CL-CLIP: CLIP-BASED CONTINUAL LEARNING FRAMEWORK WITH COST-VOLUME CATEGORY DECOUPLING FOR OBJECT DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Continual Object Detection (COD) requires a detector to acquire new categories over time while preserving previously learned ones. This goal is closely related to open-vocabulary detection, since both settings require reasoning over categories that are not fully covered by the annotations available at the current training stage. Recent CLIP-based open-vocabulary detectors have shown strong zero-shot generalization, and frameworks such as F-ViT demonstrate that vision-language pretraining can provide powerful zero-shot detection ability for unseen categories. However, real-world deployments cannot remain purely zero-shot: once these detectors are continually updated on newly introduced categories, they suffer severe catastrophic forgetting and quickly lose their previously calibrated detection ability. We therefore propose CL-CLIP, a CLIP-based COD framework that equips open-vocabulary detectors with better continual learning ability through cost-volume-guided category decoupling. Specifically, following CAT-Seg, we compute a CLIP image-text similarity cost volume, defined as dense category-wise response maps between visual tokens and class text embeddings. This zero-shot spatial prior decomposes shared region features into class-specific pathways, which are then processed by a Multi-Expert RoI head. Extensive experiments on PASCAL VOC and MS-COCO show that CL-CLIP substantially improves the F-ViT baseline under continual fine-tuning and achieves competitive performance with existing continual object detectors, especially in adapting to newly introduced categories while preserving competitive base-class performance.

1 INTRODUCTION

Continual Object Detection (COD) aims to learn novel categories from a task stream while preserving previously acquired detection ability (Li & Hoiem, 2018; Rebuffi et al., 2017; Peng et al., 2020; Feng et al., 2022). Unlike image classification, however, COD must handle multi-object scenes in which old and new categories co-occur spatially. This leads to the background relegation problem. When the model is updated on a new task, previously learned objects appear without annotations and are treated as background, rapidly eroding old-class decision boundaries and causing severe catastrophic forgetting.

Existing COD methods address this problem through knowledge distillation (Peng et al., 2020; Feng et al., 2022; He et al., 2023; Wang et al., 2025), regularization (Kirkpatrick et al., 2017; Wu et al., 2025; Luo et al., 2025), replay (Gupta et al., 2022; Liu et al., 2023; An et al., 2025), and pseudo-labeling (Yang et al., 2023). These studies provide important mechanisms for preserving previously learned categories during continual learning. In this work, we explore a different route by using open-vocabulary detection (OVD) as a starting point for COD. This route is motivated by the overlap between the two settings, which both require detectors to reason about categories that are not fully covered by the annotations available at a given training stage.

CLIP-based OVD detectors are especially relevant to this route because CLIP provides strong zero-shot ability, which has been widely used to detect categories beyond the training label set. Fine-grained CLIP variants (Zhong et al., 2022; Jing et al., 2024; Xie et al., 2025) and frameworks such as F-ViT (Wu et al., 2024) show that strong vision-language alignment can provide powerful zero-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

shot detection ability for unseen categories. However, OVD and COD differ in a crucial way. OVD relies on zero-shot transfer without updating the detector, whereas COD repeatedly fine-tunes it as new categories arrive. Once F-ViT-style frameworks are placed in this continual setting, their shared trainable heads are easily overwritten, and the old-class detection ability quickly deteriorates. Traditional remedies such as distillation or replay do not resolve this shared-head interference, and existing architecture-specific continual designs based on learnable prompts (Yi et al., 2025) or memory networks (Bhatt et al., 2025) are typically developed for DETR (Carion et al., 2020) or Grounding DINO (Liu et al., 2025) pipelines rather than for CLIP-based detectors.

To address this limitation, we propose CL-CLIP, a continual object detection framework built on category decoupling for CLIP-based open-vocabulary detectors. Inspired by CAT-Seg (Cho et al., 2024), which constructs an image-text similarity cost volume to produce category-wise spatial responses for semantic segmentation, we ask a natural question: can such responses serve as a spatial prior for previously learned classes during continual learning, thereby reducing forgetting under incomplete current-task annotations? Following this insight, CL-CLIP converts the category decomposition ability of CLIP into class-specific detection pathways, reducing direct competition between old and new categories during continual updates.

The underlying principle is that a frozen CLIP model produces category-indexed spatial responses invariant to downstream fine-tuning, serving as a *structurally stable routing signal* that decomposes shared features into class-specific pathways without being corrupted by missing-label supervision on subsequent tasks. This principle is instantiated through two core designs:

- **Cost-volume-guided category decoupling.** The CLIP cost volume provides a zero-shot spatial prior that separates shared region features into per-category pathways, keeping category responses well separated across tasks.
- **Multi-Expert RoI head with drift stabilization.** Each category is assigned a convolutional expert that processes its decoupled features and produces CLIP-aligned detection scores via per-class sigmoid classifiers. Combined with drift regularization on the shared FPN modules, this design reduces cross-class interference and better preserves old-class representations during continual updates.

To our knowledge, CL-CLIP is the first framework to systematically adapt cost-volume-based category decomposition for continual object detection, establishing a previously unexplored connection between open-vocabulary spatial priors and architectural anti-forgetting mechanisms. Through extensive experiments on PASCAL VOC (Everingham et al., 2010) and MS-COCO (Lin et al., 2014), including a systematic benchmark for forgetting across multiple CLIP variants within the F-ViT baseline, we show that CL-CLIP substantially improves continual detection performance and achieves competitive results with existing continual object detectors, with particularly strong adaptation to newly introduced categories.

2 Related works

Continual Learning. Continual learning studies how a model absorbs new knowledge from a task stream without forgetting previous tasks (Li & Hoiem, 2018; Rebuffi et al., 2017). Our work focuses on continual object detection, where the challenge is not only catastrophic forgetting but also preserving spatially localized representations while reducing region-level interference between old and new categories.

Continual Object Detection. Existing COD methods mitigate forgetting through several lines of solutions. Distillation methods (Faster-ILOD (Peng et al., 2020), GCD (Wang et al., 2025), MMA (Cermelli et al., 2022)) transfer old-task knowledge to the current model. Regularization methods (IncDet (Liu et al., 2021), GDA (Luo et al., 2025), NSGP-RePRE (Wu et al., 2025)) constrain parameters or gradients to reduce harmful updates. Replay methods (ABR (Liu et al., 2023), IOR (An et al., 2025)) preserve previous knowledge by reusing stored or generated samples. Pseudo-labeling methods (Yang et al., 2023) recover old objects that reappear without annotations. More recent methods such as ROSETTA (Yang et al., 2022) and TARL (Zhang et al., 2024) isolate task-specific parameters to reduce interference, and LEA (Song et al., 2025) learns endogenous per-task attention modules and predicts a task identity at inference to route detections. These approaches can

be characterized along two axes: *isolation granularity* (task-level prompts or branches vs. category-level pathways) and *old-class recovery strategy* (pseudo-labeling, replay or architectural separation). Most prior work operates at the task level and relies on explicit old-class recovery mechanisms. In contrast, CL-CLIP performs category-level decoupling by constructing class-specific RoI pathways processed by per-category experts, requiring neither pseudo-labels nor replay and eliminating the need for task identity at inference.

CLIP Models for Dense Prediction and Continual Adaptation. CLIP (Radford et al., 2021) has become a strong foundation for open-vocabulary dense prediction. Prior works developed improved or fine-grained CLIP variants such as RegionCLIP, SigLIP2, EVA-CLIP, FineCLIP, and FG-CLIP (Zhong et al., 2022; Tschannen et al., 2025; Sun et al., 2023; Jing et al., 2024; Xie et al., 2025), and frameworks such as F-ViT (Wu et al., 2024), the ViT-based instantiation of the frozen-VLM detector F-VLM (Kuo et al., 2023), show that a frozen vision-language backbone can support open-vocabulary detection. However, under sequential fine-tuning, F-ViT's shared detection head overwrites representations calibrated for old classes, motivating an explicitly continual architecture.

