ILD: Image-Level Labels Driven Active Learning Object Detection

Rui Tian, Jiaxuan Zhang, Yongqiang Zhang, Man Zhang, Zian Zhang, Yin Zhang, Yongqiang Li, and Wangmeng Zuo

***Abstract*—Existing SOTA methods in active learning object detection(ALOD) achieve impressive results, but they overlook two problems: 1) the requirement for instance-level labels during initialization, and 2) the constrained localization ability of the pre-trained fully supervised detector in the active learning phase. Problem 1) contradicts the fundamental purpose of active learning in balancing annotation costs and detection performance. Problem 2) arises from the fact that the active learning process relies on a single pre-trained fully supervised detector. To tackle these problems, we propose Image-level Labels Driven active learning object detection (termed as ILD). Specifically, we propose a multi-step reasoning process based on the chain-of-thought only using image-level labels, including a class-number-aware step and an iterative step, to enhance the detection ability of VLM. The detection results of the VLM and weakly supervised detector are used as pseudo ground-truth boxes to initialize a fully supervised detector during AL initialization. Thus, the initialization process of ILD eliminates the requirement for instance-level labels. In the active learning stage, we design two novel uncertainty and diversity acquisition functions to select the most informative images based on collaborative outputs from both the weakly supervised detector and the pre-trained fully supervised detector. The collaborative mechanism jointly measures the uncertainty of two detectors and the diversity of object features, thereby enhancing the localization quality. Extensive experiments demonstrate that the proposed ILD achieves state-of-the-art performance(*i.e.*, 77.5%, 25.7%, and 27.9%) on PASCAL VOC2007, MS COCO2014 and MS COCO2017 datasets, surpassing the SOTA methods by 3.4%, 1.2% and 4.7%, respectively. Our code is publicly available on https://github.com/RuiTianHIT/ILD**

***Index Terms*—Active learning object detection, weakly supervised object detection, acquisition function, large-scale vision-language pre-trained model.**

Received 30 June 2025; revised 31 August 2025 and 18 September 2025; accepted 20 September 2025. Date of publication 24 September 2025; date of current version 9 March 2026. This work was supported in part by the National Natural Science Foundation of China under Grant 62206077, in part by the Inner Mongolia Natural Science Foundation for Distinguished Young Scholars under Grant 2025JQ009, and in part by the Inner Mongolia Talent Development Project for Outstanding Young Talents. This article was recommended by Associate Editor M. Oszust. *(Rui Tian and Jiaxuan Zhang contributed equally to this work.) (Corresponding author: Yongqiang Zhang.)*

Rui Tian, Jiaxuan Zhang, Man Zhang, Zian Zhang, Yin Zhang, and Yongqiang Li are with the School of Instrumentation Science and Engineering, Harbin Institute of Technology (HIT), Harbin 150001, China (e-mail: rui.tian.hit@gmail.com; jiaxuan.zhang.hit@gmail.com; man.zhang.hit@gmail.com; sieann.hit@gmail.com; yin.zhang.hit@gmail.com; liyongqiang@hit.edu.cn).

Yongqiang Zhang is with the College of Computer Science (College of Software-College of Artificial Intelligence), Inner Mongolia University, Hohhot 010031, China (e-mail: yongqiang.zhang.hit@gmail.com).

Wangmeng Zuo is with the School of Computer Science and Technology, Harbin Institute of Technology (HIT), Harbin 150001, China (e-mail: cswmzuo@gmail.com).

Digital Object Identifier 10.1109/TCSVT.2025.3613389

I. Introduction

ACTIVE learning object detection(ALOD) [1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [11], aiming at balancing between annotation costs(*i.e.*, instance-level labels) and detection performance, samples the most informative images that cover the data probability distribution of the entire dataset and maximally improve detection performance. Existing SOTA methods in ALOD initialize a fully supervised detector using randomly selected instance-level labels during the active learning (AL) initialization. Subsequently, the initialized detector mines the most informative images for human annotation, thereby facilitating the refinement of the detector. However, the performance of ALOD critically depends on the model initialization, and ALOD only uses an initialized detector to mine the most informative images. The result is that ALOD still requires instance-level labels during AL initialization and has limited localization ability of the initialized fully supervised detector during the AL process.

First, our objective is to explore an AL initialization method that can pre-train the fully supervised detector using only image-level labels, thereby reducing annotation costs. Inspired by large-scale vision-language pre-trained models(VLMs), such as Qwen2.5-VL [14], MiniGPT-v2 [12] and OWL-ViT [15], these models can localize the desired object by category-related prompts. A straightforward solution to eliminate the requirement for instance-level annotations during AL initialization is to generate pseudo-labels via VLMs conditioned on image-level labels. Previous studies [15], [16], [17], [18] have leveraged linguistic priors(*e.g.*, scene-context prompts, category-specific descriptors) encoded in VLMs to generate pseudo-labels for task-specialized detectors. These methods facilitate knowledge distillation from VLMs to downstream detectors, enhancing their zero-shot generalization capability while eliminating the need for manual annotations. However, these methods rely on introducing learnable prompt embeddings [17], [18], instruction tuning [16], or scaling self-training [15] to unlock VLM capabilities for downstream tasks, inevitably introducing new trainable parameters. In addition, ALOWD [5] decreases the annotation requirement by designing an auxiliary image generator during AL initialization. However, the auxiliary image generator still requires instance-level labels, although fewer than 5% images are annotated by instance-level labels. In contrast, we propose a training-free multi-step reasoning process based on chain-of-thought. By utilizing only image-level labels, the multi-step reasoning

1051-8215 © 2025 IEEE. All rights reserved, including rights for text and data mining, and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: Harbin Institute of Technology. Downloaded on June 23,2026 at 09:31:57 UTC from IEEE Xplore. Restrictions apply.

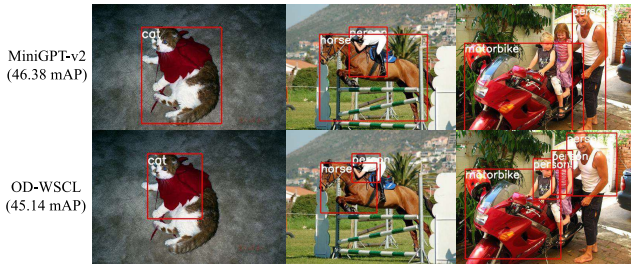

Fig. 1. The visualization results and performance in mAP of MiniGPT-v2 [12] and weakly supervised object detector OD-WSCL [13] on PASCAL VOC 2007 dataset. Note that we follow the conversation prompt template designs of MiniGPT-v2 [12] for object detection.

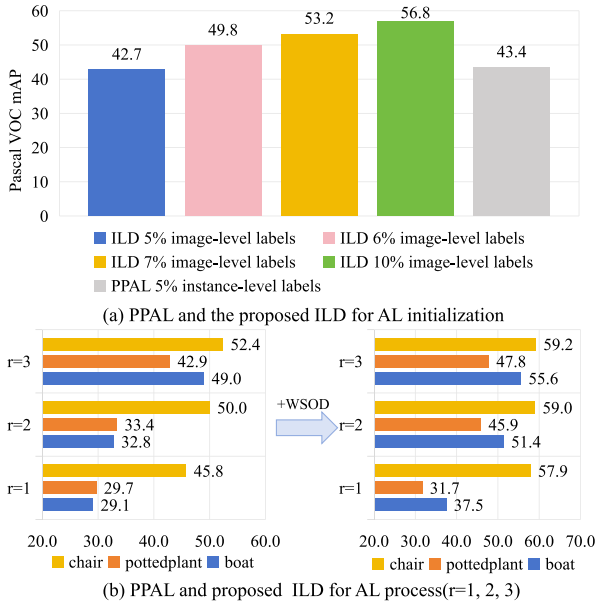

Fig. 2. The results on VOC 2007 test set of the initialized RetinaNet [20] and the first three rounds of the AL process by PPAL and ILD, where objects(chair, pottedplant, boat) with poor localization quality as examples are quantitatively shown in (b).

process activates the potential detection capabilities inherent in VLM for generating high-quality pseudo ground-truth boxes.

Given only image-level labels, we also naturally think of weakly supervised object detectors trained by image-level labels to locate objects. As shown in Figure 1, we found that the VLM detection results are achieved with high precision, while the weakly supervised detector predicts bounding boxes with high recall. Based on these foundations, we can integrate the object regions of the VLM and weakly supervised detector, and regard these regions as pseudo ground-truth boxes to initialize the fully supervised detector during AL initialization. As shown in Figure 2(a), our initialization with only 5% image-level labels achieves comparable results with the existing SOTA method PPAL [1] which requires 5% instance-level labels during AL initialization. With only a 6% image-level label ratio, our initialization surpasses PPAL based on 5% instance-level labels, eliminating the requirement of instance-level labels during AL initialization.

Second, these existing ALOD methods [1], [4], [5], [7], [8], [19] rely on designed acquisition functions and a single detector to mine the most informative images, but the localization capability of the initialized detector remains underdeveloped. As illustrated in Figure 2(b), representative examples of objects with poor localization quality (*e.g.*, chair, potted plant, boat) are quantitatively shown during the AL process. A detailed qualitative and quantitative comparison of the localization quality is presented in the Experiment Section. PPAL using a single fully supervised detector exhibits suboptimal localization capability due to insufficient pre-training, resulting in poor localization performance(*e.g.*, 29.1% AP for boat in the first round of the AL process) during the AL process. Thus, we propose a collaborative mechanism where uncertainty and diversity acquisition functions integrate results from both weakly supervised detector and pre-trained fully supervised detectors. As shown in Figure 2(b), we use the same random initialization method in AL initialization, then conduct AL process for PPAL and our proposed ILD. The localization performance during the AL process is improved by the proposed uncertainty and diversity acquisition functions based on the collaborative mechanism(*e.g.*, from 29.1% AP to 37.5% AP for boat). The uncertainty and diversity acquisition functions not only quantify the uncertainty of two detectors and the diversity of object features but also enhance the localization quality during the AL process.

Specifically, we design an Image-level Labels Driven active learning object detection method (termed as ILD) to decrease annotation costs during AL initialization and improve the localization ability of pre-trained detector during AL process. There are two issues in ILD. **(1) How to initialize a fully supervised detector without instance-level labels during AL initialization? (2) How to design the acquisition function to ensure the localization quality in the active learning phase?**

In the active learning initialization issue, we first propose a training-free multi-step reasoning process based on chain-of-thought to unlock the potential detection capabilities inherent to VLM. Inspired by the object detection process in human common sense [21] that they roughly count the number of desired objects before accurately localizing object regions and iteratively observe objects in images to prevent missing objects, a class-number-aware step and an iterative step in the multi-step reasoning process are well-designed to improve the detection performance of VLM(*e.g.*, MiniGPT-v2 [12]). Specifically, taking MiniGPT-v2 [12] as an example, we first ask the object number for each category, such as *[vqa] How many people are in the image?*[1] Then, we continue to query the spatial location of given objects with a class-number-aware step, such as *[detection] three persons*.[2] Moreover, we iteratively query the desired objects (*i.e.*, iterative step) to prevent missed objects and mine all object regions. Besides, as shown in Figure 1, we show the results of the weakly supervised detector(*i.e.*, OD-WSCL [13]), and it achieves the comparative performance(*i.e.*, 45.14% in mAP) with MiniGPT-v2

[1][vqa] is a visual question-answering task identifier of MiniGPT-v2.
[2][detection] is an object detection task identifier of MiniGPT-v2.

Authorized licensed use limited to: Harbin Institute of Technology. Downloaded on June 23,2026 at 09:31:57 UTC from IEEE Xplore. Restrictions apply.

