Just a Few Glances: Open-Set Visual Perception with Image Prompt Paradigm

Jinrong Zhang[2*], Penghui Wang[1], Chunxiao Liu[1], Wei Liu[1], Dian Jin[1], Qiong Zhang[1†‡], Erli Meng[1‡], Zhengnan Hu[1]

[1]Xiaomi AI Lab, Beijing, China
[2]Dalian University of Technology, Dalian, China
zjr15272565639@mail.dlut.edu.cn, {wangpenghui, liuchunxiao, liuwei67, jindian1, zhangqiong1, mengerli, huzhengnan}@xiaomi.com

Abstract

To break through the limitations of pre-training models on fixed categories, Open-Set Object Detection (OSOD) and Open-Set Segmentation (OSS) have attracted a surge of interest from researchers. Inspired by large language models, mainstream OSOD and OSS methods generally utilize text as a prompt, achieving remarkable performance. Following SAM paradigm, some researchers use visual prompts, such as points, boxes, and masks that cover detection or segmentation targets. Despite these two prompt paradigms exhibit excellent performance, they also reveal inherent limitations. On the one hand, it is difficult to accurately describe characteristics of specialized category using textual description. On the other hand, existing visual prompt paradigms heavily rely on multi-round human interaction, which hinders them being applied to fully automated pipeline. To address the above issues, we propose a novel prompt paradigm in OSOD and OSS, that is, **Image Prompt Paradigm**. This brand new prompt paradigm enables to detect or segment specialized categories without multi-round human intervention. To achieve this goal, the proposed image prompt paradigm uses just a few image instances as prompts, and we propose a novel framework named **MI Grounding** for this new paradigm. In this framework, high-quality image prompts are automatically encoded, selected and fused, achieving the single-stage and non-interactive inference. We conduct extensive experiments on public datasets, showing that MI Grounding achieves competitive performance on OSOD and OSS benchmarks compared to text prompt paradigm methods and visual prompt paradigm methods. Moreover, MI Grounding can greatly outperform existing method on our constructed specialized ADR50K dataset.

Introduction

To break through the limitations of pre-training models on fixed categories, Open-Set Object Detection (OSOD) and Open-Set Segmentation (OSS) have attract a surge of interest from researchers. In these fields, trained models can not only detect or segment predefined specific categories but also generalize to open scenarios, which greatly improve the ability and applicability (Li et al. 2022).

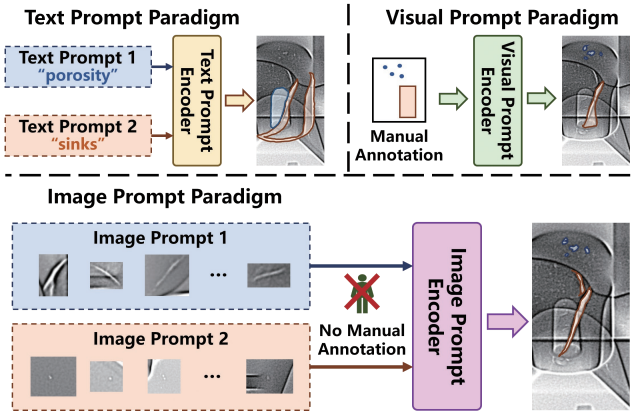

Figure 1: Image prompt paradigm vs. previous prompt paradigms. The text prompt paradigm struggles to accurately describe specialized categories. The visual prompt paradigm relies on multi-round human interaction. The proposed image prompt paradigm uses just a few image instances which can handle specialized categories without any manual annotation.

Inspired by the remarkable success achieved by foundational models (Radford et al. 2021; Li et al. 2022), mainstream OSOD and OSS methods employ a prompt as an input, which tells the model what to detect or segment in the image. Existing prompt paradigms can be mainly categorized into two types: text prompt paradigm and visual prompt paradigm. As for text prompt paradigm, users are required to provide a textual description to depict characteristics of detection or segmentation targets, and models are trained to align text prompt with visual contents in the latent space (Liu et al. 2023; Ding, Wang, and Tu 2022). Following SAM (Kirillov et al. 2023), another line of approaches employ visual prompts, such as points, boxes, and masks. The visual prompt needs to be manually designed that can locate specific targets. Such a design makes this process generally involves multi-round interaction to avoid ambiguous prompts (Kirillov et al. 2023).

However, textual and visual prompt paradigms have the

*The work was done while Jinrong Zhang was a research intern at Xiaomi AI Lab.
†These authors contributed equally.
‡Corresponding author.

following limitations. First of all, the visual feature of specialized categories are difficult to be accurately described by text, and hence hinder the application of text prompt paradigm (Li et al. 2024; Jiang et al. 2024). Second, visual prompts heavily rely on multi-round human interaction, which makes it difficult to be applied into production pipelines (Kirillov et al. 2023). As shown in Figure.1, in X-ray defect detection, we need to detect and segment specialized categories, such as "shrinkage porosity", "sinks", and "porosity". These concepts are specific to the X-ray field, which cannot reflect the visual characteristics without industrial knowledge, such as their shape, size, and texture etc. Visual prompt might alleviate this issue by providing bounding boxes of "shrinkage porosity", "sinks", and "porosity" as prompts, but it requires users to annotate or check bounding boxes to make sure they cover the target areas (Kirillov et al. 2023). These interaction processes make a single-stage, fully automated inference pipeline impossible.