Recent studies have begun to explore CLIP for continual detection. IODC (He et al., 2023) introduces CLIP through pseudo-labeling and prompt tuning, but depends on pseudo-label quality. TARL (Zhang et al., 2024) applies task-aware parameter isolation on GLIP (Li et al., 2022) but operates at the task level and relies on a grounding-based backbone. In contrast, CL-CLIP avoids pseudo-labeling and decouples at the category level, enabling per-class pathways without requiring task identity at inference. Inspired by CAT-Seg (Cho et al., 2024), designed for static open-vocabulary segmentation, we adopt its category-indexed spatial responses as a structural prior for continual detection, removing its cross-class aggregation module to avoid inter-class mixing under missing-label supervision.

3 METHODOLOGY

We present CL-CLIP, a framework that equips a CLIP-based open-vocabulary detector with continual learning ability. Starting from a frozen CLIP backbone and the F-ViT detection pipeline, we construct a Similarity Cost Volume from image-text alignment as a zero-shot spatial prior. This prior separates shared multi-scale features into class-specific pathways during sequential fine-tuning, which are processed by a Multi-Expert RoI head to reduce cross-class interference. As an auxiliary proposal-level refinement, we also use the category responses to softly guide the RPN toward likely foreground regions. Throughout the framework, only the CLIP encoders and previous-task experts in the RoI head are frozen; all other modules (cost-volume pathway, FPN, RPN, current-task experts) are trainable. The overall architecture is illustrated in Fig. 1.

Notations: Let I denote the input image and $\mathcal{F}(\cdot)$ represent the frozen CLIP vision backbone. For a given proposal b and a candidate category set $\mathcal{C}$, CL-CLIP produces class scores $S_b^{\mathcal{C}}$ through the same decoupled expert pipeline during both training and inference.

3.1 PRELIMINARIES

Continual Learning. COD requires a model to learn from a task stream $\mathcal{T} = \{\mathcal{T}_1, \ldots, \mathcal{T}_n\}$. At step t, the dataset $\mathcal{D}_t = \{\mathcal{X}_i, \mathcal{Y}_i\}$ contains annotations only for the current categories C_c, while previously learned categories C_p may still appear without labels. After training on task t, the detector is evaluated over all seen categories $C_p \cup C_c$.

F-ViT Framework. We built on F-ViT (Wu et al., 2024), the ViT-based frozen-CLIP detector introduced by CLIPSelf, which follows the F-VLM (Kuo et al., 2023) paradigm using multi-layer ViT features. Following this design, F-ViT freezes the image and text encoders and trains a standard detection head with text embeddings as classifier weights. While this provides strong zero-shot detection ability, its trainable head remains shared across categories. Under continual fine-tuning with incomplete annotations, updates for current classes can overwrite the representations previously calibrated for old classes.

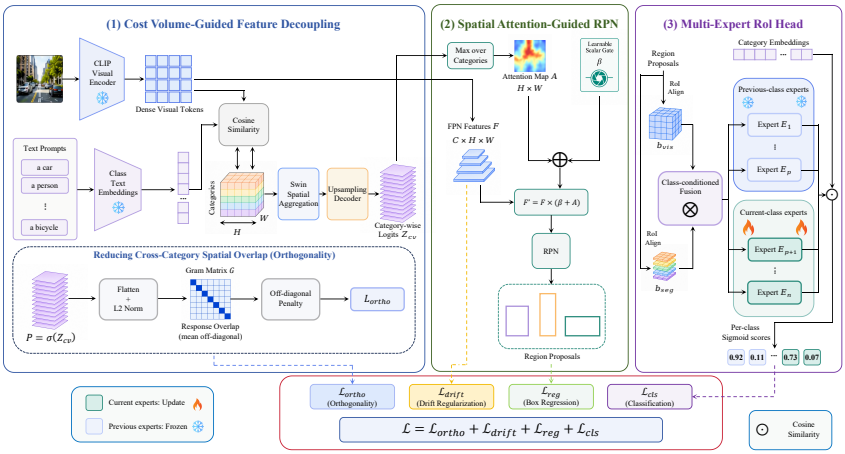

Figure 1: Overview of CL-CLIP. (1) Cost Volume-Guided Feature Decoupling constructs category-wise spatial responses from CLIP image-text similarity; (2) the auxiliary Spatial Attention-Guided RPN uses the max-pooled responses as a foreground prior; (3) Multi-Expert RoI Head assigns per-category convolutional experts with frozen/trainable separation. The orthogonality loss reduces cross-category spatial overlap via a Gram matrix penalty.

3.2 Cost Volume-Guided Feature Decoupling

The first step of CL-CLIP is to construct category-indexed spatial responses as class-specific priors for continual detection. The key property we exploit is that these responses are derived from the frozen CLIP encoders: since neither the image nor text encoder is updated during continual fine-tuning, the cost volume provides a stable per-category spatial signal that does not degrade as new tasks arrive. Following CAT-Seg (Cho et al., 2024), originally designed for open-vocabulary semantic segmentation, we compute the cosine similarity between dense CLIP visual tokens and category-specific text embeddings to obtain a Similarity Cost Volume. This cost volume is refined by Swin-based spatial aggregation within each category response and an upsampling decoder, producing category-wise logits maps $\mathbf{Z}_{cv}$:

$$\mathbf{Z}_{cv} = Cost(\mathcal{F}(I)). \tag{1}$$

We adapt CAT-Seg for continual detection with two key modifications. First, whereas CAT-Seg uses the cost volume as *final dense predictions* for segmentation, we use it as a *spatial prior* that gates region features for downstream proposal-based detection. Second, we remove CAT-Seg's class aggregation module, which re-introduces cross-category coupling and propagates forgetting signals from unannotated old objects that are treated as background. The resulting category-wise responses are later used by the RoI head to construct class-conditioned RoI representations (detailed in Sec. 3.4).

Reducing Cross-Category Spatial Overlap. Because the response maps are learned without dense mask supervision, different categories may co-activate on the same spatial regions. We introduce a soft orthogonality constraint. Let $\mathbf{P} = \sigma(\mathbf{Z}_{cv}) \in \mathbb{R}^{K\times H\times W}$ denote the response probabilities of categories. We flatten and apply ℓ_2 normalization over the $N = H \times W$ spatial positions to obtain $\tilde{\mathbf{P}} \in \mathbb{R}^{K\times N}$, then compute the Gram matrix $\mathbf{G} = \tilde{\mathbf{P}}\tilde{\mathbf{P}}^\top$ to measure spatial co-activation across class channels, and penalize its off-diagonal entries:

$$\mathcal{L}_{ortho} = \frac{1}{N_b \cdot K(K-1)} \sum_{n=1}^{N_b} \sum_{i\neq j} (\mathbf{G}_{n,i,j})^2. \tag{2}$$

We use the mean squared off-diagonal value as a quantitative measure of cross-category spatial entanglement, which we call **Response Overlap**; Tab. 5 analyzes how this metric changes across decoupling variants.

3.3 Auxiliary Cost-Volume Guidance for the RPN

As a minor proposal-level refinement, we use the Similarity Cost Volume to guide the RPN (Ren et al., 2015), whose objectness targets are otherwise defined only by current-task annotations and may therefore suppress old objects. We take the maximum response over the category dimension to obtain a class-agnostic attention map.

Rather than replacing FPN (Lin et al., 2017) features with this prior, we use a residual gating form:

$$\mathbf{F}' = \mathbf{F} \times (\beta + \mathbf{A}), \tag{3}$$

where $\mathbf{F}$ is the FPN feature, β is a learnable scalar gate initialized to 1, and $\mathbf{A}$ is the resized attention map. This softly biases the RPN toward regions with high CLIP-text response while preserving the original features. Detailed proposal-recall comparisons are provided in Appendix C.