(*i.e.*, 46.38% in mAP). Based on these experimental results, we conclude that VLM(*e.g.*, MiniGPT-v2 [12]) detects the tight bounding box with high precision but missing objects with low recall, and WSOD(*e.g.*, OD-WSCL [13]) achieves a high recall rate but highlighting discriminative object parts with low precision. Therefore, we choose VLM combined with WSOD to initialize a fully supervised detector in AL initialization.

In terms of designing the acquisition function during the AL process, we quantitatively reveal a limitation of the constrained localization capability of the pre-trained single detector during AL process as shown in Figure 2(b). To this end, we design uncertainty and diversity acquisition functions based on a collaborative mechanism to enhance the localization ability. In the collaborative mechanism, uncertainty and diversity acquisition functions integrate results from both the weakly supervised detector and the pre-trained fully supervised detector. For uncertainty, we utilize entropy to measure the classification confidence and localization difficulty of weakly supervised detector(*e.g.*, OD-WSCL [13]) and fully supervised detector(*e.g.*, RetinaNet [20]). Given a budget b, we select b images with the greatest uncertainty as the candidate set from the unlabeled set. For diversity, we record the maximum similarity for each object with same category and sum up the recorded similarities for all categories as the similarity between images. Besides, the similarity between images is used as the initial condition in the clustering process, and we select each cluster centroid as the informative image from the candidate set. Finally, these informative images annotated by humans with instance-level labels are utilized to further refine the fully supervised detector. In summary, we make the following contributions:

- We propose an ILD framework to eliminate the requirement of instance-level labels during the AL initialization while ensuring localization quality throughout the AL process.
- During AL initialization, a train-free multi-step reasoning process is proposed to deeply mine the detection ability of the VLM. The detection results of the enhanced VLM and weakly supervised detector are used as the pseudo ground-truth boxes to initialize a fully supervised detector, significantly reducing reliance on instance-level labels. During the active learning phase, uncertainty and diversity acquisition functions are proposed to leverage collaborative outputs from both the weakly supervised detector and the pre-trained fully supervised detector to select the most informative images, further enhancing localization quality.
- Extensive experiments demonstrate the effectiveness of our proposed ILD, and we achieve 77.5% mAP, 25.7% AP and 27.9% AP on VOC 2007 test set, COCO2014 val set and COCO 2017 val set, surpassing the SOTA methods by 3.4%, 1.2% and 4.7%, respectively.

II. Related Work

A. Active Learning for Object Detection

Active learning object detection [1], [2], [4], [6], [7], [22] aim to balance annotation costs and detection performance by selecting the most informative images, mainly considering the uncertainty and diversity of images. For example, Yang et al. [1] propose an uncertainty- and diversity-based sampling strategy to select images with difficult objects as informative images. Choi et al. [2] rely on mixture density networks to estimate the uncertainty for further annotating informative images. The performance of active learning depends on initialization, *e.g.*, Choi et al. [2] pre-train a fully supervised detector using 2,000 images with instance-level labels, which still requires a large annotation effort. Meanwhile, a few methods [4], [5], [23] mix a weakly supervised detector, instance-level or image-level labels in active learning. For example, Yuting et al. [5] propose a new auxiliary image generator strategy that utilizes an extremely small labeled set coupled with a large number of image-level labels to initialize a fully supervised detector. However, the auxiliary image generator still requires instance-level labels, although fewer than 5% images are annotated by instance-level labels. Vo et al. [4] design a successive strategy to extract the most informative images instead of acquisition functions, lacking quantification of detector uncertainty and sampled data diversity. Sartor et al. [23] using image-level labels and isolation forest design a bayesian active learning framework to directly output anomaly probability for anomaly detection. The performance of this framework is limited by the representation ability of the isolation forest.

The above ALOD research often focuses on AL initialization through randomly selected images and AL process via well-designed acquisition functions to identify the most informative images. When the initial annotated set exhibits distributional bias(*e.g.*, limited scene or few objects for special class), this bias propagates through subsequent active learning rounds, resulting in suboptimal sample selections with low diversity denoted as multi-peak distribution collapse [24]. From an annotation cost perspective, image-level labels are significantly easier to acquire, possibly even obtainable from the web. Under comparable annotation costs, we can obtain more labeled images with image-level labels than instance-level labels. Compared with previous works, our ILD strictly follows the weakly supervised paradigm using only image-level labels during AL initialization. Then, a collaborative mechanism where uncertainty and diversity acquisition functions integrate results from both weakly supervised and pre-trained fully supervised detector enhances localization ability during the AL process. To the best of our knowledge, our method is the first to initialize the fully supervised detector using only image-level labels and to integrate VLM with CoT and WSOD into ALOD.

B. Large-Scale Vision-Language Models

With the success of large-scale language models such as BERT [25] and GPT-3 [26], the initial explorations of the vision-language model [12], [27], [28] are started. For example, MiniGPT-v2 [12] relies on LLaMA-2 [29] language tokens to perform various vision-language tasks using proper instructional tuning. GPT4V [28] illustrates the ability to understand pixel space and perform advanced multimodal tasks for customizing arbitrary vision-language cases.

Authorized licensed use limited to: Harbin Institute of Technology. Downloaded on June 23,2026 at 09:31:57 UTC from IEEE Xplore. Restrictions apply.

Recently, with the advancement of VLMs, several works [15], [16], [17], [30] focus on using the robust prior knowledge and generalization ability of VLM to generate pseudo labels for task-specialized detectors. For example, Du et al. [16] propose the language model instruction strategy to transfer prior knowledge from VLMs to detectors using the relationships between visual concepts. Chen et al. [17] design a visually-enriched and spatially-localized captioning framework by introducing trainable prompt embeddings to effectively generate region-text labels. Matthias et al. [15] utilize OWL-ViT CLIP-L/14 [31] to scale up the detection data with self-training and generate pseudo ground-truth boxes in the image-text pairs for existing detectors. Liang et al. [30] design an automatic data engine based on VLM to query and label pseudo ground-truth boxes for open-vocabulary object detection.

To further improve the inference ability of VLM, Chain-of-Thought(CoT) allows models to decompose multi-step problems into intermediate steps [32], [33], [34], [35], [36], [37], [38], [39]. For example, Mondal et al. [39] propose a KAM-CoT framework that integrates CoT reasoning, knowledge graphs, and multiple modalities to generate contextual understanding and provide more informed answers for multimodal tasks. Kim et al. [38] design a multi-spectral CoT prompting strategy that combines the RGB image to produce final detection scores that are compatible with multi-spectral pedestrian detectors.

Compared with previous work [15], [16], [17], they utilize learnable prompt embeddings [17], instruction tuning [16], or scaling self-training [15] to unlock the capabilities of VLM and facilitate knowledge distillation from VLMs to downstream detectors, which inevitably introduces new trainable parameters. Moreover, the work [30] directly uses VLM to predict results as pseudo labels on the novel data, neglecting the potential detection ability of VLM. We propose a training-free multi-step reasoning process driven by image-level labels to activate the detection ability inherent to VLM. Then, we leverage the enhanced VLM to generate pseudo ground-truth boxes during AL initialization, eliminating the requirement for instance-level annotations.

C. Weakly Supervised Object Detection

Recent advances in weakly supervised object detection (WSOD) have been made by obtaining high-quality region proposals [40], [41], [42], [43], [44], [45], [46], [47], [48]. These methods aim to generate fewer region proposals with higher recall to reduce the difficulty of finding the correct proposals and to enhance the localization ability without instance-level labels. For example, Zhu et al. [46] propose an automated prompting pipeline that leverages classification cues to guide SAM in generating regional proposals. Yasuki and Taki [42] explore the capabilities of large kernel CNNs with large effective receptive fields to localize the object regions as region proposals for WSOD.

Existing SOTA weakly supervised detectors prevent missed objects without instance-level label supervision based on region proposals. Utilizing thousands of region proposals or high-quality region proposals in each image, WSOD covers most object instances despite highlighting the most discriminative regions. Meanwhile, image-level labels are significantly easier to acquire, even obtainable from the web. We maximize the use of efficient annotation in WSOD and make up for the problem of missing objects in VLM. VLM detects the tight bounding box with high precision but missing objects with low recall, and WSOD achieves a high recall rate but highlighting discriminative object parts with low precision. Therefore, we combine VLM and WSOD to initialize a fully supervised detector using only image-level labels during AL initialization. In addition, the weakly supervised detector and the pre-trained fully supervised detector are utilized to improve the localization quality during the AL process.

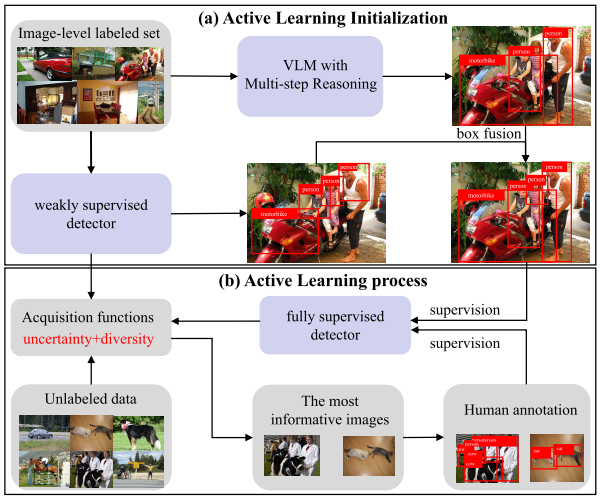

Fig. 3. The framework of our proposed ILD for active learning object detection, where VLM with multi-step reasoning is detailed in Figure 4. During active learning initialization, an off-the-shelf WSOD and a VLM with multi-step reasoning predict the detection results, and the detection results are used to initialize a fully supervised detector. In the active learning process, uncertainty and diversity acquisition functions leverage collaborative outputs from both the weakly supervised detector and the pre-trained fully supervised detector to select the most informative images. Then, the selected images annotated by humans are utilized to further refine the pre-trained detector.

III. Proposed Method

In this section, we first define the preliminaries and problem statements of our proposed ILD framework. Then, AL initialization with multi-step reasoning is described in section III-B, and well-designed uncertainty and diversity acquisition functions using the results of the weakly supervised detector and initialized detector during AL are described in Section III-C. The pipeline of our proposed method is shown in Figure 3.

A. Preliminaries and Problem Statement

Following the paradigm of ALOD, WSOD and VLM, there is a training set $X_T = \{x_i\}_{i \in N_t}$ with size N_t and a validation set $X_V = \{x_j\}_{j \in N_v}$ with size N_v. The weakly supervised detector M^w(*e.g.*, OD-WSCL [13]) is learned from image-level labels $y = [y_1, y_2, \ldots, y_c] \in \mathbb{R}^{C \times 1}$, where C denotes the number of image classes. The large-scale vision-language pre-trained model M^{vl}(*e.g.*, MiniGPT-v2 [12]) localizes the desired objects

Authorized licensed use limited to: Harbin Institute of Technology. Downloaded on June 23,2026 at 09:31:57 UTC from IEEE Xplore. Restrictions apply.

by category-related prompt templates and task identifiers. The fully supervised detector(*e.g.*, RetinaNet [20]) M^f is initialized from the detection results of M^w and M^{vl} and further refined by selecting the most informative images annotated with instance-level labels from X_T based on active learning (AL). We first define the set with image-level labels as X_{ilb}. Then, we define the set without instance-level labels as X_U^r and the set with instance-level labels as X_L^r, where $X_U^r \cup X_L^r = X_T$ and r denotes the active learning round. In round $r > 0$ with the active learning process, we gradually select the most informative image set X_Q^r to annotate instance-level labels. Then, we get the unlabeled instance-level set $X_U^{r+1} = X_U^r - X_Q^r$ and the labeled instance-level set $X_L^{r+1} = X_L^r + X_Q^r$, and the fully supervised detector is refined by X_L^{r+1} during r round of AL process.

