arXiv:2406.04675v1 [cs.CV] 7 Jun 2024

OVMR: Open-Vocabulary Recognition with Multi-Modal References

Zehong Ma[1] Shiliang Zhang[1‡] Longhui Wei[2] Qi Tian[2]
[1]National Key Laboratory for Multimedia Information Processing,
School of Computer Science, Peking University, [2]Huawei Inc.
zehongma@stu.pku.edu.cn, slzhang.jdl@pku.edu.cn, weilh2568@gmail.com, tian.qi1@huawei.com

Abstract

The challenge of open-vocabulary recognition lies in the model has no clue of new categories it is applied to. Existing works have proposed different methods to embed category cues into the model, e.g.*, through few-shot fine-tuning, providing category names or textual descriptions to Vision-Language Models. Fine-tuning is time-consuming and degrades the generalization capability. Textual descriptions could be ambiguous and fail to depict visual details. This paper tackles open-vocabulary recognition from a different perspective by referring to multi-modal clues composed of textual descriptions and exemplar images. Our method, named OVMR, adopts two innovative components to pursue a more robust category cues embedding. A multi-modal classifier is first generated by dynamically complementing textual descriptions with image exemplars. A preference-based refinement module is hence applied to fuse uni-modal and multi-modal classifiers, with the aim to alleviate issues of low-quality exemplar images or textual descriptions. The proposed OVMR is a plug-and-play module, and works well with exemplar images randomly crawled from the Internet. Extensive experiments have demonstrated the promising performance of OVMR,* e.g.*, it outperforms existing methods across various scenarios and setups. Codes are publicly available at https://github.com/Zehong-Ma/OVMR.*

1. Introduction

Open-vocabulary recognition aims to recognize unseen objects beyond the training set. It is challenging because the model has no clue of new categories in the testing set. Besides efforts on pre-training models having strong generalization capability [33, 36, 41], recent works have developed more lightweight strategies by embedding novel category clues into pre-trained backbone models [18, 59]. Among those works, a popular strategy is fine-tuning a generalizable model on a small task-specific dataset [48, 58, 59].

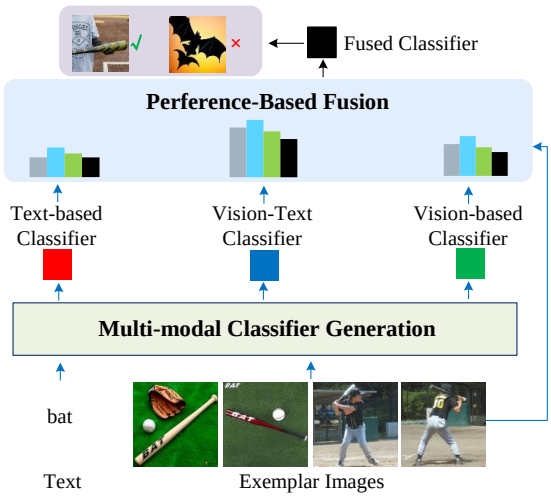


Figure 1. Illustration of the pipeline of our OVMR. It refers to textual description and exemplar images to generate classifiers for novel categories. The textual description could be ambiguous and fail to depict visual details. The exemplar images show diversified qualities. OVMR effectively complements visual and textual features and fuses classifiers to alleviate issues of low-quality exemplar images or textual description.

This few-shot fine-tuning strategy is effective in optimizing task-specific parameters, but is time-consuming, inflexible, and degrades the generalization capabilities.

Another line of research is leveraging the strong generalization capability of Vision-Language Models (VLMs) by providing images or textual descriptions as clues of novel categories [17, 54]. Some works take text embeddings extracted from textual descriptions as the classifier for novel categories [48, 62]. Textual descriptions can be ambiguous and lack detailed descriptions to visual cues. For example, the word "bat" could refer to either a piece of sports equipment or an animal. Those issues scarify the discriminative power of resulting classifiers. Collecting exemplar images could be another option for providing category cues as shown in previous works [17, 50]. However, image samples could show diversified qualities easily affected by issues of domain gaps, cluttered backgrounds, *etc*.

‡Corresponding author.

This paper tackles open-vocabulary recognition from a different perspective by referring to multi-modal clues composed of textual descriptions and exemplar images. In other words, by feeding both text descriptions and exemplar images into the VLM, we aim to mine complementary cues of text and images to learn more robust classifiers for novel categories. As illustrated in Fig. 1, during this procedure, the text modality is expected to provide generalizable semantic cues, while the exemplar images are analyzed to extract visual details, which are critical to the discriminative power of resulting classifiers. To alleviate the negative effects of low-quality text or image exemplars, we also evaluate the performance of those uni-modal and multi-modal classifiers to adaptively generate the final classifier. We name the proposed method as OVMR.

As shown in Fig. 1, OVMR takes textual descriptions and multiple exemplar images depicting a novel category as input. It incorporates two modules to generate the final classifiers. The first module dynamically fuses visual exemplars and textual descriptions to generate the multi-modal classifier. Specifically, it utilizes a lightweight visual token generator to extract visual tokens from given exemplars. Subsequently, the language encoder adaptively fuses the visual and textual tokens by inferring the contextual relationships between them [28, 51]. This multi-modal classifier generation module is efficient thanks to its lightweight structure. It does not need to train class-specific parameters, hence ensuring its good generalizability and scalability to the classes in the wild. This module leads to two uni-modal classifiers, and one multi-modal classifier.

OVMR hence generates the final classifier by fusing the above three classifiers in Fig. 1. To alleviate the negative effects of low-quality classifiers, it presents a dynamic fusion strategy by evaluating their performance. As multiple exemplar images are provided, we leverage them as a validation set to test the performance of each classifier. The preference-based fusion module uses their performance as the clue of learning fusion weights. As shown in Fig. 1, this procedure simulates the testing stage by leveraging exemplar images as the testing set. It effectively guarantees the robustness of the final fused classifier.

We have conducted extensive experiments to test the performance of OVMR. As shown in experiments, it outperforms recent open-vocabulary methods by clear margins on 11 image classification datasets, and the LVIS object detection dataset. OVMR also performs better than closely related works that simply apply naive average fusion [17] and text-guided fusion [50] for novel classifier generation. Our contributions can be summarized into three aspects:

- We present a flexible plug-and-play module to embed clues of novel classes into VLMs to boost their capabilities in open-vocabulary recognition tasks. Complementing multi-modal clues brings substantial advantages over solely relying on vision or textual cues.
- Our OVMR presents a novel pipeline to generate robust classifiers from two-modality inputs. It adaptively fuses text and vision cues to generate multi-modal classifiers, and further proposes a parameter-free fusion module to alleviate negative effects of the low-quality modality.
- Extensive experiments demonstrate the superior performance of our method in both open-vocabulary classification and detection tasks, showcasing the potential of our method in open-vocabulary recognition.

2. Related Work

This work is closely related to open-vocabulary classification and detection. This section briefly reviews recent works in those two lines and discusses our differences with them.

2.1. Open-Vocabulary Classification

Existing open-vocabulary classification methods can be summarized into three categories, *i.e.*, pre-training, prompt learning, and few-shot adaption methods, respectively.

Pre-training Methods. Many pre-training efforts have been made to enhance the capabilities of VLMs in open-vocabulary classification, including large curated datasets[10, 37] and enhanced training strategies [24, 27, 41, 46]. They need to retrain the model from scratch which consumes considerable time, samples, and annotations.