3.4 Multi-Expert RoI Head

We keep the standard box regression branch but replace the shared RoI classifier with category-specific experts. For each proposal b, RoI-Align extracts a visual feature $b_{vis} \in \mathbb{R}^{d\times d_1\times d_2}$ from the FPN and a per-category spatial gate $b_{seg} \in \mathbb{R}^{K\times d_1\times d_2}$ from the sigmoid-activated cost-volume maps, where each slice $b_{seg}[c] \in \mathbb{R}^{d_1\times d_2}$ is a spatial attention mask for category c. Broadcasting $b_{seg}[c]$ across channels yields the class-conditioned feature $b_{dec}[c] = b_{vis} \odot b_{seg}[c]$, so that each of the K slices in $b_{dec} \in \mathbb{R}^{K\times d\times d_1\times d_2}$ retains only the spatially relevant information for its category. Each category c is assigned a convolutional expert $\mathcal{E}_c$ that processes its slice $b_{dec}[c]$. Following Sylph (Yin et al., 2022), we replace the shared softmax classifier with per-class sigmoid classifiers, but differ in that Sylph dynamically generates expert weights from few-shot support images over shared RoI features, whereas our experts are directly learned modules that operate on already spatially-gated, class-specific features. Once a category's expert is trained, it is frozen in subsequent tasks, providing architectural protection against forgetting. We compute CLIP-aligned scores:

$$s_b^c = \tau \cdot \cos\big(\mathcal{E}_c(b_{dec}[c]),\, \mathbf{t}_c\big), \tag{4}$$

where $\mathbf{t}_c$ is the CLIP text embedding for category c and τ is the logit scale. At each task t, the cost volume and expert pipeline operate over all seen categories $C_p \cup C_c$, so that frozen old-class experts continue to receive their gated inputs. The classification loss $\mathcal{L}_{cls}$ is Binary Cross-Entropy with Logits applied to $\{s_b^c\}_{c\in C_c}$; proposals matching previous categories are excluded from the classification loss rather than treated as negatives. At inference, scoring is applied to all seen categories $C_p \cup C_c$, and the detection score is $S_b^c = \sigma(s_b^c)$. Only current categories' experts are updated; previous categories' experts are frozen. Tab. 3 examines the necessity of the decoupled expert pathway, while Appendix C compares alternative expert structures; a t-SNE visualization is provided in Fig. 4.

Because all experts share a fixed logit scale τ and align to frozen CLIP text embeddings through bounded cosine similarity, old and new experts produce comparable score magnitudes without post-hoc calibration.

Since the decoupled RoI features are still fused with FPN representations, we apply EWC (Kirkpatrick et al., 2017) drift regularization ($\mathcal{L}_{drift}$) on the lateral and top-down fusion convolutions of the FPN. The diagonal Fisher information is computed after each task from the total loss gradients, and the drift loss penalizes parameter deviations weighted by importance. Together with the L1 box regression loss ($\mathcal{L}_{reg}$), the full training objective is:

$$\mathcal{L} = \mathcal{L}_{cls} + \mathcal{L}_{reg} + \mathcal{L}_{ortho} + \lambda\mathcal{L}_{drift}. \tag{5}$$

4 Experiments

4.1 Datasets and Evaluation Protocols

We evaluate CL-CLIP on PASCAL VOC (Everingham et al., 2010) and MS-COCO (Lin et al., 2014). For VOC, we use the standard two-step splits 10+10, 15+5, and 19+1, where $A + B$ denotes A classes in the first task and B newly introduced classes in the second task. For COCO, we mainly use the 4-task protocol (Gupta et al., 2022) to evaluate long-horizon forgetting, with additional 40+40 and 70+10 two-step results deferred to Appendix A. We report mAP@50 for previously learned classes (mAP@P), current-task classes (mAP@C), and all seen classes (mAP@A).

4.2 Baseline Construction and Pre-Evaluation

Before evaluating CL-CLIP, we first test whether CLIP-based open-vocabulary detectors can naturally handle continual object detection. We build this diagnostic baseline on F-ViT (Wu et al., 2024) by sequentially fine-tuning its detector head on current-task categories while evaluating old categories through mAP@P. This setting directly measures whether the frozen CLIP prior can preserve previous categories after the shared trainable head is updated on new tasks.

Table 1: Performance of various CLIP models on the 4-tasks MS-COCO benchmark

Model	ViT	**Task1**	**Task2**			**Task3**			**Task4**		
		mAP@C	mAP@C	mAP@P	mAP@A	mAP@C	mAP@P	mAP@A	mAP@C	mAP@P	mAP@A
FineCLIP	ViT-B/16	64.3	44.5	**26.2**	**35.8**	**46.1**	**19.1**	**28.1**	38.4	**17.4**	**22.6**
FG-CLIP		60.3	43.2	19.3	31.8	37.0	15.6	22.7	33.2	13.3	18.3
EVA-CLIP		61.8	43.8	5.0	25.3	39.9	10.5	20.3	35.4	8.5	15.2
FineCLIP	ViT-L/14	73.7	56.2	**29.7**	**43.6**	57.4	**30.8**	**39.7**	50.7	**27.4**	**33.2**
FG-CLIP		72.6	55.4	13.3	35.4	53.4	19.0	30.5	48.3	18.1	25.6
EVA-CLIP		70.0	53.5	15.0	35.2	48.2	16.0	26.7	44.0	14.9	22.2
SigLIP2	So/14	**65.9**	**48.4**	7.3	28.9	45.5	9.2	21.3	**40.4**	8.2	16.3
SigLIP2	So/16	**76.5**	**59.7**	8.6	35.4	**58.7**	13.7	28.7	**55.8**	13.0	23.7

Using this protocol, we benchmark EVA-CLIP(Sun et al., 2023), SigLIP2(Tschannen et al., 2025), FineCLIP(Jing et al., 2024), FG-CLIP(Xie et al., 2025). Tab. 1 shows that strong CLIP representations do not by themselves solve continual detection. Several variants keep high current-task accuracy but lose old-class performance almost completely, indicating that open-vocabulary generalization and continual retention are not equivalent once the detector is sequentially fine-tuned.

FineCLIP provides the most stable retention and the best mAP@A in most stages, so we use it as the default CLIP backbone. Even this strongest baseline still suffers substantial old-class degradation, motivating an explicitly continual architecture rather than a direct reuse of an OVD detector. Appendix A reports the same pattern on additional two-step COCO and VOC settings.

4.3 Implementation Details

We implement CL-CLIP with PyTorch (Paszke et al., 2019) and MMDetection (Chen et al., 2019). Following CAT-Seg (Cho et al., 2024), we set the main hyperparameters of the cost-volume pathway to their default values: $d_F = 128$ is the feature dimension of the cost-aggregation features, $N_B = 2$ is the number of spatial-aggregation (Swin) blocks, and $N_U = 2$ is the number of up-sampling decoder blocks. Note that CL-CLIP removes CAT-Seg's class-aggregation module (Sec. 3.2), so only spatial aggregation and the up-sampling decoder remain in our cost-volume pathway. The total batch size is 16 for PASCAL VOC and 32 for MS-COCO. We use FineCLIP ViT-B/16 as the default frozen backbone. For the COCO 4-task protocol we follow the class split of OW-DETR (Gupta et al., 2022); for VOC we use the standard 10+10, 15+5, and 19+1 splits. Multi-seed variance is reported in Appendix F. Complete optimization settings are deferred to Appendix B.

4.4 Results and Analysis

Evaluating performance on PASCAL VOC. Tab. 2 summarizes the results on three standard two-step VOC splits. The prior methods are OW-DETR (Gupta et al., 2022), ABR (Liu et al., 2023), MMA (Cermelli et al., 2022), NSGP-RePRE (Wu et al., 2025), ROSETTA (Yang et al., 2022), LEA (Song et al., 2025), GCD (Wang et al., 2025), TARL (Zhang et al., 2024), and IODC (He et al., 2023).

Under FineCLIP ViT-B/16, CL-CLIP improves all-class mAP over the corresponding frozen-CLIP baseline by 7.4, 9.6, and 9.2 points on the 10+10, 15+5, and 19+1 splits, respectively. The same advantage persists when both methods use original CLIP ResNet-50, with gains of 5.7, 12.2, and 11.1 points. In particular, the ResNet-50 old-class AP improves by 4.3, 15.2, and 11.2 points. CL-CLIP improves old-class and all-class AP over the corresponding baseline with both encoders.