B. AL initialization with Multi-step Reasoning

Considering that the performance of active learning depends on initialization, we regard the detection results of weakly supervised detector(*e.g.*, OD-WSCL) and large-scale vision-language pre-trained model(*e.g.*, MiniGPT-v2) as pseudo ground-truth boxes to initialize a fully supervised detector(*e.g.*, RetinaNet). For the weakly supervised detector, it relies on region proposal generation methods(*e.g.*, Selective Search [43] and Multiscale Combinatorial Grouping [44]) to predict object regions $\widetilde{y}_w = M^w(R, I)$ with a high recall rate, where M^w denotes weakly supervised detector, R and I denote region proposals and input images. For VLM, taking miniGPT-v2 as an example, it adopts unique identifiers(*e.g.*, [detection]) and category-related prompts to infer object regions $y = M^{vl}(I, T)$, where M^{vl} denotes the VLM model, and T denotes category-related prompts. However, as shown in Figure 1, MiniGPT-v2 directly localizes the desired objects using category-related prompts, resulting in missed objects.

To further improve the detection ability of VLM and the performance of AL initialization, we propose a multi-step reasoning process based on the chain-of-thought(CoT), which includes a class-number-aware step and an iterative step, as shown in Figure 4. In the class-number-aware step, imitating the object detection process in human common sense [21], we first utilize the task identifier [*vqa*] to ask MiniGPT-v2 about the number of each class in the image based on image-level labels (*e.g.*, [vqa] How many people are in the image?), which can be formulated as:

$$n = M^{vl}(I, T_1), \tag{1}$$

where T_1 denotes the prompt with the task identifier [*vqa*] for reasoning class number, n denotes the answer from MiniGPT-v2 about the class number. Based on the number of each class, MiniGPT-v2 is further used to localize objects by combining with the task identifier [*detection*], category-related prompt and class number (*e.g.*, [detection] three (3) people), which can be expressed by eq. (2):

$$\hat{y} = M^{vl}(I, T_2, n), \tag{2}$$

where $\hat{y}$ denotes the predictions of MiniGPT-v2, T_2 is the prompt with task identifier [*detection*] for localization.

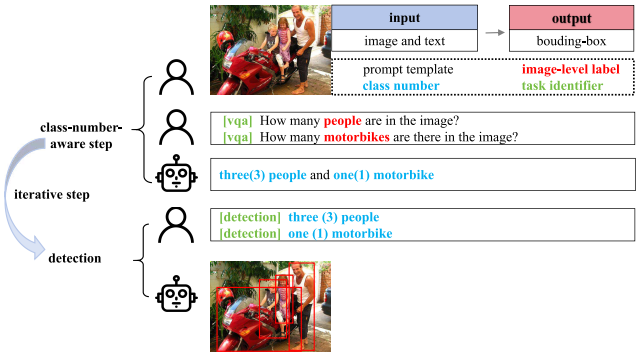

Fig. 4. The multi-step reasoning process based on chain-of-thought, including a class-number-aware step and an iterative step. We imitate the process of object detection by the human in common sense [21]. Taking MiniGPT-v2 as an example, we first ask MiniGPT-v2 about the number of different classes in the image using the task identifier [vqa]. Then, MiniGPT-v2 further detects the desired object by combining the task identifier [detection]. Finally, we iteratively ask the class number and object location to prevent missed objects. The prompt template, image-level label, class number and task identifier are labeled by black, red, blue and green fonts respectively.

Then, we iteratively ask the class number and detect object locations (called iterative step) to prevent missing objects, which is essentially a recursive process formulated as:

$$\hat{y}_m = M^{vl}(I, T_2, n_m), \qquad m = 1, 2, \ldots, M \tag{3}$$

where M is the maximum iteration steps, n_m denotes the answer of MiniGPT-v2 about class number in the m-th step. $\hat{y}_m$ indicates the m-th predictions of MiniGPT-v2. We set M to 2 as the stopping criterion detailed in the hyperparameter ablation studies IV-E.

Thanks to the multi-step reasoning process, the detection ability of VLM is improved by a large margin. Benefit from the high recall rate of the weakly supervised detector and the high precision rate of VLM, we construct a simple and effective box fusion strategy to mine the pseudo ground-truth boxes for active learning initialization. Specifically, we select all detection results of VLM and the detection results of weakly supervised detector whose confidence score over a threshold value α and IoU less than a threshold β with the detection results of VLM as the pseudo ground-truths y, which can be formulated as:

$$\mathrm{y} = \hat{y}_m \cup \eta_1 \cdot \widetilde{y}_w, \tag{4}$$

$$\eta_1 = \begin{cases} 1, & \text{if } \mathrm{C}(\widetilde{y}_w) > \alpha \text{ and } \mathrm{IoU}(\hat{y}_m, \widetilde{y}_w) < \beta \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

where $C(\cdot)$ denotes the confidence score. We set the confidence threshold α and IoU threshold β to 0.95 and 0.5, as detailed in the hyperparameter ablation studies IV-E.

C. Uncertainty and Diversity Acquisition Functions

After initializing the fully supervised detector M^f using the obtained pseudo ground-truth boxes, the detector has preliminary detection ability. So far, we follow the weak supervision paradigm, *i.e.*, only accessing image-level labels, eliminating the requirement of annotation compared to existing ALOD methods [1], [2], [7]. Following SOTA method

Authorized licensed use limited to: Harbin Institute of Technology. Downloaded on June 23,2026 at 09:31:57 UTC from IEEE Xplore. Restrictions apply.

PPAL [1] in ALOD, it utilizes the information entropy and image similarity as acquisition functions to measure the uncertainty and diversity between images. We focus on improving the information entropy and image similarity to further enhance the localization quality of the pre-trained detector during the AL process. We propose a collaborative mechanism where uncertainty and diversity acquisition functions integrate results from both the weakly supervised detector M^w and pre-trained fully supervised detector M^f. The uncertainty and diversity acquisition functions not only quantify the uncertainty of two detectors and the diversity of object features but also enhance the localization quality of the pre-trained detector. Finally, the informative images X_Q^r in the unlabeled dataset X_U^r are selected using uncertainty and diversity acquisition functions to further refine the initialized detector M^f.

1) Uncertainty Acquisition Function: Different object categories exhibit their inherent difficulty for classification, such as the ambiguity of visual features between birds and airplanes in the distant image. The class-wise difficulty facilitates the selection of challenging categories critical for improving detector performance. Otherwise, the AL process may repeatedly select easily classifiable samples. Moreover, class-wise difficulty is updated by an exponential moving average(EMA), which reflects both instance-wise difficulty for each same-category instance in the previous training batch(long-term) and recent training batch(short-term). Thus, we first introduce the image-wise uncertainty acquisition function in eq. (6), which is based on the category-related weight in eq. (7) and the class-wise difficulty in eq. (8), to select informative images.

$$U(I) = \sum_{j=1}^{N_I} w_{c(j)} \cdot \sum_{i=1}^{C} -p_{ij} \cdot log(p_{ij}), \tag{6}$$

$$w_c = 1 + \lambda \cdot \gamma log(1 + (e^{1/\gamma} - 1) \cdot g_c), \tag{7}$$

$$g_c^k \leftarrow m_c^{k-1} g_c^{k-1} + (1 - m_c^{k-1}) \frac{1}{N_c^k} \sum_{j=1}^{N_c^k} d_j, \tag{8}$$

where U denotes the uncertainty acquisition function, N_I and C denote the number of detected objects by initialized M^f from image I and the classification number respectively, and p_{ij} denotes the predicted probability of the category i corresponding to object j. $w_{c(j)}$ indicates the category-related weight for the object j corresponding to the category c obtained by class-wise difficulty g_c. w_c is the category-related weight corresponding to category c, which does not contain the mean of the object. The class-wise difficulty g_c is introduced and updated by the instance-level difficulty d using an exponential moving average(EMA) during training, where k and N_c^k denote the iteration and the number of objects corresponding to category c in the training batch in the k-th iteration. The g_c^k is updated in each iteration. The λ, γ and m are hyper-parameters set as in PPAL [1].

Note that in eq. (8) the instance-wise difficulty d is related to the classification probability and localization quality, which are usually calculated from the fully supervised detector initialized by randomly selected images with instance-level labels as in PPAL [1]. However, the localization quality of initialized fully supervised detector M_f is limited as shown in Figure 2, because the AL process relies on a single fully supervised detector trained by randomly selected images with instance-level labels or obtained pseudo ground-truth boxes that are not necessarily informative images. To overcome this issue, we collaboratively use the results of weakly supervised detector M^w and fully supervised detector M^f to obtain instance-wise difficulty for better measuring the instance-wise uncertainty. Specifically, we choose the maximum IoU between $\widetilde{y}_w$, y_f and y_{gt} to calculate the instance-wise difficulty d:

$$d = 1 - P(y_f|y_{gt})^{\rho} \cdot max(IoU(y_f, y_{gt}), IoU(\widetilde{y}_w, y_{gt}))^{1-\rho}, \tag{9}$$

where d indicates instance-wise uncertainty, y_f and $\widetilde{y}_w$ denote the predicted boxes from M^f and M^w respectively y_{gt}, is the pseudo ground-truth boxes obtained AL initialization and the instance-level label annotated by AL. $P(\cdot)$ indicates the classification probability, ρ is a hyper-parameter with a value range of $0 \leq \rho \leq 1$.

Here, we provide a detailed case to clarify the workflow of the uncertainty acquisition function in the AL process. First, during the AL process, we first calculate the instance-wise difficulty d. Given a ground truth box y_{gt} in X_L^r, the corresponding detection results y_f and $\widetilde{y}_w$ are predicted by M^f and M^w, and the classification probability is recorded as $P(y_f|y_{gt})$. Following eq. (9), the instance-level difficulty d for each instance y_{gt} can be calculated. Then, g_c^0 and m_c^0 of each category c are initialized with 1 and 0.99. We average the instance-wise difficulty of the same-category objects in the training batch and continuously update g_c^k using m_c^{k-1} following an exponential moving average. When there are no objects of class i in the training batch, we decrease m_c^k by multiplying m_c^0(*i.e.*, $m_c^k = m_c^{k-1} \cdot m_c^0$), else m_c^k is numerically set to m_c^0. Until the training iteration is complete, we obtain the class-wise difficulty g_c for each category and the category-related weight w_c following eq. 8 and eq. 7. In addition, for each image of X_U^r, we calculate the image-wise uncertainty acquisition by weighting the entropy of all detected objects based on Eq. 6. Note that when $r = 1$, y_{gt} represents the pseudo ground-truth boxes obtained during initialization; when $r > 1$, y_{gt} corresponds to a combination of pseudo ground-truth boxes and ground-truth boxes. Finally, given a budget b, we select b the most uncertain images (called the candidate set $\mathbb{C}$) based on the above uncertainty acquisition function from the unlabeled set X_U^r.

2) Diversity Acquisition Function: After measuring the image-wise uncertainty, we further select informative images X_Q^r from the candidate set based on the diversity acquisition function, which is implemented by minimizing the similarity between images. In the diversity acquisition function, we still integrate the weakly supervised detector M^w and the initialized fully supervised detector M^f together to measure the similarity between images.