In this work, we establish a novel visual perception paradigm, i.e. **Image Prompt Paradigm**, which completely abandons traditional text prompts and visual prompts, achieving a single-stage and fully automated inference. Inspired by the fact that humans can quickly grasp the characteristics of a specific category after taking **just a few glances** at its instances, the proposed image prompt paradigm utilizes just a few image instances of target as prompts. These instances are automatically constructed and calculated by our proposed **MI Grounding** framework, which uses **m**ultiple **i**mages as prompts. To bridge the gap between specialized categories and visual content, MI Grounding introduce an image prompts selection encoder, which can encode, select and integrate image prompts. The encoder module possesses extensive prior knowledge at the visual level, and can extract inherent distinctive semantic information of image prompts. The encoded image prompts are then selected and integrated to highlight high-quality image prompts automatically. After aligning the image prompts with the predicted objects, MI Grounding is learned to handle specialized categories that are difficult to describe using text, and achieves single-stage and non-interactive inference. Extensive experiments have shown that the proposed image prompt paradigm and MI Grounding achieve excellent detection and segmentation performance.

Concretely, our contributions can be summarized as follows:

- We propose a novel visual perception paradigm: Image Prompt Paradigm. Different from existing text and visual prompt paradigm, this paradigm uses just a few image instances as prompts, which can understand specialized categories that are hard to describe by text in a single-stage and non-interactive manner.
- We propose a novel framework named MI Grounding tailored for the proposed image prompt paradigm. MI Grounding utilizes just a few image prompts to perform Open-Set Object Detection and Open-Set Segmentation, and propose an image prompt selection encoder to select and integrate high-quality prompts.
- Our approach achieves competitive performance on several datasets compared with mainstream Open-Set Object Detection and Open-Set Object Segmentation methods, which show the effectiveness of our proposed image prompt paradigm. To further demonstrate the superiority, we constructed a specialized ADR50K dataset, which contains a rich set of X-ray defect detection data. Experiments demonstrate that our approach can greatly improve the performance on this specialized dataset.

Related Works

Visual Perception Based on Text Prompt Paradigm. With the widespread application of foundational methods like CLIP (Radford et al. 2021) and BERT (Devlin 2018), open-vocabulary object detection and segmentation methods have achieved remarkable success in the general visual perception field. Researchers find that object detection and segmentation can be expressed as an alignment between text prompts and visual context information. Based on the concept above, these methods have made significant breakthroughs in zero-shot and few-shot learning. Grounding DINO (Liu et al. 2023) extends the training strategy of GLIP (Li et al. 2022) to DINO (Zhang et al. 2022a), achieving strong open-set detection capabilities. DetCLIP (Yao et al. 2022) and RegionCLIP (Zhong et al. 2022) utilize image-text pairs with pseudo boxes to expand region knowledge, thereby improving open-set performance. These text prompt paradigm methods rely on text encoders, like BERT, to model text queries. However, due to the ambiguity caused by the high information density of text and the potential mismatch between text descriptions and complex visual scenes, visual perception methods based on the text prompt paradigm have inherent limitations (Li et al. 2024).

Visual Perception Based on Visual Prompt Paradigm. Researchers have begun exploring alternative prompt paradigms. SAM (Kirillov et al. 2023) pioneer an interactive open-set segmentation approach, introducing a novel prompt paradigm that includes boxes, points, masks, or lines covering the target. Subsequent researchers define this approach as the visual prompt paradigm and further explore its potential (Li et al. 2024). Semantic-SAM (Li et al. 2023) achieves semantic awareness by training on decoupled objects and parts classifications integrated from multiple datasets. Painter (Wang et al. 2023a) and SegGPT (Wang et al. 2023b) adopt a generalist strategy to tackle diverse segmentation tasks, conceptualizing segmentation as an in-context coloring problem. DINOv (Li et al. 2024) proposes a general contextual visual prompt framework, using visual context to understand new categories.

Visual Perception with Image and Text Prompt. To further enhance model performance, some researchers have introduced additional target images to augment text prompts. MQ-Det (Xu et al. 2024) uses cross-attention and weighted addition to integrate image features into the text prompt, significantly improving model performance. However, when MQ-Det uses only images as prompts, the model's performance is poor. This indicates that MQ-Det has not fully exploited the potential of image prompts. In such methods, image prompts merely enhance the prompt features rather than