Among the methods listed in Tab. 2, CL-CLIP with FineCLIP ViT-B/16 achieves the highest all-class mAP on the 10+10 and 15+5 splits (74.7 and 77.4), and the highest new-class AP on all three splits (84.5, 74.5, and 84.8). Compared with LEA (Song et al., 2025), its all-class mAP is higher by

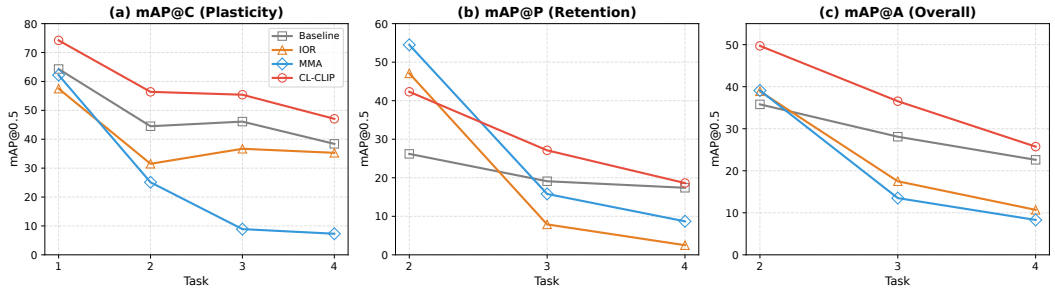

Figure 2: Performance on the 4-task MS-COCO benchmark (mAP@0.5). CL-CLIP maintains stronger plasticity and overall accuracy across the full task sequence compared with reproduced baselines.

2.1 and 0.7 points on 10+10 and 15+5, respectively. LEA achieves higher old-class AP on all three splits and higher all-class mAP on 19+1 (81.8 vs. 77.5). CL-CLIP achieves higher new-class AP on all three splits, including 84.8 vs. 32.2 on 19+1. It also consistently obtains higher new-class and all-class AP than TARL (Zhang et al., 2024) and ROSETTA (Yang et al., 2022).

Table 2: PASCAL VOC results under three two-step continual-learning splits (mAP@0.5). Prior-method results use their original architectures; the bottom block isolates the effect of CL-CLIP through matched CLIP-backbone comparisons under the same training protocol.

Method	**Detector / encoder**	**10+10**			**15+5**			**19+1**		
		1-10	**11-20**	**1-20**	**1-15**	**16-20**	**1-20**	**1-19**	**20**	**1-20**
Prior COD methods										
OW-DETR	ResNet-50	63.5	67.9	65.7	72.2	59.8	69.4	71.1	62	70.2
ABR	ResNet-50	71.2	72.8	72.0	73.0	65.1	71.0	71.0	69.7	70.9
MMA	ResNet-50	69.3	63.9	66.6	73.0	60.5	69.9	71.1	63.4	70.7
NSGP-RePRE	ResNet-50	**75.3**	72.7	74	77.5	61.8	73.6	76.3	69	76
ROSETTA	ResNet-50	67.6	66.0	66.8	71.7	61.6	69.2	69.7	68.3	69.6
LEA	ResNet-50	74.8	70.8	72.6	**84.8**	52.4	76.7	**84.4**	32.2	**81.8**
GCD	Swin-T	67.3	71.4	69.4	69.5	58.8	66.8	69.8	74.6	70.1
TARL	Swin-T	71.2	70.0	70.6	73.6	60.2	70.3	73.2	66.5	72.9
IODC	R50 + CLIP-B/32	66.1	69.2	67.6	73.1	61.2	70.2	72.8	62.5	72.3
Matched CLIP-backbone comparisons										
Baseline	FineCLIP ViT-B/16	55.8	78.8	67.3	65.9	73.4	67.8	68.0	74.3	68.3
CL-CLIP (Ours)		64.8	**84.5**	**74.7**	78.3	**74.5**	**77.4**	77.1	**84.8**	77.5
Baseline	Original CLIP RN50	55.3	70.6	62.9	58.6	63.1	59.7	58.7	60.2	58.8
CL-CLIP (Ours)		59.6	77.6	68.6	73.8	66.4	71.9	69.9	71.1	69.9

Evaluating performance on MS-COCO. The 4-task MS-COCO protocol has not been widely adopted by COD methods, most of which only report two-step COCO results. To examine long-horizon forgetting, we reproduce IOR (An et al., 2025) and MMA (Cermelli et al., 2022) with their standard Faster R-CNN + ResNet-50 setup. The class splits, task order, training schedule, and evaluation protocol are otherwise matched to ours; CL-CLIP and its baseline use FineCLIP ViT-B/16. Fig. 2 plots mAP across the four sequential tasks.

CL-CLIP improves the FineCLIP/F-ViT baseline throughout the stream, raising mAP@A from 35.8 to 49.7 (Task2), 28.1 to 36.5 (Task3), and 22.6 to 25.7 (Task4), while maintaining a more balanced stability-plasticity trade-off than IOR and MMA across the full sequence. Appendix F confirms the same trend under the stricter mAP@[.5:.95] metric.

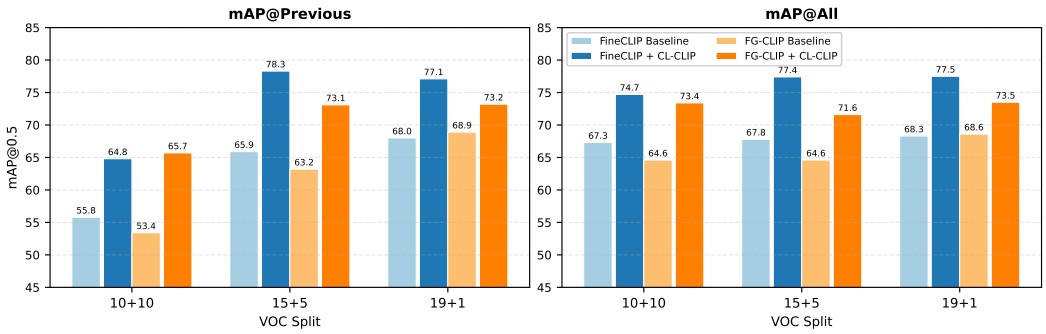

Figure 3: Generalization of CL-CLIP across CLIP backbones on PASCAL VOC (mAP@0.5). Left: previous-task classes; Right: all classes. CL-CLIP consistently improves retention and overall accuracy on both FineCLIP and FG-CLIP.

4.5 Ablation Study

Generalization across CLIP backbones. To verify that CL-CLIP is not tied to a specific CLIP variant, we evaluate it with FG-CLIP (Xie et al., 2025) ViT-B/16 in addition to the default FineCLIP backbone. Fig. 3 shows that CL-CLIP consistently improves both backbones across all three VOC splits, confirming that the category decoupling design generalizes across different CLIP representations.

Core mechanism ablation. Tab. 3 evaluates the two core anti-forgetting mechanisms: the decoupled expert pathway and EWC-style drift regularization. The pathway ablation jointly replaces cost-volume gating and per-category experts with the shared F-ViT head. Replacing this pathway with the shared F-ViT head while retaining drift regularization sharply reduces both plasticity and retention, confirming that drift regularization alone cannot substitute for class-specific pathways. Removing drift regularization preserves plasticity but causes severe retention collapse (mAP@P drops from 42.3 to 14.9 on Task2). The full configuration achieves the highest old-class and all-class AP among the evaluated variants at Tasks 2–4.

Table 3: Core mechanism ablation on the 4-task MS-COCO benchmark: the decoupled expert pathway and drift regularization. Cost-volume gating and per-category experts are ablated jointly because together they instantiate category decoupling.

Components		**Task2**			**Task3**			**Task4**		
Decoupled expert pathway	Drift Reg.	mAP@C	mAP@P	mAP@A	mAP@C	mAP@P	mAP@A	mAP@C	mAP@P	mAP@A
✓	✓	56.4	**42.3**	**49.7**	55.4	**27.1**	**36.5**	47.1	**18.6**	**25.7**
	✓	50.4	20.1	36.0	49.6	20.4	30.2	43.8	15.8	22.8
✓		**56.6**	14.9	36.8	**55.9**	10.6	25.7	**50.2**	3.9	15.4

Secondary architectural choices. After establishing the necessity of the decoupled expert pathway and drift regularization in Tab. 3, we use Tab. 4 to examine two secondary design choices. Reintroducing cross-class aggregation marginally changes current-class AP but severely damages old-class retention: on Task4, mAP@P falls from 18.6 to 5.4 and mAP@A drops from 25.7 to 16.5. This confirms that cross-category coupling is harmful under missing-label supervision.

