	**55.55**	**56.67**	**55.16**

Baselines. We conduct our experiment based on four state-of-the-art open-vocabulary object detectors: RegionCLIP [54], OVMR [33], YOLO-World [8] and GDINO [32]. Based on the object detectors, we implement different domain adaptive prompt learning methods including CoOP [55], DAPL [14], DAPro [27], and multi-source domain adaptive prompt learning methods including MPA [7], SAP-CLIP [49], APNE-CLIP [50] and POND [45] as well as lower bound (LB) which adapts pre-trained open-vocabulary object detector to the target domain with handcraft prompt " A photo of a [CLASS]".

5.1 Main Results

In Tab. 1, we provide the quantitative performance of our proposed framework MAP on four open-vocabulary object detectors [54, 33, 8, 32], and compared with the state-of-the-art domain adaptive prompt learning methods [56, 14, 27] and multi-source domain adaptive prompt learning methods [7, 50, 49, 45] on the art image. First, we observe that MAP significantly outperforms LB (i.e., the original object detector applied directly to the target domain without adaptation), as LB fails to address the substantial domain gap present in the target domain. Second, MSPL methods, including MAP, demonstrate a clear advantage over traditional DAPL methods that rely on a single source domain, highlighting the benefit of leveraging multiple source domains for more robust domain adaptation. Third, MAP exhibits a clear advantage over other MSPL methods, particularly on novel classes, achieving nearly a 5% improvement over the next-best MSPL approach, demonstrating the effectiveness of the novel mask prompt.

In table 2, we report the mAP results across the four domains in the DWD benchmark. Due to space limitations, APbase and APnovel are provided in the appendix. MAP consistently outperforms both DAPL and MSPL baselines across all domains and backbone detectors. The detailed results in the appendix further demonstrate that the detection of novel classes is significantly improved, highlighting the effectiveness of the proposed novel mask prompt.

Table 2: Domain adaptation results (mAP) on DWD. Following typical setting [47], *Day Clear* is used as source domain. The rest three are added as source domains for building a multi-source setting for each target domain for multi-source domain adaptation methods.

Backbone	RegionCLIP [54]				OVMR [33]				YOLO-World [8]				GDINO [32]			
Methods	NC	DR	NR	DF	NC	DR	NR	DF	NC	DR	NR	DF	NC	DR	NR	DF
LB	40.32	29.54	26.02	33.23	41.42	30.38	26.56	33.75	41.54	30.75	26.78	34.53	42.18	33.22	29.25	37.03
CoOP [56]	42.24	31.75	28.15	35.48	43.22	32.44	28.43	35.93	43.65	32.58	28.75	36.37	44.53	33.55	29.55	37.26
DAPL [14]	42.65	31.84	28.47	35.65	43.38	32.69	28.87	36.23	43.98	32.87	29.13	36.84	44.96	33.97	29.98	38.05
DAPro [27]	42.43	31.86	28.17	35.53	42.81	32.62	28.66	35.98	43.84	33.46	28.81	37.27	44.54	34.18	29.55	38.07
MPA* [7]	43.58	33.04	29.35	36.75	43.99	33.75	29.81	37.13	45.03	34.52	29.94	38.39	45.72	35.30	30.67	39.24
SAP* [49]	43.62	33.09	29.39	36.72	43.98	33.76	29.85	37.17	45.06	38.42	29.98	38.55	45.77	35.33	30.69	39.26
APNE* [50]	43.65	33.14	29.13	36.78	44.05	33.82	29.93	37.24	45.13	34.62	30.05	38.48	45.84	35.40	30.74	39.33
POND* [45]	43.63	33.12	29.10	36.76	44.03	33.80	29.90	37.21	45.12	34.60	30.01	38.45	45.80	35.37	30.71	39.29
MAP*(Ours)	**44.84**	**34.31**	**30.56**	**38.06**	**45.42**	**34.85**	**30.92**	**38.47**	**46.22**	**35.94**	**31.45**	**39.93**	**47.13**	**37.05**	**32.27**	**40.97**

5.2 Ablation Study

In this section, we present the ablative studies to verify the effectiveness of MAP, including different components of the prompt design (table 3), and different architectures. Additional ablatives about the impact of hyperparameters and other tasks are presented in the Appendix.

Effectiveness of each loss functions. In Table 3, we present an ablation study demonstrating that each loss component contributes to improved cross-domain open-vocabulary adaptation. Introducing a learnable class prompt via $\mathcal{L}^s_{\text{class}}$ improves performance by nearly 2% compared to fixed prompts, confirming the benefit of adapting category semantics. Incrementally adding more components yields steady improvements, with the combination of both class- and domain-level objectives for source and target domains leading to an overall gain of about 10%. While using only class or only domain prompts improves performance, the effect is less pronounced. Incorporating the novel class loss $\mathcal{L}_{\text{novel}}$ further boosts novel class detection by an additional $+1.8\%$, validating the effectiveness of explicitly modeling unseen categories. Finally, adding the orthogonality regularization $\mathcal{L}$orth contributes another $+1.5\%$, showing that encouraging disentanglement between class and domain prompts further enhances generalization.

Table 3: Ablative analysis each loss function on Clipart dataset.

$\mathcal{L}^s_{\text{class}}$	$\mathcal{L}^s_{\text{domain}}$	$\mathcal{L}^t_{\text{class}}$	$\mathcal{L}^t_{\text{domain}}$	$\mathcal{L}_{\text{novel}}$	$\mathcal{L}_{\text{orth}}$	AP	AP_{base}	AP_{novel}
✗	✗	✗	✗	✗	✗	43.45	43.94	40.79
✓	✗	✗	✗	✗	✗	45.42	45.76	40.94
✓	✓	✗	✗	✗	✗	45.55	45.88	41.57
✓	✓	✓	✗	✗	✗	46.12	46.51	42.49
✓	✓	✓	✓	✗	✗	53.24	53.87	52.25
✓	✗	✓	✗	✗	✗	44.18	44.48	39.69
✗	✓	✗	✓	✗	✗	45.08	45.25	40.32
✓	✓	✓	✓	✓	✗	54.07	53.60	54.06
✓	✓	✓	✓	✓	✓	55.55	56.67	55.16

The influence of λ. We conduct the experiment on Day Foggy dataset, where λ range from 0.25 to 1.50. We present the mAP with different λ values in the appendix (Figure 7), which indicates that too small λ such as 0.25 lead to a sub-optimal performance since the target data is not fully utilized. While setting $\lambda = 1.0$, the learning from both source and target is balanced and lead to the optimal performance. In fig. 2, we present some qualitative results of using some of the λ values. A small λ value such as 0.5 misclassified `bike` as `Motor`, and miss some smaller objects due to insufficient utilization of target information such as unlabeled novel class images. λ = 0.75 is able to detect `Bike` but still not able to detect smaller objects. Similarly, λ = 1.25 failed to detect smaller objects or separates multiple `Persons` since it puts too much focus on learning the target knowledge. While setting λ = 1.00, the learning reaches a balance and lead to optimal performance.

Computational Efficiency Analysis. Table 4 compares the inference speed and model size of our method with existing MSPL-based approaches. For a fair comparison, all methods are evaluated under identical inference settings. As shown, our method achieves the fastest inference speed of 74.1 FPS and the smallest parameter size of 37M, demonstrating superior computational efficiency and a reduced model footprint.

Table 4: Time and Memory Complexity

	MPA*	SAP*	APNE*	POND*	**MAP***(Ours)
Inference Time (FPS)	63.2	69.7	60.5	52.0	74.1
Parameter Size (M)	45	40	46	48	37



(a) λ=0.50 (b) λ=0.75 (c) λ=1.0 (d) λ=1.25

Figure 2: Detection results with different λ's. The second row shows different λ's influence to the detection of the novel class. (Traffic light in the illustrative figure.)

6 Conclusion

In this work, we tackle the unique challenges of cross-domain open-vocabulary learning, where domain shift and category shift are intricately entangled. To address these challenges, we propose MAP, a parameter-efficient Multi-source prompt learning framework. MAP unifies Multi-Source Prompt Learning (MSPL) and Unsupervised Target Prompt Learning (UTPL) to effectively leverage diverse knowledge from multiple source domains and adapt to the unlabeled target domain via learnable prompts. We further provide a theoretical analysis of the proposed prompts in terms of their fidelity and distinction properties. Extensive experiments on cross-domain benchmarks demonstrate the effectiveness of our approach. In future work, we would work on developing techniques for the interpretability and explainability of the learned prompts for better model understanding.

References

[1] Relja Arandjelović, Alex Andonian, Arthur Mensch, Olivier J Hénaff, Jean-Baptiste Alayrac, and Andrew Zisserman. Three ways to improve feature alignment for open vocabulary detection. *arXiv preprint arXiv:2303.13518*, 2023.

[2] Shuanghao Bai, Min Zhang, Wanqi Zhou, Siteng Huang, Zhirong Luan, Donglin Wang, and Badong Chen. Prompt-based distribution alignment for unsupervised domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 729–737, 2024.

[3] Ankan Bansal, Karan Sikka, Gaurav Sharma, Rama Chellappa, and Ajay Divakaran. Zero-shot object detection. In *Proceedings of the European conference on computer vision (ECCV)*, pages 384–400, 2018.

[4] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 79:151–175, 2010.

[5] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18392–18402, 2023.

[6] Qinglong Cao, Zhengqin Xu, Yuntian Chen, Chao Ma, and Xiaokang Yang. Domain-controlled prompt learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 936–944, 2024.

[7] Haoran Chen, Xintong Han, Zuxuan Wu, and Yu-Gang Jiang. Multi-prompt alignment for multi-source unsupervised domain adaptation. *Advances in Neural Information Processing Systems*, 36:74127–74139, 2023.

[8] Tianheng Cheng, Lin Song, Yixiao Ge, Wenyu Liu, Xinggang Wang, and Ying Shan. Yolo-world: Real-time open-vocabulary object detection. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2024.

[9] Junhyeong Cho, Gilhyun Nam, Sungyeon Kim, Hunmin Yang, and Suha Kwak. Promptstyler: Prompt-driven style generation for source-free domain generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15702–15712, 2023.

[10] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016.

[11] Mohammad Fahes, Tuan-Hung Vu, Andrei Bursuc, Patrick Pérez, and Raoul de Charette. Poda: Prompt-driven zero-shot domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18623–18633, 2023.

[12] Zhen Fang, Jie Lu, Feng Liu, Junyu Xuan, and Guangquan Zhang. Open set domain adaptation: Theoretical bound and algorithm. *IEEE transactions on neural networks and learning systems*, 32(10):4309–4322, 2020.

[13] Chengjian Feng, Yujie Zhong, Zequn Jie, Xiangxiang Chu, Haibing Ren, Xiaolin Wei, Weidi Xie, and Lin Ma. Promptdet: Towards open-vocabulary detection using uncurated images. In *European conference on computer vision*, pages 701–717. Springer, 2022.