Specifically, for two images I_a and I_b, the weakly supervised detector M^w and the fully supervised detector M^f predict the object triplets, termed as O'_a, O'_b and O''_a, O''_b that includes the object regions, object categories and its confidence, and visual feature of each object, respectively. Here, O'_a and O''_a denote the predictions of M^w and M^f in images I_a. The O'_b and O''_b

Authorized licensed use limited to: Harbin Institute of Technology. Downloaded on June 23,2026 at 09:31:57 UTC from IEEE Xplore. Restrictions apply.

in the images I_b have meanings similar to the above. Then, we use the similarity of O'_a, O'_b and O''_a, O''_b as a proxy for the similarity of I_a and I_b, which is computed by matching each object in the current image with its most similar counterpart in other images defined as:

$$S(I_a, I_b) = \frac{1}{4}(S'(O'_a, O'_b) + S'(O'_b, O'_a) + S''(O''_a, O''_b) + S''(O''_b, O''_a)), \tag{10}$$

$$S^{'or''}(O_a, O_b) = \frac{1}{\sum t_{a,i}} \sum_{i=1}^{M_a} t_{a,i} \cdot max(sim(o_{a,i}, O_b)), \tag{11}$$

$$S^{'or''}(O_b, O_a) = \frac{1}{\sum t_{b,i}} \sum_{i=1}^{M_b} t_{b,i} \cdot max(sim(o_{b,i}, O_a)), \tag{12}$$

where $sim(o_{a,i}, O_b)$ denotes the cosine similarity between the visual feature of object i in image I_a and each object visual feature in image I_b with the same category. M_a and $t_{a,i}$ denote the object number and category confidence of object i in the image I_a, respectively. The eq. (12) is similar to eq. (11). Note that the cosine similarity function $sim(\cdot)$ is set to 0 when the category of objects in the image I_a does not exist in the image I_b, otherwise we calculate the cosine similarity between the feature of the object of the same category in image I_a and image I_b. We use $o_{a,i}$ and O_b as an example to formalize the cosine similarity function:

$$sim(o_{a,i}, O_b) = \begin{cases} 0 & if\ c_{a,i} \notin c_{b,*} \\ \dfrac{o_{a,i} \cdot o_{b,j}}{\|o_{a,i}\| \cdot \|o_{b,j}\|} & if\ c_{a,i} \in c_{b,*} \\ and\ c_{a,i} = c_{b,j} & \end{cases} \tag{13}$$

where $c_{b,*}$ denotes the list of all predicted categories in image b. $c_{a,i}$ and $c_{b,j}$ denote the category of objects i and j in the image I_a and I_b. $o_{a,i}$ and $o_{b,j}$ indicate the visual features of object i and object j in the image I_a and the image I_b.

Based on the calculated similarity above between I_a and I_b, we mine the q images from the candidate set $\mathbb{C}$ that exhibit low similarity with each other as cluster centroids. Our objective is to maximize the inter-cluster distance with the minimum image similarity:

$$X_Q^r = arg \max_{X_Q^r \subseteq \mathbb{C}} \min_{\substack{a \neq b \\ I_a, I_b \in \mathbb{C}}} S(I_a, I_b) \qquad s.t. |X_Q^r| = q \tag{14}$$

where X_Q^r denotes the set of informative images, $S(I_a, I_b)$ is calculated by eq. 10. All images in the candidate set $\mathbb{C}$ are clustered into q clusters based on cosine similarity and Euclidean distance. During the cluster process, we update the elements in the X_Q^r using cluster centroids. When the cluster process is finished, the X_Q^r is determined as the set of informative images, which are annotated by humans with instance-level labels to further refine the detector M^f. To significantly enhance the understanding, a comprehensive algorithm diagram that contains the overall workflow of ILD as shown in algorithm 1. Meanwhile, the explanation of the key symbols is described in Table I.

Algorithm 1 Pseudo Code of ILD

Require: pre-trained vision-language models M^{vl},
pre-trained weakly supervised detector M^w,
fully supervised detector M^f,
the set without instance-level labels X_U^r,
the set with instance-level labels X_L^r,
the set with image-level labels X_{ilb}

```
//AL initialization
for I to X_ilb do
    M^vl and M^f predict detection results ŷ_m and ỹ_w on I.
    obtain pseudo ground-truth boxes y by eq. 4 and eq. 5.
    y is saved in temp list y_temp.
end for
M^f is initialized by y_temp.
pseudo ground-truth boxes y in y_temp as y_gt.
//AL process
for r to 6 do
    predict y_f and ỹ_w by M^f and M^w corresponding to y_gt.
    record classification probability P(y_f|y_gt).
    calculate instance-wise difficulty d by eq. 9.
    update class-wise difficulty g and obtain category-related
    weight w by eq. 8 and eq. 7.
    for I to X_U^r do
        calculate the image uncertainty by eq. 6.
    end for
    select b uncertainty images to append to ℂ.
    mine q images from ℂ that exhibit low similarity with
    each other using eq. 10.
    select the set of informative images X_Q^r by eq. 14.
    X_U^{r+1} = X_U^r − X_Q^r, X_L^{r+1} = X_L^r + X_Q^r, X_L^{r+1} refine M^f,
    y_gt = X_Q^r ∪ y_gt
end for
```

Ensure: M^f

IV. Experiments

In this section, we experimentally validate our ILD framework and analyze the effectiveness of each component, where includes some ablation studies and comparisons against other state-of-the-art detectors on PASCAL VOC 2007 dataset [49] and MS COCO 2014&2017 datasets [50]. Then, some qualitative results of our ILD are shown. Moreover, we further explore the influence of hyper-parameters in ILD.

A. Datasets and Evaluation Metrics

We evaluate the proposed ILD method on three widely used PASCAL VOC 2007 [51] and MS COCO [50] 2014&2017 datasets. The VOC 2007 and VOC 2012 datasets contain 5011 images and 11540 images on trainval set respectively, and VOC 2007 dataset contains 4952 images on test set from 20 object categories. We use VOC 2007+2012 trainval set for training and use VOC 2007 test set for testing. The MS COCO 2017 contains around 118000 training images and 5000 validation images from 80 categories, and we train ILD on train set and test ILD on validation set. On COCO 2014 dataset, we train ILD with the train split (82783 images) and evaluate it on the validation split (40504 images). For evaluation metrics, it complies with the PASCAL criterion on VOC datasets, and a positive detection has an IoU>0.5 with the ground-truth box. On MS COCO dataset, the performance

Authorized licensed use limited to: Harbin Institute of Technology. Downloaded on June 23,2026 at 09:31:57 UTC from IEEE Xplore. Restrictions apply.

TABLE I
THE EXPLANATION OF THE KEY SYMBOLS IN ILD

symbols	explanation
α	confidence score threshold value in box fusion strategy
β	IoU threshold value in box fusion strategy
U_I	image-wise uncertainty acquisition function
$w_{c(j)}$	category-related weight for object j corresponding to category c
w_c	category-related weight of category c
g_c^k	class-wise difficulty of category i in k-th iteration
O_*	object triplets
$o_{*,i}$	visual feature for object i in image I_*
$c_{*,i}$	predicted category for object i in image I_*
$t_{*,i}$	category confidence for object i in image I_*
$S(I_a, I_b)$	similarity of I_a and I_b

of all experiments follows the standard COCO-style precision metrics, *i.e.* AP(IoU range of 0.5:0.95:0.05), AP50(IoU@0.5) and AP50(IoU@0.75) are reported in our paper.

B. Implementation Details

In our ILD framework, we select OD-WSCL [13] or OICR [52] as the weakly supervised detector, and use MiniGPT-v2 [12], Qwen2.5-VL [14] or OWL-ViT [15] as the VLM in the multi-step reasoning stage during AL initialization. In the active learning stage, we use RetinaNet [20] or Deformable-DETR [53] as the fully supervised detector. All experimental setups are consistent with these methods unless specifically emphasized.

In the AL initialization with multi-step reasoning process, for the training set of VOC 2007+2012 and MS COCO 2014/2017 datasets, we select the top 5000 images with image-level labels to generate pseudo ground-truth boxes and use them to initialize the fully supervised detector. In the multi-step reasoning process, we set the iterative number $M = 2$ to prevent missing objects. Moreover, we set $\alpha = 0.95$ and $\beta = 0.5$ in eq. (5) as the confidence threshold and IoU threshold to fuse detection results from WSOD and VLM as pseudo ground-truth boxes.

In the AL stage, following PPAL [1], we set $\rho = 0.6$ in eq. (9) to calculate the instance-wise difficulty, and set the total AL rounds r to 6 and 4 for PASCAL VOC and MS COCO datasets, respectively. Meanwhile, the number of candidate images b and the cluster centroid q is set to 1656 and 414. The base EMA momentum m_0 in eq. (8) is set to 0.99 and the γ and λ in eq. (7) are set to 0.3 and 0.2 respectively. We use the PyTorch deep learning framework [54] and 8 NVIDIA RTX 3090 GPUs to train our model.

C. Main Results

To validate the effectiveness of our proposed method, we compare our ILD with well-known fully supervised detectors (*i.e.*, RetinaNet [20] and Deformable-DETR [53]) on PASCAL VOC 2007 and MS COCO 2017 datasets. As shown in Table II, we explore initialized fully supervised by a% images with instance-level labels, and select c% the most informative images with instance-level labels to refine the detector in each round until X_L^r contains b% images. The a%:b%:c% denotes instance-level label proportion during AL process, such as $a\% = 0\%, b = 15\%, c\% = 2.5\%$. Our proposed ILD framework combining with RetinaNet [20] achieves 77.5%, 43.2% and 26.6% in mAP, AP50 and AP(0.5:0.95:0.05) on VOC 2007 test set and MS COCO 2017 val set with only accessing 15% and 8% instance-level labels, which are comparable to the performance of RetinaNet [20] under the fully supervised setting. Moreover, to validate the generalization of our proposed method, we integrate Transformer-based Deformable-DETR [53] into ILD, we achieve 72.9% in mAP with 15% instance-level labels, which is close to 78.5% in fully supervised manner. For a fair comparison, we randomly select 15% and 8% images with instance-level labels to train RetinaNet [20] and Deformable-DETR [53]) on VOC 2007 and COCO 2017 dataset, and ILD outperforms them by about 10% absolutely, as shown in the 3rd, 4th, 5th and 6th rows of Table II. Meanwhile, as shown in the 7th row of Table II, we explore OICR and MiniGPT-v2 to collaboratively conduct active learning for RetinaNet. The detection performance of ILD integrated with the weakly supervised detector OICR and OD-WSCL on VOC and COCO achieves a comparable result(*i.e.*, 77.5% *vs.* 77.2%, 43.2% *vs.* 42.1%, 26.6% *vs.* 26.0%). The above comparisons demonstrate the effectiveness of our ILD, which is insensitive to weakly supervised detectors.

Moreover, we quantitatively show the detection performance of PPAL and our proposed ILD under consistent AL initialization, including random initialization and distribution bias initialization. Random initialization means that the fully supervised detector is initialized by randomly selected images with instance-level labels, and distribution bias initialization means that the AL initialization set only contains minimal instance-level labels for people, dogs, and cats. As shown in Table III, we can observe those container-like objects(*e.g.*, boat, potted-plant, chair, handbag, sink, bench) are difficult to locate by PPAL during the AL process under random initialization on VOC and COCO datasets. Based on the collaborative mechanism of WSOD and pre-trained fully supervised detector in uncertainty and diversity acquisition functions, ILD improves the performance of these difficult-to-localize objects in each round of active learning, further enhancing the localization quality of the fully supervised detector. Furthermore, under the distribution bias of AL initialization, the detection performance of people, dogs, and cats is significantly decreased during AL process(*e.g.*, decrease 40.2%, 30.8%, 30.4% in AP for person, dog, cat in the first round of AL process). This performance degradation becomes especially evident during the early phase of AL process(*i.e.*, r=1, 2, 3). Under such a distribution bias, our ILD still enhances localization quality during the AL process. To further demonstrate the robustness of our ILD, we conduct a comprehensive analysis for the improvement of our ILD under the setting of distribution bias initialization. During the AL initialization, distribution bias initialization leads the optimization process of the fully supervised detector to primarily rely on gradient updates from those majority categories. For rare categories (*e.g.*, person, dog, and cat in Table III), the fully supervised detector struggles to capture their diverse visual representation and tends to confuse their visual representation with the background or visually similar categories. Thus, those informative images mined

Authorized licensed use limited to: Harbin Institute of Technology. Downloaded on June 23,2026 at 09:31:57 UTC from IEEE Xplore. Restrictions apply.