Prompt Learning Methods. To efficiently strengthen the capabilities of VLM in classification, various prompt learning methods have been proposed. CoOp [59] learns static contextual tokens from a few-shot dataset but tends to overfit to the training classes, degrading performance on unseen classes. To mitigate this, CoCoOp [58] acquires dynamic instance-specific tokens from the input image, aiming to improve the classification of unseen classes. MaPLe [18] endeavors to learn multi-modal prompt tokens across different layers for both vision and language branches. These methods require fine-tuning on each downstream dataset, tending to overfit seen classes and lacking the generalization capability as the one in VLMs.

Few-shot Adaptation Methods. The few-shot classification consists of a training phase where a model is learned on a relatively large dataset and an adaptation phase where the learned model is adapted to previously unseen tasks with limited labeled samples. Under this framework, methods can be roughly divided into two groups: meta-learning methods and non-meta-learning methods. A recent work [29] reveals that the training and adaptation phases in few-shot image classification are completely disentangled. Besides, it also demonstrates the visual backbone pretrained with CLIP's training algorithm has superior performance than previous few-shot training algorithms. In our work, we take the pre-trained visual backbone of CLIP as the base model and evaluate different adaptation

methods on top of it. The adaptation methods encompass the ones from training-free methods including MatchingNet [43], Nearest Centroid Classifier(PN) [38], and the ones from training-based methods including MAML [9], Logistic Regression [42], Cosine Classifier [2], URL [26], and CEPA [15].

2.2. Open-Vocabulary Detection

Much recent work aims to transfer the open-vocabulary capabilities of VLMs to object detection [12, 55, 57]. Techniques including knowledge distillation [12] and prompt optimization [5, 48] have been used to train an open-vocabulary detector with the pre-trained VLMs. Weak-labeling and Pseudo-boxes methods [8, 11, 25, 57, 62] have also been proposed to enhance the object-level recognition ability of VLMs. In addition, some works add new detection heads on the top of the pre-trained visual backbone of VLMs or SAM [14], either by keeping the backbone frozen [23] or finetunable [21, 32]. Recently, pre-training the vision-language models for open-vocabulary detection is a new direction. GLIP [25, 56] and DetCLIP [52, 53] train on a combination of a detection, grounding, and caption data to learn the word-region alignment. RO-ViT [21] proposes pretraining region-aware positional embeddings to enhance VLM's capability in dense prediction tasks.

In addition, recent MM-OVOD [17] and MQ-Det [50] introduce exemplar images to enhance the text classifier for open-vocabulary detection. However, MM-OVOD takes two modalities equally and directly calculates the arithmetic mean of the newly learned vision-based classifier and the existing textual classifier in VLMs to obtain the multimodal classifier. MQ-Det uses textual features as queries to extract information from exemplar images and refine the original text classifier with cross-attention mechanisms. This is based on the assumption that the textual modality is more important. However, influenced by the quality of exemplars and text, the preference for the two modalities should be dynamic across different categories.

2.3. Differences with Previous Works

Our OVMR method presents several differences with previous open vocabulary classification and detection methods. First, unlike traditional pre-training methods, which require considerable resources, OVMR involves a lightweight visual token generator pre-trained on a smaller dataset. This enables an efficient integration of new category cues into the model without the need for fully retraining it. Second, our approach effectively circumvents the overfitting issues inherent in prompt learning methods, as it does not learn class-specific parameters. Additionally, the plug-and-play property allows it to transfer seamlessly to various tasks after pre-training. Third, OVMR utilizes the strong generalization capabilities of language models to adaptively fuse multi-modal cues. In contrast to methods like MM-OVOD and MQ-Det, which treat modalities equally or prioritize text, OVMR further dynamically integrates uni-modality classifiers and the multi-modal classifier by evaluating their performance. This two-stage classifier generation pipeline is more robust to scenarios with low-quality exemplars or textual descriptions, making OVMR perform substantially better in various tasks.

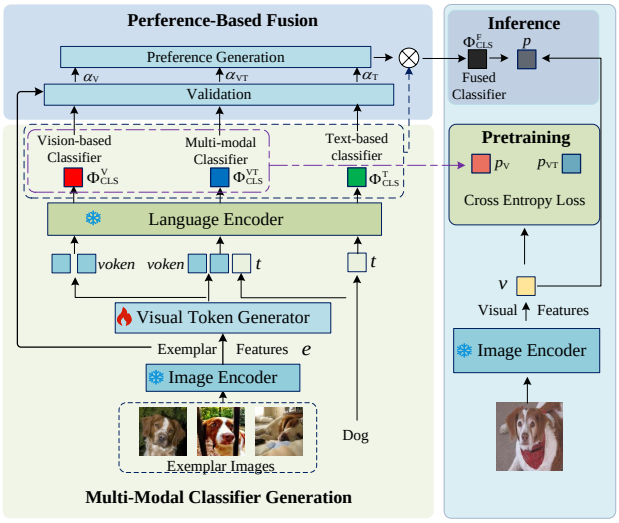


Figure 2. Illustration of the pipeline for novel classifier generation (left) and image classification (right).

3. Methodology

As illustrated in Fig. 2, our OVMR is composed of two principal modules. The first is a multi-modal classifier generation module. This module leverages a generalizable language encoder for dynamically integrating text and visual exemplars. It also includes a newly pre-trained visual token generator that embeds the exemplar images into the language space. The second module is a test-time preference-based fusion module, which does not introduce any trainable parameters. We will introduce the multi-modal classifier generation and the pre-training of the visual token generator in Sec. 3.1 and the test-time preference-based fusion in Sec. 3.2. Sec. 3.3 discusses the adaption of OVMR to the open-vocabulary detection task.

3.1. Multi-modal Classifier Generation

The multi-modal classifier generation module aims to generate a multi-modal classifier by adaptively fusing visual exemplars and textual descriptions. For a new category of interest C_i, we denote its visual exemplars, target images, and textual tokens as $E_i \in \mathbb{R}^{M\times H\times W\times 3}$, $V_i \in \mathbb{R}^{N\times H\times W\times 3}$ and $t_i \in \mathbb{R}^{L_i\times d}$, respectively, where M is the number of exemplar images, N is the number of target images, L_i is the length of textual tokens belonging to C_i category and d

is the hidden dimension of token embeddings. Furthermore, the visual exemplars E_i and target images V_i are encoded into exemplar features e_i and visual features v_i by the CLIP image encoder $\Phi_{\text{CLIP-V}}$.

In order to utilize the language encoder to generate a good multi-modal classifier, a key prerequisite is to extract robust visual tokens that the language encoder can understand from visual exemplars. These visual tokens need to accurately represent the class-discriminative visual details. The visual token generator consists of P class-agnostic learnable queries $q \in \mathbb{R}^{P\times d}$ and four layers of transformer blocks with global self-attention. Leveraging the self-attention interaction between the learnable queries and exemplar features, the learnable queries q adaptively extract visual tokens of the C_i category from exemplar features e_i. This process can be formulated as:

$$\text{voken}_i = \Phi_{\text{VOK}}([q, e_i]), \tag{1}$$

where $[\cdot]$ represents word-level concatenation, $\text{voken}_i \in \mathbb{R}^{P\times d}$ is the corresponding output of query tokens q.

Then, we use the language encoder to analyze the relationships between visual and textual tokens and adaptively generate the weight w^i_{VT} of the multi-modal classifier for the C_i category as:

$$w^i_{\text{VT}} = \Phi_{\text{CLIP-T}}([\text{voken}_i, t_i]). \tag{2}$$

The multi-modal classifier $\Phi^{\text{VT}}_{\text{CLS}}$ thus can be represented as:

$$\Phi^{\text{VT}}_{\text{CLS}}(v) = \frac{\exp(\text{sim}(v, w_{\text{VT}})/\tau_t)}{\sum_{k=1}^{C}\exp(\text{sim}(v, w^k_{\text{VT}})/\tau_t)}, \tag{3}$$

where $\text{sim}(\cdot,\cdot)$ denotes cosine similarity and τ_t is the temperature.