[14] Chunjiang Ge, Rui Huang, Mixue Xie, Zihang Lai, Shiji Song, Shuang Li, and Gao Huang. Domain adaptation via prompt learning. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[15] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *International Journal of Robotics Research (IJRR)*, 2013.

[16] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013.

[17] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921*, 2021.

[18] Weizhen He, Weijie Chen, Binbin Chen, Shicai Yang, Di Xie, Luojun Lin, Donglian Qi, and Yueting Zhuang. Unsupervised prompt tuning for text-driven object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2651–2661, 2023.

[19] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, pages 1501–1510, 2017.

[20] Naoto Inoue, Ryosuke Furuta, Toshihiko Yamasaki, and Kiyoharu Aizawa. Cross-domain weakly-supervised object detection through progressive domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5001–5009, 2018.

[21] Khawar Islam, Muhammad Zaigham Zaheer, Arif Mahmood, and Karthik Nandakumar. Diffusemix: Label-preserving data augmentation with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27621–27630, 2024.

[22] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021.

[23] Beomsu Kim, Gihyun Kwon, Kwanyoung Kim, and Jong Chul Ye. Unpaired image-to-image translation via neural schrödinger bridge. In *International Conference on Learning Representations*, volume 2024, pages 19312–19331, 2024.

[24] Gihyun Kwon and Jong Chul Ye. Clipstyler: Image style transfer with a single text condition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18062–18071, 2022.

[25] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*, 2021.

[26] Aristotelis Leventidis, Laura Di Rocco, Wolfgang Gatterbauer, Renée J Miller, and Mirek Riedewald. Domainnet: Homograph detection for data lake disambiguation. *arXiv preprint arXiv:2103.09940*, 2021.

[27] Jiaming Li, Jiacheng Zhang, Jichang Li, Ge Li, Si Liu, Liang Lin, and Guanbin Li. Learning background prompts to discover implicit knowledge for open vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16678–16687, 2024.

[28] Chuang Lin, Peize Sun, Yi Jiang, Ping Luo, Lizhen Qu, Gholamreza Haffari, Zehuan Yuan, and Jianfei Cai. Learning object-language alignments for open-vocabulary object detection. *arXiv preprint arXiv:2211.14843*, 2022.

[29] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

[30] Kuanghong Liu, Jin Wang, Kangjian He, Dan Xu, and Xuejie Zhang. Vision-aware multimodal prompt tuning for uploadable multi-source few-shot domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 18897–18905, 2025.

[31] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. *Advances in neural information processing systems*, 30, 2017.

[32] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, pages 38–55. Springer, 2024.

[33] Zehong Ma, Shiliang Zhang, Longhui Wei, and Qi Tian. Ovmr: Open-vocabulary recognition with multi-modal references. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16571–16581, 2024.

[34] M Minderer, A Gritsenko, A Stone, M Neumann, D Weissenborn, A Dosovitskiy, A Mahendran, A Arnab, M Dehghani, Z Shen, et al. Simple open-vocabulary object detection with vision transformers. arxiv 2022. In *European conference on computer vision*. Springer, 2022.

[35] Joshua Niemeijer, Manuel Schwonberg, Jan-Aike Termöhlen, Nico M Schmidt, and Tim Fingscheidt. Generalization by adaptation: Diffusion-based domain extension for domain-generalized semantic segmentation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2830–2840, 2024.

[36] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[37] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Semantic foggy scene understanding with synthetic data. *International Journal of Computer Vision*, 126:973–992, 2018.

[38] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Acdc: The adverse conditions dataset with correspondences for semantic driving scene understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10765–10775, 2021.

[39] Taylor Shin, Yasaman Razeghi, Robert L Logan IV, Eric Wallace, and Sameer Singh. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. *arXiv preprint arXiv:2010.15980*, 2020.

[40] Manli Shu, Weili Nie, De-An Huang, Zhiding Yu, Tom Goldstein, Anima Anandkumar, and Chaowei Xiao. Test-time prompt tuning for zero-shot generalization in vision-language models. *Advances in Neural Information Processing Systems*, 35:14274–14289, 2022.

[41] Mainak Singha, Ankit Jha, Shirsha Bose, Ashwin Nair, Moloud Abdar, and Biplab Banerjee. Unknown prompt, the only lacuna: Unveiling clip's potential for open domain generalization. *arXiv preprint arXiv:2404.00710*, 2024.

[42] Mainak Singha, Harsh Pal, Ankit Jha, and Biplab Banerjee. Ad-clip: Adapting domains in prompt space using clip. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4355–4364, 2023.

[43] Ananda Padhmanabhan Suresh, Sanjana Jain, Pavit Noinongyao, Ankush Ganguly, Ukrit Watchareeruetai, and Aubin Samacoits. Fastclipstyler: Optimisation-free text-based image style transfer using style representations. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 7316–7325, 2024.

[44] Tung-Long Vuong, Hoang Phan, Vy Vo, Anh Bui, Thanh-Toan Do, Trung Le, and Dinh Phung. Preserving clusters in prompt learning for unsupervised domain adaptation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19974–19984, 2025.

[45] Junxiang Wang, Guangji Bai, Wei Cheng, Zhengzhang Chen, Liang Zhao, and Haifeng Chen. Pond: Multi-source time series domain adaptation with information-aware prompt tuning. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 3140–3151, 2024.

[46] Liyuan Wang, Yan Jin, Zhen Chen, Jinlin Wu, Mengke Li, Yang Lu, and Hanzi Wang. Transitive vision-language prompt learning for domain generalization. *arXiv preprint arXiv:2404.18758*, 2024.

[47] Aming Wu and Cheng Deng. Single-domain generalized object detection in urban scene via cyclic-disentangled self-distillation. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 847–856, 2022.

[48] Serin Yang, Hyunmin Hwang, and Jong Chul Ye. Zero-shot contrastive loss for text-guided diffusion image style transfer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22873–22882, 2023.

[49] Yuxiang Yang, Yun Hou, Lu Wen, Pinxian Zeng, and Yan Wang. Semantic-aware adaptive prompt learning for universal multi-source domain adaptation. *IEEE Signal Processing Letters*, 2024.

[50] Yuxiang Yang, Lu Wen, Yuanyuan Xu, Jiliu Zhou, and Yan Wang. Adaptive prompt learning with negative textual semantics and uncertainty modeling for universal multi-source domain adaptation. In *2024 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2024.

[51] Lewei Yao, Jianhua Han, Youpeng Wen, Xiaodan Liang, Dan Xu, Wei Zhang, Zhenguo Li, Chunjing Xu, and Hang Xu. Detclip: Dictionary-enriched visual-concept paralleled pre-training for open-world detection. *Advances in Neural Information Processing Systems*, 35:9125–9138, 2022.

[52] Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. Open-vocabulary object detection using captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14393–14402, 2021.

[53] Cairong Zhao, Yubin Wang, Xinyang Jiang, Yifei Shen, Kaitao Song, Dongsheng Li, and Duoqian Miao. Learning domain invariant prompt for vision-language models. *IEEE Transactions on Image Processing*, 2024.

[54] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. Regionclip: Region-based language-image pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16793–16803, 2022.

[55] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[56] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision (IJCV)*, 2022.

[57] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In *European Conference on Computer Vision*, pages 350–368. Springer, 2022.

[58] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.

Appendix

Appendix Contents

A Training Algorithm and Notations **16**

B Implementation Details **16**
Model Architecture. . . . 16
Implementation details. . . . 16
Text-guide Synthesis. . . . 16

C Detailed Illustration of Loss Functions **17**
C.1 Illustration of MSPL Learning Objectives 17
C.2 Illustration of UTPL Learning Objectives 19

D Additional Experimental Results **19**
D.1 AP_{base} and AP_{novel} results of the DWD benchmark 19
D.2 Ablation Study 19
Target model initialization. . . . 19
Heterogeneous source model architecture. . . . 20
Impact of different augmentation methods. . . . 20
Integrating LLMs in OVOD. . . . 21
D.3 Impact of Hyperparameters 21
The influence of M_1, M_2, M_3 21
The influence of λ 21
The influence of η. . . . 21
D.4 Additional Ablations 22
Impact of different components. . . . 22
Comparison with unsupervised prompt learning methods. . . . 22
Performance of MAP on classification or segmentation tasks. . . . 23
Additional experiments for single-source methods with combined source domains. . . . 23
Diffusion-based augmentation. . . . 23
D.5 Visualization of Prompt Embeddings 23
D.6 Visualization of Class Embeddings 24
D.7 Additional datasets 24
D.8 Additional Metric 25

E Theoretical Justification **25**
E.1 Proof of theorem 4.3. . . . 27
E.2 Proof of Proposition 4.4 28
E.3 Proof of Proposition 4.5 28
E.4 Proof of Proposition 4.6 28
E.5 Proof of Proposition 4.7 29
E.6 Proof of Proposition 4.8 29
E.7 Generalization Guarantee 30

F Potential Societal Impact and Limitations **31**

G Source Code **32**

A Training Algorithm and Notations

The notations used in the main paper are shown in table 5. And the learning process of MAP is presented in algorithm 1.

Table 5: Summary of Major Notations

Notation	Description
s	s-th source domain
t	Target domain
N	Number of source domains
K_s	Length of label space in domain d
K_t	Length of label space in target domain t
$\mathbb{D}^s$	s-th source domain
$\mathbb{D}^t$	target domain t
$\mathcal{D}^s$	Dataset of s-th source domain $\mathbb{D}^s$
$\mathcal{D}^t$	Dataset of target domain $\mathbb{D}^t$
I_i^s	The i-th image of s-th source domain
y_i^s	The i-th label of s-th source domain
b_i^s	The bounding box of the i-th image of s-th source domain
c_i^s	The class label of the i-th image of s-th source domain
I_i^t	The i-th image of the target domain t
$\mathbb{C}^s$	Label space of s-th domain, $\lvert\mathbb{C}^s\rvert = K_s$
$\mathbb{C}^t$	Label space of target domain t
$\mathbb{C}^{\text{base}}$	Base classes
$\mathbb{C}^{\text{novel}}$	Novel classes
$\mathbb{V}^c$	The class-specific domain-invariant tokens for class c
$\mathbb{V}^d$	The domain-specific class-agnostic tokens for domain s_d
$[\text{“}\texttt{domain}_s\text{”}]$	Handcrafted textual domain descriptions related to domain s
M_1	Length of the learnable tokens of $\mathbb{V}^c$
M_2	Length of the learnable tokens of $\mathbb{V}^d$
M_3	Length of the learnable tokens of $\mathbb{M}$
$\mathbf{pt}_c^s$	The prompt for class c in source domain s
$\mathbf{pt}_c^t$	The prompt for class c in target domain t
$f(\cdot)$	Visual encoder
$g(\cdot)$	Text encoder
$\texttt{cos}(\cdot,\cdot)$	cosine similarity

B Implementation Details

Model Architecture. In this work, we tested four open-vocabulary object detectors. Each source model is initialized with a pre-trained open-vocabulary object detector. They are then fine-tuned using their own labeled source datasets, and adapted to the target domain via MAP. During pre-training, the text encoders are kept frozen, while the visual encoders are fine-tuned using the labeled source data. During adaptation, each fine-tuned visual encoder is used to process source images from its corresponding domain. The target visual encoder is initialized as a vanilla visual encoder.