Removing Spatial-RPN changes all-class mAP from 49.7/36.5/25.7 to 49.3/36.2/26.5 over Tasks 2–4, with both configurations reaching 46.5 average incremental mAP. Category decoupling, per-category experts, and drift regularization remain unchanged in this comparison. Appendix C reports the proposal-recall analysis

Category response overlap. To examine cross-category spatial entanglement, Tab. 5 reports the Response Overlap metric. Without the orthogonality loss, overlap reaches 0.9945 because the trainable aggregation modules, optimized solely by the detection loss, converge toward a category-agnostic foreground pattern that erases the per-class separation present in the raw CLIP similarities. The

Table 4: Analysis of secondary architectural choices on the 4-task MS-COCO benchmark. All variants retain cost-volume category decoupling, per-category convolutional experts, and drift regularization. Average incremental mAP averages the accuracy over all four task endpoints.

Variant	Task1	Task2			Task3			Task4			Avg. Inc. mAP
	mAP@C	mAP@C	mAP@P	mAP@A	mAP@C	mAP@P	mAP@A	mAP@C	mAP@P	mAP@A	
Full CL-CLIP	**74.2**	56.4	**42.3**	**49.7**	55.4	**27.1**	**36.5**	47.1	18.6	25.7	**46.5**
w/ cross-class aggregation	**74.2**	**56.5**	4.1	31.6	**55.5**	10.2	25.3	**49.8**	5.4	16.5	36.9
w/o auxiliary Spatial-RPN	74.0	55.9	42.0	49.3	55.0	26.8	36.2	48.4	**19.2**	**26.5**	**46.5**

orthogonality loss counteracts this by penalizing spatial co-activation, reducing overlap to 0.0017 and improving average old-class AP from 23.5 to 29.3. The class aggregation variant shows moderate overlap (0.0112) but much lower retention (Avg. mAP@P 6.6), confirming that cross-category coupling harms old-class preservation through a different mechanism.

Table 5: Response Overlap and average detection performance over Task2–Task4 of 4-task MS-COCO. Response Overlap is the mean squared off-diagonal Gram value of sigmoid response maps after spatial ℓ_2 normalization (lower is better).

Variant	Avg. mAP@P	Avg. mAP@C	Avg. mAP@A	Response Overlap ↓
w/ class aggregation	6.6	53.9	24.5	0.0112
w/o ortho loss	23.5	53.7	34.4	0.9945
Full CL-CLIP	29.3	53.0	37.3	0.0017

Auxiliary loss ablation. Tab. 11 (Appendix E) studies the auxiliary losses. Drift regularization is critical for FPN stability; combined with the orthogonality loss it yields the best overall accuracy, as the two losses complement each other in stabilizing shared features and maintaining separable cost-volume responses.

Implementation choices. Appendix C compares alternative regularization strategies and expert structures, supporting our use of EWC and convolutional experts.

5 Conclusion

We studied continual object detection from the perspective of CLIP-based open-vocabulary detection. Our baseline analysis shows that the zero-shot advantage of CLIP-based detectors does not automatically translate into continual learning ability. CL-CLIP addresses this by using a CLIP-text similarity cost volume as a zero-shot spatial prior to build class-specific detection pathways, reducing cross-class interference. Experiments on PASCAL VOC and MS-COCO confirm substantially improvements over the F-ViT baseline and competitive performance compared with existing COD methods, suggesting that category decoupling is a promising direction for CLIP-based continual detection.

Limitations. Evaluation covers PASCAL VOC and MS-COCO, with multi-seed results reported for VOC 10+10. Generalization to larger vocabularies and longer task sequences remains untested. Computation grows with the number of seen classes because all seen-category experts participate in each forward pass. Across the four-task COCO sequence, inference FLOPs increase from 289.5 G to 594.7 G and throughput decreases from 19.9 to 8.6 FPS. Freezing old experts limits their parameter updates but does not eliminate their forward computation; detailed cost measurements are provided in Appendix F. The current classifier covers categories introduced in the task stream, leaving detection of categories without trained experts outside the evaluated setting.

AI use statement

We used generative AI tools to assist with writing, literature discovery, and drafting portions of this paper, and code development. Writing assistance included polishing English expression, improving clarity, and refining the organization of the text. For literature discovery, AI tools assisted in identifying potentially relevant prior work for consideration. We also used AI tools to prepare initial

drafts of portions of the manuscript based on author-provided descriptions, experimental results, and revision instructions. In addition, AI tools assisted in writing portions of the code used in this work. All AI-assisted content, including text and code, has been manually reviewed and verified by the authors. The authors retain responsibility for the final manuscript, including the accuracy of the claims and citations and the interpretation of the experimental results.

Ethics statement

This work investigates continual object detection using existing public benchmarks, PASCAL VOC and MS-COCO, and pretrained vision-language models. It does not involve recruiting human participations or collecting new personal data.

Reproducibility statement

To support reproducibility, we describe the model architecture, training objectives, and parameter-freezing strategy in Sec. 3. Sec. 4 specifies the benchmarks, evaluation protocols, baseline construction, and main implementation settings. Appendix B provides optimization hyperparameters, hardware details, and the class splits and task order used in our experiments. Additional ablations are reported in Sec. 4.5 and Appendices C and E. We also report results across three random seeds on the VOC 10+10 split in Tab. 14 to assess variability on a representative setting.

References

Zijia An, Boyu Diao, Libo Huang, Ruiqi Liu, Zhulin An, and Yongjun Xu. Ior: Inversed objects replay for incremental object detection. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2025. doi: 10.1109/ICASSP49660.2025.10888892.

Gaurav Bhatt, James Ross, and Leonid Sigal. Preventing catastrophic forgetting through memory networks in continuous detection. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision – ECCV 2024*, pp. 442–458, Cham, 2025. Springer Nature Switzerland.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (eds.), *Computer Vision – ECCV 2020*, pp. 213–229, Cham, 2020. Springer International Publishing. ISBN 978-3-030-58452-8.

Fabio Cermelli, Antonino Geraci, Dario Fontanel, and Barbara Caputo. Modeling missing annotations for incremental learning in object detection. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 3699–3709, 2022. doi: 10.1109/CVPRW56347.2022.00414.

Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, Zheng Zhang, Dazhi Cheng, Chenchen Zhu, Tianheng Cheng, Qijie Zhao, Buyu Li, Xin Lu, Rui Zhu, Yue Wu, Jifeng Dai, Jingdong Wang, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. Mmdetection: Open mmlab detection toolbox and benchmark. *CoRR*, abs/1906.07155, 2019.

Seokju Cho, Heeseong Shin, Sunghwan Hong, Anurag Arnab, Paul Hongsuck Seo, and Seungryong Kim. Cat-seg: Cost aggregation for open-vocabulary semantic segmentation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4113–4123, 2024. doi: 10.1109/CVPR52733.2024.00394.

Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. The Pascal Visual Object Classes (VOC) Challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010. doi: 10.1007/s11263-009-0275-4.

Tao Feng, Mang Wang, and Hangjie Yuan. Overcoming catastrophic forgetting in incremental object detection via elastic response distillation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9417–9426, 2022. doi: 10.1109/CVPR52688.2022.00921.

Akshita Gupta, Sanath Narayan, K J Joseph, Salman Khan, Fahad Shahbaz Khan, and Mubarak Shah. Ow-detr: Open-world detection transformer. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9225–9234, 2022. doi: 10.1109/CVPR52688.2022.00902.

Yupeng He, Ziyue Huang, Qingjie Liu, and Yunhong Wang. Incremental object detection with CLIP. *CoRR*, abs/2310.08815, 2023.

Dong Jing, Xiaolong He, Yutian Luo, Nanyi Fei, Guoxing Yang, Wei Wei, Huiwen Zhao, and Zhiwu Lu. Fineclip: Self-distilled region-based clip for better fine-grained understanding. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 27896–27918. Curran Associates, Inc., 2024. doi: 10.52202/079017-0875.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017.

