Task-Agnostic Incremental Vision-Language Object Detection via Prompt Augmentation and Distribution-Aware Fusion

Yonghan Jiang[1], Zhengyuan Xie[1], Wenchu Liu[1], Linlan Huang[1], Fei Yang[1,3], and Xialei Liu[1,2,3]*

[1] VCIP, CS, Nankai University, Tianjin, China
[2] AAIS, Nankai University, Tianjin, China
[3] NKIARI, Shenzhen Futian, Shenzhen, China
{yonghanjiang, xiezhengyuan, wenchuliu, huanglinlan}@mail.nankai.edu.cn
{feiyang, xialei}@nankai.edu.cn

Abstract. Incremental Vision-Language Object Detection (IVLOD) empowers pre-trained open-vocabulary detectors to continuously learn new visual concepts from sequential tasks without forgetting foundational knowledge. However, existing methods predominantly rely on oracle task identities during inference, severely limiting their open-world practicality. In this paper, we formalize Task-Agnostic IVLOD (TA-IVLOD), a more realistic setting where such priors are unavailable, requiring the model to simultaneously recognize classes from all learned tasks. This unconstrained setting exposes severe performance degradation due to cross-task semantic interference and parameter conflicts. To tackle these challenges, we propose TADA, a novel modular framework. Specifically, TADA introduces Stochastic Prompt Augmentation to mitigate semantic interference via training-time noise injection, and Test-Time Distribution-Aware Fusion to dynamically weight class-specific LoRA experts, effectively resolving parameter conflicts. Extensive evaluations demonstrate that our method significantly outperforms baselines in standard IVLOD and yields substantial improvements in the rigorous TA-IVLOD setting on the ODinW-13 benchmark, while effectively preserving zero-shot generalization on the MS COCO dataset. The code is available at `https://github.com/yonghanjiang/TADA`.

Keywords: Continual Learning · Open-Vocabulary Object Detection

1 Introduction

Object detection [3, 39] has long been a fundamental task in computer vision. Recently, Vision-Language Object Detection (VLOD), powered by large-scale vision-language pre-training [22], has enabled Open-Vocabulary Object Detection (OVOD) [9, 19, 27, 28, 40], allowing detectors to localize objects beyond predefined categories through textual queries. Despite their strong zero-shot generalization, these models often suffer from substantial performance degradation

* Corresponding author.

when deployed in specialized downstream domains whose visual distributions differ from web-scale pre-training data.

To bridge this gap, Incremental Vision-Language Object Detection (IVLOD) has been introduced to continuously adapt OVOD models to novel tasks while mitigating catastrophic forgetting [5]. To achieve efficient continual learning, recent vision-language frameworks have explored various mechanisms to preserve prior knowledge [4, 13, 14, 23, 38, 44]. A common paradigm involves decoupling the acquisition of new concepts from the preservation of old capabilities, often by isolating task-specific parameters within the network. Such structural modifications allow for domain-specific updates without altering the core knowledge stored in the backbone, ensuring that the model remains robust on previously learned tasks while accommodating new information.

While current modular adaptation strategies (e.g., ZiRa [5] and DitHub [2]) successfully balance plasticity and stability, they fundamentally rely on a highly restrictive assumption: the Task-Incremental setting [24, 29]. By depending on oracle task information during inference, these methods assume the task identity of each test sample is provided a priori. Consequently, they evaluate the model using a constrained prompt comprising only the classes from that specific task, simplifying the problem into processing isolated, single-task prompts. This reliance severely limits their applicability in open-world scenarios where task boundaries are unknown, forcing the model to confront an overwhelming semantic space of all encountered classes simultaneously.

To align with the demands of real-world applications, we formally introduce a more realistic and challenging setting: Task-Agnostic Incremental Vision-Language Object Detection (TA-IVLOD). Under this setting, the model operates entirely without task identity priors during inference. However, this paradigm shift exposes a fundamental bottleneck we term Prompt Interference. Since mainstream OVOD architectures rely on deep cross-modal fusion, transitioning from single-task prompts to cross-task prompts naturally causes severe attention dilution. Crucially, the sequential nature of continual learning severely exacerbates this vulnerability. During training, the model optimizes cross-modal alignments for disjoint subsets of classes in complete isolation. Consequently, when these independently learned concepts are concatenated into a cross-task prompt during task-agnostic inference, the visual attention allocated to different semantic classes inevitably collides. Lacking the ability to coordinate attention across concurrently presented, heterogeneous classes, this semantic conflict directly trans-

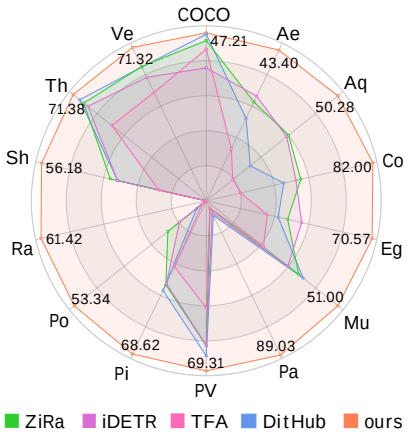


Fig. 1: Our model achieves superior performance on 14 datasets under the TA-IVLOD setting.

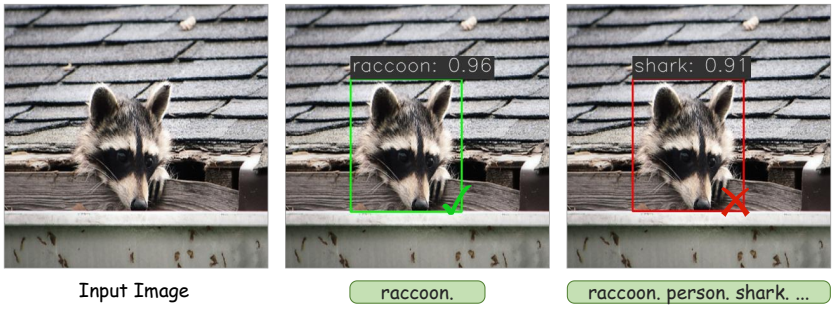


Fig. 2: Comparison of detection results. (Middle) Single-task prompt in the standard IVLOD setting. (Right) Cross-task prompt in the TA-IVLOD setting.

lates into parameter conflict and catastrophic performance degradation, as illustrated in Fig. 2.