Implementation details. We adopt the open-vocabulary object detector backbone to encode the input images and their corresponding text encoder to encode the input text prompts. The length of learnable tokens M_1, M_2 and M_3 are fixed as 8. The hyperparameter λ is set to 1. We use SGD optimizer with learning rate equals to 0.002. Following [14], we randomly initialize each prompt with a zero-mean Gaussian distribution with a standard deviation of 0.02. The training of the prompts and the inference are conducted on one NVIDIA V100 GPU. The augmented images are generated with the pre-trained CLIPstyler model [24].

Text-guide Synthesis. Text-guide Synthesis aim to generated images with different styles given an input base image and a text description. CLIPstyler [24] design a modulation of the style of content images only with a single text condition using the pre-trained text-image embedding

Algorithm 1 Training Procedure of MAP

INPUT Source datasets $\{\mathcal{D}^s\}_{s=1}^N$, target dataset $\mathcal{D}^t$, domain style text prompt $\{T_s\}_{d=1}^N$, and target domain style text prompt T_t
Initialize source and target prompts as "A photo of [CLS]"
Obtain source prompts and target prompt embedding using text encoder $g(\cdot)$: $\{\{\mathbf{pt}_c^s\}_{s=1}^N\}_{c=1}^K$, and $\{\mathbf{pt}_c^t\}_{c=1}^K$
while not converged **do**
  for $d \in \{1, ..., N\}$ **do**
    Sample a source batch $\mathcal{B}_s \sim \mathcal{D}^s$
    Compute $\mathcal{L}_{\text{domain}}^s$ and $\mathcal{L}_{\text{class}}^s$ using eq. (5) and eq. (3), respectively
  end for
  Sample a target batch $\mathcal{B}_t \sim \mathcal{D}^t$
  for each image I_t in $\mathcal{B}_t$ **do**
    for each style description in $\{T_s\}_{d=1}^N$ **do**
      Augment I_t with T_s and T_t to obtain $\mathcal{A}_s(I_t)$
    end for
  end for
  Compute $\mathcal{L}_{\text{domain}}^t$ and $\mathcal{L}_{\text{class}}^t$ using eq. (9) and eq. (8), respectively
  Compute $\mathcal{L}_{\text{novel}}$ using eq. (11)
  Compute $\mathcal{L}_{\text{orth}}$ using eq. (12)
  Compute overall loss and update $\mathbf{pt}^s$ and $\mathbf{pt}^t$
end while

model of CLIP, and propose a patch-wise text-image matching loss with multiview augmentations for realistic texture transfer. PODA [11] propose Prompt-driven Instance Normalization (PIN) to learn style statistics based on CLIP feature alignment which are later used for image augmentation. PromptStyler [9] simulates various distribution shifts in the joint space by synthesizing diverse styles via prompts without using any images to deal with source-free domain generalization. In this work, we use CLIPstyler [24] to generate augmentations of unlabeled target images to learn target prompts. Formally, CLIPStyler aims to transfer the semantic style of target text t_{sty} to the content image I_c through the pre-trained text-image embedding model CLIP without a style image I_s to as a reference. Given the semantic text style of the style target and the input content t_{sty} and t_{src}, CLIPStyler transforms the content image I_c to stylized image I_{cs}. In our setting, we treat the unlabeled target image as the content image I_c, the target domain semantic description as t_{src}, and the source domain semantic description as t_{sty}. With multiple source domains available, we augment the target image with multiple styles. The corresponding domain descriptions are shown in Tab. 6. The augmentation results are shown in Fig. 3.

Table 6: Domain descriptions of DWD

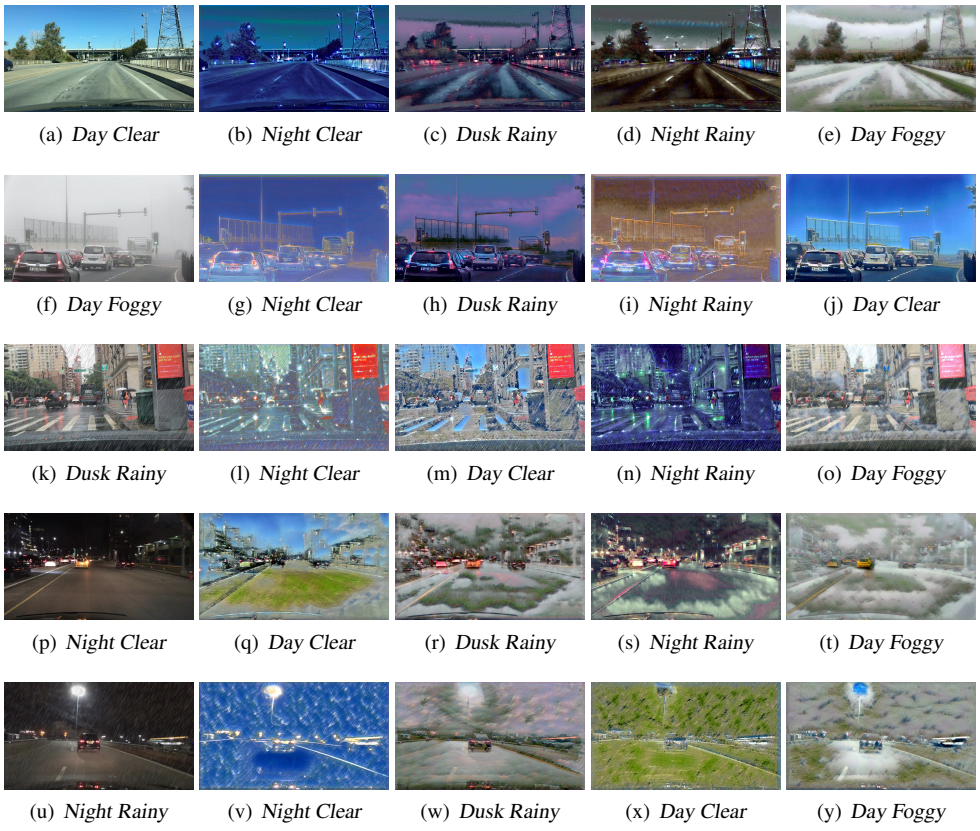
Domain Name	Domain Description
Day Clear	A driving in a clear day photo
Day Foggy	A driving in a foggy day photo
Night rainy	A driving in a rainy night photo
Night clear	A driving in a clear night photo
Dusk rainy	A driving in a rainy dusk photo

C Detailed Illustration of Loss Functions

In this section, we provide a detailed illustration of the learning objectives used in MAP through accompanying figures.

C.1 Illustration of MSPL Learning Objectives

In fig. 4, we present figurative illustration for $\mathcal{L}_{\text{class}}^s$ and $\mathcal{L}_{\text{domain}}^s$. To differentiate $\mathcal{L}_{\text{class}}^s$ and $\mathcal{L}_{\text{domain}}^s$, we use “Src Prompt 1 Emb" and “Src Prompt N Emb" to specify different source prompt embeddings outputed by the text encoder. As shown in fig. 4, both $\mathcal{L}_{\text{class}}^s$ and $\mathcal{L}_{\text{domain}}^s$Both losses take a visual embedding (omitted in the figure for clarity, please refer to fig. 1) and the corresponding source



(a) *Day Clear* (b) *Night Clear* (c) *Dusk Rainy* (d) *Night Rainy* (e) *Day Foggy*

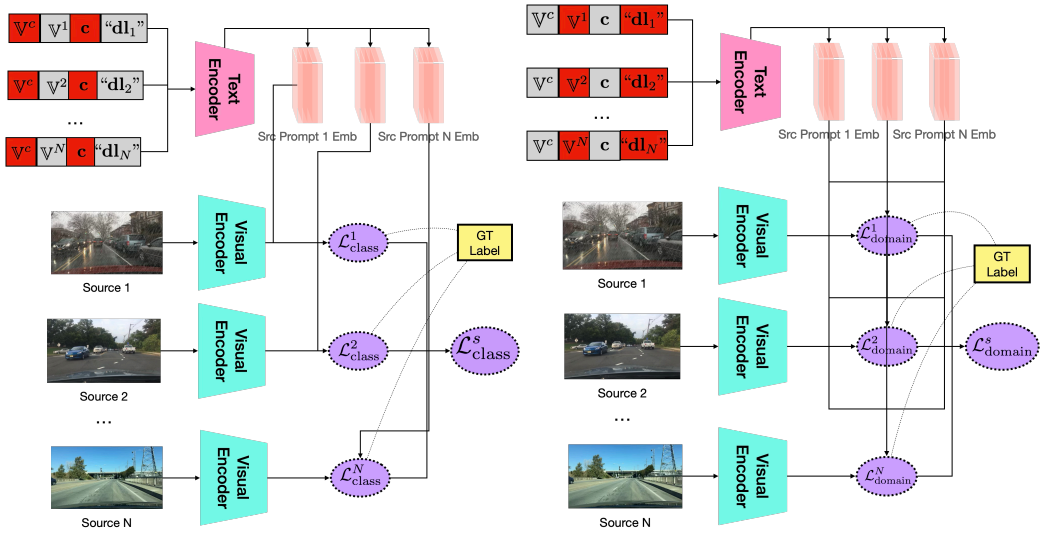
(f) *Day Foggy* (g) *Night Clear* (h) *Dusk Rainy* (i) *Night Rainy* (j) *Day Clear*

(k) *Dusk Rainy* (l) *Night Clear* (m) *Day Clear* (n) *Night Rainy* (o) *Day Foggy*

(p) *Night Clear* (q) *Day Clear* (r) *Dusk Rainy* (s) *Night Rainy* (t) *Day Foggy*

(u) *Night Rainy* (v) *Night Clear* (w) *Dusk Rainy* (x) *Day Clear* (y) *Day Foggy*

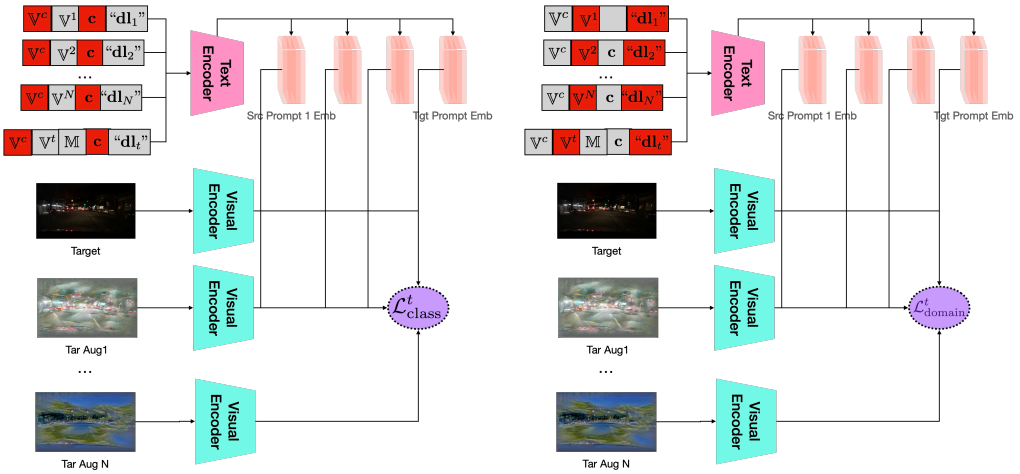
Figure 3: Stylized images of different styles with image sampled from different domains of DWD dataset wit descriptions shown in Tab. 6.