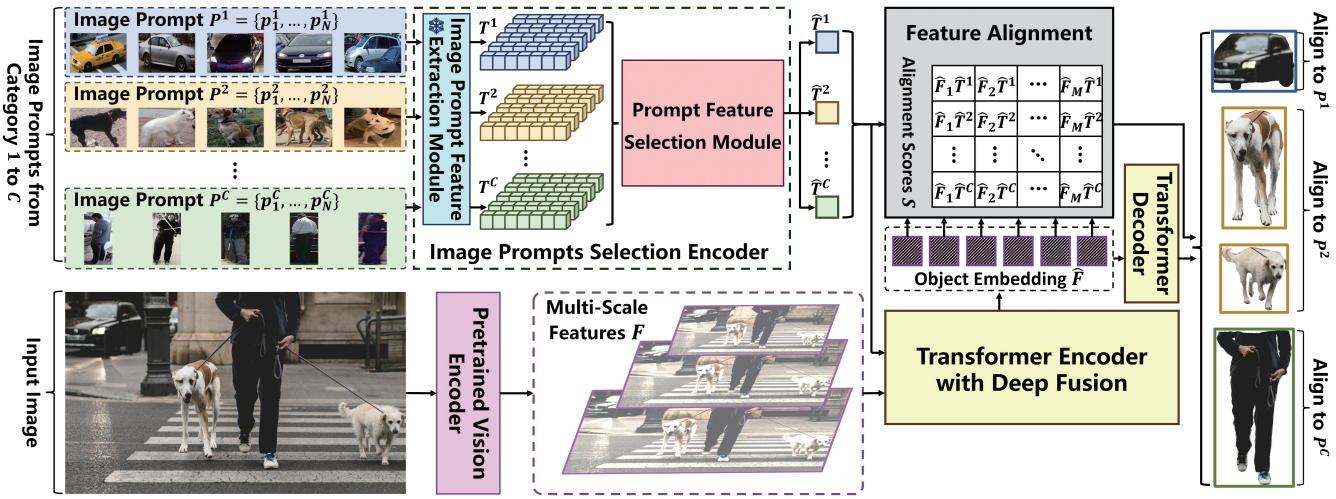

Figure 2: The overall framework of MI Grounding. Image prompts are encoded, selected, and integrated through the image prompts selection encoder (IPS encoder) to obtain category-specific prompt features. These prompt features are then deeply fused and aligned with multi-scale features from the input image to achieve open-set visual perception.

lead the inference process.

Method

Our goal is to establish an image prompt paradigm, where the model completes open-set detection and segmentation by just taking a few glances at images of objects similar to the detection target. We first introduce how image prompts are constructed during the training and testing process. Then, we detail our proposed MI Grounding, including the model design and training strategy.

Image Prompt Paradigm

In order to eliminate the tedious interaction process similar to the visual prompt paradigm and ensure that the data of the detection target is not leaked, we build an image prompt library using the training split of the dataset. Specifically, we crop instance targets from the original images based on their detection box labels and store them categorized by their class labels. The detailed process of extracting image prompts from the original images is as follows:

$$p = Crop(I, L_{box}) = I[y : y + h_p, x : x + w_p], \quad (1)$$

where $I \in \mathbb{R}^{3\times H\times W}$ represents the original large image, and $p \in \mathbb{R}^{3\times h_p\times w_p}$ is the instance target image obtained by cropping. H and W are the height and width of the original image. $L_{box} = \{x, y, w_p, h_p\}$ is the bounding box label, with w_p and h_p being the width and height of the corresponding instance target's bounding box, and x and y being the coordinates of the top-left corner of the bounding box. To ensure that target data from the test set is not leaked and enhance the model's robustness, we use only instance target images cropped from the training set as image prompts during both training and testing.

MI Grounding

As shown in Fig. 2, MI Grounding consists of an image prompts selection encoder (IPS encoder), a vision encoder, a transformer encoder with deep fusion following GLIP (Li et al. 2022), and a transformer decoder. The IPS encoder extracts, selects, and integrates features from the image prompts, while the vision encoder extracts features from the input image. In the image prompt paradigm, the model uses a set of instance images $P^c = \{p_1^c, \ldots, p_N^c\}$ of a specified category c as prompts. The goal of MI Grounding is to detect and segment objects of the corresponding category from the input image I based on P^c.

Image Prompts Selection Encoder. The IPS encoder consists of an image prompt feature extraction module and a prompt feature selection module. As for image prompt, how to extract their features to handle specialized categories is a crucial problem. Inspired by the text prompt that using pre-trained text encoders to fully utilize the latent semantic information, we learn that the critical point is to extract features that can distinguish the detection target from other instances. As a result, we employ pre-trained ViT as image prompt feature extractor, since it show good clustering properties, indicating pre-trained ViT contains semantic information useful for open-set visual perception. In the image prompt feature extractor, we compute features for all prompt images P^c and aggregate them into a prompt feature matrix:

$$T^c = STACK(ViT(p_1^c, \ldots, p_N^c)), \quad (2)$$

where $STACK(\cdot)$ stands for feature stacking along the prompt quantity dimension, and $ViT(\cdot)$ represents a frozen vision transformer backbone. $T^c \in \mathbb{R}^{N\times D}$ denotes the prompt feature matrix composed of N image prompts for a specified category c, where D represents the feature dimension.