TABLE II

THE MAP(VOC), AP50(COCO) AND AP(COCO) PERFORMANCE OF OUR PROPOSED ILD AND OTHERFULLY SUPERVISED DETECTORS ON VOC2007 TEST SET AND MS COCO 2017 VAL SET UNDER THE SETTING OF DIFFERENT INSTANCE-LEVEL LABEL PROPORTIONS, THE PROPORTION DENOTES INSTANCE-LEVEL LABEL PROPORTION IN DATASETS.† DENOTES THAT OICR AND MINIGPT-V2 COLLABORATIVELY CONDUCT ACTIVE LEARNING FOR RETINANET.* MEANS TRAINING ON 07+12 TRAINVAL SET.A%:B%:C% OF INSTANCE-LEVEL LABEL PROPORTION DENOTES THE DETECTOR IS INITIALIZED BY A% IMAGES WITH INSTANCE-LEVEL LABELS, AND SELECT C% THE MOST INFORMATIVE IMAGES WITH INSTANCE-LEVEL LABELS TO REFINE THE DETECTOR IN EACH ROUND UNTIL X_L^r CONTAINS B% IMAGES. THUS, THE TOTAL ACTIVE LEARNING ROUNDS ARE $(b-a)\%/c\%$

Method	VOC		COCO		
	proportion	mAP	proportion	AP50	AP(0.5:0.95:0.05)
RetinaNet [20]	100%	81.6*	100%	54.9	36.0
Deformable-DETR [53]	100%	78.5*	100%	65.2	46.2
RetinaNet [20]	15%(random)	63.8*	8%(random)	34.1	20.2
Deformable-DETR [53]	15%(random)	62.4*	8%(random)	39.3	24.4
ILD with RetinaNet (Ours)	0%:15%:2.5%	77.5*	0%:8%:2%	43.2	26.6
ILD with Deformable-DETR (Ours)	0%:15%:2.5%	72.9*	0%:8%:2%	45.6	27.9
ILD with RetinaNet (Ours)†	0%:15%:2.5%	77.2*	0%:8%:2%	42.1	26.0

TABLE III

THE LOCALIZATION ABILITY OF PPAL AND ILD DURING AL PROCESS, WHERE DIFFICULT-TO-LOCATE OBJECTS ARE LISTED. THE RANDOM INITIALIZATION DENOTES THE FULLY SUPERVISED DETECTOR IS INITIALIZED BY RANDOMLY SELECTED IMAGES WITH INSTANCE-LEVEL LABELS, AND DISTRIBUTION BIAS INITIALIZATION MEANS THAT AL INITIALIZATION SET ONLY CONTAINS MINIMAL INSTANCE-LEVEL LABELS FOR PEOPLE, DOGS, AND CATS

Method	round	random initialization		distribution bias initialization	Method	random initialization		distribution bias initialization
		VOC	**COCO**	**VOC**		**VOC**	**COCO**	**VOC**
PPAL	r=1	boat=29.1 pottedplant=29.7 chair=45.8	handbag=2.5 sink=9.4 bench=5.8	person=32.5(↓40.2) dog=40.7(↓30.8) cat=46.1(↓30.4)	ILD	boat=37.5 pottedplant=31.7 chair=57.9	handbag=3.2 sink=9.8 bench=7.4	person=40.1 dog=42.5 cat=48.7
	r=2	boat=32.8 pottedplant=33.4 chair=50.0	handbag=3.5 sink=10.2 bench=8.0	person=56.2(↓20.2) dog=60.0(↓16.8) cat=62.5(↓17.3)		boat=51.4 pottedlant=45.9 chair=59.0	handbag=4.9 sink=11.8 bench=9.0	person=59.2 dog=62.0 cat=68.5
	r=3	boat=49.0 pottedplant=42.9 chair=52.4	handbag=4.8 sink=10.0 bench=7.7	person=68.9(↓8.8) dog=69.0(↓7.7) cat=72.7(↓8.8)		boat=55.6 pottedplant=47.8 chair=59.2	handbag=5.7 sink=13.1 bench=8.0	person=69.0 dog=70.6 cat=72.6
	r=4	boat=48.9 pottedplant=44.1 chair=54.9	handbag=5.3 sink=12.4 bench=9.2	person=74.5(↓4.4) dog=71.2(↓7.1) cat=75.5(↓7.5)		boat=56.0 pottedplant=48.8 chair=61.4	handbag=6.5 sink=13.7 bench=9.9	person=74.9 dog=71.0 cat=75.4
	r=5	boat=50.2 pottedplant=46.5 chair=55.9		person=75.0(↓4.6) dog=72.8(↓7.5) cat=80.4(↓3.0)		boat=57.4 pottedplant=50.2 chair=62.0		person=76.2 dog=73.1 cat=80.8
	r=6	boat=54.0 pottedplant=48.4 chair=57.0		person=75.8(↓4.9) dog=72.7(↓7.7) cat=80.8(↓3.3)		boat=57.9 pottedplant=50.5 chair=62.2		person=76.0 dog=72.5 cat=81.1

by acquisition functions in PPAL [1] are unreliable during the AL process. By our designed collaborative mechanism, WSOD captures most regions of each object, thereby enabling more accurate instance-wise difficulty d in Eq. 9 and image similarity S in Eq. 11. The collaborative mechanism in our acquisition functions effectively helps in correcting the bias of the pre-trained fully supervised detector introduced during AL initialization with distribution bias.

To demonstrate the generalizability and adaptability of the proposed ILD, we integrate various VLMs(*i.e.*, MiniGPT-v2 [12], OWL-ViT [15], Qwen2.5-VL [14]) into our ILD framework. As shown in Figure 5, the performance of various VLMs integrated into the ILD framework yields comparable results, where MiniGPT-v2 exhibits slightly better performance within ILD(*i.e.*, ILD with MiniGPT-v2, Qwen2.5-VL and OWL-ViT achieve 77.5%, 75.5% and 75.3% in mAP, respectively). The reason is that Qwen2.5-VL and OWL-ViT encounter a higher missed detection rate compared to MiniGPT-v2 during AL initialization, leading to slightly lower initial performance. Consequently, the overall performance of the AL process is marginally lower than that achieved using ILD with MiniGPT-v2.

Meanwhile, we also make comparisons of runtime / memory efficiency with other AL pipelines, as shown in Table IV. Compared to baseline PPAL [1], we increase 1.5G, 10G, 11G memory for OWL-ViT [15], MiniGPT-v2 [12], Qwen2.5-VL [14] within ILD. In our multi-step reasoning process based on Chain-of-Thought during AL initialization, generating pseudo ground-truth boxes constitutes the primary runtime overhead. Our ILD with OWL-ViT [15],

Authorized licensed use limited to: Harbin Institute of Technology. Downloaded on June 23,2026 at 09:31:57 UTC from IEEE Xplore. Restrictions apply.

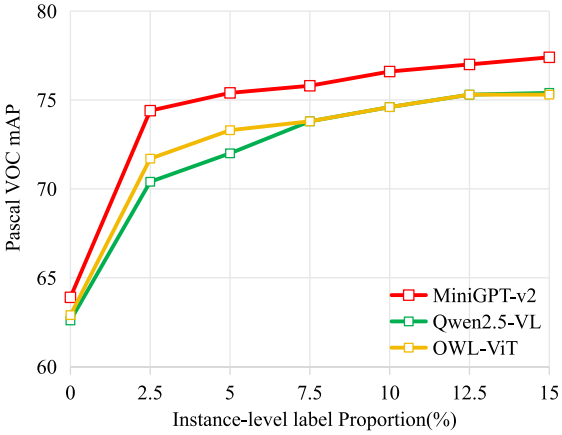

Fig. 5. The ILD performance using various VLMs, including MiniGPT-v2 [12], OWL-ViT [15], Qwen2.5-VL [14].

TABLE IV
COMPARISONS OF RUNTIME/MEMORY EFFICIENCY WITH OTHER AL METHODS

Method	Memory	AL initialization Runtime
VDPL-MIOC [55]	1.8 GB per GPU	—
PPAL(baseline) [1]	2.0 GB per GPU	—
ILD with OWL-ViT [15]	3.5 GB per GPU	80ms
ILD with MiniGPT-v2 [12]	12.0 GB per GPU	1200ms
ILD with Qwen2.5-VL [14]	13.0 GB per GPU	540ms

MiniGPT-v2 [12], and Qwen2.5-VL [14] exhibits an average inference latency of ~80 ms, ~1200 ms, and ~540 ms per image, computed on the first 100 images of the VOC 2007 test set with 8 NVIDIA RTX 3090 GPUs. The "AL initialization Runtime" in Table IV represents the computational overhead introduced during the AL initialization phase. Both VDPL-MIOC [55] and PPAL [1] utilize randomly selected images with instance-level labels to initialize the fully supervised detector (*i.e.*, RetinaNet [20]). In contrast to our AL initialization method with VLM, the paradigm of random selection in VDPL-MIOC [55] and PPAL [1] introduce no additional hardware computational cost. Therefore, the "AL initialization Runtime" for VDPL-MIOC [55] and PPAL [1] is denoted by "—" in Table IV. Notably, the inference computational overhead of fully supervised detectors remains unchanged.

D. State-of-the-Art Comparison

We compare our proposed ILD with other active learning object detection methods [1], [2], [6], [7], [8], [10], [55], [56], [57], [58], [59] on VOC2007 test and MS COCO 2017 val sets, and the comparison results are shown in Table V. We can observe that we surpass all previous approaches. For example, our ILD exceeds the previous SOTA method EDL-HUA [10] by 3.4% (74.1% *vs.* 77.5%) in mAP on VOC2007 test while using fewer instance-level labels(15% *vs.* 20%).

As shown in Figure 6(a), we visualize the performance of ILD and previous SOTA methods during AL step by step on VOC 2007 test set. Note that previous methods access instance-level labels to initialize the detector, while we do not access any instance-level labels (*i.e.*, only using image-level labels to generate pseudo bounding-boxes). Thus, we start the AL process with the 0% instance-level label proportion. The results show that ILD is obviously better than the previous methods. For example, benefiting from the well-designed AL initialization method, ILD with RetinaNet [20] and Deformable-DETR [53] achieve a notable performance 62.9% and 56.2% in mAP with 0% instance-level label proportion. Moreover, our ILD with RetinaNet using only 2.5% instance-level labels surpasses previous SOTA methods using 20% instance-level labels.