(a) Learning class-specific tokens via $\mathcal{L}^s_{\text{class}}$. (b) Learning domain-specific tokens via $\mathcal{L}^s_{\text{domain}}$.

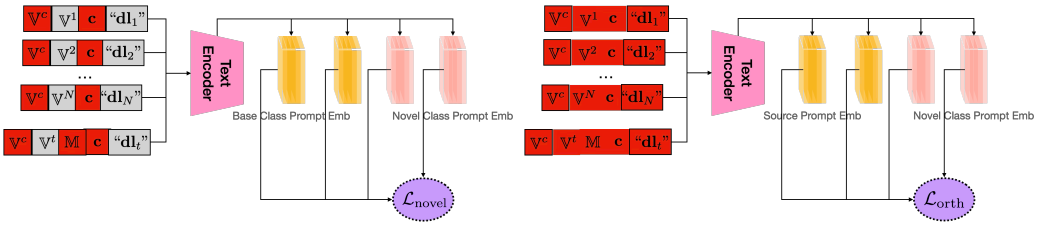
Figure 4: Illustration of source prompt learning objectives: $\mathcal{L}^s_{\text{class}}$ and $\mathcal{L}^s_{\text{domain}}$.

prompt as input. The key difference is that $\mathcal{L}^s_{\text{domain}}$ incorporates prompt embeddings from other source domains in addition to its own, promoting the learning of domain-aware representations.



(a) Class-specific tokens (b) Domain-specific tokens

Figure 5: Illustration of learning target domain prompts.



(a) Novel mask (b) Prompt Orthogonality

Figure 6: Illustration of class and domain orthogonality loss.

C.2 Illustration of UTPL Learning Objectives

In fig. 5, we illustrate the loss functions of learning target class-specific and domain specific tokens. Both are learned in an unsupervised manner with augmented views and source prompts involved. In fig. 6, we demonstrate the loss functions of the novel class mask and the prompt orthogonality.

D Additional Experimental Results

In this section, we present additional AP_{base} and AP_{novel} results of the DWD benchmark in table 7 and table 8. We present ablative for hyper-parameters including λ, M_1, M_2 and M_3 in appendix D.3. The visualization of prompt embeddings are shown in appendix D.5. Additional experiments on datasets including Cityscapes, FoggyCityscapes, and KITTI are presented in appendix D.7.

D.1 AP_{base} and AP_{novel} results of the DWD benchmark

In table 7 and table 8, we present the AP_{base} and AP_{novel}. We observe that MAP has ranked the top across different object detector backbones and different domains. Similarly, MSPL methods performs better than SSPL methods due to the utilization of multiple source models. But our model outperform other MSPL methods since other methods only focus on addressing domain shift while neglecting the category shift.

D.2 Ablation Study

Target model initialization. For simplicity, our previous experiments initialized the target model using a vanilla visual encoder. In this ablation study, we investigate the effect of initializing the target

Table 7: Domain adaptation results (AP_{base}) on DWD with the same setting as table 2.

Backbone	RegionCLIP [54]				OVMR [33]				YOLO-World [8]				GDINO [32]			
Methods	NC	DR	NR	DF	NC	DR	NR	DF	NC	DR	NR	DF	NC	DR	NR	DF
LB	44.56	33.25	29.45	36.79	44.98	34.56	30.11	36.98	45.15	34.68	30.09	37.88	45.92	37.05	32.82	40.82
CoOP [56]	45.78	34.66	31.59	38.67	46.54	35.92	31.74	38.55	46.83	35.62	31.58	39.75	46.84	37.12	33.04	40.95
DAPL [14]	45.57	34.94	31.82	37.74	47.05	37.01	31.95	39.52	47.05	36.02	33.05	40.03	48.06	37.05	33.08	41.97
DAPro [27]	45.56	34.89	31.22	38.62	45.78	35.69	31.75	39.02	46.95	36.58	31.95	40.32	47.68	37.22	33.69	41.15
MPA* [7]	46.89	36.45	32.58	39.98	47.15	36.87	32.96	40.25	48.24	37.85	33.04	41.52	48.88	38.65	33.84	42.41
SAP* [49]	46.78	36.12	32.45	39.84	47.01	36.84	32.89	40.20	48.15	37.54	33.09	41.74	48.95	38.51	33.79	42.42
APNE* [50]	47.14	37.25	33.08	40.82	48.08	37.85	33.95	41.25	49.12	38.70	34.11	42.52	49.85	39.45	34.90	43.38
POND* [45]	47.42	37.32	33.12	40.85	48.11	37.90	33.98	41.41	49.18	38.65	34.15	42.63	49.90	39.46	34.89	43.40
MAP*(Ours)	**47.45**	**37.38**	**33.18**	**40.92**	**48.16**	**37.98**	**34.09**	**41.49**	**49.24**	**38.77**	**34.22**	**42.78**	**50.02**	**39.58**	**34.95**	**43.51**

Table 8: Domain adaptation results (AP_{novel}) on DWD with the same setting as table 2.

Backbone	RegionCLIP [54]				OVMR [33]				YOLO-World [8]				GDINO [32]			
Methods	NC	DR	NR	DF	NC	DR	NR	DF	NC	DR	NR	DF	NC	DR	NR	DF
LB	36.95	25.98	22.96	29.93	38.05	27.15	23.35	30.52	38.39	27.50	23.62	30.98	38.95	30.18	26.19	34.02
CoOP [56]	38.79	27.89	24.29	31.79	39.54	28.82	24.56	32.12	39.85	28.78	24.91	32.59	40.84	29.78	26.89	35.82
DAPL [14]	38.85	27.94	24.35	31.86	39.62	28.88	24.61	32.18	39.95	28.88	24.98	32.68	40.95	29.87	26.99	35.94
DAPro [27]	38.86	27.96	24.37	31.89	39.65	28.94	24.68	32.22	39.98	28.91	25.04	32.72	40.79	30.36	27.05	35.96
MPA* [7]	40.44	29.95	26.18	33.55	40.89	30.54	26.65	34.02	41.94	31.39	26.58	34.89	42.61	33.21	27.58	36.19
SAP* [49]	40.68	30.14	26.37	33.81	41.15	30.76	26.88	34.26	42.19	31.67	26.74	35.15	42.82	33.44	27.80	36.41
APNE* [50]	40.84	30.32	26.55	34.03	41.35	30.98	27.02	34.47	42.41	31.89	26.95	35.38	43.04	33.65	28.02	36.65
POND* [45]	41.25	30.78	26.96	34.51	41.78	31.42	27.48	34.90	42.85	32.31	27.38	35.80	43.49	34.09	28.44	37.10
MAP*(Ours)	**42.75**	**32.19**	**28.42**	**35.83**	**43.22**	**32.66**	**28.87**	**36.31**	**44.15**	**33.79**	**29.36**	**37.90**	**45.09**	**35.02**	**30.50**	**38.24**

model with a source model fine-tuned on one of the source domains. This approach can fully leverage prior knowledge from a particular source domain and may accelerate adaptation, especially when the target domain shares visual or semantic similarity with that source. However, such initialization may also introduce domain bias, potentially leading to overfitting to the source characteristics and suboptimal performance when the source-target domain gap is large.

To study this effect, we perform experiments on the Clipart1k target domain, using three initialization strategies: a vanilla object detector, a source model fine-tuned on Comic2k, and a source model fine-tuned on Watercolor2k. Results across four backbone detectors (RegionCLIP, OVMR, YOLO-World, and GDINO) are shown in table 9. We observe that initializing with a Comic2k or Watercolor2k model generally improves AP slightly over the vanilla initialization, especially when the backbone benefits from domain-specific knowledge without suffering from strong domain bias. This highlights the trade-off between leveraging source-specific priors and maintaining generalization.

Table 9: Domain adaptation results (APs) on Clipart1k using different pre-trained models to initialize the target model.

Backbone	RegionCLIP [54]			OVMR [33]			YOLO-World [8]			GDINO [32]		
Methods	AP	AP_{base}	AP_{novel}	AP	AP_{base}	AP_{novel}	AP	AP_{base}	AP_{novel}	AP	AP_{base}	AP_{novel}
Vanilla	52.54	53.25	52.18	55.35	56.42	55.14	54.86	**55.98**	54.42	55.55	56.67	55.16
Comic2k	52.69	53.45	**52.36**	**55.99**	**56.66**	**55.70**	**54.89**	55.67	**54.55**	56.28	56.71	56.09
Watercolor2k	**52.90**	**54.80**	52.08	55.79	56.06	55.67	54.79	55.69	54.41	**56.40**	**56.83**	**56.22**

Heterogeneous source model architecture. For simplicity, prior experiments used the same object detector architecture across all source models. However, our framework naturally extends to heterogeneous source backbones, as source models remain frozen during adaptation. In this ablation study, we explore this flexibility by employing a combination of four distinct backbone detectors: RegionCLIP, OVMR, YOLO-World, and GDINO, each pre-trained on a different source domain—Night Rainy, Dusk Rainy, Day Clear, and Night Clear, respectively. We adapt these models to the target domain Day Foggy. The resulting performance achieves an AP of 39.40%, with APbase at 40.89% and APnovel at 38.77%. While this is slightly lower than using a single strong backbone such as YOLO-World or GDINO alone—due to the relatively weaker performance of RegionCLIP and OVMR—it demonstrates that heterogeneous backbone integration is feasible and effective. Overall, this experiment confirms that MAP is compatible with multi-source models of varying architectures, enabling greater flexibility in practical deployment.