To address these challenges, we present TADA (Task-Agnostic Distribution-Aware Adaptation), a modular framework for task-agnostic inference. Rather than relying on naive parameter merging or constrained training setups, our approach addresses two key challenges: prompt interference and parameter conflict. Specifically, we introduce Stochastic Prompt Augmentation (SPA) to proactively simulate semantic noise during optimization, and propose Test-Time Distribution-Aware Fusion (TTDF) to dynamically align visual feature distributions at inference. By synergizing these two modules, our framework enables the detector to maintain robust cross-modal alignments. As visually summarized in Fig. 1, our method delivers superior detection performance in the TA-IVLOD setting.

The main contributions of this paper are summarized as follows:

- We introduce TA-IVLOD, a realistic and challenging setting devoid of task oracle priors.
- We design Stochastic Prompt Augmentation to mitigate cross-task semantic interference via training-time noise injection.
- We propose Test-Time Distribution-Aware Fusion to dynamically weight class-specific experts, effectively resolving parameter conflicts.
- Extensive experiments on the ODinW-13 benchmark demonstrate substantial performance gains in TA-IVLOD and consistent improvements in standard IVLOD.

2 Related Work

Incremental Vision-Language Object Detection. Vision-Language Object Detection (VLOD) [48] aims to generalize detection to unseen classes via cross-modal alignment, effectively overcoming the bottleneck of fixed category sets. In this domain, pioneering works such as ViLD [9] distill knowledge from CLIP [31] to align region proposals with text embeddings. Subsequently, GLIP [19] reformulates detection as phrase grounding for deeper fusion, and Grounding

DINO [22] leverages Transformer architectures to significantly boost open-set accuracy. However, these models primarily assume a static vocabulary, struggling to adapt to dynamic data streams.

To address this, Incremental Vision-Language Object Detection (IVLOD) [5] requires models to sequentially learn new classes while mitigating catastrophic forgetting of base classes and preserving zero-shot generalization. This rigorous tri-objective optimization distinguishes IVLOD from conventional incremental learning, which typically neglects open-world capabilities. Recent methods tackle this via modular adaptation. ZiRa [5] incorporates zero-interference loss and re-parameterization, while DitHub [2] proposes a scalable LoRA-based framework for flexible class extension. However, their reliance on task-specific oracle prompts during inference severely restricts their applicability in unconstrained environments. To overcome this limitation, we formalize TA-IVLOD, a challenging new paradigm designed for true open-world scenarios where task boundaries are entirely unknown.

Parameter Fusion and LoRA Integration. Recent works [15, 18, 21, 30, 46] have explored parameter-efficient fine-tuning (PEFT) methods, such as Adapter Tuning [10] and Prompt Learning [46, 47], to adapt pre-trained models without updating all parameters. Among these, Low-Rank Adaptation (LoRA) [11] has gained widespread adoption due to its competitive efficiency. However, maintaining separate LoRA modules in incremental learning leads to linear growth in inference latency and memory consumption. To mitigate this, LoRA fusion techniques [12, 35, 37, 45, 46] merge multiple low-rank matrices into a single adapter via weight averaging [1] or orthogonal projection [45].

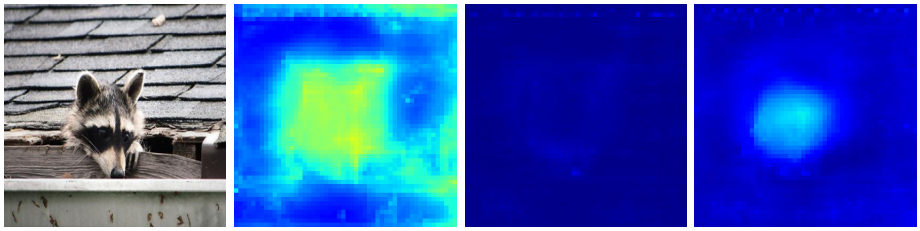
In the context of IVLOD, DitHub [2] adopts this paradigm by statically averaging the $\mathbf{A}$ matrices of LoRA adapters across observed classes while sharing a common $\mathbf{B}$ matrix. While such static averaging preserves general knowledge within isolated, single-task environments, it inevitably triggers severe **parameter conflicts** when processing the unified, cross-task prompts required for task-agnostic inference. In contrast, our approach dynamically weights class-specific parameters by aligning them with the input visual distribution at test time. This distribution-aware strategy effectively resolves parameter conflicts and ensures robust cross-modal alignment without requiring oracle task identities.

3 Task-Agnostic IVLOD

3.1 Problem Formulation

Following the standard incremental learning paradigm, we formulate the training process as a sequence of N tasks, denoted as $\mathcal{T} = \{T_1, T_2, \ldots, T_N\}$. Each task T_k is associated with a specific label set $\mathcal{C}_k$. The goal is to sequentially learn these tasks while mitigating catastrophic forgetting.

Existing IVLOD approaches predominantly adopt an evaluation protocol that implicitly aligns with the Task-Incremental setting. Under this scenario, task information is implicitly exploited during prompt construction. In particular, for a test image belonging to task T_k, the query prompt $\mathcal{P}_I$ is restricted



(a) Image **(b)** "raccoon." **(c)** "raccoon. person. shark. ..."

Fig. 3: Visualization of image feature attention towards the correct target class ("raccoon"). (a) Original image. (b) Attention under a single-task prompt. (c) Attention under a cross-task prompt. While the baseline (left) fails to focus on the correct class due to semantic interference, our proposed method (right) accurately attends to the target object.

strictly to the classes within that specific task: $\mathcal{P}_I^{(k)} = \{c \mid c \in \mathcal{C}_k\}$. This implicit task prior simplifies the problem by partitioning the global label space into disjoint subspaces, thereby artificially circumventing the interference from classes across different tasks.

To align with realistic open-world scenarios, we introduce the TA-IVLOD setting. In this challenging setup, no task priors are provided during inference. Consequently, instead of utilizing constrained, task-specific queries, the model is required to use a unified query prompt encompassing the union of all learned classes: $\mathcal{P}_{TA} = \bigcup_{i=1}^{N} \mathcal{C}_i$. Furthermore, unlike the standard IVLOD protocol that evaluates each task i independently on its corresponding isolated test set $\mathcal{D}_{test}^{(i)}$, TA-IVLOD mandates a joint evaluation over a comprehensive global test set formed by uniting all encountered task domains: $\mathcal{D}_{test}^{TA} = \bigcup_{i=1}^{N} \mathcal{D}_{test}^{(i)}$. Consequently, the model must localize and classify objects within a significantly expanded semantic candidate space in a single inference pass. This evaluation protocol introduces a high volume of negative samples for each task—specifically images devoid of its target classes—thereby increasing the risk of cross-task false positives and providing a more rigorous assessment of open-world robustness.