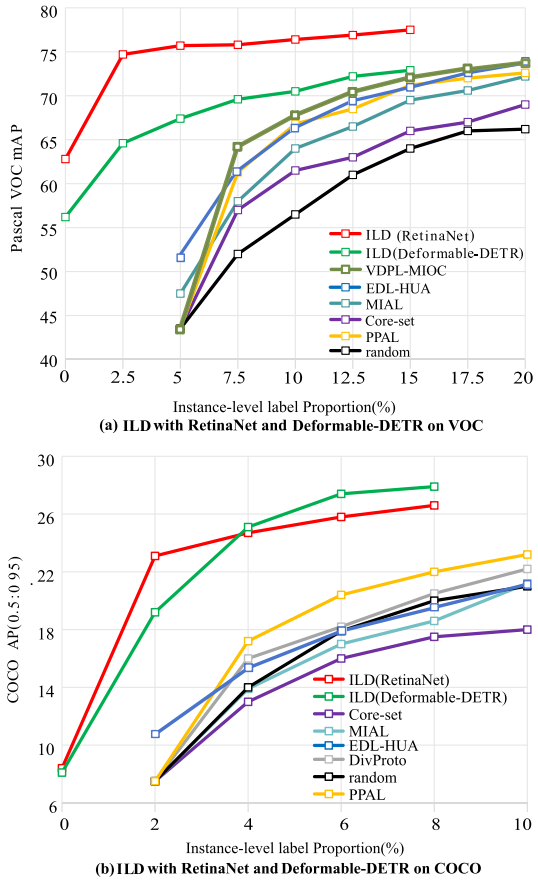

Fig. 6. Comparison of ILD and previous SOTA active learning based detection methods on VOC2007 and COCO2017 datasets.

To demonstrate the robustness of ILD, we also evaluate it on COCO 2017 val set. The results are shown in Table V, similar to the above experiments, we outperform all previous methods. Compared with the previous SOTA approach PPAL [1], ILD improves AP50 from 38.7% to 45.6% and AP(0.5:0.95:0.05) from 23.2% to 27.9%. As shown in Table VI, ILD exceeds the previous state-of-the-art method ALWOD [5] by 3.6% (39.8% vs. 43.4%) in AP50 and 1.2% (24.5% vs. 25.7%) in AP on COCO 2014 val set. The comparison on MS COCO 2014&2017 datasets further proves that our AL initialization and uncertainty & diversity acquisition functions are effective for ALOD. In Figure 6(b), we visualize ILD and previous methods during AL step by step on COCO 2017 val set. Similarly, ILD with RetinaNet [20] and Deformable-DETR

Authorized licensed use limited to: Harbin Institute of Technology. Downloaded on June 23,2026 at 09:31:57 UTC from IEEE Xplore. Restrictions apply.

TABLE V

THE MAP(VOC), AP50(COCO) AND AP(COCO) PERFORMANCE OF OUR PROPOSED ILD COMPARED WITH OTHER SOTA METHODS ON VOC 2007 TEST AND COCO 2017 VAL SET. THE PROPORTION DENOTES INSTANCE-LEVEL LABEL PROPORTION IN DATASETS

Method	VOC		COCO		
	proportion	mAP	proportion	AP50	AP(0.5:0.95:0.05)
MC-DropOut [56]	12%:42%:6%	74.4*	6.25%:12.5%:1.25%	31.5	18.3
NALAE [6]	12%:42%:6%	74.9*	6.25%:12.5%:1.25%	32.8	19.7
PM [2]	12%:42%:6%	73.2*	6.25%:12.5%:1.25%	32.3	18.9
DivProto [7]	6%:60%:6%	75.5*	2%:10%:2%	34.9	22.2
IAU [57]	10%:60%:10%	68.9	4.2%:8.4%:0.84%	32.5	19.6
Core-set [58]	5%:20%:2.5%	70.2*	2%:10%:2%	30.5	18.0
MIAL [8]	5%:20%:2.5%	72.1*	2%:10%:2%	31.0	20.8
MIDL [59]	5%:20%:2.5%	74.0*	2%:10%:2%	33.9	21.5
EDL-HUA [10]	5%:20%:2.5%	74.1*	2%:10%:2%	32.7	21.2
VDPL-MIOC [55]	5%:20%:2.5%	73.2*	4%:10%:1%	33.1	21.8
PPAL [1]	5%:20%:2.5%	72.6*	2%:10%:2%	38.7	23.2
ILD with RetinaNet (Ours)	0%:15%:2.5%	**77.5***	0%:8%:2%	43.2	26.6
ILD with Deformable-DETR (Ours)	0%:15%:2.5%	72.9*	0%:8%:2%	**45.6**	**27.9**

TABLE VI

THE AP50(COCO) AND AP(COCO) PERFORMANCE OF OUR PROPOSED ILD COMPARED WITH OTHER SOTA METHODS ON COCO 2014 VAL SET

Method	COCO		
	instance-level label proportion	AP50	AP(0.5:0.95:0.05)
BiB [4]	1%:5%:1%	33.2	16.5
ALWOD [5]	1%:5%:1%	39.8	24.5
ILD with RetinaNet(Ours)	0%:4%:1%	**43.4**	**25.7**

TABLE VII

THE MAP AND AP50 OF MINIGPT-V2 WITH AND WITHOUT CLASS-NUMBER-AWARE STEP AND ITERATIVE STEP ON VOC 2007 TEST AND COCO 2017 VAL SET

Method	**class-number-aware step**	**iterative step**	**VOC-mAP(%)**	**COCO-AP50(%)**
MiniGPT-v2	✗	✗	46.4	27.9
	✓	✗	50.4	28.5
	✓	✓	**51.5**	**29.0**

TABLE VIII

ABLATION RESULTS OF ILD WITH RETINANET ON VOC2007 TEST SET

random init.	**ILD init. VLM**	**ILD init. WSOD**	**AL w. DCUS & CCMS in [1]**	**AL w. Unce. & Dive. in ILD**	**mAP(%)**
✓	✗	✗	✗	✗	43.4
✗	✓	✗	✗	✗	60.5
✗	✗	✓	✗	✗	44.6
✗	✓	✓	✗	✗	62.9
✗	✓	✓	✓	✗	76.6
✗	✓	✓	✗	✓	**77.5**

[53] only accessing 2% and 4% instance-level labels surpasses previous SOTA methods using 10% instance-level labels.

Finally, as shown in Table VII, we validate the enhanced detection performance of MiniGPT-v2 [12] by multi-step reasoning process.. We can observe that the class-number-aware step improves mAP from 46.4% to 50.4% and from 27.9% to 28.5% on VOC 2007 test set and COCO 2017 val set. By further introducing the iterative step, the detection performance of MiniGPT-v2 [12] increases from 50.4% to 51.5% and from 28.5% to 29.0%. Based on enhanced detection ability, we provide high-quality pseudo ground-truth boxes for AL initialization.

E. Ablation Studies

1) Effectiveness of AL initialization using ILD: As shown in the 1st and 4th rows of Table VIII, we initialize RetinaNet [20] using randomly selective instance-level labels as well as ILD and compare the performance of these two initialization methods. Surprisingly, we can observe that AL initialization by ILD with multi-step reasoning and WSOD brings 19.5% improvement (from 43.4% to 62.9%) in mAP. As mentioned earlier, the reason is that AL initialization with ILD introduces the accurate category number and prevents missing objects in these pseudo ground-truth boxes.

To further demonstrate the effectiveness of our AL initialization(*i.e.*, ILD init. in Table VIII), we explore the improvement of our AL initialization with and without VLM and WSOD as shown in the 2nd, 3rd and 4th rows of Table VIII. The detection results obtained from the VLM and WSOD are used to independently initialize the RetinaNet [20] detector, achieving 60.5% and 44.6% mAP during AL initialization. Furthermore, when these detection results are together utilized as pseudo ground-truth boxes to initialize the RetinaNet [20] detector during AL initialization, a superior performance gain is achieved (*i.e.*, 62.9% mAP), surpassing the results obtained by using either VLM or WSOD alone. The experimental results further demonstrate the effectiveness of our AL initialization(*i.e.*, VLM and WSOD jointly initialize the fully supervised detector).

2) Effectiveness of AL with Uncertainty & Diversity acquisition functions: To prove the effectiveness of AL with uncertainty & diversity acquisition functions in ILD, we explore different methods to obtain instance-level labeled images using our designed uncertainty & diversity acquisition

Authorized licensed use limited to: Harbin Institute of Technology. Downloaded on June 23,2026 at 09:31:57 UTC from IEEE Xplore. Restrictions apply.

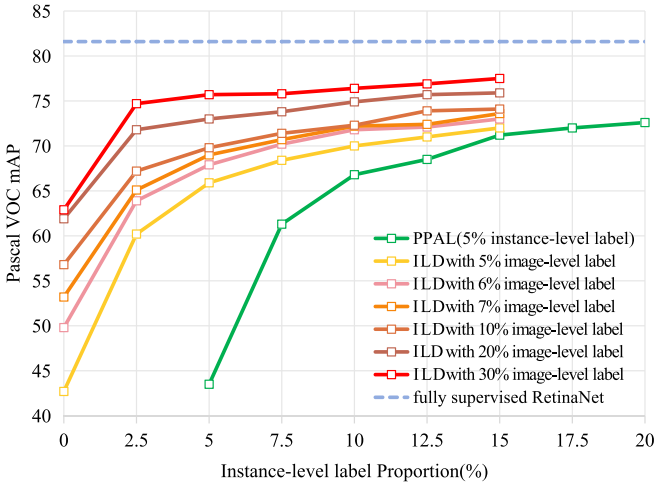

Fig. 7. The influence of image-level label proportions in active learning initialization on VOC 2007 test set.

TABLE IX
INFLUENCE OF IMAGE-LEVEL LABEL PROPORTIONS DURING AL INITIALIZATION ON COCO 2017 VAL SET

Label Type	Label Proportion	AP(%)	AP50(%)	AP75(%)
image-level	8%	8.4	16.4	7.8
image-level	4%	8.4	16.8	7.4
image-level	2%	2.2	4.8	1.8
instance-level	2%	7.4	15.9	6.5

function('AL w. Uncer. and Diver.') and another acquisition function in PPAL [1]('AL w. DCUS & CCMS'). As shown in the 5th and 6th rows of Table VIII, we employ AL initialization with multi-step reasoning to initialize PPAL and ILD. ILD with uncertainty & diversity acquisition functions achieves 77.5% in mAP on VOC 2007 test set, which outperforms PPAL with DCUS & CCMS acquisition functions by 0.9% in mAP. The reason for this improvement is that our well-designed two novel uncertainty and diversity acquisition functions based on the results of the weakly supervised detector and initialized detector mines more informative images compared to the acquisition functions in [1].

3) Influence of image-level label proportions in AL initialization: We explore the influence of different image-level label proportions in the AL initialization on detection performance on VOC 2007 test set and COCO 2017 val set. As shown in Figure 7, we can observe that ILD with 5% image-level labels (yellow line) in AL initialization achieves competitive performance compared to PPAL [1] which uses 5% instance-level labels during the AL process. In addition, ILD with 6% image-level labels achieves superior performance(73.0%) in mAP compared to PPAL(72.6%). As the proportions of image-level labels gradually increase in AL initialization, the detection performance of ILD is close to fully supervised RetinaNet [20]. We believe that increasing the proportion of image-level labels can further improve detection performance but it increases the annotation effort.

Table IX shows the influence of the image-level label proportions on the RetinaNet detector during AL initialization on COCO 2017 val set. We can observe that the performance of AL initialization increases from 2.2% to 8.4% when using various image-level label proportions(*i.e.*, 2%, 4% and 8%). We achieve the highest AL initialization performance(*i.e.*, 8.4% AP) when the image-level label proportion is set to 4% and 8%. We believe that increasing the proportion of image-level labels can further improve initialization performance, but it increases the annotation effort of image-level labels. Our proposed ILD with 4% image-level labels surpasses the previous AL initialization method(*i.e.*, 2% instance-level label) by 1.0% in mAP (*i.e.*, 8.4% *vs.* 7.4%). So, we utilize the 4% image-level label instead of the 2% instance-level label to perform AL initialization. Based on the above experimental results, to balance the AL initialization performance and annotation effort, we set the image-level label proportion to 4% in all experiments on COCO 2017 dataset.