Impact of different augmentation methods. We evaluate several augmentation strategies, including CLIPstyler [24], CycleGAN [58], UNIT [31], and UNSB [23]. As shown in Table 10, CycleGAN achieves the best overall AP with a slight gain of $+0.23\%$ over our baseline, suggesting that richer pixel-level stylization can marginally enhance domain realism. However, the improvement is limited,

and text-guided augmentation remains competitive while being more lightweight, controllable, and free from style image requirements.

Table 10: Performance comparison of different methods.

Metric	CLIPstyler	CycleGAN	UNIT	UNSB
AP	55.55	55.78	54.89	55.42
AP_{base}	56.67	57.23	56.45	56.79
AP_{novel}	55.16	55.01	54.88	54.95

Integrating LLMs in OVOD. In this ablation, we examine the impact of leveraging more powerful pre-trained large language models (LLMs). Using YOLO-World as a baseline, we replace the original CLIP text encoder with the stronger LLaVA-OneVision-0.5b-ov [29]. Since the detector and the language model were pre-trained independently, we follow YOLO-World's pre-training strategy and re-train the detector for 5 epochs to align its visual features with the new text encoder. As shown in Table 11, this substitution yields consistent gains in detection accuracy, demonstrating that stronger LLM/MLLM backbones can significantly enhance cross-domain open-vocabulary detection. This result highlights the potential of integrating state-of-the-art multimodal language models into OVD frameworks to further push performance boundaries.

Table 11: Comparison between different vision-language models on cross-domain open-vocabulary detection.

Method	AP	AP_{base}	AP_{novel}
YOLO-World with CLIP (LB)	41.96	42.87	39.83
YOLO-World with LLaVA	42.88	43.85	40.70
MAP with CLIP	54.86	55.98	54.42
MAP with LLaVA (Ours)	55.19	56.45	54.68

D.3 Impact of Hyperparameters

In this section, we perform the experiments of different choices of the hyperparameters λ, M_1, M_2, M_3, and η to explore their impact.

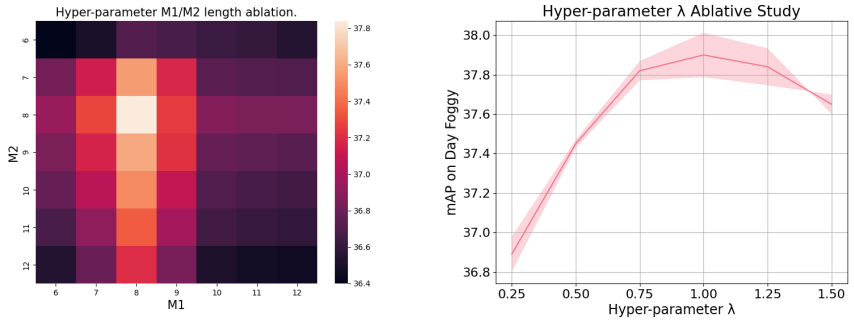
The influence of M_1, M_2, M_3 We conduct experiments on the *Day Foggy* dataset using G-DINO as the model backbone, varying M_1 and M_2 from 6 to 12. Figure 7(a) presents a heatmap of mAP scores for different values of M_1 and M_2. The results indicate that values that are too small or too large result in sub-optimal performance, likely due to underfitting or overfitting. The best performance is achieved when both M_1 and M_2 are set to 8. To evaluate the influence of M_3, we fix M_1 and M_2 at 8 and vary M_3 from 6 to 12. The results show that the optimal performance is achieved when $M_3 = 9$. However, variations in M_3 do not significantly affect performance on the Day Foggy domain when M_1 and M_2 are fixed.

Table 12: Impact of the hyperparameter M_3

Length	6	7	8	9	10	11	12
AP	40.41	40.55	40.85	**40.97**	40.95	40.82	40.69
AP_{base}	43.04	43.19	43.37	**43.51**	43.48	43.30	43.17
AP_{novel}	37.92	38.05	38.13	**38.24**	38.22	38.14	38.04

The influence of λ We conduct the experiment on *Day Foggy* dataset, where λ range from 0.25 to 1.50. We present the mAP with different λ values in fig. 7(b), which indicates that too small λ such as 0.25 lead to a sub-optimal performance since the target data is not fully utilized. While setting $\lambda = 1.0$, the learning from both source and target is balanced and lead to the optimal performance.

The influence of η. We evaluate the effect of η on the Clipart dataset with values ranging from 0.001 to 0.30, and report the results in Table 13. The best performance is obtained at $\eta = 0.10$, indicating that a small but non-zero regularization from the orthogonality term enhances performance. However, when η becomes too large, the orthogonality term dominates training and significantly degrades both base and novel class accuracy.



(a) Heatmap mAP on *Day Foggy* in terms of different lengths of the learnable tokens of $\mathbb{V}^c$ (M_1) and $\mathbb{V}^d$ (M_2).

(b) Multi-source test mAP on *Day Foggy* with λ ablations.

Figure 7: Impact of Hyperparameters.

Table 13: Impact of the hyperparameter η.

Metric	0.001	0.01	0.10	0.30
AP	55.47	55.50	**55.55**	52.13
AP_{base}	56.60	56.63	**56.67**	52.98
AP_{novel}	55.05	55.07	**55.16**	51.69

D.4 Additional Ablations

Impact of different components. We conduct three extra ablations to verify the effectiveness of text-guided style augmentation, the entropy minimization objective and the domain description respectively by (1) Removing text-guided style augmentation. We replace the text-guided source-style augmentation with generic augmentations (random cropping, translation, rotation, Gaussian noise). Because these augmentations do not preserve source-domain style, the domain gap between the source model and the augmented views increases, leading to a noticeable performance drop compared to using text-guided style augmentation. (2) Removing entropy minimization and consistency. We remove the target-domain losses includes $\mathcal{L}^t_{\text{class}}$, and $\mathcal{L}^t_{\text{domain}}$, and train UTPL only with $\mathcal{L}_{\text{novel}}$ $\mathcal{L}_{\text{orth}}$. Without entropy minimization and consistency, the target prompt cannot learn meaningful class-specific or domain-specific tokens. As a result, the remaining regularization terms are insufficient to guide the target prompt, and performance degrades significantly. (3) Replacing domain-specific descriptions with generic ones. Previously we use domain-specific descriptions such as "comic" (Comic2k) or "watercolor" (Watercolor2k). We replace them with generic descriptions: "art." These generic descriptions encode little domain information, so the target prompt is less domain-aware, which reduces adaptation performance.

Table 14: Ablation study results.

Method Variant	AP	AP_{base}	AP_{novel}
remove text guided style augmentation	46.52	48.24	44.35
remove entropy minimization and consistency	41.88	42.63	39.59
generic domain descriptions	50.92	49.77	48.65

Comparison with unsupervised prompt learning methods. To provide a more comprehensive evaluation, we have included comparisons with recent few-shot and unsupervised prompt tuning methods that are closely related to our setting [40, 18, 44]. As shown in Table 15, these approaches represent strong and up-to-date baselines in unsupervised prompt tuning. Their performance is generally in line with other single-source prompt learning approaches. In contrast, our method explicitly leverages multiple heterogeneous source domains and integrates them through the meta-learned aggregation mechanism, which leads to substantially better performance across base and novel categories.

Table 15: Comparison with recent unsupervised prompt tuning baselines.

Method	AP	AP_{base}	AP_{novel}
TPT [40]	46.37	47.45	45.52
UPT [18]	44.33	44.95	43.61
M-CRPL [44]	40.28	41.35	38.84
MAP	**55.55**	**56.67**	**55.16**

Performance of MAP on classification or segmentation tasks. To further support the generality of the framework, we additionally conduct experiments on classification (DomainNet [26], Table 16) and semantic segmentation (ACDC [38], Table 17), confirming that the same MAP formulation transfers well to other tasks without architectural changes. For both tasks, we follow the open-vocabulary protocol by removing a subset of classes from the source domain to ensure no label overlap with the target domain.

Table 16: Classification Accuracy across DomainNet domains.

Domain	AP	AP_{base}	AP_{novel}
Clipart	92.88	94.45	91.52
Infograph	90.67	91.23	90.12
Painting	88.45	89.65	88.01
Quickdraw	89.98	88.79	90.24
Real	87.65	88.42	86.98
Sketch	91.23	91.88	90.47

Table 17: Segmentation Performance across ACDC domains.

Domain	AP	AP_{base}	AP_{novel}
Foggy	33.7	34.8	30.5
Nighttime	33.1	33.9	30.4
Rainy	31.5	32.4	28.2
Snowy	31.2	32.2	28.1

Additional experiments for single-source methods with combined source domains. For fairness comparison, we conducted additional experiments where all single-source methods are trained on the combined source domains, treating them as a single source. The results are presented in Table 18. These results show that incorporating information from multiple source domains indeed boosts the performance of SSDA methods. This further reinforces the importance of properly integrating heterogeneous sources, something that our proposed meta-learned aggregation framework is explicitly designed to address.

Diffusion-based augmentation. Our choice of CLIPStyler employs lightweight, controllable text-driven style augmentation. Here we consider diffusion-based style transfer methods. We conducted additional experiments using diffusion-based augmentation [21, 35, 48, 5]. The results are summarized in Table 19. The results show that diffusion-based augmentations produce higher-quality stylized images and generally lead to improved adaptation performance, further confirming that style-based augmentation effectively reduces the domain gap. This also demonstrates that our framework is flexible and can directly benefit from more advanced generative models as they continue to improve.

D.5 Visualization of Prompt Embeddings

In this ablation, we visualized the different components of the learned prompts. We first randomly select 75 sample prompts that summarize each pairing input image. By reducing their dimensionality into a 2-d plane with T-SNE, we visualize each textual prompt embedding from every domain along with the domain invariant and domain specific tokens generated in our proposed method, as shown in Figs 8, 9 and 10. We choose to visualize the embedding of the class `car` to show the domain shift effect and our method's capability to decompose such effect into a respective domain invariant token and a domain specific embedding token for each input image prompt. In Fig. 8, we first plot the

Table 18: Single-source methods trained on the combined source domains.

Method	AP	AP_{base}	AP_{novel}
LB	46.14	47.08	43.85
CoOP	45.24	46.08	43.21
DAPL	46.54	47.58	44.62
DAPro	47.85	48.88	45.49

Table 19: Comparison of diffusion-based and style-transfer augmentations.

Method	**AP**	**AP_base**	**AP_novel**
Diffusemix [2]	53.54	54.28	53.08
DIDEX [3]	48.54	48.99	47.62
ZeCon [4]	49.54	49.99	48.62
InstructPix2Pix [5]	53.78	55.18	53.70
PODA [6]	50.62	50.87	49.49

image embeddings from *Day Clear* (i.e., source) and *Day Foggy* (i.e., target) with red and blue dots. The domain-invariant tokens are denoted with orange stars. We also plot the template embedding with handcrafted prompts such as "a photo of car in the road, in the street", etc and denote them as magenta pentagon. From Fig. 8 red circle grouped dots, we find that the model discovered domain invariant embedding resides between the source and target domains, and locate very near to the template. Similar finding can be seen in Fig. 9 with *Night Clear* and *Night Rainy* domains. All above findings confirm that our method could summarize effective domain invariant embeddings that grasp the major objects of the image. In Fig. 10, we plot the domain-specific tokens with different colors of stars: blue, green, red and purple. We also plot the handcrafted template in orange pentagon. From Fig. 10, we can also see that the domain-specific tokens reside near to each domain's shifted embedding while be far away from the template, for all four data domains. We have highlighted two groups with blue and red circles in Fig. 10 for better illustration.