3.2 Challenges

Prompt Interference. As incremental tasks accumulate, the target class set continuously expands. To ensure real-world practicality, the model must perform unified inference across all learned classes simultaneously, without relying on task oracle priors. This fundamentally differs from the sparse, task-specific prompts used during training, creating a severe training-inference prompt discrepancy. In OVOD architectures, the text prompt serves as an implicit conditional adapter that dictates the alignment of visual features. Subjecting the model to an extensive, consolidated inference prompt triggers Prompt Interference, where the cluttered textual context tends to dilute the attention weights assigned to key semantic tokens. This attention dilution and contextual misalignment introduce

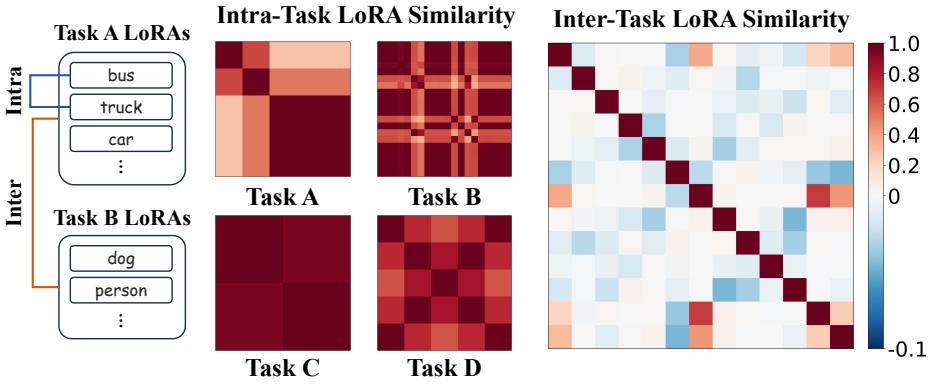


Fig. 4: Cosine similarity matrices of class-specific LoRA experts. The left panels show the intra-task similarity among LoRA modules learned for different classes within the same task. The right panel shows the inter-task similarity matrix across all 13 tasks, which is computed by averaging the class-specific LoRA modules within each task beforehand.

substantial semantic noise, preventing the visual features from accurately attending to their corresponding target tokens. To visualize this, we extract the image-to-text attention weights corresponding to the ground-truth class (*e.g.*, "raccoon") from the final cross-attention layer of the Grounding DINO feature enhancer. As shown in Fig. 3, single-task prompts yield focused and well-aligned attention (Fig. 3b), whereas cross-task prompts cause severe attention diffusion and misalignment (Fig. 3c, left). This demonstrates that prompt expansion acts as a strong distractor, overwhelming the target semantics and degrading visual-textual alignment.

Parameter Conflict. Crucially, the sequential nature of continual learning severely amplifies the aforementioned prompt-induced interference. Because the model is strictly optimized on disjoint tasks with task-local objectives and without access to data from other tasks, the learning process inherently lacks cross-task coordination. This vulnerability is particularly evident in representative modular frameworks, such as DitHub, which allocate a distinct LoRA expert for each individual class. Driven by the isolated training paradigm, these independently learned experts inevitably follow divergent optimization trajectories. When these uncoordinated representations are forced into a unified semantic space to process cross-task prompts during task-agnostic inference, they suffer from intrinsic structural conflicts and pronounced representational misalignment. To empirically validate this, we visualize the cosine similarity among these class-specific LoRA experts. Prior studies [36] establish that merging LoRA modules with low directional similarity significantly exacerbates performance degradation. As illustrated in Fig. 4, experts optimized within the same task maintain high correlation (left panels), ensuring that intra-task fusion introduces minimal parameter conflict. Conversely, the inter-task similarity across different tasks

is significantly lower (right panel). Consequently, fusing these highly divergent modules under the unconstrained TA-IVLOD setting leads to severe parameter conflicts, exposing a fundamental weakness of existing incremental adaptation schemes.

4 Method

Overview. Figure 5 illustrates the overall architecture of our method, built upon the pre-trained open-vocabulary detector Grounding DINO [22]. To address the task-agnostic incremental learning challenge, our framework is decoupled into two distinct phases. During the training phase (Fig. 5, left), we maintain class-specific LoRA experts and a shared general knowledge module. To proactively align the training distribution with the complex inference scenario, we introduce a Stochastic Prompt Augmentation strategy that constructs cross-task and noisy text prompts dynamically during training. During the task-agnostic inference phase (Fig. 5, right), we propose a Test-Time Distribution-Aware Fusion mechanism. Instead of relying on uniform static averaging, this mechanism dynamically weights and aggregates the uncoordinated LoRA experts based on the probabilistic correspondence between the input visual features and the class mean, thereby mitigating parameter conflicts.

4.1 Stochastic Prompt Augmentation

Revisiting Modular Adaptation. To enable incremental adaptation without catastrophic forgetting, we build upon the modular framework of DitHub [2], which conceptualizes incremental learning as the progressive expansion of a Low-Rank Adaptation (LoRA) library. Rather than employing standard LoRA updates, this framework decouples the adaptation into a shared projection matrix $\mathbf{B}$ and a set of class-specific experts $\{\mathbf{A}_c\}$. For a given class c, the adapted weight $\mathbf{W}$ is formulated as:

$$\mathbf{W} = \mathbf{W}_0 + \mathbf{B}\mathbf{A}_c, \tag{1}$$

where $\mathbf{W}_0$ denotes the frozen pre-trained weights of the base model, and $\mathbf{W}$ represents the final adapted weights for the target layer. During training, only the expert $\mathbf{A}_c$ corresponding to the ground-truth class is activated and updated. This strict parameter isolation effectively circumvents catastrophic forgetting, while the shared matrix $\mathbf{B}$ is continuously optimized across all tasks to maintain generalizable representations.