Pre-training of the Visual Token Generator. The visual token generator is the only trainable component in our OVMR, which extract class-discriminative visual information from exemplar images and affects the performance of the multimodal classifier. Thus it's important to make sure that the visual tokens can encompass the class-discriminative visual details as much as possible. We separately input the visual tokens into the language encoder to generate the vision-based classifier $\Phi^{\text{V}}_{\text{CLS}}$, and then optimize the vision-based classifier with the multi-modal classifier together. The prediction probability of target images v over the vision-based and multi-modal classifiers can be computed as:

$$p_{\text{V}} = \Phi^{\text{V}}_{\text{CLS}}(v), \quad p_{\text{VT}} = \Phi^{\text{VT}}_{\text{CLS}}(v). \tag{4}$$

The overall pre-training objective of the visual token generator can be represented by:

$$\mathcal{L} = \text{CE}(p_{\text{V}}, \mathcal{T}) + \text{CE}(p_{\text{VT}}, \mathcal{T}), \tag{5}$$

where $\text{CE}(\cdot)$ denotes the cross-entropy loss, $\mathcal{T}$ are the ground-truth labels of target images.

To ensure the generalizability of the visual token generator, we design an effective pre-training strategy. During each training iteration, We randomly sample K-shot images for each category from the ImageNet21k-OVR, a subset of ImageNet21k detailed in Sec. 4.1. From these images, M images are randomly selected as visual exemplars for each category, while the remaining N=K-M images are used as target images. The number of exemplars M is varied randomly within the range $[K/4, 3K/4]$ to simulate diverse scenarios encountered in practical applications. It is important to ensure that there is no overlap between exemplar and target images, which guides the model to learn class-discriminative features rather than instance-specific details. Additionally, we have incorporated techniques such as random path dropout in the attention layer and channel-wise dropout in each transformer block of the visual token generator to further enhance its generalizability.

3.2. Preference-Based Fusion

The preference-based fusion aims to simulate the test-time validation results on exemplar images to measure the preference of text-based, vision-based, and multi-modal classifiers. It hence generates a more powerful fused classifier based on the estimated preference. And the text-based classifier $\Phi^{\text{T}}_{\text{CLS}}$ can be acquired conveniently by feeding textual tokens into a language encoder.

The preference-based fusion process begins by validating different classifiers on exemplar images. Similar to Eq. (4), by inputting the exemplar features e into various classifiers, we obtain the prediction probabilities of exemplar images over different classifiers. These probabilities, $\hat{p}_{\text{VT}}$, $\hat{p}_{\text{V}}$, and $\hat{p}_{\text{T}}$, correspond to the multi-modal, vision-based, and text-based classifiers, respectively. Then, based on these prediction probabilities, we can get each category's preferences for different classifiers:

$$\begin{aligned} \alpha_{\text{T}} &= \Omega(\hat{p}_{\text{T}}, \mathcal{T}_E), \\ \alpha_{\text{V}} &= \Omega(\hat{p}_{\text{V}}, \mathcal{T}_E), \\ \alpha_{\text{VT}} &= \Omega(\hat{p}_{\text{VT}}, \mathcal{T}_E), \end{aligned} \tag{6}$$

where $\alpha_{\text{T}}, \alpha_{\text{V}}, \alpha_{\text{VT}} \in \mathbb{R}^{|C|\times 1}$ denotes the preferences for text-based, vision-based, and multi-modal classifiers, separately. $\mathcal{T}_E$ are the ground-truth labels of exemplar images. Ω denotes the chosen evaluation metric, which is the F1 score [39] in open-vocabulary classification tasks. We select the F1 score because it stably reflects the quality of a classifier by comprehensively considering both precision and recall. The preferences for different classifiers can be denoted as:

$$[\hat{\alpha}_{\text{V}}, \hat{\alpha}_{\text{VT}}, \hat{\alpha}_{\text{T}}] = \sigma(\tau_p[\alpha_{\text{V}}, \alpha_{\text{VT}}, \alpha_{\text{T}}]), \tag{7}$$

where σ denotes softmax function, τ_p is the temperature for preference generation, and $[\cdot]$ represent concatenation at the last dimension.

Then the fused classifier of the multi-modal classifier and uni-modal classifiers can be formulated as:

$$\Phi^F_{\text{CLS}} = \hat{\alpha}_{\text{V}} \cdot \Phi^{\text{V}}_{\text{CLS}} + \hat{\alpha}_{\text{VT}} \cdot \Phi^{\text{VT}}_{\text{CLS}} + \hat{\alpha}_{\text{T}} \cdot \Phi^{\text{T}}_{\text{CLS}}. \quad (8)$$

We hence can get the final prediction probability p of the target images over the fused classifier as follows:

$$p = \Phi^F_{\text{CLS}}(v). \quad (9)$$

The preference-based fusion effectively leverages exemplar images to boost the recognition ability of VLMs, without introducing any trainable parameters.

3.3. Adaptation to Open-Vocabulary Detection

Our method is not limited to the open-vocabulary classification. It can be flexibly adapted to other open vocabulary recognition tasks like detection. In this work, we make use of a popular multi-stage detector based on CenterNet2 [61] as done in Detic [62] and MM-OVOD [17]. For simplicity, we consider the two-stage variant of this classifier. Using $\{b_j, p^j\}_{j=1}^H$ to denote its output, and H as the number of proposals, it can be formulated as:

$$\{r_j\}_{j=1}^H = \Phi_{\text{ROI}} \circ \Phi_{\text{PG}} \circ \Phi_{\text{ENC}}(I), \quad (10)$$

$$\{b_j, p^j\}_{j=1}^H = \{\Phi_{\text{BBOX}}(r_j), \Phi_{\text{CLS}} \circ \Phi_{\text{PROJ}}(r_j)\}_{j=1}^H, \quad (11)$$

where each input image I is first sequentially processed by a set of operations: a trainable image encoder (Φ_{ENC}), a proposal generator (Φ_{PG}), a region-of-interest (RoI) feature pooling module (Φ_{ROI}), finally yielding a set of RoI features $\{r_j\}_{j=1}^H$. The RoI features are processed by a bounding box module (Φ_{BBOX}) to infer position of objects, $\{b_j\}_{j=1}^H$.

Additionally, the RoI features are processed by a classification module, consisting of a linear projection (Φ_{PROJ}), and multi-modal classifiers $\left(\Phi_{\text{CLS}} \in \{\Phi^{\text{VT}}_{\text{CLS}}, \Phi^{\text{V}}_{\text{CLS}}, \Phi^{\text{T}}_{\text{CLS}}\}\right)$, yielding prediction probabilities of RoI features, $\{p^j\}_{j=1}^H$, where $p^j \in \{p^j_{\text{VT}}, p^j_{\text{V}}, p^j_{\text{T}}\}$.

During training, the multi-modal, vision-based, and text-based classifiers are kept frozen while other components in the detector are trainable. We incorporate a sigmoid cross entropy loss for each classifier. In preference-based fusion, the recently trained detector, along with the multi-modal classifier, is utilized to identify the most accurate pseudo box within the ground-truth label for each exemplar image. Subsequently, with these pseudo annotations, we calculate the Average Precision (AP) for each class, which aids in computing the preference for different classifiers.