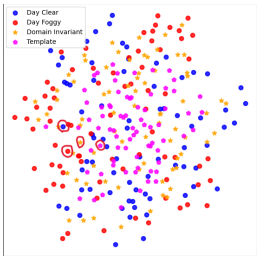


Figure 8: Day Clear versus Day Foggy Textual Embedding and Domain Invariant Tokens.

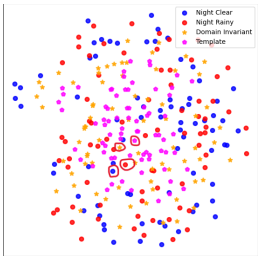


Figure 9: Night Clear versus Night Rainy Textual Embedding and Domain Invariant Tokens.

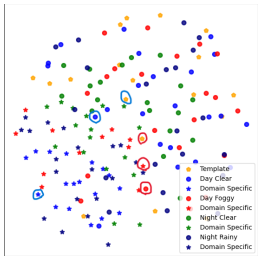


Figure 10: All domains textual embedding and their respective Domain Specific Tokens.

D.6 Visualization of Class Embeddings

In fig. 11, we present the t-SNE visualization of base and novel classes on the Clipart dataset. Aside from a few outliers, the base and novel classes are well separated, indicating that the novel mask loss encourages the novel classes to explore a distinct feature space.

D.7 Additional datasets

We create a new multi-source dataset by simply utilizing **Cityscapes**[2] [10], **Foggy-Cityscapes**[3] [37], and **KITTI**[4] [16] for further evaluation. Cityscapes consist of 2,975 training images and 500 testing

[2] https://github.com/tiancity-NJU/da-faster-rcnn-PyTorch
[3] https://www.cityscapes-dataset.com/downloads/
[4] http://www.cvlibs.net/datasets/kitti/

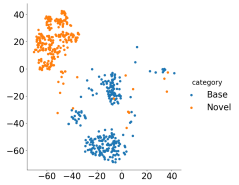


Figure 11: t-SNE of base and novel class in Clipart.

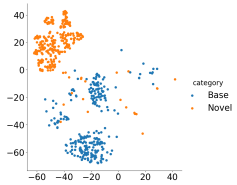


Figure 12: t-SNE of base and novel class in DWD Day Foggy.

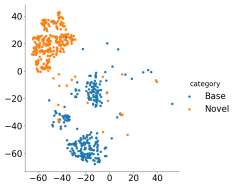


Figure 13: t-SNE of base and novel class in Foggycityscapes.

images, have a total of 8 categories captured under normal weather. Foggy-Cityscapes applies images of Cityscapes to simulate foggy as well as inherits the annotations of Cityscapes. KITTI contains 7,481 urban images which are different from Cityscapes. The Cityscapes and KITTI are used as source domains, while FoggyCityscapes is the target domain. To augment the target images with the source domain styles, we use "A photo of driving in the city in a foggy daytime" as source description and "A photo of driving in the city during daytime" and "A photo of driving in the urban city" as target descriptions. For open-vocabulary object detection, we reserve `Rider, Bike, Motor` as novel classes. The performance is presented in Tab. 20. It's evident that our method excels in the single-source domain adaptation methods, attaining state-of-the-art results. Furthermore, when compared to other multi-source domain adaptive techniques, the advantage of our approach is evident across novel and base classes.

Table 20: Domain adaptation results (APs) on Foggycityscapes. For multi-source domain adaptation methods (highlighted with *), Cityscapes and KITTI are used jointly as source domains. For single-source domain adaptation methods, either Cityscapes or KITTI is used as the source domain, and only the best performance is reported. LB denotes Lower Bound, where the backbone detector is directly applied to the target domain without adaptation.

Backbone	RegionCLIP [54]			OVMR [33]			YOLO-World [8]			GDINO [32]		
Methods	AP	AP_{base}	AP_{novel}	AP	AP_{base}	AP_{novel}	AP	AP_{base}	AP_{novel}	AP	AP_{base}	AP_{novel}
LB	40.15	41.87	38.06	41.87	42.34	39.59	40.95	41.89	38.86	42.07	42.95	39.80
CoOP [56]	40.25	41.97	38.18	42.05	42.57	39.74	41.11	42.03	38.96	42.20	43.07	40.05
DAPL [14]	40.59	42.26	38.49	42.38	42.86	40.06	41.56	42.40	39.44	42.49	43.49	40.49
DAPro [27]	41.44	42.15	39.39	43.34	43.78	40.97	42.50	43.37	40.39	43.42	44.42	41.40
MPA* [7]	44.15	45.14	42.33	46.15	46.64	43.79	45.52	46.35	43.32	46.36	47.29	44.23
SAP* [49]	45.26	47.05	43.24	47.02	47.58	44.75	46.32	47.28	44.33	47.28	48.29	45.24
APNE* [50]	45.78	47.60	43.76	47.38	48.12	45.30	46.86	47.80	44.78	47.85	48.72	45.66
POND* [45]	48.50	50.31	46.42	50.08	50.82	48.78	49.55	50.44	47.45	50.48	51.44	48.41
MAP*(Ours)	**52.02**	**52.76**	**51.69**	**54.87**	**55.89**	**54.62**	**54.15**	**55.26**	**53.74**	**54.88**	**55.98**	**54.39**

D.8 Additional Metric

In this section, we provide AP AP@[.5:.95] for the GDINO backbone on Clipart1k. As shown below, the performance gain of our method is preserved under the stricter COCO-style metric.

E Theoretical Justification

Before providing the proofs of the propositions, we first provide the following Lemmas.

Lemma E.1 (Cross-Entropy Maximizes a Lower Bound of Mutual Information)**.** *Let* $\mathbf{pt}$ *denote a prompt representation and let* $q(y|\mathbf{pt})$ *be the predicted class distribution. Then minimizing the cross-entropy loss*

$$\mathcal{L}_{\text{CE}} = -\mathbb{E}_{(x,y)} \log q(y|\mathbf{pt})$$

maximizes a lower bound of the mutual information $I(\mathbf{pt}; Y)$ *between the prompt representation and the class label.*

Proof. For a prompt $\mathbf{pt}$, suppose the classifier predicts $q(y|\mathbf{pt})$. The cross-entropy loss satisfies

$$\mathcal{L}_{\text{CE}} = H(Y|\mathbf{PT}) + D_{\text{KL}}\big(p(Y|\mathbf{PT})\|q(Y|\mathbf{PT})\big), \tag{15}$$

Table 21: Comparison of methods on AP, AP_{base}, and AP_{novel}.

Method	AP	AP_{base}	AP_{novel}
LB	35.47	36.52	33.48
CoOP	35.67	36.78	33.77
DAPL	35.96	37.21	34.18
DAPro	36.84	38.12	35.05
MPA	39.52	40.68	37.65
SAP	40.62	41.79	38.73
APNE	42.85	43.54	41.29
POND	46.17	46.95	45.08
MAP	**50.29**	**50.88**	**48.79**

where $H(\cdot)$ denotes entropy and $D_{\mathrm{KL}}(\cdot\|\cdot)$ is the Kullback–Leibler divergence.

Since

$$D_{\mathrm{KL}} \geq 0,$$

we have

$$H(Y|\mathbf{PT}) \leq \mathcal{L}_{\mathrm{CE}}. \tag{16}$$

By the definition of mutual information,

$$I(\mathbf{PT}; Y) = H(Y) - H(Y|\mathbf{PT}), \tag{17}$$

where $H(Y)$ is constant for a fixed dataset. Therefore,

$$I(\mathbf{PT}; Y) \geq H(Y) - \mathcal{L}_{\mathrm{CE}}. \tag{18}$$

Hence minimizing the cross-entropy loss maximizes a lower bound of the mutual information between the prompt and the class label. □

Lemma E.2 (Entropy Minimization Maximizes a Lower Bound of Mutual Information)**.** *Let $\hat{Y}$ denote the predicted label for an unlabeled sample $\mathbf{x}$. Under the assumption that the marginal entropy $H(\hat{Y})$ is approximately constant during optimization, minimizing the prediction entropy*

$$\mathcal{L}_{\mathrm{ent}} = H(\hat{Y}|\mathbf{x})$$

maximizes a lower bound of the mutual information $I(\mathbf{x}; \hat{Y})$.

Proof. The mutual information between the input image and its prediction is

$$I(\mathbf{x}; \hat{Y}) = H(\hat{Y}) - H(\hat{Y}|\mathbf{x}). \tag{19}$$

The entropy minimization objective is exactly

$$\mathcal{L}_{\mathrm{ent}} = H(\hat{Y}|\mathbf{x}). \tag{20}$$

Assuming the marginal entropy $H(\hat{Y})$ remains approximately constant during optimization, we obtain

$$I(\mathbf{x}; \hat{Y}) = H(\hat{Y}) - \mathcal{L}_{\mathrm{ent}}. \tag{21}$$

Therefore,

$$I(\mathbf{x}; \hat{Y}) \geq H(\hat{Y}) - \mathcal{L}_{\mathrm{ent}}, \tag{22}$$

which shows that minimizing the entropy objective maximizes a lower bound of the mutual information. □

Lemma E.3 (Orthogonality Promotes Prompt Distinction)**.** *Assume prompt embeddings are ℓ_2-normalized. Minimizing the orthogonality loss*

$$\mathcal{L}_{\mathrm{orth}} = \sum_{i \neq j} (\mathbf{pt}_i^\top \mathbf{pt}_j)^2$$

reduces the mutual information between different prompt embeddings, thereby increasing prompt distinction.

Proof. Since all prompt embeddings are normalized, their inner product equals cosine similarity,

$$\mathbf{pt}_i^\top \mathbf{pt}_j = \cos(\theta_{ij}).$$

Minimizing

$$(\mathbf{pt}_i^\top \mathbf{pt}_j)^2$$

forces

$$\cos(\theta_{ij}) \to 0,$$

making different prompt embeddings nearly orthogonal.

Assuming the prompt embeddings follow an approximately Gaussian distribution, the mutual information between two prompt embeddings can be expressed as

$$I(\mathbf{pt}_i; \mathbf{pt}_j) = -\frac{1}{2}\log(1 - \rho_{ij}^2),$$

where ρ_{ij} denotes their correlation coefficient.

