CoOVD: Continual Open-Vocabulary Object Detection with Novel Concept Injection

Yupeng Zhang[1,2] **Ruize Han**[3] **Yuzhong Feng**[1] **Zixin Ren**[1] **Yuntong Tian**[4,1] **Liang Wan**[1,2]*

[1]College of Intelligence and Computing, Tianjin University.
[2]Key Research Center for Surface Monitoring and Analysis of Relics, State Administration of Cultural Heritage.
[3]Faculty of Computer Science and Artificial Intelligence, Shenzhen University of Advanced Technology.
[4] School of Computer Science and Software Engineering, Southwest Petroleum University

{zhangyupeng, yzfeng, renzixin, lwan}@tju.edu.cn
hanruize@suat-sz.edu.cn, tianyuntong@swpu.edu.cn

Abstract

Open-vocabulary object detection (OVD) has made significant progress, enabling detectors to generalize from seen to unseen categories. However, real-world category spaces continually evolve, and existing OVD models still struggle with newly emerging concepts, while repeated full retraining is prohibitively expensive. To this end, we introduce a new task setting, termed Continual OVD with Novel Concept Injection (CoOVD), where models sequentially learn incoming novel concept groups while preserving prior concepts and original open-vocabulary knowledge, along with a new benchmark, Novel-114. Our key observation is that pretrained visual encoders often already perceive and represent many novel concepts, and the main bottleneck lies in the lack of stable semantic alignment between visual representations and textual concepts. Based on this, we propose NoInDet, an efficient continual injection framework without additional parameters. NoInDet freezes the visual encoder, preserves the text representation space using only texts of common concepts and previously injected concepts, and injects novel concepts by updating only a small subset of text-branch parameters beneficial to novel concept learning. Extensive experiments show that NoInDet effectively learns novel concepts, preserves old knowledge, and consistently outperforms existing continual learning methods for VLMs without introducing additional parameters. Novel-114 and NoInDet are available on Hugging Face.

1 Introduction

In recent years, open-vocabulary object detection (OVD) has extended object detection beyond fixed category sets by leveraging the textual semantic space learned from large-scale vision-language pre-training and the ability to recombine known semantic concepts. However, real-world category spaces evolve continuously. New concepts emerge across scenarios, domains, and tasks, including long-tail categories with scarce data, specialized terms, and domain-specific objects. While existing OVD methods can generalize to conventional unseen categories inferred from known semantics, they still struggle to establish reliable vision-language alignment for **genuinely novel concepts**, leading to degraded detection performance. When novel concepts emerge continually, full retraining can absorb new knowledge and preserve existing capabilities, but requires large-scale historical image-text data

*Corresponding author.

and prohibitive training costs, which further grow as concepts accumulate. Therefore, enabling open-vocabulary detectors to continually learn novel concepts without repeated full retraining, while preserving both learned concepts and original open-vocabulary knowledge, becomes an important and practical problem.

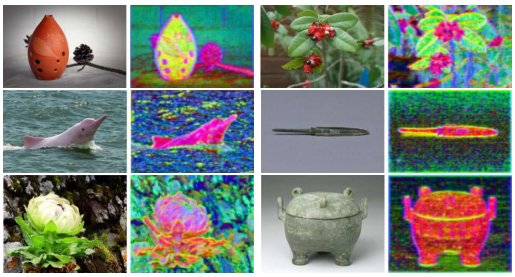


Figure 1: Visualization of features from the pretrained visual backbone. Backbone outputs are projected to RGB space using PCA. Although these concepts are poorly recognized by the initial detector, the backbone highlights object regions and produces coherent responses within semantic areas.

Based on the above observations, we introduce a new task, **Continual Open-Vocabulary Object Detection with Novel Concept Injection (CoOVD)**, which aims to enable open-vocabulary detectors to continually absorb novel concepts after deployment in a parameter-efficient and progressive manner, without repeated full retraining. In this setting, the model sequentially learns incoming groups of novel concepts and is evaluated at each stage on current concept acquisition, previously injected concept retention, and preservation of original open-vocabulary knowledge. Unlike ZiRa [1], which focuses on incremental adaptation to different downstream tasks while preserving zero-shot generalization, CoOVD emphasizes concept-level continual learning, aiming to inject novel concepts while retaining both previously injected concepts and the original open-vocabulary knowledge. Moreover, existing continual learning benchmarks for VLM-based object detection are insufficient for evaluating genuine novel concept injection. For example, ODinW-13 [2], used in ZiRa [1], mainly introduces common categories already likely covered by large-scale vision-language pre-training. As a result, gains on such benchmarks may reflect better use of existing semantic priors rather than true acquisition of novel concepts. To address this, we construct **Novel-114**, a benchmark with 114 more challenging novel concepts organized into 7 scenario-based continual injection groups, enabling systematic evaluation of both novel concept acquisition and retention of previously injected concepts.