However, as analyzed in Sec. 3.2, a key limitation arises during task-agnostic evaluation. The model is optimized with isolated, single-task prompts, but evaluated with a unified cross-task prompt containing all learned classes. To address this mismatch between training and inference, we propose **Stochastic Prompt Augmentation (SPA).** By stochastically constructing augmented training prompts, SPA simulates the prompt composition encountered during task-agnostic inference, thereby improving robustness to cross-task prompt interference. Specifically, SPA introduces two types of distractors:

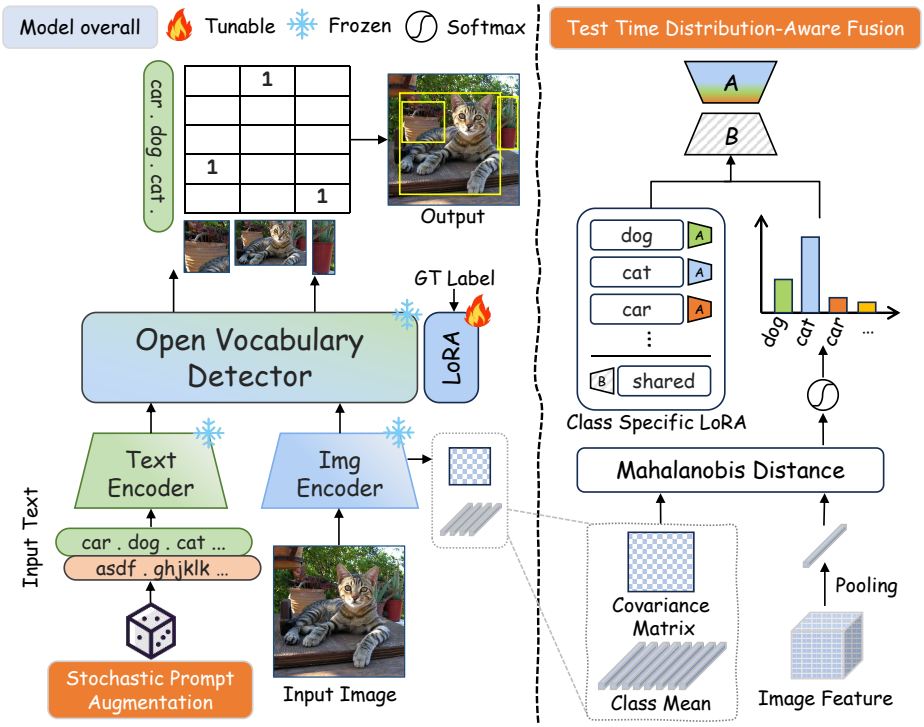


Fig. 5: Overview of our proposed framework. (Left) The training architecture, incorporating Stochastic Prompt Augmentation (SPA) to simulate cross-task semantic interference during optimization. (Right) The Test-Time Distribution-Aware Fusion (TTDF) module, utilized exclusively during inference to dynamically weight and fuse class-specific LoRA experts.

Historical Replay Tokens. To alleviate the model's sensitivity to cross-task interference, we introduce a dynamically sampled number of classes from previous tasks. This forces the model to discriminate the current objects from previously learned concepts within the unified textual context.

Synthetic Noise Tokens. To fortify the model against the semantic interference of unseen classes from future incremental tasks and unknown concepts inherent to open-vocabulary queries, we inject a variable quantity of randomly generated strings. These synthetic tokens act as surrogates for out-of-vocabulary distractors. Crucially, by randomizing not only the textual content but also the exact number of injected tokens, we force the model to flexibly adapt to heterogeneous prompt structures, ensuring robust feature alignment regardless of unknown textual interference.

Formally, let $\mathcal{C}_{cur}$ denote the class set of the current task. During each training iteration, the augmented prompt set $\mathcal{P}_{train}$ is constructed as:

$$\mathcal{P}_{train} = \mathcal{C}_{cur} \cup \mathcal{C}'_{prev} \cup \mathcal{C}_{noise}, \tag{2}$$

where $\mathcal{C}'_{prev}$ and $\mathcal{C}_{noise}$ denote the stochastically sampled past classes and synthetic noise strings, respectively. These additional tokens are introduced for multimodal image–text fusion to enhance robustness to semantic interference and are therefore excluded from object matching and loss computation.

4.2 Test-Time Distribution-Aware Fusion

Inference via Merging. During inference, DitHub dynamically retrieves the corresponding LoRA modules from its library according to the classes specified in the input prompt. The final adaptation is obtained by merging these modules, directly through a naive arithmetic mean operation:

$$\mathbf{W} = \mathbf{W}_0 + \mathbf{B}\left(\frac{1}{|\mathcal{P}|}\sum_{c\in\mathcal{P}}\mathbf{A}_c\right), \tag{3}$$

where $\mathcal{P}$ denotes the set of target classes present in the current inference prompt.

While the static parameter averaging employed by DitHub proves effective for isolated tasks—where intra-task experts share consistent optimization trajectories—indiscriminately merging LoRA modules across heterogeneous tasks inherently disrupts learned weights and triggers severe parameter conflicts. Motivated by this critical vulnerability, we propose a **Test-Time Distribution-Aware Fusion (TTDF)** mechanism. Rather than treating all class-specific experts equally, TTDF dynamically aggregates parameters by evaluating the likelihood that an input image aligns with the learned feature distribution of each expert. In particular, we maintain a probabilistic profile (comprising the mean and covariance) for each LoRA module, enabling us to precisely measure the visual correspondence and selectively activate the most relevant experts during inference.

Online Distribution Modeling. Formally, for each class-specific LoRA module, we track a class mean vector $\boldsymbol{\mu}_c$ (as the retrieval key) and a counter t_c. To characterize the global intra-class variance, we additionally estimate a shared scatter matrix $\mathbf{S}$. All these statistics are computed in a streaming manner during training via an online update rule.

Given an input image $\mathbf{x}$ processed by the image encoder $f(\cdot)$ with ground-truth classes $\mathcal{C}_{gt} = \{c_1, \ldots, c_m\}$, for each class $c \in \mathcal{C}_{gt}$, we update the statistics as follows:

$$\boldsymbol{\mu}_c^{(t)} = \boldsymbol{\mu}_c^{(t-1)} + \frac{1}{t_c}(f(\mathbf{x}) - \boldsymbol{\mu}_c^{(t-1)}), \tag{4}$$

$$\mathbf{S}^{(t)} = \mathbf{S}^{(t-1)} + (f(\mathbf{x}) - \boldsymbol{\mu}_c^{(t-1)})(f(\mathbf{x}) - \boldsymbol{\mu}_c^{(t)})^\top. \tag{5}$$