4. Experiments

4.1. Datasets

Classification: Following prompt learning methods [18, 58], we use the 11 image classification datasets, which cover a diverse set of recognition tasks. Specifically, the benchmark includes ImageNet [4] and Caltech101 [7] for classification on generic objects; OxfordPets [35], StanfordCars(Cars) [22], Flowers102 [34], Food101 [1] and FGVCAircraft(Aircraft) [31] for fine-grained classification; SUN397 [49] for scene recognition; UCF101 [40] for action recognition; DTD [3] for texture classification; and finally EuroSAT [16] for satellite imagery recognition.

Detection: In the open-vocabulary detection, the LVIS [13] dataset is used following previous works [12, 62], which contains 100K images with 1,203 classes. The classes are divided into three groups, namely frequent, common and rare, based on the number of training images. We treat 337 rare classes as novel classes and use the frequent and common classes as base classes for training. When using image classification data as extra weak supervision, we use the subset of categories in ImageNet21k that overlap with the LVIS vocabulary and denote this subset as IN-L, as in Detic [62].

Pre-training Set: We construct the pre-training dataset based on ImageNet21k. To prevent data leakage, we remove any overlapping categories present in both the 11 classification datasets and the LVIS dataset from ImageNet21k. Additionally, we limit the number of images per category to 64 to improve pre-training efficiency. As a result, we create a 64-shot subset of ImageNet21k, termed ImageNet21k-OVR. This subset contains 18,631 categories and encompasses a total of 1.1 million images, which is much smaller than the dataset used for the VLM's pre-training.

4.2. Implementation Details

Open-Vocabulary Classification. Following prompt learning methods [18, 58], we select ViT-B/16 of CLIP as our base model and pre-train a plug-and-play visual token generator to enhance its recognition capability. The sample number K of each class is set to 8 and we sample 192 classes per batch, which results in a total batch size of 1536. We pre-train the visual token generator on ImageNet21k-OVR for 30 epochs in 12 hours on a single 3090 GPU. We adopt an Adam optimizer and a cosine learning rate scheduler, where the learning rate is set to 0.0002. The number P of visual tokens is set to 2. The τ_p in preference-based fusion is set to 10.

Open-Vocabulary Detection. The architecture and training recipe is almost identical to that in Detic [62] and MM-OVOD [17], using the CenterNet2 model with a ResNet-50 or Swin-B backbone pre-trained on ImageNet21k-P. It's worth noting that we re-implement a memory-efficient ver-

Table 1. Open-Vocabulary Classification Results in Prompt Learning Setup. Our method achieves comparable performance to the SoTA method using the same 16-shot exemplar images without any extra training. †: integrating CoOP's prompt tokens into our method.

Methods	Extra Training	ImageNet	Caltech101	OxfordPets	Cars	Flowers102	Food101	Aircraft	SUN397	DTD	EuroSAT	UCF101	Average
CLIP [36]	×	72.43	96.84	91.17	63.37	72.08	90.10	27.19	69.36	53.24	56.48	70.53	69.34
CoOp [59]	✓	76.47	98.00	93.67	78.12	97.60	88.33	40.44	80.60	79.44	92.19	84.69	82.69
CoCoOp [58]	✓	75.98	97.96	95.20	70.49	94.87	90.70	33.41	79.74	77.01	87.49	82.33	80.47
MaPLe [18]	✓	76.66	97.74	**95.43**	72.94	95.92	**90.71**	37.44	80.82	80.36	**94.07**	83.00	82.28
OVMR	×	76.77	98.00	94.97	73.93	97.83	89.93	40.37	81.83	77.10	90.00	85.03	82.34
OVMR†	✓	**76.87**	**98.27**	94.23	**79.67**	**98.53**	89.13	**43.07**	**82.40**	**80.37**	93.07	**85.30**	**83.72**

Table 2. Open-Vocabulary Classification Results in Few-shot Setup. Our method shows superior performance on average.

Methods	Extra Training	ImageNet	Caltech101	OxfordPets	Cars	Flowers102	Food101	Aircraft	SUN397	DTD	EuroSAT	UCF101	Average
MAML [9]	✓	56.93	92.97	33.97	59.10	76.50	67.83	26.07	41.67	62.60	83.80	69.17	60.96
MatchingNet [43]	×	55.87	95.57	75.53	58.17	91.83	79.97	33.60	68.17	66.33	82.73	75.43	71.20
LR [42]	✓	63.67	97.00	84.60	66.07	94.90	84.70	37.93	75.37	74.43	88.40	80.57	77.06
PN [38]	×	65.77	96.13	83.37	76.30	96.13	84.93	40.43	76.30	72.20	86.87	80.53	77.32
CEPA [15]	✓	69.00	97.10	88.10	71.70	96.10	85.20	39.10	78.10	73.70	90.70	80.40	79.02
CC [2]	✓	70.83	97.23	86.53	75.27	96.97	87.07	41.30	79.53	76.83	91.07	83.00	80.51
URL [26]	✓	72.07	97.77	89.27	**78.17**	97.23	87.40	**44.53**	79.97	**80.77**	**92.00**	82.13	82.13
OVMR	×	**76.77**	**98.00**	**94.97**	73.93	**97.83**	**89.93**	40.37	**81.83**	77.10	90.00	**85.03**	**82.34**

sion of Detic by limiting the maximum number of ground-truth boxes per image to 10 following [44, 60] in order to reproduce Detic's experiments on four 24GB 3090 GPUs. Based on the memory-efficient version, we replace the original classifier with our proposed multi-modal classifiers and further introduce the preference-based fusion. Besides, for a fair comparison, following previous works [62], we pre-train a new visual token generator for ViT-B/32 of CLIP on ImageNet21k-OVR with LVIS base categories. In system-level implementation, we take Swin-B as our backbone and train the detector with additional image-labeled data(IN-L) following Detic [62].

Training and Test Setups: In open-vocabulary classification, we select the first half of the categories in the classification dataset as base classes and conduct comparisons on base classes for a fair comparison with existing methods. In the prompt learning setup, during inference, we utilize the same 16-shot training images used in previous methods to serve as exemplar images. Prompt learning methods need to fine-tune for each downstream dataset using these 16-shot images, while our OVMR does not require any extra training. The test set is also kept the same as prompt learning methods. In the traditional few-shot setup, following the recent work [29], we take the visual encoder of CLIP as the base model and evaluate different few-shot adaptation algorithms in the same 16-shot setup as prompt learning methods. In open-vocabulary detection, the exemplar images are the same as those in MM-OVOD [17], which provides 5-shot exemplar images for each category. The main evaluation metric is the mask AP averaged over the "rare" classes, which is denoted as mask AP_r.

4.3. Comparison with Recent Works

Comparison with Prompt Learning Methods. In Tab. 1, we compare our method with existing prompt learning methods on 11 classification datasets. Our methods yield performance comparable to the state-of-the-art method CoOp without any extra training in downstream tasks. It's worth noting that the pre-training dataset, ImageNet21k-OVR, used for the visual token generator has been curated to exclude the categories from the 11 classification datasets and LVIS. The categories in these evaluation datasets are only seen at test time. Providing a few exemplar images during inference, our method manages to match the performance of existing prompt learning methods. Further, by integrating task-specific prompt tokens used in CoOp into our approach and fine-tuning these tokens under the same conditions as CoOp, we observe an enhanced performance, which surpasses CoOp by an average accuracy of 1.03% across the 11 datasets. These results demonstrate that our method can not only enhance open-vocabulary classification without additional training but also be compatible with other optimization-based prompt learning approaches.

Comparison with Traditional Few-shot Methods. We compare our method with few-shot methods in Tab. 2. Our training-free method exhibits superior performance over the state-of-the-art training-based method URL, both on average and across the majority of classification datasets. Particularly in datasets encompassing a broad range of categories, our method significantly outperforms URL, as evidenced by a notable margin of 4.7% in ImageNet. Conversely, in datasets with a few classes, like EuroSAT, URL shows superior performance. This is because supervised fine-tuning tends to be more effective in easy scenarios with

Table 3. LVIS Open-Vocabulary Detection. ‡: reporting box AP. "OVMR_T": pure text-based classifier.