In prompt feature selection module, it is worth to highlight that the quality of image prompts have a significant im-

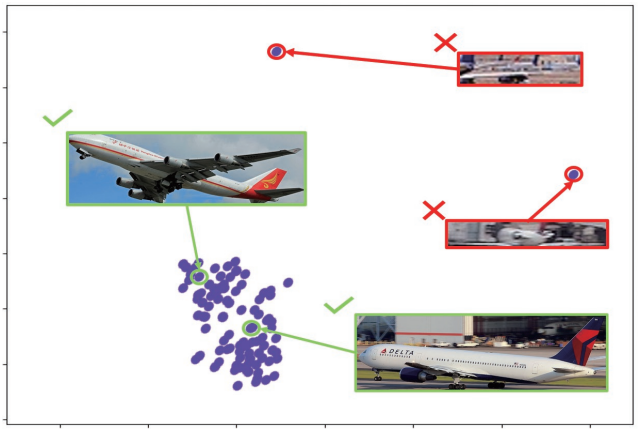

Figure 3: Quality of image prompts. Green indicates good image prompts, while red indicates poor ones.

pact on detection and segmentation. Directly integrating all the image prompts of the same category will lead to unstable performance due to the low information density of images. As shown in Fig. 3, Despite most image prompt features exhibit good clustering properties, there still exist a few outliers caused by instances that are hard to recognize. These outliers will reduce the distinctiveness of the semantic information in image prompts. We observe that high-quality image prompt features within the same category tend to be highly similar, while low-quality ones always show significant differences.

Inspired by the above observation, we develop a prompt feature selection module based on self-attention (PFSM) to leverage the correlation between prompt features, reducing the impact of poor-quality image prompts, as shown in Fig 4. The overall process is illustrated in EQ. 3, where θ represents the learnable parameters of $PFSM(\cdot)$:

$$\hat{T}^c = PFSM(T^c, \theta). \tag{3}$$

In PFSM, we first use self-attention to calculate the correlation between the N image prompts:

$$Q = MLP_1(T^c), K = MLP_2(T^c), V = MLP_3(T^c), \tag{4}$$

$$A = softmax(QK^T \Big/ \sqrt{D_1}), \tag{5}$$

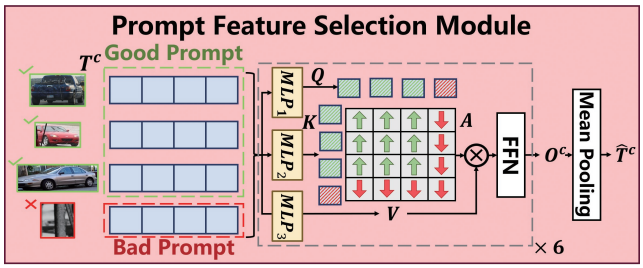

Figure 4: The overall framework of PFSM.

where $MLP(\cdot)$ is a fully connected network for feature dimension adjustment, and $Q, K, V \in \mathbb{R}^{N \times D_1}$ are the query, key, and value needed for self-attention. $A \in \mathbb{R}^{N \times N}$ is the correlation matrix between the N image prompts. As analyzed earlier, the stronger the correlation with other prompt features, the more accurate the semantic information it contains. Conversely, weaker correlations suggest a higher likelihood of being an outlier. We assign higher weights to prompt features with more accurate semantic information:

$$O^c = FFN(AT^c), \tag{6}$$

where $O^c \in \mathbb{R}^{N \times D_2}$ represents the enhanced image prompt features, $FFN(\cdot)$ is a feed-forward layer, and D_2 is the transformed feature dimension. Finally, we apply average pooling to reduce the dimensionality of O^c along the prompt quantity dimension:

$$\hat{T}^c = MeanPooling(O^c + Linear(T^c)), \tag{7}$$

where $\hat{T}^c \in \mathbb{R}^{1 \times D_2}$ represents the final image prompt feature for category c. For all categories $\{1, 2, \ldots, C\}$, $PFSM(\cdot)$ uses the same θ to obtain $\hat{T} = \{\hat{T}^1, \hat{T}^2, \ldots, \hat{T}^C\}$.

Vision Encoder. To enhance the model's robustness to targets of different scales, we use a vision transformer backbone to construct the vision encoder, retaining the features from different layers as multi-scale features of the input image. The multi-scale features are defined as $F = \{f_1, \ldots, f_L\}$, where f_i represents the features from the i-th layer of the vison encoder.

Transformer Encoder with Deep Fusion. To reduce the difficulty of feature alignment, we fuse the enhanced image prompt features with the input image features by referencing the cross-modality interaction method from language-vision models. Specifically, we use a multi-scale deformable cross-attention (Zhu et al. 2020) to fuse the prompt features $\hat{T}$ with the multi-scale image features F, resulting in the object embeddings $\hat{F}$:

$$\hat{F} = MSDeformAttn(F, \hat{T}). \tag{8}$$

Region-Level Feature Alignment. Inspired by the image-text alignment in the text prompt paradigm, we achieve region-level classification feature alignment between image prompts and predicted objects in the image prompt paradigm. Specifically, we directly compute the alignment scores $S \in \mathbb{R}^{M \times C}$ between the prompt features $\hat{T}$ and the object embeddings $\hat{F}$:

$$S = \hat{F} \cdot \hat{T}, \tag{9}$$

where M is the predefined number of object embeddings. Finally, we use a transformer decoder to decode the object embeddings $\hat{F}$ into bounding box labels and mask labels.