The shared covariance matrix $\boldsymbol{\Sigma}$ is then derived by normalizing the scatter matrix:

$$\boldsymbol{\Sigma} = \frac{1}{\sum t_c - 1}\mathbf{S}. \tag{6}$$

Mahalanobis Distance-based Reweighting. During task-agnostic inference, for a test image $\mathbf{x}$, we compute the relevance score of each LoRA module using the Mahalanobis distance [8, 17]. Unlike standard Euclidean or cosine metrics that treat all feature dimensions isotropically, the Mahalanobis distance leverages the estimated covariance to capture the true anisotropic distribution of the feature space. This effectively downweights noisy dimensions and accounts for feature correlations:

$$d_c(\mathbf{x}) = (f(\mathbf{x}) - \boldsymbol{\mu}_c)\boldsymbol{\Sigma}^{-1}(f(\mathbf{x}) - \boldsymbol{\mu}_c)^\top. \tag{7}$$

Since a smaller distance indicates a higher likelihood, we convert these distances into normalized importance weights using the Softmax function with a temperature $\boldsymbol{\tau}$:

$$w_c(\mathbf{x}) = \frac{e^{-d_c(\mathbf{x})/\boldsymbol{\tau}}}{\sum_{j\in\mathcal{P}_{TA}} e^{-d_j(\mathbf{x})/\boldsymbol{\tau}}}, \tag{8}$$

where $\mathcal{P}_{TA}$ denotes the set of all learned classes. The temperature $\boldsymbol{\tau}$ explicitly controls the sharpness of this weight distribution.

Finally, the adapted weights are computed by dynamically fusing the class-specific $\mathbf{A}$ matrices based on the input-specific weights $w_c(\mathbf{x})$, while sharing the $\mathbf{B}$ matrix:

$$\mathbf{W} = \mathbf{W}_0 + \mathbf{B}\left(\sum_{c\in\mathcal{P}_{TA}} w_c(\mathbf{x})\mathbf{A}_c\right). \tag{9}$$

This mechanism ensures that the final adaptation is dominated heavily by the most relevant experts while suppressing the noisy contributions of unrelated classes, effectively resolving the parameter conflict in the TA-IVLOD setting.

5 Experiments

5.1 Experimental Setup

Evaluation settings. To comprehensively evaluate our proposed framework, we conduct experiments under two distinct protocols. First, to align with realistic open-world scenarios, we evaluate under our proposed TA-IVLOD setting. In this challenging setup, the model must evaluate all test samples using a unified, cross-task prompt encompassing all encountered classes, without any task identity priors. Second, we use the standard IVLOD setting, where the model relies on a task oracle during inference to construct sparse, task-specific prompts.

Following previous works [2, 5], we conduct experiments on the Object Detection in the Wild (ODinW [19]) benchmark and the MS COCO [20] dataset. ODinW is a challenging benchmark designed to test model performance under significant domain shifts in real-world scenarios. We primarily employ **ODinW-13**, which comprises 13 diverse sub-datasets spanning from traditional natural images (e.g., Pascal VOC [7]) to highly specialized domains (e.g., Thermal imagery, Aerial Maritime Drones, and Aquariums). Specifically, each sub-dataset is

Table 1: Quantitative comparison of mAP performance on ODinW-13 under the TA-IVLOD settings. Avg represents the average mAP across all 13 sequentially learned downstream tasks, while zero-shot generalization performance is reported in the Z_{coco} column. **G-Dino** represents the zero-shot performance of Grounding DINO. **Bold** and underline indicate the best and second-best results, respectively.

Method	Z_{coco}	Avg	Ae	Aq	Co	Eg	Mu	Pa	Pv	Pi	Po	Ra	Sh	Th	Ve
G-Dino	48.4	24.0	15.9	11.5	16.3	28.1	21.0	5.2	46.7	36.4	2.4	0.0	20.4	55.3	53.0
TFA	42.7	22.0	15.0	10.1	16.9	25.9	22.2	4.1	43.6	29.5	1.6	0.0	16.2	50.7	50.4
iDETR	37.4	34.0	30.1	30.7	44.1	40.6	31.6	7.0	58.9	37.0	10.8	0.2	30.3	63.5	57.4
ZiRa	45.1	35.0	28.5	31.4	46.5	34.6	35.8	4.6	59.0	37.6	15.5	0.1	32.7	65.7	62.4
DitHub	46.8	32.5	23.8	16.8	38.2	30.6	37.5	8.5	63.2	40.2	3.8	0.0	30.2	67.9	62.2
ours	**47.2**	**64.5**	**43.4**	**50.3**	**82.0**	**70.6**	**51.0**	**89.0**	**69.3**	**68.6**	**53.3**	**61.4**	**56.2**	**71.4**	**71.3**

formulated as an individual task within the incremental learning process, where the model sequentially learns across these 13 tasks.

For quantitative evaluation, we adopt mean Average Precision (mAP), measured after the model finishes sequential training on all sub-datasets. We report two primary metrics: **Avg** and $\mathbf{Z_{coco}}$. Specifically, Avg represents the average mAP over all downstream tasks in ODinW-13. Under the TA-IVLOD protocol, following the established evaluation paradigm of traditional incremental object detection [16, 25, 26, 41, 42], the individual test sets of all 13 sub-datasets are merged into a unified test set. On this merged test set, an AP is computed for each unique class. For classes shared across multiple tasks, a global AP is computed and applied consistently to the respective task-level mAP calculations.

In addition, we report Z_{coco}, the zero-shot mAP on the MS COCO dataset. Since COCO covers a broad and diverse visual domain, this metric evaluates whether the open-vocabulary detector preserves its inherent zero-shot generalization ability after sequential fine-tuning.

We compare our framework against several state-of-the-art methods. Our primary baselines are ZiRa [5] and DitHub [2], which are specifically tailored for IVLOD. Additionally, we benchmark against methods designed for Incremental Learning (TFA [33], iDETR [6]).

Implementation details. Unless otherwise specified, we employ Grounding DINO with a Swin-Tiny backbone as our default Vision-Language Object Detection Model. This base model is pre-trained on Objects365 [32], GoldG [19], and Cap4M [19]. To enable incremental learning, we build our pipeline upon the DitHub framework. During sequential training, we follow its LoRA-based optimization strategy and corresponding hyperparameter settings to cultivate class-specific experts for all newly encountered classes.

5.2 Comparison Results

All reported quantitative results are averaged over three independent runs with different random seeds to ensure statistical reliability. Prior baselines, including

Table 2: Quantitative comparison of mAP performance on ODinW-13 under the standard IVLOD setting. **G-Dino** represents the zero-shot performance of Grounding DINO. **Bold** and underline indicate the best and second-best results, respectively.