Method	Detector Backbone	Extra Data	AP_r	AP
ResNet50 Comparison:				
ViLD-Ens [12]	RN50	×	16.6	25.5
OV-DETR [54]	RN50	×	17.4	26.6
Detic [62]	RN50	×	17.8	26.8
F-VLM [23]	RN50	×	18.6	24.2
PromptDet [8]	RN50	×	19.0	21.4
BARON [47]	RN50	×	19.2	26.5
DetPro [6]	RN50	×	19.8	25.9
MM-OVOD$_\text{T}$ [17]	RN50	×	19.3	30.3
MM-OVOD [17]	RN50	×	19.3	30.6
OVMR_T	RN50	×	19.0	29.6
OVMR	RN50	×	**21.2**	30.0
System-level Comparison:				
RegionCLIP [57]	RN50x4(87M)	✓	22.0	32.3
CondHead [45]	RN50x4(87M)	×	24.4	32.0
OWL-ViT‡ [32]	ViT-L/14(303M)	✓	25.6	34.7
F-VLM [23]	RN50x4(87M)	×	26.3	28.5
RO-ViT [21]	ViT-B/16(86M)	✓	28.0	30.2
CORA‡ [48]	RN50x4(87M)	✓	28.1	-
CFM-ViT [19]	ViT-B/16(86M)	✓	28.8	32.0
CoDet [30]	Swin-B(88M)	✓	29.4	39.2
RO-ViT [21]	ViT-L/16(303M)	✓	32.1	34.0
DITO [20]	ViT-B/16(86M)	✓	32.5	34.0
Detic [62]	Swin-B(88M)	✓	33.8	40.7
OVMR_T	Swin-B(88M)	✓	33.3	40.8
OVMR	Swin-B(88M)	✓	**34.4**	40.9

fewer classes. In contrast with the training-free method PN, our approach shows a substantial performance gain in complex datasets with hundreds of classes. For instance, on ImageNet, we observe an impressive improvement of 11.0%.

Comparison with Open-Vocabulary Detection Methods. Tab. 3 presents our open-vocabulary detection results on LVIS. In the ResNet50 comparison, our OVMR, which utilizes multi-modal classifiers and preference-based fusion, outperforms the existing approach MM-OVOD using the same exemplar images by 1.9% mask AP_r on novel categories. In system-level comparison, follow Detic [62], OVMR achieves a superior performance with additional weak supervision from IN-L, which outperforms recent pretraining-based method DITO by 1.9% on mask AP_r on novel categories. Using only the text classifier, we achieve a mask AP_r of 33.3%, which is marginally lower by 0.5% compared to the 33.8% mask AP_r reported by Detic. However, our multi-modal version, OVMR, reaches 34.4% mask AP_r, surpassing the original Detic by 0.6% in mask AP_r. Moreover, OVMR_T and OVMR share the same detector parameters. When there are no exemplar images, our method can still work with only textual descriptions.

4.4. Ablation Study

Effectiveness of Our Proposed Components. In Tab. 4, we systematically evaluate each component of our proposal through ablation studies on 11 classification tasks. Our baseline method, OVMR_T, employs only a text-based classifier. In contrast, OVMR_V utilizes solely a vision-based classifier, while OVMR_VT integrates a multi-modal classifier. A comparative analysis of OVMR_VT against OVMR_T and OVMR_V reveals the superior performance of the multi-modal classifier. OVMR_VT^- takes K sampled images both as exemplars and target images and removes the dropout operations. The performance disparity between OVMR_VT and OVMR_VT^- underscores the effectiveness of our pre-training strategy depicted in Sec. 3.1. As demonstrated in $\text{OVMR}_{\text{T+V, T+VT, V+VT}}$ and OVMR, the preference-based fusion of different classifiers further elevates performance, achieving a best average accuracy of 82.34% across 11 datasets when fusing all three classifiers. This marks a significant 15.54% improvement over the baseline text-based classifier. Parallel ablation studies and consistent results for our components in open-vocabulary detection are presented in Tab. 5.

Table 4. Ablation study of each component in open-vocabulary classification. "T": text-based classifier. "V": vision-based classifier. "PS": pre-training strategy of the visual token generator. "VT": multi-modal classifier. "Fusion": preference-based fusion.

Methods	T	V	PS	VT	Fusion	Average
OVMR_T	✓					66.80
OVMR_V		✓	✓			80.28
OVMR_VT^-				✓		77.86
OVMR_VT			✓	✓		80.99
$\text{OVMR}_\text{T+V}$	✓	✓			✓	81.71
$\text{OVMR}_\text{T+VT}$	✓		✓	✓	✓	81.30
$\text{OVMR}_\text{V+VT}$		✓	✓	✓	✓	81.96
OVMR	✓	✓	✓	✓	✓	**82.34**

Table 5. Ablation study of each component in open-vocabulary detection. AP_r is reported.

Backbone	OVMR_T	OVMR_V	OVMR_VT	OVMR
RN50	19.0	15.9	21.0	**21.2**
Swin-B	33.3	31.9	33.8	**34.4**

Superiority of Our Proposed Fusion. Our fusion method is compared against static fusion with arithmetic mean and text-based dynamic fusion, as detailed in Tab. 6. Our dynamic fusion method outperforms the static fusion method MM-OVOD, achieving a 2.4% improvement despite MM-OVOD's better vision-based classifier performance. Text-guided dynamic fusion method MQ-Det performs the worst, as the quality of the text query may be unreliable and this fusion largely disrupts the original features of CLIP. The results in Tab. 6 demonstrate the superiority of our fusion method. In Tab. 7, we further compare our fusion method with MM-OVOD and MQ-Det in open-vocabulary detection. When using the same 5-shot exemplar images. Our OVMR outperforms these methods in mask AP_r by a notable margin.

Evaluation Metric in Preference Generation. In Tab. 8, directly averaging the predictions of three classifiers with static "Mean" can improve the accuracy to 81.45%, which

Table 6. Comparative analysis of fusion methods in classification. "CLIP-F": Forming a vision-based classifier by averaging CLIP's visual features of exemplar images and statically averaging it with text-based classifier. "MM-OVOD": Pre-training a discriminative vision-based classifier and statically averaging it with text-based classifier. "MQ-Det": Utilizing the text feature as a query to dynamically aggregate related visual details from exemplar images.

Fusion Methods	Dynamic	Text	Vision	Multi-Modal
CLIP-F	×	66.80	77.32	79.92
MM-OVOD [17]	×	66.80	80.44	79.94
MQ-Det [50]	✓	66.80	-	62.60
OVMR	✓	66.80	80.28	**82.34**

Table 7. Comparison with other methods using exemplar images in open-vocabulary detection.

Fusion Methods	Dynamic	Backbone	AP_r	mAP
MM-OVOD [17]	×	ResNet-50	19.3	30.6
MQ-Det [50]	✓	Swin-T	15.4	22.6
OVMR	✓	ResNet-50	**21.2**	30.0

Table 8. Comparison between different evaluation metrics in open-vocabulary classification. "Mean": treating three classifiers as equally important, assigning them identical preferences.

$OVMR_{VT}$	Mean	Precision	Recall	F1
80.99	81.45	82.03	81.70	**82.34**

shows that uni-modal classifiers contain some distinct features that are beneficial for classification. Besides, the accuracy of the F1 metric is superior to other evaluation metrics, because recall and precision cannot comprehensively evaluate the quality of a classifier.