Training Strategy and Optimization Objective

Image Prompt Training Strategy. To enhance the model's generalization ability, we use a random image prompt strategy. During training, we randomly sample N cropped instance images as image prompts for each category, updating

Method	Training Data	Prompt Paradigm	COCO (out-domain)	LVIS-1203 (out-domain)				ODinW-35 (out-domain)
			AP^b	AP^b	AP^b_f	AP^b_c	AP^b_r	AP^b
GLIP T (Li et al. 2022)	O365+GoldG+...	Text	46.7	17.2	25.5	12.5	10.1	19.6
GLIP L (Li et al. 2022)	FourODs+GoldG+...		49.8	26.9	**35.4**	23.3	17.1	23.4
Grounding DINO T (Liu et al. 2023)	O365+GoldG+...		48.4	-	-	-	-	22.3
Grounding DINO L (Liu et al. 2023)	O365+GoldG+...		52.5	-	-	-	-	**26.1**
DINOv T (Li et al. 2024)	COCO+SA-1B	Visual	-	-	-	-	-	14.9
DINOv L (Li et al. 2024)	COCO+SA-1B		-	-	-	-	-	15.7
MI Grounding-D (ours)	O365	Image	**53.7**	**27.4**	32.4	**25.8**	**19.9**	21.5

Table 1: Object Detection with MI Grounding-D. Bold and underline denote the best and second-best results in each column. AP^b represents the average precision for object detection. AP^b_f, AP^b_c, and AP^b_r represent the average precision for frequent, common, and rare classes, respectively.

Method	Training Data	Prompt Paradigm	COCO (in-domain)		ADE20K (out-domain)		SegInW (out-domain)	
			AP^m	AP^b	AP^m	AP^b	AP^m_{avg}	AP^m_{med}
GLIPv2 H (Zhang et al. 2022b)	COCO+O365+...	Text	48.9	-	-	-	-	-
MaskCLIP L (Ding, Wang, and Tu 2022)	YFCC100M		-	-	15.1	6.0	23.7	-
FC-CLIP L (Yu et al. 2024)	COCO		44.6	-	16.8	-	-	-
OpenSeed T (Zhang et al. 2023)	COCO+O365		47.6	52.0	14.1	17.0	33.9	21.5
X-Decoder T (Zou et al. 2023)	COCO+CC3M+...		40.5	43.6	9.8	-	22.7	15.2
X-Decoder L (Zou et al. 2023)	COCO+CC3M+...		46.7	-	13.1	-	36.1	38.7
OpenSeed L (Zhang et al. 2023)	COCO+O365		**53.2**	**58.2**	15.0	17.7	36.1	38.7
DINOv T (Li et al. 2024)	COCO+SA-1B	Visual	41.5	45.2	11.4	12.8	39.5	41.6
DINOv L (Li et al. 2024)	COCO+SA-1B		50.4	54.2	15.1	14.3	40.6	**44.6**
MI Grounding-S (ours)	COCO+LVIS	Image	46.1	54.7	**21.0**	**25.3**	**46.9**	41.3

Table 2: Instance Segmentation with MI Grounding-S. Bold and underline denote the best and second-best results in each column. AP^b represents the average precision for object detection. AP^m represents the average precision for instance segmentation.

them every iteration. The random sampling allows the model to adapt to cross-domain image prompts. The frequent updates help the model learn and adjust to more complex prompts. It's important to note that during testing, the model also uses only instance images cropped from the training set as prompts. This not only prevents data leakage from the test set but also demonstrates the model's generalization ability to cross-domain image prompts.

Optimization Objective. Since our model directly predicts the target's class, box, and mask in an end-to-end manner, the loss function $\mathcal{L}$ of MI Grounding consists of classification loss $\mathcal{L}_{class}$, localization losses $\mathcal{L}_{L1}$ and $\mathcal{L}_{GIoU}$, and segmentation loss $\mathcal{L}_{mask}$:

$$\mathcal{L} = \mathcal{L}_{class} + \mathcal{L}_{L1} + \mathcal{L}_{GIoU} + \mathcal{L}_{mask}. \tag{10}$$

For the classification loss, we use a contrastive loss (Radford et al. 2021) to calculate the difference between the predicted target and the image prompt features for open-set classification. For the localization loss, we apply L1 loss (Ren et al. 2015) for regressing the bounding box coordinates and GIoU loss (Rezatofighi et al. 2019) to enhance convergence stability. In the segmentation loss, $\mathcal{L}_{mask}$ is a cross-entropy loss for mask segmentation.