Method	Z_{coco}	Avg	Ae	Aq	Co	Eg	Mu	Pa	Pv	Pi	Po	Ra	Sh	Th	Ve
G-Dino	48.4	52.9	19.2	20.4	80.0	59.7	33.6	81.1	56.4	67.6	27.2	68.4	38.7	74.1	61.3
TFA	42.7	50.1	18.0	19.5	76.7	52.3	34.5	79.7	52.6	64.3	21.4	67.5	32.2	74.3	58.4
iDETR	37.4	58.4	35.1	40.3	79.3	59.7	47.6	82.6	62.6	66.0	39.3	67.4	46.9	75.1	57.0
ZiRa	45.1	60.6	34.6	43.1	82.0	63.9	48.7	82.4	64.5	69.3	46.4	68.9	44.8	77.3	62.2
DitHub	46.8	66.3	35.9	51.9	83.2	**71.6**	**56.4**	87.4	70.0	70.6	54.7	73.4	55.2	82.9	69.2
ours	**47.2**	**67.5**	**36.9**	**54.7**	**85.1**	71.3	53.6	**89.2**	**71.0**	**71.8**	**55.1**	**74.8**	**58.7**	**84.6**	**71.0**

Table 3: Quantitative results of few-shot (1, 5, and 10 shots) incremental object detection on ODinW-13. Avg_I and Avg_{TA} denote the average mAP across all 13 downstream tasks under IVLOD and TA-IVLOD settings, respectively.

Method	1 shot			5 shot			10 shot		
	Z_{coco}	Avg_I	Avg_{TA}	Z_{coco}	Avg_I	Avg_{TA}	Z_{coco}	Avg_I	Avg_{TA}
TFA	45.1	50.5	23.0	44.4	51.0	22.9	44.7	50.9	22.9
iDETR	45.6	52.7	27.7	41.8	55.5	29.3	42.7	56.8	31.7
ZiRa	46.5	54.1	27.1	45.2	56.9	32.1	45.7	57.1	30.1
DitHub	45.9	54.9	26.2	44.1	59.8	28.3	45.2	60.5	30.3
ours	46.7	56.2	33.7	44.1	61.2	45.7	45.5	61.0	53.9

TFA, iDETR, ZiRa, and DitHub, are re-implemented under our unified evaluation protocol. The discrepancies between our reproduced results and those reported in [5] are discussed in detail in the supplementary material.

Performance in the TA-IVLOD Setting. As shown in Tab. 1, our proposed method demonstrates robust performance against cross-task semantic interference, yielding a state-of-the-art average mAP of 64.5 in the challenging TA-IVLOD setting. This corresponds to a +32.0 absolute improvement over DitHub. This substantial gain indicates that the synergistic design of SPA and TTDF effectively alleviates parameter conflicts and cross-task semantic interference.

In contrast, prior methods experience significant performance degradation under this unconstrained setting, often approaching near-zero performance on certain tasks (*e.g.*, Raccoon (Ra) and Packages (Pa)). This behavior is largely attributed to the synthetic long-tailed distribution introduced by aggregating test sets from distinct tasks. Without explicit task boundaries, models optimized solely with single-task prompts struggle to suppress irrelevant cross-task distractors. Consequently, for small-scale tasks with limited positive instances, increased false-positive rates substantially degrade the precision metric.

Performance in the Standard IVLOD Setting. Beyond the improvements observed in the task-agnostic scenario, our approach maintains strong detec-

Table 4: Ablation study of TTDF and SPA on ODinW-13.

	TTDF	SPA	Z_{coco}	Avg_{TA}
baseline			46.8	32.5
ours	✓		47.4	45.9
	✓	✓	47.2	64.5

Table 5: Ablation study of different prompt augmentation strategies.

	Z_{coco}	Avg_{TA}
Learned classes	47.0	57.2
Noise tokens	47.4	60.3
COCO	46.0	58.2
Learned class + Noise tokens	47.2	64.5

tion performance under the standard IVLOD setting. As detailed in Tab. 2, our method achieves an average mAP of 67.5, outperforming DitHub by +1.2 mAP. This consistent improvement indicates that our TTDF module effectively mitigates parameter conflicts among class-specific experts even when oracle task identities are provided during inference.

Robustness Across Few-Shot Regimes. To further assess generalization, we evaluate its performance under strictly limited data conditions. As shown in Tab. 3, our approach consistently outperforms existing baselines across 1-shot, 5-shot, and 10-shot configurations. Notably, the performance advantage is maintained in both the standard IVLOD (Avg_I) and the TA-IVLOD (Avg_{TA}) settings, underscoring the general efficacy and stability of our proposed modules across varying data scales.

5.3 Model Analysis and Ablation Studies

Module Ablation. We investigate the individual and combined efficacy of our core modules on the ODinW-13 benchmark, as detailed in Tab. 4. Compared to the baseline (DitHub), which suffers significant performance degradation in the task-agnostic setting, incorporating TTDF yields a notable improvement of +12.7 mAP, demonstrating its effectiveness in mitigating dynamic parameter conflicts. Furthermore, integrating Stochastic Prompt Augmentation (SPA) on top of TTDF brings an additional gain of +18.4 mAP. This demonstrates that synergizing train-time textual augmentation with test-time dynamic routing facilitates more discriminative target identification under complex semantic interference, ultimately improving the overall performance to 64.5 mAP.

Prompt Augmentation Strategies. Table 5 compares four augmentation strategies utilizing different distractor types: historical learned classes, synthetic noise tokens, external COCO classes, and our combined approach. Specifically, while learned classes provide a basic level of robustness, the gains are constrained by the limited diversity of previously seen concepts. Although introducing COCO classes enhances task-agnostic resistance in downstream tasks, it leads to a performance decline in the zero-shot detection of those specific classes due to the semantic confusion induced during training. Employing any real-world classes as distractors inherently impairs the model's ability to recognize those specific classes, as it is forced to treat valid semantic concepts as negative interference. In contrast, synthetic noise tokens offer a unique advantage by providing textual

Table 6: Comparison of different distance metrics for TTDF.

	Z_{coco}	Avg_{TA}
Euclidean Distance	45.9	51.3
Mahalanobis Distance	47.2	64.5

Table 7: Comparison with task-agnostic baselines under IOD and TA-IVLOD settings.

	G-Dino	GCD	DitHub	ours
IOD (40-40)	48.4	55.2	53.2	54.5
TA-IVLOD	24.0	29.0	32.5	64.5

complexity without concrete semantic meaning, thus preserving the zero-shot capabilities of real-world classes. By combining historical classes with synthetic noise, our method successfully emulates the complex, varied-length prompts encountered during actual inference. This synergy between realistic distractors and neutral noise yields the best performance across all metrics.