Since the cosine similarity is proportional to the correlation of normalized embeddings,

$$\rho_{ij} \approx \mathbf{pt}_i^\top \mathbf{pt}_j,$$

minimizing the orthogonality loss reduces ρ_{ij} toward zero, yielding

$$I(\mathbf{pt}_i; \mathbf{pt}_j) \to 0.$$

Therefore, the orthogonality loss explicitly encourages different prompts to encode non-redundant information and thus promotes prompt distinction. □

E.1 Proof of theorem 4.3.

Proof. In the unsupervised setting, where ground-truth labels y are unavailable, we cannot directly evaluate the mutual information $\text{MI}(\mathbf{pt}; y)$. Instead, we approximate *fidelity* by quantifying the information shared between the input region $\mathbf{x}$ and the predicted label $\hat{y}$, i.e., $\text{MI}(\mathbf{x}; \hat{y})$, where $\hat{y}$ is the model's output given prompt $\mathbf{pt}$. We argue that under the assumption of an approximately constant $H(\hat{y})$, low entropy of the prediction distribution $p(\hat{y}|\mathbf{x}, \mathbf{pt})$ implies high mutual information between the input $\mathbf{x}$ and the predicted label $\hat{y}$, thus serving as a proxy for high fidelity.

Formally, the mutual information between $\mathbf{x}$ and $\hat{y}$ is given by:

$$\text{MI}(x; \hat{y}) = H(\hat{y}) - H(\hat{y}|\mathbf{x}), \tag{23}$$

where $H(\hat{y})$ is the marginal entropy over predictions across the dataset, and $H(\hat{y}|\mathbf{x})$ is the conditional entropy of predictions for a given input $\mathbf{x}$.

Minimizing the entropy of the prediction distribution:

$$\mathcal{L}_{\text{entropy}} = -\sum_k p(\hat{y} = k|\mathbf{x}, \mathbf{pt}) \log p(\hat{y} = k|\mathbf{x}, \mathbf{pt}) \tag{24}$$

effectively reduces $H(\hat{y}|x)$. If $H(\hat{y})$ is kept high or remains approximately constant (e.g., via diverse predictions across the dataset), then the mutual information $\text{MI}(\mathbf{x}; \hat{y})$ increases:

$$\text{Low } H(\hat{y}|\mathbf{x}) \Rightarrow \text{High MI}(\mathbf{x}; \hat{y}). \tag{25}$$

Hence, minimizing entropy encourages confident and consistent predictions, which implies that the prompt encodes meaningful and class-relevant information. This serves as an effective surrogate for *high fidelity* in the absence of label supervision:

$$\boxed{\text{Low entropy} \Rightarrow \text{High mutual information } I(\mathbf{x}; \hat{y}) \Rightarrow \text{High unsupervised fidelity}}. \tag{26}$$

□

E.2 Proof of Proposition 4.4

Proof. The source class-specific prompt is optimized by minimizing the classification loss

$$\mathcal{L}^s_{\text{class}} = -\mathbb{E}_{\mathcal{D}^s} \log p^s_{\text{class}}(\hat{y} = k|\mathbf{r}^s), \tag{27}$$

where the prediction probability is defined in Eq. (4).

By Lemma E.1, minimizing the cross-entropy loss maximizes a lower bound of the mutual information between the learned class prompt $\mathbf{pt}^s_k$ and its corresponding class label,

$$I(\mathbf{pt}^s_k; Y) \geq H(Y) - \mathcal{L}^s_{\text{class}}. \tag{28}$$

Therefore, optimizing $\mathcal{L}^s_{\text{class}}$ encourages the prompt to preserve semantic information that is highly correlated with the class labels, thereby achieving high class fidelity.

Meanwhile, the orthogonality regularization

$$\mathcal{L}_{\text{orth}} = \sum_{i \neq j} (g(\mathbf{pt}^i)^\top g(\mathbf{pt}^j))^2 \tag{29}$$

is minimized jointly during training. According to Lemma E.3, minimizing $\mathcal{L}_{\text{orth}}$ reduces the mutual information between prompts from different classes,

$$I(\mathbf{pt}_i, \mathbf{pt}_j) \to 0, \qquad i \neq j, \tag{30}$$

thus encouraging each class prompt to encode distinct semantic information.

Therefore, jointly minimizing $\mathcal{L}^s_{\text{class}}$ and $\mathcal{L}_{\text{orth}}$ produces source class prompts with high fidelity and strong inter-class distinction. □

E.3 Proof of Proposition 4.5

Proof. The source domain-specific prompts are optimized by minimizing

$$\mathcal{L}^s_{\text{domain}} = -\mathbb{E}_{\mathcal{D}^s} \log p^s_{\text{domain}}(\hat{y} = k|\mathbf{r}^s), \tag{31}$$

whose prediction probability is defined in Eq. (6).

Unlike the class-specific loss, the denominator of Eq. (6) contains all class–domain prompt pairs. Consequently, minimizing $\mathcal{L}^s_{\text{domain}}$ encourages the visual feature to be aligned with the correct class–domain prompt while suppressing its similarity to prompts from other domains.

Since the objective is again a cross-entropy loss, Lemma E.1 gives

$$I(\mathbf{pt}^s; D) \geq H(D) - \mathcal{L}^s_{\text{domain}}, \tag{32}$$

where D denotes the domain label. Thus minimizing $\mathcal{L}^s_{\text{domain}}$ maximizes a lower bound of the mutual information between the prompt and its corresponding domain, leading to high domain fidelity.

Furthermore, minimizing $\mathcal{L}_{\text{orth}}$ reduces the mutual information among prompts belonging to different domains according to Lemma E.3, thereby encouraging domain-specific prompts to encode complementary and non-overlapping representations.

Therefore, the source domain prompts achieve high domain fidelity and strong inter-domain distinction. □

E.4 Proof of Proposition 4.6

Proof. The target class-specific prompt is optimized using the entropy minimization objective

$$\mathcal{L}^t_{\text{class}} = -\sum_k \tilde{p}(k) \log \tilde{p}(k), \tag{33}$$

where $\tilde{p}$ denotes the averaged prediction probability over the original target image and its style-transferred augmentations.

By Lemma E.2, minimizing the entropy objective maximizes a lower bound of the mutual information between the target image and the predicted class,

$$I(\mathbf{x}; \hat{Y}) \geq H(\hat{Y}) - \mathcal{L}^t_{\text{class}}. \tag{34}$$

Consequently, the learned prompt is encouraged to preserve semantic information that produces confident and consistent predictions across different augmented views, thereby satisfying the proposed definition of unsupervised fidelity.

Moreover, the optimization is performed on the averaged prediction over multiple domain-style augmentations,

$$\tilde{p} = \frac{1}{N+1} \sum_{s=0}^{N} p(\hat{y} | \mathcal{A}_s(\mathbf{r}^t)). \tag{35}$$

Since all augmented views correspond to the same semantic content, minimizing the entropy forces the learned class prompt to produce consistent predictions regardless of domain style, thereby encouraging the prompt to encode domain-invariant semantic information.

Finally, $\mathcal{L}_{\text{orth}}$ reduces the mutual information between different class prompts according to Lemma E.3, thus increasing inter-class distinction.

Therefore, the learned target class prompts possess high unsupervised fidelity together with strong semantic distinction. □

E.5 Proof of Proposition 4.7

Proof. The target domain-specific prompt is optimized through the entropy loss

$$\mathcal{L}^t_{\text{domain}}, \tag{36}$$

which averages predictions over all stylized target views and performs entropy minimization over all class–domain prompt combinations.

By Lemma E.2, minimizing $\mathcal{L}^t_{\text{domain}}$ maximizes a lower bound of the mutual information between the target image and its predicted domain-aware representation, thereby improving unsupervised fidelity.

Unlike $\mathcal{L}^t_{\text{class}}$, the probability formulation in Eq. (10) contains all source and target domain prompts in the denominator. Consequently, entropy minimization encourages the correct target class–domain prompt to dominate the prediction while suppressing mismatched domain prompts. This improves the discriminability of the learned target-domain prompt.

Finally, $\mathcal{L}_{\text{orth}}$ reduces the mutual information between prompts from different domains according to Lemma E.3, thus explicitly promoting inter-domain distinction.

Therefore, the target domain prompt simultaneously achieves high unsupervised fidelity and strong inter-domain distinction. □

E.6 Proof of Proposition 4.8

Proof. The novel class mask is learned by minimizing

$$\mathcal{L}_{\text{novel}} = \sum_{c \in \mathbb{C}^{\text{base}}} \sum_{k \in \mathbb{C}^{\text{novel}}} (g(\mathbf{pt}_c)^\top g(\mathbf{pt}_k))^2. \tag{37}$$

The objective directly penalizes the squared inner product between base and novel prompt embeddings. Since the prompt embeddings are normalized, minimizing $\mathcal{L}_{\text{novel}}$ forces

$$g(\mathbf{pt}_c)^\top g(\mathbf{pt}_k) \to 0, \tag{38}$$

which makes the base and novel prompt embeddings approximately orthogonal.

By Lemma E.3, orthogonal prompt embeddings exhibit low mutual information. Hence,

$$I(\mathbf{pt}_c; \mathbf{pt}_k) \to 0, \tag{39}$$

implying that the learned novel prompts encode information that is distinct from the base classes.

Therefore, minimizing $\mathcal{L}_{\text{novel}}$ explicitly encourages semantic separation between base and novel categories and increases their inter-class distinction. □

E.7 Generalization Guarantee

In this section, we explore the generalization error of open-vocabulary domain adaptation under multi-source setting. Given N source domains, weighted by $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, ..., \alpha_N)$, $\sum_{i=1}^{N} \alpha_i = 1$, we examine the convex combination of training error $\hat{err}_{\boldsymbol{\alpha}}(h)$ from each source domain following the setting in [4], where $h \sim \mathcal{H}$ is a hypothesis in the hypothesis space $\mathcal{H}$. Denoting the empirical error in the target domain as $\hat{err}_t(h)$, the generalization error of multi-source open-vocabulary domain adaptation is bounded by the following theorem E.4.