Experiments

Datasets and Settings

In our experiments, we provide two sets of model parameters: MI Grounding-S for segmentation and MI Grounding-D for object detection. In MI Grounding-S, we use only the COCO (Lin et al. 2014) and LVIS (Gupta, Dollar, and Girshick 2019) datasets for joint training and test on the COCO, ADE20K (Zhou et al. 2017), and SegInW (Zou et al. 2023) datasets. In MI Grounding-D, we use only the Objects365 (Shao et al. 2019) dataset for training and test on the COCO, LVIS, and ODinW (Li et al. 2022) datasets. In both MI Grounding-S and MI Grounding-D, we use ViT-L as the vision backbone. We use 8 as the number of image prompts in our method, as discussed in the ablation study.

Comparison to Prior Works

To explore the generalization ability of the image prompt paradigm and MI Grounding, we test our model on multiple

datasets across different domains. It's important to note that we train MI Grounding-D on Objects365 for 32 A100 days and MI Grounding-S on COCO+LVIS for 16 A100 days. Our training data and duration are significantly less than most methods in Table. 1 and Table. 2. For example, GLIP L is trained on the FourODs GoldG, and Cap24M datasets for 600 V100 days (Li et al. 2022).

Object Detection with MI Grounding-D. In Table. 1, we test on well-established benchmarks, including common object detection datasets like COCO, long-tailed datasets like LVIS, and complex cross-domain datasets like ODinW. MI Grounding-D demonstrates strong performance in out-of-domain scenarios. Notably, MI Grounding-D leads by 2.8% in AP for rare categories on LVIS, further highlighting the generalization ability of the image prompt paradigm.

Instance Segmentation with MI Grounding-S. As shown in Table. 2, we test MI Grounding-S on multiple datasets under both in-domain and out-domain conditions. Notably, in out-domain scenarios, MI Grounding-S achieves a significant advantage of 4.2% on ADE20K and 6.3% on SegInW. SegInW is a complex cross-domain dataset containing 25 different sub-datasets, and the performance advantage on this dataset underscores the generalization ability of the image prompt paradigm.

Ablation

Effectiveness of Prompt Feature Selection Module. To demonstrate the effectiveness of the prompt feature selection module (PFSM) based on self-attention in the image prompts selection encoder, we replace it with three other modules: a fully connected network, a convolutional neural network, and mean pooling. The results in Table. 3 are obtained from training and testing only on COCO. In the fully connected network and convolutional network, we use supervised neural networks to reduce the dimensionality of the N image prompt features. In the mean pooling, we directly take the mean of the N image prompt features to obtain the final prompt feature. As shown in Table. 3, our proposed PFSM proves to be the most effective among the various strategies.

Prompt Feature Calculation	COCO			
	AP^b	AP^b_{med}	AP^m	AP^m_{med}
Mean Pooling	56.5	60.9	48.4	52.6
FC	54.5	59.9	47.2	51.2
CNN	58.0	62.4	49.7	53.4
PFSM	**61.7**	**67.2**	**53.1**	**57.8**

Table 3: Ablation study of PFSM. FC represents the fully connected network, and CNN represents the convolutional neural network.

Impact of Image Prompt Update Frequency. Even within the same category, instance images can be highly diverse. To help the model adapt to this diversity, we increase the frequency of image prompt updates, allowing the model to learn from a wider range of image prompts during training. As shown in Table. 4, we demonstrate the impact of update frequency on model performance. We gradually increase the update frequency from once every 200 iterations to once per iteration, and the model's performance improve accordingly. Finally, we set the model to update the image prompts once per iteration.

Prompt Update Frequency	COCO			
	AP^b	AP^b_{med}	AP^m	AP^{med}_m
200 Iterations	57.9	63.4	49.7	54.4
150 Iterations	58.8	64.1	50.6	55.3
100 Iterations	60.58	65.87	52.2	57.0
50 Iterations	59.2	64.5	50.7	55.2
1 Iteration	**61.7**	**67.2**	**53.1**	**57.8**

Table 4: Ablation study of image prompt update frequency.

Impact of Image Prompt Quantity. As previously mentioned, due to the low information density of images, a single image prompt often fails to fully convey the semantic information of the target class. As the number of image prompts increases, the semantic representation of the class becomes more complete, leading to higher-quality open-set visual perception. In Table. 5, we perform an ablation study on the number of image prompts. As the number of image prompts increases, the model's performance gradually improves. However, when the number of image prompts exceeds 8, there is no significant performance gain. More seriously, with the number of image prompt increases, the computational cost of the model increases. Therefore, the optimal setting of image prompt quantity is 8.

Image Prompt Quantity	COCO			
	AP^b	AP^b_{med}	AP^m	AP^m_{med}
1	44.6	48.5	38.5	41.4
3	56.2	61.4	48.3	52.9
5	59.4	64.5	51.3	55.6
8	**61.7**	**67.2**	**53.1**	**57.8**
12	61.4	66.6	52.7	57.2
20	61.6	66.9	52.8	57.4

Table 5: Ablation study of image prompt quantity. Model performance improves as the number of image prompts increases. After reaching 8 prompts, there is no significant further improvement.