Mahalanobis Distance. Table 6 compares the effectiveness of different distance metrics within our TTDF module. As illustrated, employing Mahalanobis distance yields substantial improvements over the standard Euclidean distance. The effectiveness of Mahalanobis distance stems from its ability to account for the underlying data distribution and feature correlations. Euclidean distance assumes an isotropic spherical distribution, which often fails to capture true semantic proximity in high-dimensional latent spaces characterized by varying scales and feature inter-dependencies. By contrast, Mahalanobis distance effectively normalizes these variances and incorporates the covariance structure of the feature clusters. This allows the TTDF module to more accurately identify and retrieve the most relevant class-specific experts during inference, thereby substantially mitigating parameter conflicts and enhancing the overall discriminative power of the model.

Generalization to Incremental Object Detection. Conventional Incremental Object Detection (IOD) aims to incrementally learn new object categories while retaining previously learned ones, and has been widely studied in class-incremental detection scenarios [6, 16, 34, 41, 43]. Unlike our TA-IVLOD setting, which additionally involves open-vocabulary detectors, heterogeneous downstream domains, and unified cross-task prompts without task-specific priors, conventional IOD is typically evaluated within a fixed dataset and a closed category space. To further validate the generalizability of our method beyond the IVLOD benchmark, we additionally conduct experiments under the conventional Incremental Object Detection (IOD) setting on MS COCO. Specifically, the 80 COCO categories are split into two sequential learning stages, with 40 classes introduced at each stage. In this setting, images may naturally contain objects from different incremental stages, requiring the detector to handle mixed-category scenes after sequential learning. As shown in Tab. 7, our method remains competitive with the recent task-agnostic IOD method GCD [34] under the conventional IOD protocol, while outperforming other baselines. More importantly, our method exhibits a clear advantage over GCD under the more challenging TA-IVLOD setting, where diverse task domains are evaluated jointly without task-specific priors. This demonstrates that our framework not only gen-

eralizes well to conventional IOD scenarios, but also remains effective in more heterogeneous task-agnostic environments.

6 Conclusion

In this paper, we identify a critical limitation in current IVLOD research: the over-reliance on oracle task identities, which restricts practical open-world deployment. To address this, we formalize TA-IVLOD, a more realistic setting where task boundaries are unknown during inference. We demonstrate that existing methods suffer from severe performance degradation in this setting due to prompt-induced semantic interference.

To overcome these challenges, we propose a modular LoRA-based framework featuring a Test-Time Distribution-Aware Fusion mechanism and a Stochastic Prompt Augmentation strategy. These components effectively mitigate parameter conflicts and enhance cross-modal robustness against cross-task, unified prompts. Extensive evaluations on the ODinW-13 benchmark prove that our method not only excels in standard IVLOD but also yields substantial improvements in the challenging TA-IVLOD setting, while effectively preserving pre-trained zero-shot capabilities.

Despite its effectiveness, the dynamic nature of our test-time fusion introduces a slight computational overhead during inference. Additionally, while our current fusion mechanism relies on image-level features for dynamic weighting, exploring instance-level features presents a promising avenue for future work.

Acknowledgements

This work is funded by NSFC (NO. 62206135, 62225604), Young Elite Scientists Sponsorship Program by CAST (2023QNRC001), the Fundamental Research Funds for the Central Universities (Nankai University, 070-63263251), Guangdong Basic and Applied Basic Research Foundation (No. 2026A1515011147), Shenzhen Science and Technology Program (JCYJ20240813114237048).

References

1. Asadi, N., Beitollahi, M., Khalil, Y., Li, Y., Zhang, G., Chen, X.: Does combining parameter-efficient modules improve few-shot transfer accuracy? Artificial Intelligence for Engineering **1**(1), 7–18 (2025)
2. Cappellino, C., Mancusi, G., Mosconi, M., Porrello, A., Calderara, S., Cucchiara, R.: Dithub: A modular framework for incremental open-vocabulary object detection. arXiv preprint arXiv:2503.09271 (2025)
3. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European conference on computer vision. pp. 213–229. Springer (2020)
4. Chen, X., Zhang, J., Wang, X., Zhang, N., Wu, T., Wang, Y., Wang, Y., Chen, H.: Continual multimodal knowledge graph construction. arXiv preprint arXiv:2305.08698 (2023)

5. Deng, J., Zhang, H., Ding, K., Hu, J., Zhang, X., Wang, Y.: Zero-shot generalizable incremental learning for vision-language object detection. Advances in Neural Information Processing Systems **37**, 136679–136700 (2024)
6. Dong, N., Zhang, Y., Ding, M., Lee, G.H.: Incremental-detr: Incremental few-shot object detection via self-supervised learning. Proceedings of the AAAI Conference on Artificial Intelligence **37**(1), 543–551 (2023)
7. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. International journal of computer vision **88**(2), 303–338 (2010)
8. Goswami, D., Liu, Y., Twardowski, B., Van De Weijer, J.: Fecam: Exploiting the heterogeneity of class distributions in exemplar-free continual learning. Advances in Neural Information Processing Systems **36**, 6582–6595 (2023)
9. Gu, X., Lin, T.Y., Kuo, W., Cui, Y.: Open-vocabulary object detection via vision and language knowledge distillation. In: International Conference on Learning Representations (2022)
10. Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., Attariyan, M., Gelly, S.: Parameter-efficient transfer learning for nlp. In: International conference on machine learning. pp. 2790–2799. PMLR (2019)
11. Hu, E.J., yelong shen, Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-rank adaptation of large language models. In: International Conference on Learning Representations (2022)
12. Huang, C., Liu, Q., Lin, B.Y., Pang, T., Du, C., Lin, M.: Lorahub: Efficient cross-task generalization via dynamic lora composition. arXiv preprint arXiv:2307.13269 (2023)
13. Huang, L., Cao, X., Lu, H., Liu, X.: Class-incremental learning with clip: Adaptive representation adjustment and parameter fusion. In: European Conference on Computer Vision. pp. 214–231. Springer (2024)
14. Huang, L., Cao, X., Lu, H., Meng, Y., Yang, F., Liu, X.: Mind the gap: Preserving and compensating for the modality gap in clip-based continual learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3777–3786 (2025)
15. Jia, M., Tang, L., Chen, B.C., Cardie, C., Belongie, S., Hariharan, B., Lim, S.N.: Visual prompt tuning. In: European conference on computer vision. pp. 709–727. Springer (2022)
16. Kim, J., Ku, Y., Kim, J., Cha, J., Baek, S.: Vlm-pl: Advanced pseudo labeling approach for class incremental object detection via vision-language model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4170–4181 (2024)
17. Lee, K., Lee, K., Lee, H., Shin, J.: A simple unified framework for detecting out-of-distribution samples and adversarial attacks. Advances in neural information processing systems **31** (2018)
18. Li, J., Huang, Y., Wu, M., Zhang, B., Ji, X., Zhang, C.: Clip-sp: Vision-language model with adaptive prompting for scene parsing. Computational Visual Media **10**(4), 741–752 (2024)
19. Li, L.H., Zhang, P., Zhang, H., Yang, J., Li, C., Zhong, Y., Wang, L., Yuan, L., Zhang, L., Hwang, J.N., et al.: Grounded language-image pre-training. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10965–10975 (2022)
20. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014)