TABLE X
THE MAP PERFORMANCE OF DIFFERENT CONFIDENCE THRESHOLD α IN THE FUSION STRATEGY ON VOC 2007 TEST SET

α	mAP(%)
without WSOD	60.5
0.5	58.4
0.6	59.7
0.7	59.9
0.8	61.0
0.9	62.5
0.95	**62.9**

TABLE XI
THE MAP PERFORMANCE OF DIFFERENT IOU THRESHOLD β IN THE FUSION STRATEGY ON VOC 2007 TEST SET

β	mAP(%)
0.1	60.7
0.3	61.1
0.5	**62.9**
0.7	62.0
0.9	61.8

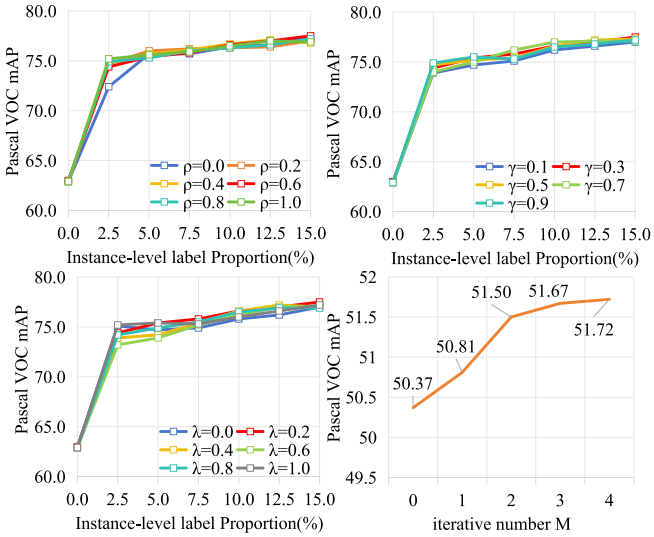

Fig. 8. Detection performance of ILD under various ρ, λ, γ in eq. 7 and eq. 9 and M in eq. 3 on VOC 2007 test set.

Authorized licensed use limited to: Harbin Institute of Technology. Downloaded on June 23,2026 at 09:31:57 UTC from IEEE Xplore. Restrictions apply.

Fig. 9. Qualitative results of our ILD on VOC2007 test set and COCO 2017 val set. The detection results predicted ILD above the dotted line (red and green bounding boxes denote our predictions and ground-truth boxes). The pseudo ground-truth boxes generated by ILD below the dotted line (red and green bounding boxes denote our generated pseudo ground-truth boxes and ground-truth boxes. yellow and blue bounding boxes indicate WSOD and VLM predictions in our box fusion strategy).

4) Influence of confidence threshold α and IoU threshold β in the fusion strategy: We report the influence of the confidence threshold α and IoU threshold β in the box fusion strategy as shown in Table X and Table XI. The confidence threshold α and IoU threshold β directly control the performance of initialized RetinaNet by selecting detection results from the weakly supervised detector OD-WSCL as pseudo ground-truth boxes. Table X shows the comparison of mAP performance using different threshold values α(α = 0.5, 0.6, 0.7, 0.8, 0.9, 0.95) on VOC 2007 test set. We can observe that using the detection results of the weakly supervised detector OD-WSCL as pseudo ground-truth boxes improves the detection performance from 60.5% to 62.9%, and RetinaNet detector achieves the highest mAP(*i.e.*, 62.9%) when the threshold value α is set to 0.95. In our box fusion strategy, bounding boxes with high confidence scores from WSOD typically cover complete objects or most regions of objects, and those boxes with lower confidence scores frequently contain misclassified regions, which aligns with the consistent conclusion in [46]. Thus, high-confidence bounding boxes of WSOD(*e.g.*, $\alpha > 0.9$) as pseudo-ground truth boxes are beneficial in AL initialization, ensuring robust generalization to novel datasets.

Moreover, we also compare different IoU threshold values β (β = 0.1, 0.3, 0.5, 0.7, 0.9) on VOC 2007 test set as shown Table XI. The optimal AL initialization (*i.e.*, 62.9%) is achieved by our bounding box fusion strategy when β is set to 0.5 in RetinaNet. Meanwhile, we can observe that AL initialization performance degrades when β is either too small (β = 0.1, 0.3) or too large (β = 0.7, 0.9). This reason is that a smaller β overlooks adjacent objects detected by VLM and WSOD, while a larger β obtains multiple bounding boxes of the same objects from VLM and WSOD. In particular, the NMS operation in RetinaNet prevents performance degradation caused by larger β, thus comparable performance is achieved even with higher β (*e.g.*, β = 0.7, 0.9). In our paper, we set α = 0.95 and β = 0.5. Our box fusion strategy can be configured with a larger α (*e.g.*, $\alpha > 0.9$) and an appropriate β (*e.g.*, β near 0.5), ensuring robust generalization to novel datasets.

Influence of ρ, λ, γ and M during AL process. We explore the influence of ρ, λ, γ in eq. 7 and eq. 9 during AL process. As shown in Figure 8, we can observe that the detection performance of ILD is insensitive for different ρ, λ and γ. Compared with the different settings of ρ, λ, and γ, the ρ, λ and γ are set to 0.6, 0.2 and 0.3, achieving the highest mAP(*i.e.*, 77.5%). Moreover, we conduct ablation experiments to select a suitable iterative number M in eq. 3 in multi-step reasoning process. The fourth sub-figure in Figure 8 shows the comparison results of different iterative

Authorized licensed use limited to: Harbin Institute of Technology. Downloaded on June 23,2026 at 09:31:57 UTC from IEEE Xplore. Restrictions apply.

numbers on VOC 2007 test set. It can be observed that the initial performance in mAP(*i.e.*, MiniGPT-v2 only with class-number-aware step) is 50.37%. By increasing the number of iterations, the detection performance increases substantially compared to the case where $M = 0$. However, continue to increase the iterative number when $M \geq 2$, the change of detection performance in not obvious(*i.e.*, from 51.50% to 51.72% in mAP). Meanwhile, as the iterative number is increased, the computational cost of MiniGPT-v2 keeps on rising. The reason is that the computational complexity is directly proportional to the iterative number M. Therefore, to balance detection performance and computational costs, we set M to 2 for all experiments in our paper.

F. Qualitative Results

The Figure 9 above the dotted line shows some detection results from ILD on VOC 2007 test set and COCO 2014/2017 val set. From these qualitative results, we observe that ILD detects the whole object region even though some instances are occluded and dense, *e.g.*, there are bottles on a diningtable and people sitting around it as shown in the 1st row of Figure 9. Moreover, we also visualize some failure cases with incorrect categories as shown in the third row of Figure 9. The main reason is that these object regions with incorrectly identified categories by VLM are regarded as pseudo ground-truth boxes during AL initialization. We will focus on solving this problem in the future. Meanwhile, we also show some cases of pseudo ground-truth boxes obtained by our ILD as shown in Figure 9 below the dotted line. These pseudo ground-truth boxes containing the complete object region or most of the object region are regarded as the foundation of active learning. For example, benefiting from the well-designed multi-step reasoning process and box fusion strategy, despite the presence of multiple same-class objects in the image or a tiny sheep on the hillside, our method captures each sheep without missed objects. Additionally, we also list that the pottedplant and boat tend to some failure pseudo ground-truth boxes as shown in the last row of Figure 9. The pottedplant exhibits significant visual similarity to the background, resulting in missing objects within the MiniGPT-v2. Meanwhile, weakly supervised detectors achieve a low AP(24.3%) in pottedplant category reported in [13], highlighting the discriminative region(*i.e.* flowerpot). In addition, the background contains textured regions resembling boats, which may lead WSOD to misclassify these regions as boats and generate erroneous pseudo-ground truth boxes. Under these conditions, our box fusion strategy struggles to capture complete and accurate pottedplants and boats, resulting in low-quality pseudo ground-truth boxes. We will further optimize our box fusion strategy in the future to make it more robust to various types of objects in various scenarios.

V. Conclusion

In this paper, we propose an Image-Level labels Driven active learning method for object detection (ILD). During AL initialization, the multi-step reasoning process based on CoT, including a class-number-aware step and an iterative step, is proposed to enhance the detection ability of VLM, and we regard detection results of the weakly supervised detector and VLM as pseudo ground-truth boxes to initialize a fully supervised detector. Based on the paradigm, the initialization process in ILD eliminates the requirement for instance-level labels. Moreover, we design uncertainty and diversity acquisition functions based on collaborative outputs from both the weakly supervised detector and the pre-trained fully supervised detector. The collaborative mechanism prevents the reliance on a single pre-trained fully supervised detector to enhance the localization quality of the pre-trained detector. Finally, quantitative and qualitative experimental results demonstrate the effectiveness of our ILD.

In the future, we will utilize uncertainty and diversity acquisition functions to adapt to a semi-supervised detector. The two functions will select images with high-quality pseudo instance-level labels. At the same time, it will mine the most informative images with ground-truth instance-level labels. Together, these two sets of images will be used to further refine the fully supervised detector initialized by ILD.

Limitations. ILD relies on high-quality pseudo ground-truth boxes to initialize the fully supervised detector. Once encountering rare objects or datasets with distribution differences(*e.g.*, fire, platypus, cartoon) that VLMs have never known before, ILD still has the potential ability to localize the object by using prompt engineering such as ***[detection] It with a tail like a withered leaf***, but it may detect multiple parts of same object.

References

[1] C. Yang, L. Huang, and E. J. Crowley, "Plug and play active learning for object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 17784–17793.

[2] J. Choi, I. Elezi, H.-J. Lee, C. Farabet, and J. M. Alvarez, "Active learning for deep object detection via probabilistic modeling," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 10244–10253.

[3] R. Mao, S. K. Maharana, R. Iyer, and Y. Guo, "STONE: A submodular optimization framework for active 3D object detection," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, vol. 37, 2024, pp. 48292–48308.

[4] H. V. Vo, O. Siméoni, S. Gidaris, A. Bursuc, and P. Pérez, "Active learning strategies for weakly-supervised object detection," in *Proc. Eur. Conf. Comput. Vis.*Cham, Switzerland: Springer, 2022, pp. 211–230.

[5] W. Yuting, V. Ilic, and J. Li, "ALWOD: Active learning for weakly-supervised object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2023, pp. 6459–6469.

[6] I. Elezi, Z. Yu, A. Anandkumar, L. Leal-Taixé, and J. M. Alvarez, "Not all labels are equal: Rationalizing the labeling costs for training object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 14472–14481.

[7] J. Wu, J. Chen, and D. Huang, "Entropy-based active learning for object detection with progressive diversity constraint," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 9387–9396.

[8] T. Yuan et al., "Multiple instance active learning for object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 5330–5339.

[9] S. Hwang, S. Lee, H. Kim, M. Oh, J. Ok, and S. Kwak, "Active learning for semantic segmentation with multi-class label query," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, vol. 36, 2023, pp. 27020–27039.

[10] Y. Park, W. Choi, S. Kim, D.-J. Han, and J. Moon, "Active learning for object detection with evidential deep learning and hierarchical uncertainty aggregation," in *Proc. 11th Int. Conf. Learn. Represent.*, 2023, pp. 1–20.

Authorized licensed use limited to: Harbin Institute of Technology. Downloaded on June 23,2026 at 09:31:57 UTC from IEEE Xplore. Restrictions apply.

[11] S. Sun, H. Xu, Y. Li, P. Li, B. Sheng, and X. Lin, "FastAL: Fast evaluation module for efficient dynamic deep active learning using broad learning system," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 2, pp. 815–827, Feb. 2024.

[12] J. Chen et al., "MiniGPT-V2: Large language model as a unified interface for vision-language multi-task learning," 2023, *arXiv:2310.09478*.