ADR50K Dataset

To further demonstrate the advantages of the image prompt paradigm, we create the Automatic Defect Recognition dataset (ADR50K). In ADR50K, we collect more than 50,000 X-ray images of defect inspections. We provide classification annotations for three types of defects using specialized terminology: "sink", "shrinkage porosity", and "porosity". Additionally, we provide detection and segmentation annotations for all defects. The relevant details of the ADR50K dataset are shown in Table. 7.

Method	Prompt Paradigm	Object Detection				Instance Segmentation			
		AP_{50}	AP_{sink}	$AP_{shrinkage}$	$AP_{porosity}$	AP_{50}	AP_{sink}	$AP_{shrinkage}$	$AP_{porosity}$
Grounding DINO L	Text	43.4	43.1	18.2	17.5	-	-	-	-
DINOv L	Visual	16.5	21.6	1.403	0.1	15.4	13.8	1.0	0.11
MI Grounding (ours)	Image	**50.2**	**49.6**	**19.4**	**25.8**	**50.4**	**45.0**	**15.8**	**27.9**

Table 6: Comparative experiments on ADR50K. AP_{50} represents the average precision for all categories at an IoU threshold of 0.50. AP_{sink}, $AP_{shrinkage}$, and $AP_{porosity}$ represent the average precision for "sink", "shrinkage porosity", and "porosity", respectively.

Data Split	Number of Images	Number of Instances		
		Categories		Total
Train	51572	"sink"	167886	196673
		"shrinkage porosity"	24320	
		"porosity"	4467	
Test	7696	"sink"	20946	24437
		"shrinkage porosity"	2955	
		"porosity"	536	

Table 7: ADR50K Dataset Details. We allocated approximately 10% of the data to the test set, with the remaining data used for the training set.

Unlike conventional object detection and segmentation datasets, the detection targets and class names in the ADR50K dataset are highly specialized and even misleading. As shown in Figure 5, we present some example images of the three defect types in the dataset. The ADR50K dataset poses three main challenges for open-set visual perception methods. **(1) Difficulty in aligning category names with instances.** In typical scenarios, "sink" usually refers to a basin, commonly found in kitchens or bathrooms, where water is supplied through a faucet and drains away. However, in the ADR50K dataset, "sink" refers to a slender indentation that is completely different from a basin. **(2) Confusion between category names.** The terms "shrinkage porosity" and "porosity" seem to be same category as their textual name look similar with each other, but they refer to entirely different types of defects. In the ADR50K dataset, "shrinkage porosity" refers to a sheet-like shallow depression, while "porosity" refers to small round pits. **(3) Confusion with background images.** The images in the ADR50K dataset contain numerous shadows caused by overlapping structural components, which are not defects. These shadows can easily be mistaken for defects that need to be detected.

We compared our proposed image prompt method with text prompt and visual prompt methods on the ADR50K dataset. As shown in Table. 6, MI Grounding outperforms Grounding DINO L (Liu et al. 2023) and DINOv L (Li et al. 2024) in both object detection and instance segmentation tasks. In MI Grounding, we use instance images of defects as prompts instead of potentially misleading text. Additionally, unlike visual prompt methods, MI Grounding in the image prompt paradigm does not require a separate prompt for each target instance.

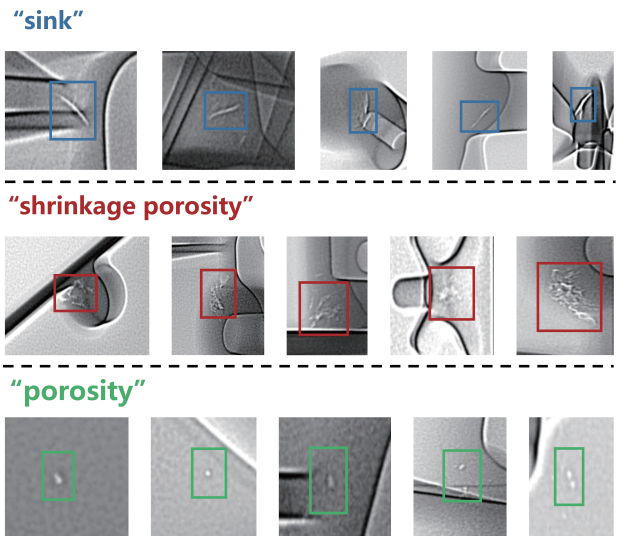

Figure 5: Examples of specialized categories in ADR50K dataset. The text denotes the category name of defects, and the areas within bounding boxs denote the corresponding category region.