Although continual learning for VLMs has been studied, existing methods still have clear limitations. 1) Prompt-based methods [3] often store novel knowledge in external prompts rather than truly teaching the model novel concepts, and require these prompts to be preserved and managed as concepts accumulate. 2) Adapter-based methods [4, 5] and mixture-of-experts methods [6] improve adaptation, but introduce extra parameters and continuously enlarge the model. 3) Parameter-efficient fine-tuning [7] and reparameterization [1] may also disrupt the original semantic space during novel concept injection, leading to unstable old knowledge representations. **Existing methods rely on external modules or parameter adaptation for short-term adjustment, rather than truly and stably internalizing novel concepts into the model, making it difficult to balance novel concept learning and old knowledge retention.**

More importantly, as shown in Fig. 1, pretrained visual encoders exhibit strong perception and representation ability for many novel concepts. Since novel objects are often composed of familiar visual primitives, such as colors, materials, textures, shapes, and local structures, the encoder can map them into a reasonable visual representation space even if they are not explicitly seen during training, as illustrated in Fig. 2. Therefore, the key limitation is not the lack of visual perception of novel objects, but the absence of stable and accurate semantic alignment between their visual representations and corresponding textual concepts. **In other words, the model can 'see' novel objects but has not learned their 'names' in human semantics.** This observation further suggests that the key to continual novel concept injection is not to reshape visual repre-

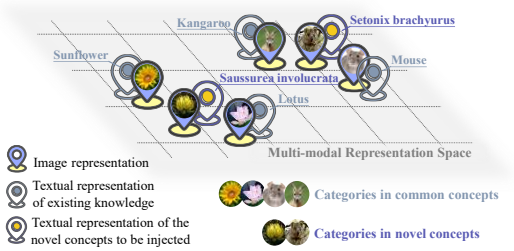


Figure 2: Illustration of the multi-modal representation space. After large-scale pre-training, the visual encoder of a VLM can represent visual cues such as color, material, texture, shape, and local structure, and map novel concepts to reasonable positions in the multi-modal space based on their correlations. For example, Setonix brachyurus shares similar fur color, texture, and overall appearance with a kangaroo, while its frontal local shape is closer to that of a mouse; therefore, its visual representation is closer to those of kangaroos and mice.

sentations [4] or jointly update both vision and language branches [1, 6, 8, 7], as in existing VLMs continual learning methods. Instead, it should establish semantic bindings between novel concepts and their corresponding visual representations from the textual side, while keeping the visual space as stable as possible. This preserves the general perception and representation ability acquired through large-scale pre-training, avoids unnecessary perturbation to complex visual patterns, and improves the stability and interpretability of novel concept learning.

Motivated by the above, we propose NoInDet, an efficient continual novel concept injection framework for OVD. NoInDet introduces no extra parameters and freezes the visual encoder, injecting novel concepts by only fine-tuning key parameters in the text branch, while using textual constraints from common concepts and previously injected concepts to maintain a stable representation space. Specifically, NoInDet first learns current novel concepts through contrastive learning on their corresponding region-text pairs, providing basic absorption and alignment of new knowledge. On this basis, to preserve the stability of the textual representation space, we design **Representation Space Stability Distillation (RSSD)**, which uses the frozen textual branch from the preceding task and performs both absolute semantic distillation and relative relational distillation on randomly sampled common concepts and stored previously injected concepts. Since text is easy to obtain and store, this strategy provides continual old-knowledge regularization at low cost. Meanwhile, to mitigate the conflict between novel concept learning and old knowledge retention, we introduce **Knowledge-aware Parameter Decoupling (KPD)**, which suppresses excessive drift of parameters sensitive to old knowledge and concentrates updates on parameters more beneficial to novel concept learning. Unlike existing continual learning methods for VLMs that rely on historical image-text data, separate modules, external prompts, or parameter adapters to carry new knowledge, NoInDet emphasizes stably injecting novel concepts into the existing semantic space without expanding the model architecture.