[13] J. Seo, W. Bae, D. J. Sutherland, and J. Noh, "Object discovery via contrastive learning for weakly supervised object detection," in *Proc. Eur. Conf. Comput. Vis.*Cham, Switzerland: Springer, 2022, pp. 312–329.

[14] S. Bai et al., "Qwen2.5-VL technical report," 2025, *arXiv:2502.13923*.

[15] M. Minderer, A. Gritsenko, and N. Houlsby, "Scaling open-vocabulary object detection," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, vol. 36, 2023, pp. 72983–73007.

[16] P. Du et al., "LaMI-DETR: Open-vocabulary detection with language model instruction," in *Proc. Eur. Conf. Comput. Vis.*Cham, Switzerland: Springer, 2024, pp. 312–328.

[17] H.-Y. Chen et al., "Contrastive localized language-image pre-training," 2024, *arXiv:2410.02746*.

[18] C. Xu, K. Xu, X. Jiang, and T. Sun, "PLOVAD: Prompting vision-language models for open vocabulary video anomaly detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 35, no. 6, pp. 5925–5938, Jun. 2025.

[19] D. Li, Z. Wang, Y. Chen, R. Jiang, W. Ding, and M. Okumura, "A survey on deep active learning: Recent advances and new frontiers," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 36, no. 4, pp. 5879–5899, Apr. 2025.

[20] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dolláar, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.

[21] Y. Xu and F. Song, "Learning spatial similarity distribution for few-shot object counting," in *Proc. Int. Joint Conf. Artif. Intell.*, 2024, pp. 1507–1515.

[22] J. Guo et al., "Semi-supervised active learning for semi-supervised models: Exploit adversarial examples with graph-based virtual labels," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 2876–2885.

[23] D. Sartor, T. Barbariol, and G. A. Susto, "Bayesian active learning isolation forest (B-ALIF): A weakly supervised strategy for anomaly detection," *Eng. Appl. Artif. Intell.*, vol. 130, pp. 107671–107684, Apr. 2024.

[24] C. J. Burke, P. N. Tobler, M. Baddeley, and W. Schultz, "Neural mechanisms of observational learning," *Proc. Nat. Acad. Sci. USA*, vol. 107, no. 32, pp. 14431–14436, Aug. 2010.

[25] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.

[26] T. Brown et al., "Language models are few-shot learners," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 1877–1901.

[27] J.-B. Alayrac et al., "Flamingo: A visual language model for few-shot learning," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 23716–23736.

[28] D. Zhu, J. Chen, X. Shen, X. Li, and M. Elhoseiny, "MiniGPT-4: Enhancing vision-language understanding with advanced large language models," 2023, *arXiv:2304.10592*.

[29] H. Touvron et al., "Llama 2: Open foundation and fine-tuned chat models," 2023, *arXiv:2307.09288*.

[30] M. Liang et al., "AIDE: An automatic data engine for object detection in autonomous driving," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 14695–14706.

[31] M. Minderer et al., "Simple open-vocabulary object detection with vision transformers," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 728–755.

[32] P. Qin, T. Xu, C. Zhang, H. Wang, Y. Hu, and E. Chen, "Scenario-aware multimodal chain-of-thought prompting for rationales of video social relations," *IEEE Trans. Circuits Syst. Video Technol.*, early access, 2025, doi: 10.1109/TCSVT.2025.3571422.

[33] J. Wei et al., "Chain-of-thought prompting elicits reasoning in large language models," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 24824–24837.

[34] M. D. Hoffman et al., "Training chain-of-thought via latent-variable inference," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, vol. 36, 2024, pp. 1–23.

[35] C. Cohn, N. Hutchins, T. Le, and G. Biswas, "A chain-of-thought prompting approach with llms for evaluating students' formative assessment responses in science," in *Proc. AAAI Conf. Artif. Intell.*, 2024, pp. 23182–23190.

[36] L. Wang et al., "T-SciQ: Teaching multimodal chain-of-thought reasoning via large language model signals for science question answering," in *Proc. AAAI Conf. Artif. Intell.*, vol. 38, Mar. 2024, pp. 19162–19170.

[37] J. Tang, G. Zheng, J. Yu, and S. Yang, "CoTDet: Affordance knowledge prompting for task driven object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2023, pp. 3068–3078.

[38] T. Kim, S. Chung, D. Yeom, Y. Yu, H. Gu Kim, and Y. M. Ro, "MSCoTDet: Language-driven multi-modal fusion for improved multispectral pedestrian detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 35, no. 5, pp. 5006–5021, May 2025.

[39] D. Mondal, S. Modi, S. Panda, R. Singh, and G. S. Rao, "KAM-CoT: Knowledge augmented multimodal chain-of-thoughts reasoning," in *Proc. AAAI Conf. Artif. Intell.*, Mar. 2024, vol. 38, no. 17, pp. 18798–18806.

[40] R. Xia, G. Li, Z. Huang, H. Meng, and Y. Pang, "CBASH: Combined backbone and advanced selection heads with object semantic proposals for weakly supervised object detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 10, pp. 6502–6514, Oct. 2022.

[41] J. Lin, Y. Shen, B. Wang, S. Lin, K. Li, and L. Cao, "Weakly supervised open-vocabulary object detection," in *Proc. AAAI Conf. Artif. Intell.*, Mar. 2024, vol. 38, no. 4, pp. 3404–3412.

[42] S. Yasuki and M. Taki, "CAM back again: Large kernel CNNs from a weakly supervised object localization perspective," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 341–351.

[43] J. R. Uijlings, K. E. Van De Sande, T. Gevers, and A. W. Smeulders, "Selective search for object recognition," *Int. J. Comput. Vis.*, vol. 104, pp. 154–171, Apr. 2013.

[44] P. Arbeláez, J. Pont-Tuset, J. Barron, F. Marques, and J. Malik, "Multiscale combinatorial grouping," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 328–335.

[45] A. Kirillov et al., "Segment anything," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2023, pp. 4015–4026.

[46] L. Zhu, J. Zhou, Y. Liu, X. Hao, W. Liu, and X. Wang, "WeakSAM: Segment anything meets weakly-supervised instance-level recognition," in *Proc. 32nd ACM Int. Conf. Multimedia*, Oct. 2024, pp. 7947–7956.

[47] H. Tang, Z. Li, D. Zhang, S. He, and J. Tang, "Divide-and-conquer: Confluent triple-flow network for RGB-T salient object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 47, no. 3, pp. 1958–1974, Mar. 2025.

[48] X. Zhang, B. Yin, Z. Lin, Q. Hou, D.-P. Fan, and M.-M. Cheng, "Referring camouflaged object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 47, no. 5, pp. 3597–3610, May 2025.

[49] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, 2010.

[50] T.-Y. Lin et al., "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*Cham, Switzerland: Springer, 2014, pp. 740–755.

[51] M. Everingham, S. A. Eslami, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes challenge: A retrospective," *Int. J. Comput. Vis.*, vol. 111, no. 1, pp. 98–136, 2015.

[52] P. Tang, X. Wang, X. Bai, and W. Liu, "Multiple instance detection network with online instance classifier refinement," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2843–2851.

[53] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable DETR: Deformable transformers for end-to-end object detection," in *Proc. Int. Conf. Learn. Represent.*, 2020, pp. 1–6.

[54] A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 8026–8037.

[55] J. Yang et al., "Active learning for object detection with vectorized dual pseudo loss and multiple instance offset constraint," *IEEE Trans. Cybern.*, vol. 55, no. 9, pp. 4427–4440, Sep. 2025.

[56] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 1050–1059.

[57] Z. Zhang et al., "Instance-aware uncertainty for active learning in object detection," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2024, pp. 298–304.

[58] S. Ozan and S. Silvio, "Active learning for convolutional neural networks: A core-set approach," in *Proc. Int. Conf. Learn. Represent.*, 2018, pp. 1–13.

[59] F. Wan et al., "Multiple instance differentiation learning for active object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 10, pp. 12133–12147, Oct. 2023.

Authorized licensed use limited to: Harbin Institute of Technology. Downloaded on June 23,2026 at 09:31:57 UTC from IEEE Xplore. Restrictions apply.

Rui Tian received the M.S. degree in instrument science and technology from Harbin Institute of Technology (HIT), Harbin, China, in 2023, where he is currently pursuing the Ph.D. degree. His research interests include computer vision, pattern recognition, machine learning, deep learning, weakly supervised object detection, depth estimation, and contrastive learning.

Jiaxuan Zhang received the bachelor's degree in measurement and control technology and instrumentation from Northeastern University in 2023. He is currently pursuing the master's degree with Harbin Institute of Technology (HIT). His research interests include computer vision, pattern recognition, machine learning, weakly-supervised object detection, visual language models, and multi-agent reinforcement learning.

Yongqiang Zhang received the M.S. and Ph.D. degrees in instrument science and technology from Harbin Institute of Technology (HIT), Harbin, China, in 2015 and 2020, respectively. He was with the King Abdullah University of Science and Technology (KAUST), as a Visiting Student, from 2017 to 2018. He is currently a Professor with the School of Computer Science (College of Software-College of Artificial Intelligence), Inner Mongolia University. His research interests include computer vision, pattern recognition, machine learning, deep learning, face detection, weakly/fully supervised object detection, activity detection, and image and video understanding in the real world.

Man Zhang received the M.S. degree in optical engineering from Harbin Institute of Technology (HIT), Harbin, China, in 2020, where he is currently pursuing the Ph.D. degree. His research interests include computer vision, pattern recognition, machine learning, deep learning, object detection, motion forecasting, and 3D occupancy prediction.

Zian Zhang received the B.S. and M.S. degrees in instrument science and technology from Harbin Institute of Technology (HIT), Harbin, China, in 2018 and 2020, respectively, where he is currently pursuing the Ph.D. degree with the School of Instrumentation Science and Engineering. His research interests include computer vision, pattern recognition, machine learning, deep learning, human pose estimation, object detection, action recognition, and image and video understanding in the monitoring systems.

Yin Zhang received the M.S. degree in instrument science and technology from Harbin Institute of Technology (HIT), Harbin, China, in 2021, where he is currently pursuing the Ph.D. degree. His research interests include computer vision, pattern recognition, machine learning, deep learning, object detection, knowledge distillation, and image super-resolution.

Yongqiang Li received the B.S., M.S., and Ph.D. degrees from the Department of Automated Testing and Control, School of Electrical Engineering and Automation, Harbin Institute of Technology (HIT), Harbin, China, in 2007, 2009, and 2013, respectively. He was a Visiting Scholar with Rensselaer Polytechnic Institute, USA, from 2010 to 2012. He is currently an Associate Professor with the School of Instrumentation Science and Engineering, HIT. He has published over 30 papers in peer-reviewed journals and conferences. His research interests include machine vision, environment perception, and autonomous driving.

Wangmeng Zuo received the Ph.D. degree in computer application technology from Harbin Institute of Technology, Harbin, China, in 2007. From 2004 to 2006, he was a Research Assistant with the Department of Computing, The Hong Kong Polytechnic University, Hong Kong. From 2009 to 2010, he was a Visiting Professor with Microsoft Research Asia. He is currently a Professor with the School of Computer Science and Technology, Harbin Institute of Technology. He has published over 100 papers in top-tier academic journals and conferences. His current research interests include image enhancement and restoration, image/video generation, and image classification. He served as an Associate Editor for IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE and a senior editor of the *Journal of Electronic Imaging*.

Authorized licensed use limited to: Harbin Institute of Technology. Downloaded on June 23,2026 at 09:31:57 UTC from IEEE Xplore. Restrictions apply.