Conclusion

In this paper, we introduce a novel visual perception paradigm called the Image Prompt Paradigm. Unlike existing text and visual prompts, this paradigm uses a few image instances as prompts, enabling it to understand specialized categories which are challenging to describe with text in a single-stage and non-interactive manner. To support this new paradigm, we present a framework named MI Grounding. MI Grounding utilizes multiple image prompts to perform Open-Set Object Detection and Open-Set Segmentation, and it includes an image prompt selection encoder designed to choose and integrate high-quality prompts effectively. Our approach achieves competitive performance across several datasets when compared to mainstream methods in Open-Set Object Detection and Open-Set Segmentation, demonstrating the effectiveness of the proposed iamge prompt paradigm. To further validate the superiority, we developed a specialized ADR50K dataset, which comprises an extensive collection of X-ray defect detection data. Experimental results show that our approach significantly enhances performance on this specialized dataset.

References

Devlin, J. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Ding, Z.; Wang, J.; and Tu, Z. 2022. Open-Vocabulary Panoptic Segmentation MaskCLIP.

Gupta, A.; Dollar, P.; and Girshick, R. 2019. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5356–5364.

Jiang, Q.; Li, F.; Zeng, Z.; Ren, T.; Liu, S.; and Zhang, L. 2024. T-Rex2: Towards Generic Object Detection via Text-Visual Prompt Synergy. *arXiv preprint arXiv:2403.14610*.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4015–4026.

Li, F.; Jiang, Q.; Zhang, H.; Ren, T.; Liu, S.; Zou, X.; Xu, H.; Li, H.; Yang, J.; Li, C.; et al. 2024. Visual in-context prompting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12861–12871.

Li, F.; Zhang, H.; Sun, P.; Zou, X.; Liu, S.; Yang, J.; Li, C.; Zhang, L.; and Gao, J. 2023. Semantic-sam: Segment and recognize anything at any granularity. *arXiv preprint arXiv:2307.04767*.

Li, L. H.; Zhang, P.; Zhang, H.; Yang, J.; Li, C.; Zhong, Y.; Wang, L.; Yuan, L.; Zhang, L.; Hwang, J.-N.; et al. 2022. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10965–10975.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, 740–755. Springer.

Liu, S.; Zeng, Z.; Ren, T.; Li, F.; Zhang, H.; Yang, J.; Li, C.; Yang, J.; Su, H.; Zhu, J.; et al. 2023. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28.

Rezatofighi, H.; Tsoi, N.; Gwak, J.; Sadeghian, A.; Reid, I.; and Savarese, S. 2019. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 658–666.

Shao, S.; Li, Z.; Zhang, T.; Peng, C.; Yu, G.; Zhang, X.; Li, J.; and Sun, J. 2019. Objects365: A large-scale, high-quality dataset for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, 8430–8439.

Wang, X.; Wang, W.; Cao, Y.; Shen, C.; and Huang, T. 2023a. Images speak in images: A generalist painter for in-context visual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6830–6839.

Wang, X.; Zhang, X.; Cao, Y.; Wang, W.; Shen, C.; and Huang, T. 2023b. Seggpt: Segmenting everything in context. *arXiv preprint arXiv:2304.03284*.

Xu, Y.; Zhang, M.; Fu, C.; Chen, P.; Yang, X.; Li, K.; and Xu, C. 2024. Multi-modal queried object detection in the wild. *Advances in Neural Information Processing Systems*, 36.

Yao, L.; Han, J.; Wen, Y.; Liang, X.; Xu, D.; Zhang, W.; Li, Z.; Xu, C.; and Xu, H. 2022. Detclip: Dictionary-enriched visual-concept paralleled pre-training for open-world detection. *Advances in Neural Information Processing Systems*, 35: 9125–9138.

Yu, Q.; He, J.; Deng, X.; Shen, X.; and Chen, L.-C. 2024. Convolutions die hard: Open-vocabulary segmentation with single frozen convolutional clip. *Advances in Neural Information Processing Systems*, 36.

Zhang, H.; Li, F.; Liu, S.; Zhang, L.; Su, H.; Zhu, J.; Ni, L. M.; and Shum, H.-Y. 2022a. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*.

Zhang, H.; Li, F.; Zou, X.; Liu, S.; Li, C.; Yang, J.; and Zhang, L. 2023. A simple framework for open-vocabulary segmentation and detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1020–1031.

Zhang, H.; Zhang, P.; Hu, X.; Chen, Y.-C.; Li, L.; Dai, X.; Wang, L.; Yuan, L.; Hwang, J.-N.; and Gao, J. 2022b. Glipv2: Unifying localization and vision-language understanding. *Advances in Neural Information Processing Systems*, 35: 36067–36080.

Zhong, Y.; Yang, J.; Zhang, P.; Li, C.; Codella, N.; Li, L. H.; Zhou, L.; Dai, X.; Yuan, L.; Li, Y.; et al. 2022. Regionclip: Region-based language-image pretraining. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16793–16803.

Zhou, B.; Zhao, H.; Puig, X.; Fidler, S.; Barriuso, A.; and Torralba, A. 2017. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 633–641.

Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; and Dai, J. 2020. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*.

Zou, X.; Dou, Z.-Y.; Yang, J.; Gan, Z.; Li, L.; Li, C.; Dai, X.; Behl, H.; Wang, J.; Yuan, L.; et al. 2023. Generalized decoding for pixel, image, and language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15116–15127.