21. Liu, S.Y., Wang, C.Y., Yin, H., Molchanov, P., Wang, Y.C.F., Cheng, K.T., Chen, M.H.: Dora: Weight-decomposed low-rank adaptation. In: Forty-first International Conference on Machine Learning (2024)
22. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Jiang, Q., Li, C., Yang, J., Su, H., et al.: Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In: European conference on computer vision. pp. 38–55. Springer (2024)
23. Liu, W., Zhu, F., Wei, L., Tian, Q.: C-clip: Multimodal continual learning for vision-language model. In: The Thirteenth International Conference on Learning Representations (2025)
24. Liu, X., Yang, H., Ravichandran, A., Bhotika, R., Soatto, S.: Multi-task incremental learning for object detection. arXiv preprint arXiv:2002.05347 (2020)
25. Liu, Y., Cong, Y., Goswami, D., Liu, X., Van De Weijer, J.: Augmented box replay: Overcoming foreground shift for incremental object detection. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 11367–11377 (2023)
26. Luo, W., Zhang, S., Cheng, D., Xing, Y., Liang, G., Wang, P., Zhang, Y.: Gradient decomposition and alignment for incremental object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4486–4495 (2025)
27. Maaz, M., Rasheed, H., Khan, S., Khan, F.S., Anwer, R.M., Yang, M.H.: Class-agnostic object detection with multi-modal transformer. In: European conference on computer vision. pp. 512–531. Springer (2022)
28. Minderer, M., Gritsenko, A., Houlsby, N.: Scaling open-vocabulary object detection. Advances in Neural Information Processing Systems **36**, 72983–73007 (2023)
29. Oren, G., Wolf, L.: In defense of the learning without forgetting for task incremental learning. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 2209–2218 (2021)
30. Peng, Z., Xu, Z., Zeng, Z., Huang, Y., Wang, Y., Shen, W.: Parameter-efficient fine-tuning in hyperspherical space for open-vocabulary semantic segmentation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 15009–15020 (2025)
31. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
32. Shao, S., Li, Z., Zhang, T., Peng, C., Yu, G., Zhang, X., Li, J., Sun, J.: Objects365: A large-scale, high-quality dataset for object detection. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 8430–8439 (2019)
33. Wang, X., Huang, T.E., Darrell, T., Gonzalez, J.E., Yu, F.: Frustratingly simple few-shot object detection. arXiv preprint arXiv:2003.06957 (2020)
34. Wang, X., Wang, Z., Lin, Z.: Gcd: Advancing vision-language models for incremental object detection via global alignment and correspondence distillation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 8015–8023 (2025)
35. Wu, X., Huang, S., Wei, F.: Mixture of lora experts. arXiv preprint arXiv:2404.13628 (2024)
36. Wu, Y., Piao, H., Huang, L.K., Wang, R., Li, W., Pfister, H., Meng, D., Ma, K., Wei, Y.: Sd-lora: Scalable decoupled low-rank adaptation for class incremental learning. arXiv preprint arXiv:2501.13198 (2025)

37. Yadav, P., Tam, D., Choshen, L., Raffel, C.A., Bansal, M.: Ties-merging: Resolving interference when merging models. Advances in neural information processing systems **36**, 7093–7115 (2023)
38. Yu, Y.C., Huang, C.P., Chen, J.J., Chang, K.P., Lai, Y.H., Yang, F.E., Wang, Y.C.F.: Select and distill: Selective dual-teacher knowledge transfer for continual learning on vision-language models. In: European Conference on Computer Vision. pp. 219–236. Springer (2024)
39. Zhang, H., Li, F., Liu, S., Zhang, L., Su, H., Zhu, J., Ni, L.M., Shum, H.Y.: Dino: Detr with improved denoising anchor boxes for end-to-end object detection. arXiv preprint arXiv:2203.03605 (2022)
40. Zhang, H., Zhang, P., Hu, X., Chen, Y.C., Li, L., Dai, X., Wang, L., Yuan, L., Hwang, J.N., Gao, J.: Glipv2: Unifying localization and vision-language understanding. Advances in Neural Information Processing Systems **35**, 36067–36080 (2022)
41. Zhang, H., Gao, B.B., Zeng, Y., Tian, X., Tan, X., Zhang, Z., Qu, Y., Liu, J., Xie, Y.: Learning task-aware language-image representation for class-incremental object detection. Proceedings of the AAAI Conference on Artificial Intelligence **38**(7), 7096–7104 (2024)
42. Zhang, J., Liu, L., Silvén, O., Pietikäinen, M., Hu, D.: Few-shot class-incremental learning for classification and object detection: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence **47**(4), 2924–2945 (2025)
43. Zhang, S., Lv, X., Xing, Y., Wu, Q., Xu, D., Zhao, C., Zhang, Y.: Yolo-iod: Towards real time incremental object detection. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40, pp. 12744–12752 (2026)
44. Zhang, X., Zhang, F., Xu, C.: Vqacl: A novel visual question answering continual learning setting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19102–19112 (2023)
45. Zheng, S., Wang, H., Huang, C., Wang, X., Chen, T., Fan, J., Hu, S., Ye, P.: Decouple and orthogonalize: A data-free framework for lora merging. arXiv preprint arXiv:2505.15875 (2025)
46. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Conditional prompt learning for vision-language models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16816–16825 (2022)
47. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. International journal of computer vision **130**(9), 2337–2348 (2022)
48. Zhu, C., Chen, L.: A survey on open-vocabulary detection and segmentation: Past, present, and future. IEEE Transactions on Pattern Analysis and Machine Intelligence **46**(12), 8954–8975 (2024)