Our contributions are summarized as follows:

- We introduce CoOVD, a continual novel concept injection task for OVD, and construct Novel-114 to evaluate concept acquisition and retention of previously injected concepts.
- We propose NoInDet, an efficient continual injection framework without additional parameters. By freezing the visual encoder, updating only key textual parameters, and using RSSD and KPD to preserve textual representation stability and mitigate novel-old knowledge conflicts, NoInDet enables injection of novel concepts into the existing semantic space.
- Experiments on Novel-114 and MS COCO show that NoInDet effectively improves novel concept learning and outperforms existing continual learning methods for VLMs in retaining both previously injected concepts and original open-vocabulary knowledge.

2 Related Works

Open-vocabulary object detection (OVD). OVD [9] extends detection to categories lacking bounding-box supervision during detector training, leveraging semantic representations from VLMs such as CLIP [10] and ALIGN [11]. ViLD [12], OADP [13], and DK-DETR [14] transfer VLM knowledge through distillation. Detic [15] and OCO [16] exploit image-level supervision and object mining, while SAS-Det [17], DST-Det [18], and OV-DQUO [19] use pseudo labels or unknown-object supervision. ProxyDet [20] synthesizes proxy novel classes through classwise mixup, whereas DetPro [21], LBP [22], and CORA [23] explore prompt learning and region adaptation. F-VLM [24] exploits frozen VLM features, while CLIPSelf [25] and DeCLIP [26] enhance dense representations for region recognition. Further developments include region-aware pretraining in RO-ViT [27] and text-conditioned detection in GLIP [28], GroundingDINO [29], LLMDet [30], YOLO-World [31], and YOLOE [32]. CLIFF [33] models diffusion-based transfer among object, image, and text latent spaces; its notion of "continual" concerns latent-space transfer rather than sequential task learning. We study continual concept injection into LLMDet while preserving its original OVD capability.

Continual learning for VLMs. Continual adaptation of VLMs requires retaining learned task knowledge and pretrained zero-shot capabilities. ZSCL [34] combines reference-data distillation with weight averaging. DIKI [3] integrates residual knowledge into a frozen backbone with distribution-aware calibration. LADA [5] learns label-specific memory units after a frozen image encoder, while MoE-Adapters [6] combines expert adapters with distribution-based selection between adapted and original predictions. C-CLIP [8] combines LoRA with contrastive knowledge consolidation,

whereas KeepLoRA [7] restricts updates through residual gradient adaptation. NoInDet addresses this balance in OVD by updating a fixed subset of shared textual parameters while freezing the visual encoder. It couples textual representation preservation with parameter-update control, without task- or class-specific adaptation modules at inference.

Continual object detection and knowledge preservation. Incremental object detection learns categories sequentially while retaining previously learned ones. CL-DETR [35] combines detector distillation with distribution-preserving exemplar replay, while RICO [36] studies domain shifts with fixed classes and joint class–domain increments. SGVF [4] uses parameter fusion for class-incremental adaptation, ZiRa [1] employs reparameterizable branches, and DitHub [37] maintains class-specific LoRA modules selected and merged according to textual queries. ZiRa and DitHub address incremental vision-language object detection (IVLOD), considering downstream adaptation and zero-shot preservation. Within this broader line, CoOVD focuses on injecting disjoint concept groups poorly recognized by the initial detector, tracking current-concept acquisition, historical-concept retention, and original OVD preservation. Novel-114 emphasizes concept acquisition while reducing confounding effects from extreme imaging conditions. Our protocol permits historical concept texts but excludes historical image replay, and NoInDet adapts shared textual parameters without an expanding expert library. Its preservation mechanisms build on classical continual learning principles but differ in their signals and constraints. LwF [38] preserves old-task outputs on current-task images; RSSD instead uses common-concept and historical-concept texts to align individual representations and pairwise semantic relations at the text-encoder and projector levels. EWC [39] uses past-task Fisher information to impose soft quadratic parameter penalties; KPD instead derives sensitivity from probe gradients of RSSD's absolute semantic distillation terms and directly masks parameter-element training gradients. RSSD and KPD share a textual preservation signal, coupling representation stability with sensitivity-guided update control for continual concept injection.