Theorem E.4. *Given N labeled source domains $S_1, ..., S_N$ and an unlabeled target domain T, let $\hat{err}_{\boldsymbol{\alpha}}(h)$ be the empirical $\boldsymbol{\alpha}$-weighted error of a hypothesis h, π^t_{novel} be the class prior probability for the novel classes in the target domain, for any $\delta \in (0, 1)$, with probability at lease $1 - \delta$,*

$$\frac{\hat{err}_t(h)}{1 - \pi^t_{novel}} \leq \hat{err}_{\boldsymbol{\alpha}}(h) + \sum_{i=1}^{N} \alpha_i (2\lambda_i + d_{\mathcal{H}\Delta\mathcal{H}}(D_i, D_T)) + \sum_{i=1}^{N} \alpha_i \left(\frac{\hat{err}^t_{novel}(h)}{1 - \pi^t_{novel}} - \hat{err}^{s_i}_{novel}(h) \right) \tag{40}$$

where $\lambda_i = \min_{h \in \mathcal{H}} \{err_t(h) + err_i(h)\}$. D_i, D_T are the domain distributions for source domain S_i and target domain T, respectively. $\hat{err}^{s_i}_{novel}(h)$ and $\hat{err}^t_{novel}(h)$ are the empirical risk of samples belong to the novel classes.

eq. (40) contains three parts, convex combination of source error, domain discrepancy, and open-vocabulary difference. Detailed proof of theorem E.4 is provided in as follows:

Proof. **Step 1.** Given a symmetric loss ℓ, with $h \in \mathcal{H}$, the expected risks of the convex combination source and target is defined as:

$$err_{\boldsymbol{\alpha}}(h) = \sum_{i=1}^{N} \alpha_i \mathbb{E}_{(\mathbf{x},\mathbf{y}) \sim D_i} \ell(h(\mathbf{x}), \mathbf{y}). \tag{41}$$

$$err_t(h) = \mathbb{E}_{(\mathbf{x},\mathbf{y}) \sim D_t} \ell(h(\mathbf{x}), \mathbf{y}) \tag{42}$$

The partial risk of known target classes is

$$err^*_t(h) = \frac{1}{1 - \pi^t_{\text{novel}}} \int_{\mathcal{X} \times \mathcal{Y}^s} \ell(h(\mathbf{x}), \mathbf{y}) dP_{X^t Y^t}(\mathbf{x}, \mathbf{y}) \tag{43}$$

where $\mathcal{X}$ is the feature space, $\mathcal{Y}^s$ is the source label space, we suppose all the source domain share the same label space $\mathcal{Y}^s$ for simplification. $P_{X^t Y^t}$ is the joint distribution of the target domain.

The partial risk for the unknown target classes is

$$err^{\text{novel}}_t(h) = \int_{\mathcal{X}} \ell(h(\mathbf{x}), \mathbf{y}_{\text{novel}}) dP_{X^t | \mathbf{y}_{\text{novel}}}(\mathbf{x}) \tag{44}$$

And we have

$$err_t(h) = \pi^t_{\text{novel}} err^{\text{novel}}_t(h) + (1 - \pi^t_{\text{novel}}) err^*_t(h) \tag{45}$$

Then

$$\frac{err_t(h)}{1 - \pi^t_{\text{novel}}} - err_{\boldsymbol{\alpha}}(h) = err^*_t(h) - err_{\boldsymbol{\alpha}}(h) + \frac{\pi^t_{\text{novel}}}{1 - \pi^t_{\text{novel}}} err^{\text{novel}}_t(h) \tag{46}$$

Step 2. Since we assume all the source domain share the same label space $\mathcal{Y}^s$, we have

$$\sum_{i=1}^{N} \alpha_i \int_{\mathcal{X} \times \mathcal{Y}^{s_i}} \ell(h(\mathbf{x}), \mathbf{y}) dP_{X^s Y^s}(\mathbf{x}, \mathbf{y}) = \int_{\mathcal{X} \times \mathcal{Y}^s} \ell(h(\mathbf{x}), \mathbf{y}) dP_{X^s Y^s}(\mathbf{x}, \mathbf{y}) \tag{47}$$

$$
\begin{aligned}
err_t^*(h) - err_{\boldsymbol{\alpha}}(h) &= \int_{\mathcal{X}\times\mathcal{Y}^t} \ell(h(\mathbf{x}),\mathbf{y}) dP_{X^tY^t|\mathcal{Y}^s}(\mathbf{x},\mathbf{y}) - \sum_{i=1}^N \alpha_i \int_{\mathcal{X}\times\mathcal{Y}^s} \ell(h(\mathbf{x}),\mathbf{y}) dP_{X^{s_i}Y^s}(\mathbf{x},\mathbf{y}) \\
&\leq err_t^*(\tilde{h}) + \int_{\mathcal{X}\times\mathcal{Y}^t} \ell(\tilde{h}(\mathbf{x}),\mathbf{y}) dP_{X^tY^t|\mathcal{Y}^s}(\mathbf{x},\mathbf{y}) \\
&+ err_{\boldsymbol{\alpha}}(\tilde{h}) - \sum_{i=1}^N \alpha_i \int_{\mathcal{X}\times\mathcal{Y}^s} \ell(\tilde{h}(\mathbf{x}),\mathbf{y}) dP_{X^{s_i}Y^s}(\mathbf{x},\mathbf{y})
\end{aligned} \tag{48}
$$

where $\tilde{h}$ is any hypothesis in $\mathcal{H}$. According to [12], we have

$$
\int_{\mathcal{X}\times\mathcal{Y}^t} \ell(h(\mathbf{x}),\tilde{h}(\mathbf{x})) dP_{X^tY^t|\mathcal{Y}^s}(\mathbf{x},\mathbf{y}) = \int_{\mathcal{X}} \ell(h(\mathbf{x}),\tilde{h}(\mathbf{x})) dP_{X^t|\mathcal{Y}^s}(\mathbf{x},\mathbf{y}) \tag{49}
$$

And for $i \in \{1,...,N\}$

$$
\int_{\mathcal{X}\times\mathcal{Y}^s} \ell(h(\mathbf{x}),\tilde{h}(\mathbf{x})) dP_{X^{s_i}Y^s}(\mathbf{x}) = \int_{\mathcal{X}} \ell(h(\mathbf{x}),\tilde{h}(\mathbf{x})) dP_{X^{s_i}}(\mathbf{x}) \tag{50}
$$

Based on above equations, we have

$$
\begin{aligned}
err_t^*(h) - err_{\boldsymbol{\alpha}}(h) &\leq err_t^*(\tilde{h}) - err_{\boldsymbol{\alpha}}(\tilde{h}) + \sum_{i=1}^N \alpha_i \Big| \int_{\mathcal{X}} \ell(h(\mathbf{x}),\tilde{h}(\mathbf{x})) dP_{X^t|\mathcal{Y}^s}(\mathbf{x},\mathbf{y}) \\
&\quad - \int_{\mathcal{X}} \ell(h(\mathbf{x}),\tilde{h}(\mathbf{x})) dP_{X^{s_i}}(\mathbf{x}) \Big| \\
&\leq err_t^*(\tilde{h}) - err_{\boldsymbol{\alpha}}(\tilde{h}) + \sum_{i=1}^N \alpha_i d_{\mathcal{H}}(P_{X^t|\mathcal{Y}^s}, P_{X^{s_i}})
\end{aligned} \tag{51}
$$

Hence, based on the definition of λ_i, and denoting $D_i = P_{X^{s_i}}$, $D_T = P_{X^t|\mathcal{Y}^s}$, we have,

$$
err_t^*(h) - err_{\boldsymbol{\alpha}}(h) \leq \sum_{i=1}^N \alpha_i (2\lambda_i + d_{\mathcal{H}\Delta\mathcal{H}}(D_i, D_T)) \tag{52}
$$

Step 3. In this step, we prove that

$$
\frac{\pi^t_{\text{novel}}}{1-\pi^t_{\text{novel}}} err^t_{\text{novel}}(h) \leq d_{\mathcal{H}\Delta\mathcal{H}}(D_i, D_T) + \frac{R^t_{\text{novel}}(h)}{1-\pi^t_{\text{novel}}} - R^t_{\text{novel}}(h) \tag{53}
$$

This step is similar the proof in [12]. Please refer to the detailed in [12].

By combining the results from Step 2 & 3, we have

$$
\begin{aligned}
\frac{\hat{err}_t(h)}{1-\pi^t_{\text{novel}}} \leq & \hat{err}_{\boldsymbol{\alpha}}(h) + \sum_{i=1}^N \alpha_i (2\lambda_i + d_{\mathcal{H}\Delta\mathcal{H}}(D_i, D_T)) \\
& + \sum_{i=1}^N \alpha_i \left(\frac{\hat{err}^t_{\text{novel}}(h)}{1-\pi^t_{\text{novel}}} - \hat{err}^{s_i}_{\text{novel}}(h) \right)
\end{aligned} \tag{54}
$$

□

F Potential Societal Impact and Limitations

Open-vocabulary domain adaptation with multiple sources has the potential to significantly expand its application in diverse settings with various constraints, such as heterogeneous and dynamic domains with scarce labels. The transformative potential of open-vocabulary models lies in their ability to generalize from limited data by leveraging pre-trained knowledge, reducing the necessity for large, domain-specific datasets. This approach allows models to remain relevant and effective in dynamic environments without extensive retraining. By minimizing the need for comprehensive data

collection and labeling, open-vocabulary domain adaptation also lowers the costs associated with model development and maintenance. In terms of potential negative societal impact, by utilizing the training and testing data from different domains, there is a possibility of data privacy concern. A potential limitation of open-vocabulary domain adaptation is that the effectiveness of open-vocabulary domain adaptation heavily relies on the quality and breadth of the pre-trained models. If these models are not sufficiently comprehensive or up-to-date, their generalization capabilities may be limited.

G Source Code

For the source code, please check `https://anonymous.4open.science/r/LEET-585B/README.md`.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction claim that the paper proposes a novel prompt learning method for cross-domain open-vocabulary object detection, which is the main contribution of this paper. The theoretical analysis has proven the upper bound of the multi-source cross-domain open-vocabulary adaptation. And the experiments conducted on the benchmark open-vocabulary cross-domain object detection datasets have confirmed the effectiveness of the proposed method.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper mentioned that previous open-vocabulary object detection methods neglect the domain shift between the source and target domain, while other domain adaptation methods did not cover category shifts. In terms of prompt learning methods, cross-domain open-vocabulary prompt learning is underexplored. Current methods only focus on cross-domain prompt learning. None of the previous methods can fully address the complicated cross-domain open-vocabulary object detection issue. The proposed method has designed a cross domain open-vocabulary prompt the adapts the model from both source domain to the target domain and base categories to novel categories in the target domain.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.

- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: All assumptions for each theoretical result are clearly stated, and referenced. The proofs are provided in the appendix or properly referenced.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The contribution is a novel prompt design for cross-domain open-vocabulary object detection. The prompt design and the corresponding loss function are provided. Code and instructions are also provided.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may

be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.

- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The paper conducts experiments on open-source benchmark object detection datasets. And the code are provided anonymously.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

Justification: The details of the experimental setting are presented in section 5 and appendix B.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: Due to the nature of object detection task, the error bars are not provided just like all other object detection works.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: The details of the computer resources are presented in section 5.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: The authors remain anonymous.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: The discussion of both potential positive societal impacts and negative societal impacts of the work are appendix F.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The paper did not release models and data that have a high risk for misuse.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.

- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The related datasets and models used in the paper are properly cited.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The code is released and properly documented anonymously.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper doesn't include crowdsourcing and research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper doesn't involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.