4.5. Analysis of Exemplar Images

The Number of Exemplar Images. As shown in Fig. 3, with increasing the shot number of exemplar images in open-vocabulary classification, the average performance of classifiers using visual clues steadily improves. However, when it comes to 16 or 32 shots, the performance of each classifier with visual clues tends to saturate, and there is almost no improvement when increased to 64 shots. It indicates that for most classification tasks, 16-shot exemplars are enough to efficiently boost the recognition capability. In addition, in the 2-shot setting, even though the performance of the vision-based classifier is close to that of the text-based classifier, a significant improvement can be achieved using our multi-modal classifier generation module, which further demonstrates the effectiveness of our method. However, there is a small drop after introducing the preference-based fusion in the 2-shot setting. It is an inherent limitation of preference-based fusion, as fewer images may not robustly evaluate each category's preference for various classifiers. But considering that it is easy to obtain multiple exemplars of one category from the Internet, we can alleviate the negative influence by collecting more exemplars.

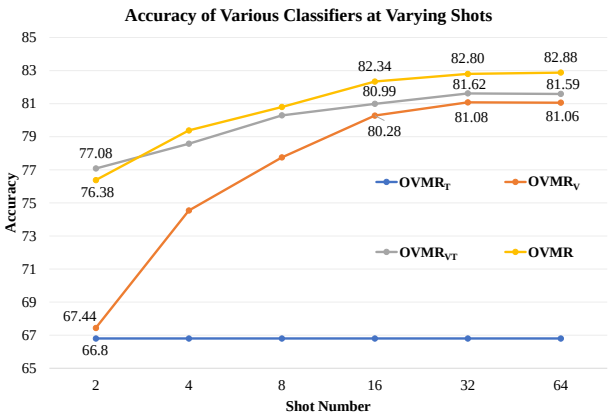


Figure 3. The accuracy of various classifiers at varying shots.

Table 9. Evaluate our method on ImageNet with exemplar images from different sources. $OVMR_T$ is the zero-shot accuracy of CLIP.

Source	$OVMR_T$	$OVMR_V$	$OVMR_{VT}$	OVMR
Internet	66.80	69.83	73.97	**74.56**
ImageNet	66.80	72.63	76.50	**76.77**

Sources of Exemplar Images. We also evaluate our OVMR on ImageNet in a 16-shot setting with the exemplar images crawled from the Internet. The results in Tab. 9 indicate that leveraging the uncurated, readily available exemplar images from the Internet can also significantly enhance open-vocabulary recognition by a notable improvement.

5. Conclusion

In this paper, we propose a plug-and-play method, OVMR, to embed multi-modal clues of novel classes into VLMs to enhance their capability in open-vocabulary recognition. It initially utilizes the multi-modal classifier generation module to embed the exemplar images into visual tokens and then adaptively fuse multi-modal cues by inferring their contextual relationships with the language encoder. To alleviate the negative effects of the low-quality modality, we further propose a parameter-free fusion module to dynamically integrate the multi-modal classifier with two uni-modal classifiers by each category's specific preference for these classifiers. Extensive experiments validate our method's superior performance in both open-vocabulary classification and detection tasks, underscoring its potential to advance open-vocabulary recognition.

Acknowledgement This work is supported in part by Natural Science Foundation of China under Grant No. U20B2052, 61936011, in part by the Okawa Foundation Research Award.

References

[1] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *ECCV*, 2014. 5

[2] Wei-Yu Chen, Yen-Cheng Liu, Zsolt Kira, Yu-Chiang Frank Wang, and Jia-Bin Huang. A closer look at few-shot classification. *arXiv preprint arXiv:1904.04232*, 2019. 3, 6

[3] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *CVPR*, 2014. 5

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 5

[5] Yu Du, Fangyun Wei, Zihe Zhang, Miaojing Shi, Yue Gao, and Guoqi Li. Learning to prompt for open-vocabulary object detection with vision-language model. In *CVPR*, 2022. 3

[6] Yu Du, Fangyun Wei, Zihe Zhang, Miaojing Shi, Yue Gao, and Guoqi Li. Learning to prompt for open-vocabulary object detection with vision-language model. In *CVPR*, 2022. 7

[7] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *CVPR-W*, 2004. 5

[8] 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. 3, 7

[9] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pages 1126–1135. PMLR, 2017. 3, 6

[10] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. Datacomp: In search of the next generation of multimodal datasets. *arXiv preprint arXiv:2304.14108*, 2023. 2

[11] Mingfei Gao, Chen Xing, Juan Carlos Niebles, Junnan Li, Ran Xu, Wenhao Liu, and Caiming Xiong. Open vocabulary object detection with pseudo bounding-box labels. In *ECCV*, 2022. 3

[12] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. In *ICLR*, 2022. 3, 5, 7

[13] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5356–5364, 2019. 5

[14] Xumeng Han, Longhui Wei, Xuehui Yu, Zhiyang Dou, Xin He, Kuiran Wang, Zhenjun Han, and Qi Tian. Boosting segment anything model towards open-vocabulary learning. *arXiv preprint arXiv:2312.03628*, 2023. 3

[15] Fusheng Hao, Fengxiang He, Liu Liu, Fuxiang Wu, Dacheng Tao, and Jun Cheng. Class-aware patch embedding adaptation for few-shot image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 18905–18915, 2023. 3, 6

[16] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2019. 5

[17] Prannay Kaul, Weidi Xie, and Andrew Zisserman. Multi-modal classifiers for open-vocabulary object detection. *arXiv preprint arXiv:2306.05493*, 2023. 1, 2, 3, 5, 6, 7, 8

[18] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19113–19122, 2023. 1, 2, 5, 6

[19] Dahun Kim, Anelia Angelova, and Weicheng Kuo. Contrastive feature masking open-vocabulary vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15602–15612, 2023. 7

[20] Dahun Kim, Anelia Angelova, and Weicheng Kuo. Detection-oriented image-text pretraining for open-vocabulary detection. *arXiv preprint arXiv:2310.00161*, 2023. 7

[21] Dahun Kim, Anelia Angelova, and Weicheng Kuo. Region-aware pretraining for open-vocabulary object detection with vision transformers. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3, 7

[22] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *ICCV-W*, 2013. 5

[23] Weicheng Kuo, Yin Cui, Xiuye Gu, AJ Piergiovanni, and Anelia Angelova. F-vlm: Open-vocabulary object detection upon frozen vision and language models. *ICLR*, 2023. 3, 7

[24] Juncheng Li, Xin He, Longhui Wei, Long Qian, Linchao Zhu, Lingxi Xie, Yueting Zhuang, Qi Tian, and Siliang Tang. Fine-grained semantically aligned vision-language pre-training. *Advances in neural information processing systems*, 35:7290–7303, 2022. 2

[25] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. Grounded language-image pre-training. In *CVPR*, 2022. 3

[26] Wei-Hong Li, Xialei Liu, and Hakan Bilen. Universal representation learning from multiple domains for few-shot classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9526–9535, 2021. 3, 6

[27] Yanghao Li, Haoqi Fan, Ronghang Hu, Christoph Feichtenhofer, and Kaiming He. Scaling language-image pre-training via masking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23390–23400, 2023. 2

[28] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023. 2

[29] Xu Luo, Hao Wu, Ji Zhang, Lianli Gao, Jing Xu, and Jingkuan Song. A closer look at few-shot classification

again. In *Proceedings of the 40th International Conference on Machine Learning*. JMLR.org, 2023. 2, 6