Weicheng Kuo, Yin Cui, Xiuye Gu, AJ Piergiovanni, and Anelia Angelova. Open-vocabulary object detection upon frozen vision and language models. In *The Eleventh International Conference on Learning Representations*, 2023.

Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. Grounded language-image pre-training. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10955–10965, 2022. doi: 10.1109/CVPR52688.2022.01069.

Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(12):2935–2947, 2018. doi: 10.1109/TPAMI.2017.2773081.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars (eds.), *Computer Vision – ECCV 2014*, pp. 740–755, Cham, 2014. Springer International Publishing.

Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 936–944, 2017. doi: 10.1109/CVPR.2017.106.

Liyang Liu, Zhanghui Kuang, Yimin Chen, Jing-Hao Xue, Wenming Yang, and Wayne Zhang. Incdet: In defense of elastic weight consolidation for incremental object detection. *IEEE Transactions on Neural Networks and Learning Systems*, 32(6):2306–2319, 2021. doi: 10.1109/TNNLS.2020.3002583.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision – ECCV 2024*, pp. 38–55, Cham, 2025. Springer Nature Switzerland. ISBN 978-3-031-72970-6.

Yuyang Liu, Yang Cong, Dipam Goswami, Xialei Liu, and Joost van de Weijer. Augmented box replay: Overcoming foreground shift for incremental object detection. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11333–11343, 2023. doi: 10.1109/ICCV51070.2023.01044.

Wenlong Luo, Shizhou Zhang, De Cheng, Yinghui Xing, Guoqiang Liang, Peng Wang, and Yanning Zhang. Gradient decomposition and alignment for incremental object detection. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4486–4495, 2025. doi: 10.1109/ICCV51701.2025.00427.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Can Peng, Kun Zhao, and Brian C. Lovell. Faster ilod: Incremental learning for object detectors based on faster rcnn. *Pattern Recognition Letters*, 140:109–115, 2020. ISSN 0167-8655. doi: https://doi.org/10.1016/j.patrec.2020.09.030.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 18–24 Jul 2021.

Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H. Lampert. icarl: Incremental classifier and representation learning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5533–5542, 2017. doi: 10.1109/CVPR.2017.587.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

Xiang Song, Yuhang He, Jingyuan Li, Qiang Wang, and Yihong Gong. Learning endogenous attention for incremental object detection. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 30354–30364, 2025. doi: 10.1109/CVPR52734.2025.02826.

Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. EVA-CLIP: improved training techniques for CLIP at scale. *CoRR*, abs/2303.15389, 2023.

Michael Tschannen, Alexey A. Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier J. Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *CoRR*, abs/2502.14786, 2025.

Xu Wang, Zilei Wang, and Zihan Lin. GCD: advancing vision-language models for incremental object detection via global alignment and correspondence distillation. In Toby Walsh, Julie Shah, and Zico Kolter (eds.), *Thirty-Ninth AAAI Conference on Artificial Intelligence, Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence, Fifteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2025, Philadelphia, PA, USA, February 25 - March 4, 2025*, pp. 8015–8023. AAAI Press, 2025.

Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss (eds.), *Computer Vision – ECCV 2018*, pp. 3–19, Cham, 2018. Springer International Publishing. ISBN 978-3-030-01234-2.

Qirui Wu, Shizhou Zhang, De Cheng, Yinghui Xing, Di Xu, Peng Wang, and Yanning Zhang. Demystifying catastrophic forgetting in two-stage incremental object detector. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, volume 267 of *Proceedings of Machine Learning Research*. PMLR / OpenReview.net, 2025.

Size Wu, Wenwei Zhang, Lumin Xu, Sheng Jin, Xiangtai Li, Wentao Liu, and Chen Change Loy. Clipself: Vision transformer distills itself for open-vocabulary dense prediction. In B. Kim,

Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 52484–52503, 2024.

Yuxin Wu and Kaiming He. Group normalization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

Chunyu Xie, Bin Wang, Fanjing Kong, Jincheng Li, Dawei Liang, Gengshen Zhang, Dawei Leng, and Yuhui Yin. FG-CLIP: fine-grained visual and textual alignment. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, volume 267 of *Proceedings of Machine Learning Research*. PMLR / OpenReview.net, 2025.

Binbin Yang, Xinchi Deng, Han Shi, Changlin Li, Gengwei Zhang, Hang Xu, Shen Zhao, Liang Lin, and Xiaodan Liang. Continual object detection via prototypical task correlation guided gating mechanism. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9245–9254, 2022. doi: 10.1109/CVPR52688.2022.00904.

Dongbao Yang, Yu Zhou, Xiaopeng Hong, Aoting Zhang, Xin Wei, Linchengxi Zeng, Zhi Qiao, and Weipinng Wang. Pseudo object replay and mining for incremental object detection. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 153–162, 2023.

Huahui Yi, Wei Xu, Ziyuan Qin, Xi Chen, Xiaohu Wu, Kang Li, and Qicheng Lao. idpa: Instance decoupled prompt attention for incremental medical object detection. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, volume 267 of *Proceedings of Machine Learning Research*. PMLR / OpenReview.net, 2025.

Li Yin, Juan M Perez-Rua, and Kevin J Liang. Sylph: A hypernetwork framework for incremental few-shot object detection. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9025–9035, 2022. doi: 10.1109/CVPR52688.2022.00883.

Hongquan Zhang, Bin-Bin Gao, Yi Zeng, Xudong Tian, Xin Tan, Zhizhong Zhang, Yanyun Qu, Jun Liu, and Yuan Xie. Learning task-aware language-image representation for class-incremental object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 7096–7104, 2024.

Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, and Jianfeng Gao. Regionclip: Region-based language-image pretraining. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16772–16782, 2022. doi: 10.1109/CVPR52688.2022.01629.

A Baseline Results on Two-Stage Benchmarks

Tab. 6 and Tab. 7 present additional baseline pre-evaluation results across CLIP variants on two-step MS-COCO and PASCAL VOC settings, respectively. We compare FineCLIP (Jing et al., 2024), FG-CLIP (Xie et al., 2025), EVA-CLIP (Sun et al., 2023), and SigLIP2 (Tschannen et al., 2025). These results complement the 4-task MS-COCO analysis in Sec. 4.2 and examine whether the same forgetting pattern holds under shorter continual sequences.

The results are consistent with the main text. FineCLIP generally provides stronger old-class retention and all-class accuracy than the other evaluated CLIP variants, which supports its use as the default backbone for CL-CLIP. However, all variants still show clear degradation on previously learned classes after sequential fine-tuning. This confirms that stronger open-vocabulary representations alone are insufficient for continual object detection.

Table 6: Performance of various CLIP models on the two-stage MS-COCO benchmark

Setting	Model	ViT	**Task1**	**Task2**		
			mAP@C	mAP@C	mAP@P	mAP@A
40+40	FineCLIP	ViT-B/16	57.2	46.9	17.1	32.0
	FG-CLIP		49.4	35.4	16.8	26.1
	EVA-CLIP		52.0	36.6	11.5	24.1
	FineCLIP	ViT-L/14	65.5	56.3	34.2	45.3
	FG-CLIP		62.2	50.7	19.5	35.1
	EVA-CLIP		60.9	46.1	15.7	30.9
	SigLIP2	So/14	55.4	42.9	12.3	27.6
	SigLIP2	So/16	66.1	56.5	14.8	35.7
70+10	FineCLIP	ViT-B/16	52.7	34.8	24.1	25.5
	FG-CLIP		43.7	25.0	21.2	21.7
	EVA-CLIP		46.1	25.9	17.2	18.3
	FineCLIP	ViT-L/14	61.1	46.6	36.9	38.1
	FG-CLIP		57.8	40.6	21.2	21.7
	EVA-CLIP		55.0	37.7	28.4	29.5
	SigLIP2	So/14	51.1	31.7	20.3	21.8
	SigLIP2	So/16	62.6	48.6	26.1	28.9

Table 7: Performance of various CLIP models on the two-stage PASCAL VOC benchmark