3 Novel-114: A Benchmark for CoOVD

Motivation. Existing continual learning benchmarks for detection VLMs are insufficient for evaluating continual novel concept injection in OVD. For example, ODinW-13 [2], used in ZiRa [1], mainly introduces common concepts covered by large-scale vision-language pre-training, and contains clear domain shifts such as infrared, grayscale, and underwater imagery. This matches ZiRa's focus on gradual adaptation to specialized scenarios, tasks, or domains. Therefore, gains on such benchmarks may reflect better use of existing semantic priors or domain adaptation, rather than genuine novel concept learning. Similarly, although LVIS [40] contains rare categories, their number is limited, and many are supplementary annotations built on COCO, often involving small objects, severe occlusion, and sparse instances. Thus, their difficulty may arise more from perceptual challenges than from concept learning. To this end, we construct Novel-114, a new benchmark for CoOVD. Unlike existing benchmarks, Novel-114 focuses on concepts insufficiently covered by vision-language pre-training corpora and common detection benchmarks, enabling stricter evaluation of whether models acquire new conceptual knowledge rather than transfer from existing semantic priors.

Design principles. In constructing Novel-114, we follow five principles: ① ***Novelty*****:** Concepts are not well covered by common detection benchmarks or large-scale pre-training corpora. ② ***Learnability*****:** Extreme small objects, severe occlusion, and harsh imaging conditions are minimized to focus evaluation on concept learning. ③ ***Continuity*****:** Concepts support multi-stage incremental injection to evaluate both knowledge acquisition and retention. ④ ***Diversity*****:** Groups show large scenario gaps and varied concept numbers, enabling evaluation under diverse learning scenarios. ⑤ ***Annotation reliability*****:** Instances are carefully annotated and manually verified to ensure clear definitions, accurate boxes, and consistent train/test labels. Based on these principles, we collect images from multiple domains and use CLIP [10] and LLMDet [30] to screen candidate categories, filtering out concepts already well covered by existing VLMs and OVD models. We construct Novel-114 with 114 categories, organized into 7 continual learning groups by scenario and semantic relevance, each corresponding to one learning stage. The group names indicate the number of concepts, e.g., Animals-25 and Clothing-6 contain 25 animal and 6 clothing concepts, respectively. We adopt a fixed heterogeneous injection order: Animals-25, Clothing-6, Cold Weapons-4, Food-28, Instruments-8, Plants-25, and Vessels-18. This order is not based on category number, visual difficulty, or semantic similarity, but alternates across natural objects, man-made objects, specialized concepts, and daily-life concepts, with group sizes ranging from 4 to 28. This design simulates realistic

scenarios where novel concepts arrive from diverse domains and batch sizes, enabling evaluation under non-curriculum, cross-scenario, and non-uniform continual learning settings.

Statistics. As shown in Table 1, Novel-114 contains 114 novel concepts organized into 7 continual learning tasks. Each task is split into training and test sets, yielding 4,424 training images with 9,293 annotations and 4,467 test images with 9,405 annotations in total. The tasks vary substantially in concept count, image scale, and object instances, forming a continual benchmark with a clearly non-uniform task distribution. Unlike conventional static detection benchmarks, Novel-114 is naturally organized for continual learning: the model learns novel concepts task by task, and is evaluated at each stage on both current concept acquisition and retention of previously injected concepts. Therefore, Novel-114 provides a more targeted and challenging benchmark for continual novel concept injection, enabling comprehensive evaluation of the balance between novel concept learning and old knowledge retention. ***More details of Novel-114 are in Appendix B.***

Table 1: Statistics of each task in the Novel-114.

Task	# Concepts	Subset	# Images	# Objects
Animals-25	25	train	1049	1334
		test	1057	1360
Clothing-6	6	train	184	211
		test	187	231
Cold Weapons-4	4	train	150	272
		test	151	247
Food-28	28	train	1407	3093
		test	1420	3188
Instruments-8	8	train	134	170
		test	139	250
Plants-25	25	train	873	3456
		test	882	3370
Vessels-18	18	train	627	757
		test	631	759
All Categories	**114**	**train**	**4424**	**9293**
		test	**4467**	**9405**