[30] Chuofan Ma, Yi Jiang, Xin Wen, Zehuan Yuan, and Xiaojuan Qi. Codet: Co-occurrence guided region-word alignment for open-vocabulary object detection. *arXiv preprint arXiv:2310.16667*, 2023. 7

[31] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. 5

[32] Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, Xiao Wang, Xiaohua Zhai, Thomas Kipf, and Neil Houlsby. Simple open-vocabulary object detection with vision transformers. In *ECCV*, 2022. 3, 7

[33] Zixuan Ni, Longhui Wei, Siliang Tang, Yueting Zhuang, and Qi Tian. Continual vision-language representation learning with off-diagonal information. In *International Conference on Machine Learning*, pages 26129–26149. PMLR, 2023. 1

[34] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *ICVGIP*, 2008. 5

[35] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *CVPR*, 2012. 5

[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. 1, 6, 2

[37] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade W Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa R Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5b: An open large-scale dataset for training next generation image-text models. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. 2

[38] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30, 2017. 3, 6

[39] Marina Sokolova and Guy Lapalme. A systematic analysis of performance measures for classification tasks. *Information processing & management*, 45(4):427–437, 2009. 4

[40] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 5

[41] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023. 1, 2

[42] Yonglong Tian, Yue Wang, Dilip Krishnan, Joshua B Tenenbaum, and Phillip Isola. Rethinking few-shot image classification: a good embedding is all you need? In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, pages 266–282. Springer, 2020. 3, 6

[43] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. *Advances in neural information processing systems*, 29, 2016. 3, 6

[44] Dongkai Wang and Shiliang Zhang. Contextual instance decoupling for instance-level human analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 6

[45] Tao Wang. Learning to detect and segment for open vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7051–7060, 2023. 7

[46] Longhui Wei, Lingxi Xie, Wengang Zhou, Houqiang Li, and Qi Tian. Mvp: Multimodality-guided visual pre-training. In *European Conference on Computer Vision*, pages 337–353. Springer, 2022. 2

[47] Size Wu, Wenwei Zhang, Sheng Jin, Wentao Liu, and Chen Change Loy. Aligning bag of regions for open-vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15254–15264, 2023. 7

[48] Xiaoshi Wu, Feng Zhu, Rui Zhao, and Hongsheng Li. Cora: Adapting clip for open-vocabulary detection with region prompting and anchor pre-matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7031–7040, 2023. 1, 3, 7

[49] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *CVPR*, 2010. 5

[50] Yifan Xu, Mengdan Zhang, Chaoyou Fu, Peixian Chen, Xiaoshan Yang, Ke Li, and Changsheng Xu. Multi-modal queried object detection in the wild. *arXiv preprint arXiv:2305.18980*, 2023. 1, 2, 3, 8

[51] Shiyu Xuan, Qingpei Guo, Ming Yang, and Shiliang Zhang. Pink: Unveiling the power of referential comprehension for multi-modal llms. *arXiv preprint arXiv:2310.00582*, 2023. 2

[52] 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. In *arXiv:2209.09407*, 2022. 3

[53] Lewei Yao, Jianhua Han, Xiaodan Liang, Dan Xu, Wei Zhang, Zhenguo Li, and Hang Xu. Detclipv2: Scalable open-vocabulary object detection pre-training via word-region alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23497–23506, 2023. 3

[54] Yuhang Zang, Wei Li, Kaiyang Zhou, Chen Huang, and Chen Change Loy. Open-vocabulary DETR with conditional matching. In *ECCV*, pages 106–122. Springer, 2022. 1, 7

[55] Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. Open-vocabulary object detection using captions. In *CVPR*, 2021. 3

[56] Haotian Zhang, Pengchuan Zhang, Xiaowei Hu, Yen-Chun Chen, Liunian Li, Xiyang Dai, Lijuan Wang, Lu Yuan, Jenq-Neng Hwang, and Jianfeng Gao. Glipv2: Unifying localization and vision-language understanding. *Advances in Neural Information Processing Systems*, 35:36067–36080, 2022. 3

[57] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, and Jianfeng Gao. Regionclip: Region-based language-image pretraining. In *CVPR*, 2022. 3, 7

[58] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16816–16825, 2022. 1, 2, 5, 6

[59] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. 1, 2, 6

[60] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. *arXiv preprint arXiv:1904.07850*, 2019. 6

[61] Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. Probabilistic two-stage detection. *arXiv preprint arXiv:2103.07461*, 2021. 5

[62] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In *ECCV*, 2022. 1, 3, 5, 6, 7

OVMR: Open-Vocabulary Recognition with Multi-Modal References

Supplementary Material

A. Ablation Study

A.1. The Number of Visual Tokens

Tab. 10 demonstrates the impact of different numbers of visual tokens on the multi-modal and fused classifiers. We observe the highest average performance for both classifiers across 11 classification datasets when P is set to 2. Consequently, the number of visual tokens is set to 2.

Table 10. Impact of different numbers of visual tokens on average accuracy across 11 classification datasets.

P	OVMR_{VT}	OVMR
1	80.69	82.16
2	**80.99**	**82.34**
4	80.95	82.23
8	80.54	82.17

A.2. Temperature in Preference Generation

Tab. 11 illustrates that setting τ_p to 10 yields the best average performance for the fused classifier. Furthermore, the performance variations under different τ_p values are not significant, suggesting that our method is insensitive to τ_p. This observation validates the robustness of our proposed preference-based fusion approach. Based on these findings, we set τ_p to 10 in our experiments.

Table 11. Average accuracy of the fused classifier with different τ_p across 11 classification datasets.

τ_p	OVMR
1	81.82
10	**82.34**
20	82.07

B. Comparison on Novel Sets

By embedding multi-modal clues of novel categories into vision-language models, the generalization ability of our method is superior to prompt learning methods on novel sets(the last half of categories) of 11 classification datasets. Further comparisons of our method with prompt learning methods across the novel sets of 11 classification datasets are illustrated in Tab. 12. It's important to note that Tab. 1 presents results on the base sets of these 11 classification datasets, conducted in the same 16-shot manner. When evaluating prompt learning methods on novel sets, they do not require additional images after fine-tuning the base categories of each dataset. Our method necessitates a few exemplar images to embed visual clues into the categories of novel sets. Using just one image, our method surpasses current state-of-the-art (SoTA) methods in average performance across 11 datasets. With merely two images, our method achieves an unprecedented average performance exceeding 80.00%. Increasing the number of images to 16 per category boosts average performance to 84.76%, significantly outperforming current SoTA methods by 9.62%. Considering the ease of collecting online data and our method's plug-and-play nature without extra training, it stands as an acceptable competitor to prompt learning methods in low-shot settings. Our method provides a generalizable and efficient approach to embedding multi-modal clues of novel classes into VLMs.

C. Analysis of Preference Weight

In this section, we delve into how the preference-based fusion module mitigates the adverse effects of low-quality text or images. This is achieved by adjusting the preference weights of different classifiers in response to the variable quality of multi-modal references.

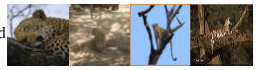
Multi-modal References in Different Qualities	Text-based Classifier	Vision-based Classifier	Multi-Modal Classifier
(a) Chain	0.25	0.25	**0.50**
(b) Balloon Flower	0.11	**0.65**	0.34
(c) Leopard 	**0.41**	0.25	0.34

Figure 4. The variation in preference weight for different classifiers corresponding to multi-modal references of various qualities. (a) The category name "Chain" and the exemplar images are both of high quality and effectively complement each other. (b) The category name "Balloon Flower" may not describe the fine-grained flower in detail and is of low quality, whereas the exemplar images accurately represent the category. (c) The exemplar images with various backgrounds, poses, and appearances are of low quality, but the common word "Leopard" clearly defines the animal.