Model	ViT	**10+10**				**15+5**				**19+1**			
		Task1	**Task2**			**Task1**	**Task2**			**Task1**	**Task2**		
		mAP@C	mAP@C	mAP@P	mAP@A	mAP@C	mAP@C	mAP@P	mAP@A	mAP@C	mAP@C	mAP@P	mAP@A
FineCLIP	ViT-B/16	80.2	78.8	55.8	67.3	81.4	73.4	65.9	67.8	79.5	74.3	68.0	68.3
FG-CLIP		76.4	75.7	53.4	64.6	77.0	69.0	63.2	64.6	76.4	62.1	68.9	68.6
EVA-CLIP		76.7	78.7	37.2	58.0	79.5	73.6	55.4	59.9	78.5	74.5	64.0	64.6
FineCLIP	ViT-L/14	84.2	82.8	62.0	72.4	84.3	78.6	74.2	75.3	83.4	80.9	73.3	73.6
FG-CLIP		84.1	82.9	53.8	68.4	84.0	76.6	73.1	74.0	83.5	74.6	76.9	76.8
EVA-CLIP		83.0	83.5	39.1	61.3	84.6	80.2	66.1	69.6	83.4	76.9	73.8	74.0
SigLIP2	So/14	81.7	81.9	48.2	65.1	82.1	76.5	60.9	64.8	81.3	78.3	67.8	68.3
SigLIP2	So/16	85.9	84.8	47.5	66.1	86.5	82.2	66.3	70.3	84.6	82.5	70.8	71.3

B Implementation Details

All experiments are conducted on 4× NVIDIA RTX 4090 GPUs. The experimental hyperparameters of our proposed method are shown in Tab. 8. Notably, with the exception of the aggregation module in CAT-Seg, Group Normalization (Wu & He, 2018) with the number of groups set to 32 is universally adopted throughout our detector head.

Task class splits. For the COCO 4-task protocol we follow the class ordering of OW-DETR (Gupta et al., 2022), partitioned into four groups of 19+21+20+20 categories: **Task 1** (1–19): airplane, bicycle, bird, boat, bus, car, cat, cow, dog, horse, motorcycle, sheep, train, elephant, bear, zebra, giraffe, truck, person; **Task 2** (20–40): traffic light, fire hydrant, stop sign, parking meter, bench, chair, dining table, potted plant, backpack, umbrella, handbag, tie, suitcase, microwave, oven, toaster, sink, refrigerator, bed, toilet, couch; **Task 3** (41–60): frisbee, skis, snowboard, sports ball, kite, baseball bat, baseball glove, skateboard, surfboard, tennis racket, banana, apple, sandwich, orange, broccoli, carrot, hot dog, pizza, donut, cake; **Task 4** (61–80): laptop, mouse, remote, keyboard, cell phone, book, clock, vase, scissors, teddy bear, hair drier, toothbrush, wine glass, cup, fork, knife, spoon, bowl, tv, bottle.

For PASCAL VOC, the 20 categories follow the alphabetical order: aeroplane, bicycle, bird, boat, bottle, bus, car, cat, chair, cow, dining table, dog, horse, motorbike, person, potted plant, sheep, sofa, train, tv monitor. The three splits are 10+10, 15+5, and 19+1, partitioned sequentially in this order.

Table 8: Implementation details of CL-CLIP

Parameter	Value
τ	50.0
Optimizer	AdamW
Learning Rate	4×10^{-4}
Weight Decay	0.1
Learning Rate Schedule	Cosine Annealing
Training Epochs	20
Warmup Epochs	4
λ (drift reg. weight)	1000
Norm Cfg	GroupNorm
Backbone	FineCLIP ViT-B/16
Batch Size (VOC / COCO)	16 / 32
Input Resolution	384×384
Random Seed	0

C Additional Ablations and Auxiliary Component Analysis

Tab. 9 provides additional analysis on two implementation choices in CL-CLIP. We compare different regularization strategies for stabilizing the FPN and evaluate alternative expert structures in the Multi-Expert RoI head. We then report a detailed proposal-recall analysis of the auxiliary Spatial-RPN.

For FPN stabilization, we compare EWC (Kirkpatrick et al., 2017) with gradient-space constraint methods including GDA (Luo et al., 2025) and NSGP (Wu et al., 2025). EWC penalizes updates to parameters that are important for previous tasks, while GDA and NSGP modify the optimization trajectory by constraining gradients during backpropagation. The upper part of Tab. 9 shows that EWC is more suitable for our FPN stabilizer. GDA and NSGP preserve high current-class AP, but their old-class AP drops sharply in later tasks, which leads to much lower overall accuracy.

We also compare two expert structures in the Multi-Expert RoI head. The convolutional expert processes the RoI feature map with a compact convolutional residual network, while the residual MLP expert first applies average pooling and then uses a 3-layer residual MLP. The lower part of Tab. 9 shows a clear advantage for convolutional experts. This suggests that preserving local spatial structure inside each RoI is important for class-specific detection.

Table 9: Ablation studies on alternative architectural design choices

Component	Methods	Task1	Task2			Task3			Task4		
		mAP@C	mAP@C	mAP@P	mAP@A	mAP@C	mAP@P	mAP@A	mAP@C	mAP@P	mAP@A
Regularization	EWC	74.2	56.4	42.3	49.7	55.4	27.1	36.5	47.1	18.6	25.7
	GDA	74.2	55.5	1.2	29.7	55.2	2.1	19.8	49.5	2.6	14.4
	NSGP	74.2	57.2	9.0	34.3	56.3	5.2	22.2	49.8	1.7	13.8
Expert Type	Conv Expert	74.2	56.4	42.3	49.7	55.4	27.1	36.5	47.1	18.6	25.7
	Residual MLP Expert	70.2	51.1	4.1	28.8	49.8	6.3	20.8	44.0	3.2	13.4

C.1 Auxiliary Spatial-RPN Analysis

Tab. 10 compares a standard RPN, CBAM-style spatial attention (Woo et al., 2018), direct cost-volume modulation, and our residual-gated cost-volume prior on VOC 10+10. Direct cost-volume modulation obtains the highest proposal recall, suggesting that the cost-volume prior improves object coverage, but its detection accuracy is lower than the residual-gated design. The residual-gated variant improves all-class mAP by 0.6 points over the standard RPN while retaining competitive recall. Together with the near-identical COCO results with and without Spatial-RPN in Tab. 4, this confirms that the module is a minor, dataset-dependent refinement rather than a source of the main continual-learning gains.

Table 10: Proposal recall analysis for the auxiliary Spatial-RPN on the PASCAL VOC 10+10 setting. R@100 is computed at IoU 0.5 using the top-100 RPN proposals per image.

RPN Variant	mAP@P	mAP@A	R@100-P	R@100-C	R@100-A
Standard RPN	62.7	74.1	0.320	**0.571**	0.475
CBAM-style spatial attention	63.3	74.4	0.327	0.570	0.478
Direct cost-volume modulation	63.4	74.2	**0.397**	0.562	**0.499**
Residual-gated cost-volume prior	**64.8**	**74.7**	0.356	0.567	0.487

D Feature-Space Visualization

To qualitatively examine how the representation space changes after continual fine-tuning, we visualize the RoI features and text embeddings with t-SNE in Fig. 4. Original FineCLIP already shows moderate semantic organization, suggesting that the frozen CLIP representation provides a useful prior but insufficient separation for continual detection. After F-ViT fine-tuning, RoI features become more dispersed and entangled, and text embeddings move away from most visual clusters. In contrast, CL-CLIP forms more compact class-wise RoI clusters with clearer separation, consistent with the intended effect of per-class experts and frozen old-class experts.

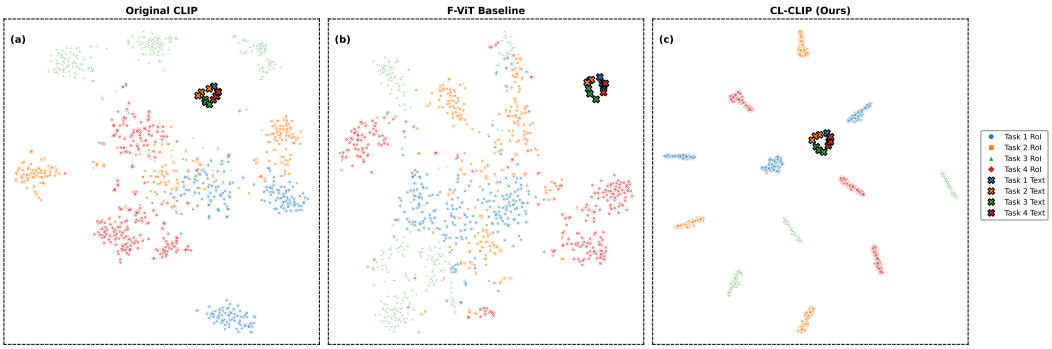

Figure 4: t-SNE visualization of RoI features and CLIP text embeddings after the 4-task MS-COCO sequence. Compared with original FineCLIP and the F-ViT baseline, CL-CLIP forms more compact and better separated RoI feature clusters, suggesting a more structured RoI-text feature space.

E Auxiliary Loss Ablation

Tab. 11 reports the full loss component ablation on the 4-task MS-COCO benchmark.

Table 11: Ablation study of loss components on the 4-task MS-COCO benchmark

Losses	Task1	Task2			Task3			Task4		
	mAP@C	mAP@C	mAP@P	mAP@A	mAP@C	mAP@P	mAP@A	mAP@C	mAP@P	mAP@A
cls+reg	**74.6**	**57.6**	5.9	33.0	**56.4**	4.7	21.9	49.9	4.4	15.8
cls+reg+drift	**74.6**	56.1	29.1	43.3	55.5	23.2	34.0	49.5	18.2	**26.0**
cls+reg+ortho	74.2	56.6	14.9	36.8	55.9	10.6	25.7	**50.2**	3.9	15.4
cls+reg+ortho+drift	74.2	56.4	**42.3**	**49.7**	55.4	**27.1**	**36.5**	47.1	**18.6**	25.7

F Additional Experiments

Stricter evaluation metric. To verify that the improvements are not limited to mAP@0.5, we report the stricter COCO-standard mAP@[.5:.95] under the same 4-task protocol. Tab. 12 shows that CL-CLIP consistently outperforms both IOR (An et al., 2025) and MMA (Cermelli et al., 2022) on

overall accuracy (mAP@A) at every stage, confirming that the gains transfer to the more demanding localization threshold.

Table 12: 4-task MS-COCO results under the strict mAP@[.5:.95] metric

Method	Task1	Task2			Task3			Task4		
	mAP@C	mAP@C	mAP@P	mAP@A	mAP@C	mAP@P	mAP@A	mAP@C	mAP@P	mAP@A
MMA	37.5	12.9	31.7	21.9	4.3	8.2	6.9	3.5	4.1	3.9
IOR	38.7	20.5	30.5	25.3	23.8	4.8	11.1	24.0	1.4	7.0
CL-CLIP (Ours)	**48.1**	**35.6**	**17.3**	**26.9**	**32.7**	**13.4**	**19.8**	**29.1**	**9.3**	**14.3**

Efficiency analysis. Fig. 5 compares the inference cost of CL-CLIP and the F-ViT baseline on the VOC 10+10 setting. CL-CLIP has more total parameters (177.48M→179.50M vs. baseline 107.26M) due to the per-category expert heads, with a slight increase as new experts are added. However, CL-CLIP achieves *lower* FLOPs than the baseline (244.45 G vs. 316.24 G at Task 1) because the cost-volume pathway is computationally lighter. As the number of seen categories grows, FLOPs increase (294.48 G at Task 2) but remain below the baseline in this 20-class setting. On larger-scale benchmarks with more categories (e.g., 80-class COCO), FLOPs grow proportionally with the number of seen-category experts, which could be addressed in future work through expert sharing or pruning strategies.

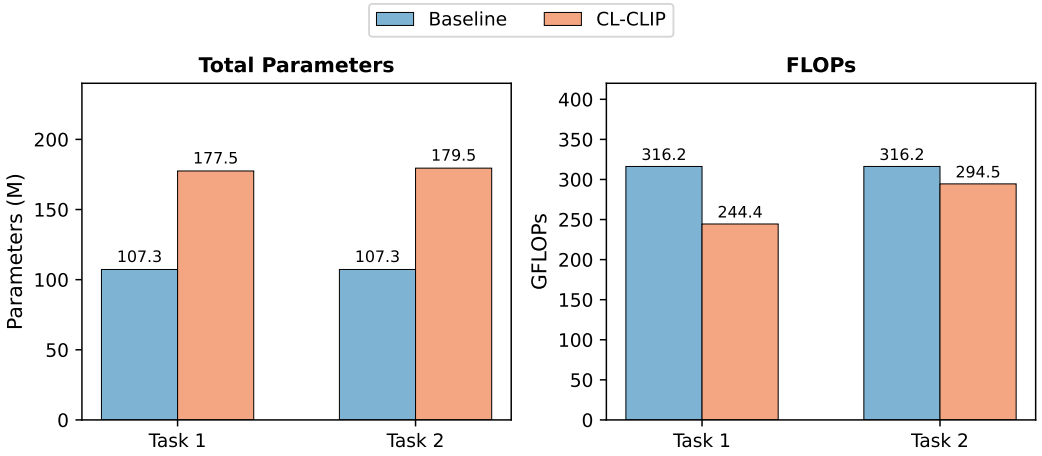

Figure 5: Inference efficiency on VOC 10+10. Params counts parameters activated during the forward pass. CL-CLIP achieves lower FLOPs than the baseline due to the lighter cost-volume pathway, with FLOPs growing moderately as more category experts become active.

Cost growth over the full COCO sequence. To quantify how the cost scales as the vocabulary grows, Tab. 13 reports parameters, trainable parameters, FLOPs, latency, FPS, and peak memory across the full 4-task COCO sequence (19→80 classes). The growth follows a simple linear model $\mathrm{Cost}(C) = \mathrm{Cost}_{\mathrm{fixed}} + C \cdot \mathrm{cost}_{\mathrm{expert}}$: each per-category expert adds only 0.20M parameters (∼0.1% of the model) and 5.0G FLOPs, while the fixed cost-volume pathway (∼68M) dominates the parameter count. Two points follow. First, *trainable* parameters stay essentially constant (∼116M) because only current-task experts are optimized and old experts are frozen. This bounds the size of the optimizer state, but it does *not* keep the training compute constant: at every iteration both the cost-volume pathway and the expert pipeline operate over all seen categories $C_p \cup C_c$, so each frozen old-class expert still contributes forward activations (and gradients through the shared features it feeds), and the per-task training FLOPs, memory, and wall-clock time therefore grow with the number of seen classes, following the same linear trend as the inference cost in Tab. 13. Second, at small class counts CL-CLIP is even cheaper than the F-ViT baseline in FLOPs (see Fig. 5), and since experts operate on spatially-gated features, the cost per image scales with the categories actually present rather than the full vocabulary. This linear growth in both training and inference is the

main scalability limitation of the design; we note expert sharing, grouping, and cost-volume-gated routing as natural mitigations for large-vocabulary deployment.

Table 13: Inference cost of CL-CLIP across the 4-task MS-COCO sequence (19→80 classes).

Task (classes)	Params	Trainable	FLOPs	Latency (ms)	FPS	Peak Mem
Task1 (19)	179.29M	115.87M	289.48G	50.36	19.86	1884MB
Task2 (40)	183.54M	116.27M	394.55G	66.67	15.00	2956MB
Task3 (60)	187.58M	116.07M	494.62G	95.08	10.52	3894MB
Task4 (80)	191.62M	116.07M	594.69G	115.84	8.63	4899MB

Multi-seed variance. To confirm reproducibility, we run CL-CLIP on the VOC 10+10 split with three random seeds. Tab. 14 reports per-seed results and the mean±std. The low variance indicates that our method is stable across different initializations.

Table 14: Multi-seed results on PASCAL VOC 10+10 (mAP@0.5)

Seed	Task1	Task2		
	mAP@C	mAP@C	mAP@P	mAP@A
0	86.2	84.5	64.8	74.7
1	86.4	84.8	63.7	74.3
2	86.4	84.7	64.9	74.8
Mean±Std	86.3±0.1	84.7±0.2	64.5±0.7	74.6±0.3