In Fig. 4(a), both the category "Chain" and its corresponding exemplar images are of high quality or complementary to each other. Consequently, this category favors the multi-modal classifier, assigning it the highest preference weight of 0.50. In Fig. 4(b), the category name

Table 12. **Open-vocabulary Classification Results on Novel Sets in Prompt Learning Setup.**

Methods	Shot Number	ImageNet	Caltech101	OxfordPets	Cars	Flowers102	Food101	Aircraft	SUN397	DTD	EuroSAT	UCF101	Average
CLIP [36]	0	68.14	94.00	97.26	74.89	77.80	91.22	36.29	75.35	59.90	64.05	77.50	74.22
CoOp [59]	0	67.88	89.81	95.29	60.40	59.67	82.26	22.30	65.89	41.18	54.74	56.05	63.22
CoCoOp [58]	0	70.43	93.81	97.69	73.59	71.75	91.29	23.71	76.86	56.00	60.04	73.45	71.69
MaPLe [18]	0	70.54	94.36	97.76	74.00	72.46	92.05	35.61	78.70	59.18	73.23	78.66	75.14
Ours	1	66.67	94.33	95.47	76.60	93.87	89.30	39.10	77.07	61.37	73.60	82.33	77.25
	2	71.83	94.17	97.50	76.90	95.53	91.03	41.40	81.00	67.57	83.83	84.20	80.45
	4	73.03	95.13	97.57	79.90	96.80	91.50	46.17	83.03	69.43	84.30	87.13	82.18
	16	74.87	96.30	97.67	86.23	97.13	91.70	52.03	84.60	74.73	89.57	87.57	84.76

"Balloon Flower" may fail to adequately describe the fine-grained characteristics of the flower in detail, reflecting its low quality, while the exemplar images depict the category more accurately. Thus, the vision-based classifier is assigned the highest preference weight, effectively mitigating the negative impact of the low-quality category name. Conversely, in Fig. 4(c) for the category "Leopard", the exemplar images are of low quality as a result of various backgrounds, poses, and appearances. In contrast, the common word "Leopard" can clearly illustrate the animal. Therefore, the text-based classifier receives the highest preference weight, compensating for the poor quality of the exemplar images.

D. Sources of Exemplar Images

In Tab. 9, we showcase the performance of our method on the base classes of ImageNet using exemplar images sourced from the Internet and ImageNet's training set. When crawling images for a given category, we initiate the process by using the category name as a search query on Google. The first 16 images returned by Google are downloaded as the exemplar images for this category. In Fig. 5, we present a set of examples crawled from the web and examples sampled from ImageNet's training set. It is evident that the images in ImageNet typically exhibit a higher diversity within the same class. The diversity includes differences in the background environment, the number of subjects, their poses, etc. Conversely, images sourced from the Internet often focus on a single subject, featuring simpler poses and backgrounds. Furthermore, images obtained from the Internet can be less reliable due to the noise and ambiguity inherent in text-based queries. For instance, a search for 'Jaguar' may yield images of either the animal or the car, as illustrated in the last row of Fig. 5. As demonstrated in Tab. 9, the use of diverse images from ImageNet's training set as exemplars results in enhanced performance. This improvement is attributed to the closer domain correlation of the test set with the training set in ImageNet, as well as the potential noise and ambiguity of web-crawled images.

E. Visualization of Detection Results

To demonstrate the effectiveness of our method in open-vocabulary detection, we separately showcase the detection results of our OVMR model using the Swin-B backbone, specifically for all categories in Fig. 6 and for the novel categories in Fig. 7, on the LVIS dataset.

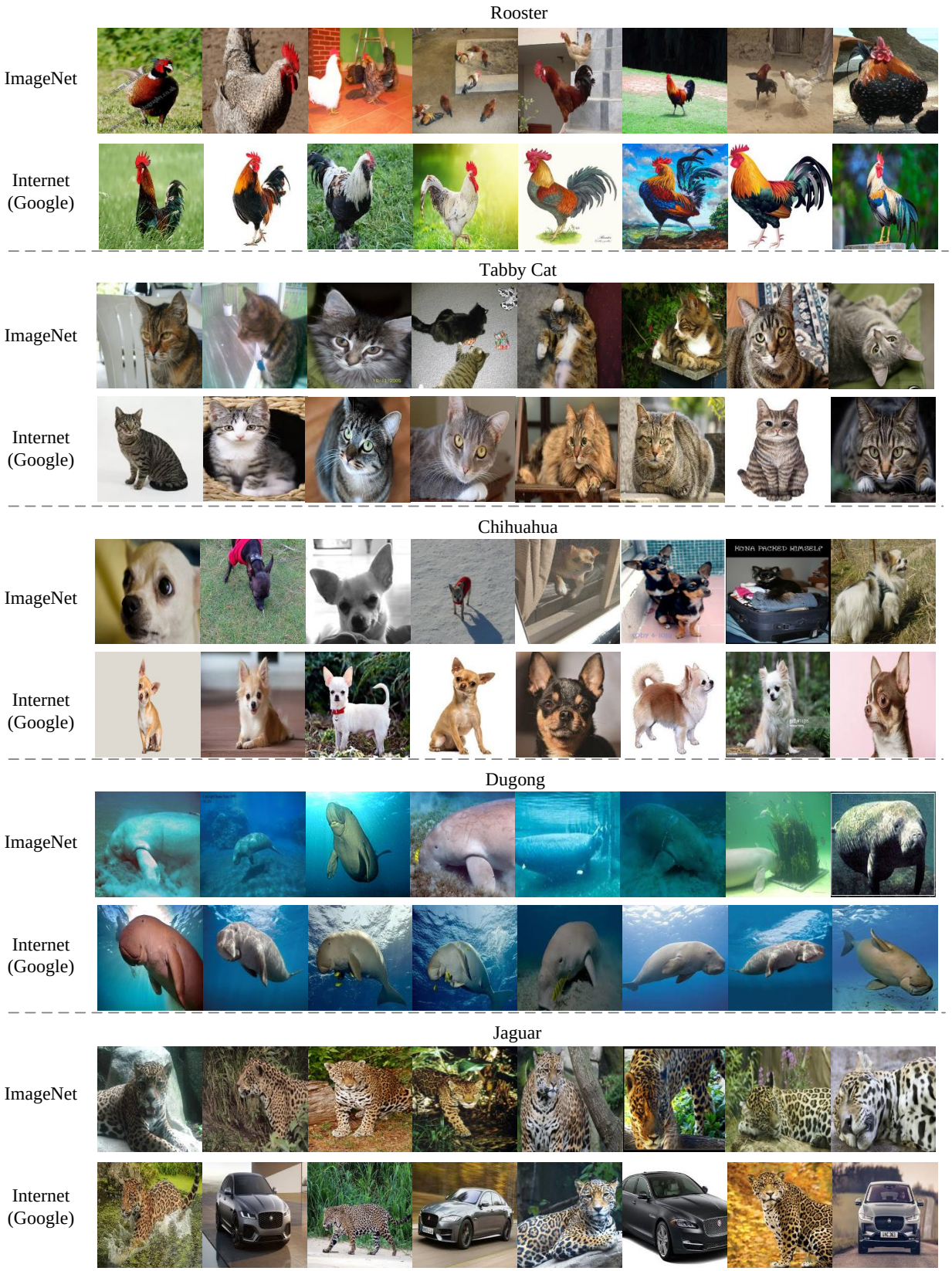


Figure 5. Exemplar images sampled from the training set of ImageNet and the Internet-crawled images.

Figure 6. Detection results for all LVIS categories.

Figure 7. Detection results for novel LVIS categories.