4 The Proposed Baseline Method

4.1 Problem Formulation

We first formalize the proposed **CoOVD** problem. At initialization, we are given an open-vocabulary detector f_{θ_0}. During continual learning, the model sequentially learns a series of novel concept sets, denoted by $\{\mathcal{C}^t\}_{t=1}^{T}$, where T is the total number of continual learning tasks, $\mathcal{C}^t$ denotes the set of novel concepts introduced at task t, and $\mathcal{C}^i \cap \mathcal{C}^j = \varnothing$ for any $i \neq j$. For each task $t \in \{1, \ldots, T\}$, the corresponding training set is defined as $\mathcal{D}^t = \{(\mathbf{I}_i^t, L_i^t)\}_{i=1}^{N_t}$, where N_t is the number of training images in task t, $\mathbf{I}_i^t$ denotes the i-th input image, and L_i^t denotes its detection annotations. Specifically, $L_i^t = \{(\mathbf{b}_{i,k}^t, \mathbf{c}_{i,k}^t)\}_{k=1}^{K_i}$, where K_i is the number of annotated objects in $\mathbf{I}_i^t$, $\mathbf{b}_{i,k}^t$ denotes the bounding box of the k-th object, and $\mathbf{c}_{i,k}^t$ denotes its concept label, with $\mathbf{c}_{i,k}^t \in \mathcal{C}^t$. At each task, the model can access only the current-task data $\mathcal{D}^t$, while image data from all previous tasks are unavailable.

During evaluation, the updated detector f_{θ_t} is required to detect objects from three sets: the current novel concepts $\mathcal{C}^t$, the previously learned concepts $\mathcal{C}^{\text{his},t} = \bigcup_{k=1}^{t-1} \mathcal{C}^k$, and the original open-vocabulary concept set $\mathcal{C}^{\text{ovd}}$, where $\mathcal{C}^{\text{his},t}$ denotes all novel categories introduced before task t, and $\mathcal{C}^{\text{ovd}}$ denotes the concepts space supported by the initial detector. Therefore, CoOVD requires the model to continually incorporate novel concepts while preserving detection performance on both previously learned concepts and the original open-vocabulary categories.

4.2 Overview of the Method

Our goal is to continually inject novel concepts into an open-vocabulary detector LLMDet in a parameter-efficient and stable manner. As shown in Fig. 3, NoInDet freezes the visual encoder of a pre-trained open-vocabulary detector and updates only a small subset of key textual parameters. For each task, it learns current concepts via contrastive learning on region-text pairs, while using sampled common concepts and stored historical concept texts as lightweight old-knowledge supervision. We introduce **Representation Space Stability Distillation (RSSD)** to preserve semantic consistency and structural stability in the textual embedding space, and **Knowledge-aware Parameter Decoupling (KPD)** to suppress drift in old-knowledge-sensitive parameters. Thus, NoInDet enables continual novel concept injection without additional parameters or historical image replay.

4.3 Representation Space Stability Distillation (RSSD)

In continual novel concept learning, new-task training can shift the textual representation space, causing old concept forgetting and weakening the open-vocabulary foundation. To address this, we propose **RSSD**, which preserves textual-space stability via lightweight stored textual knowledge distillation without historical images.

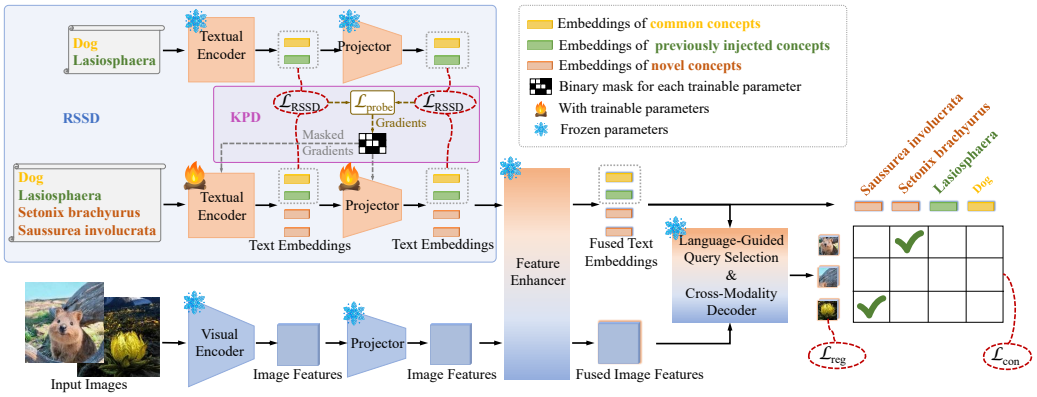


Figure 3: Overview of NoInDet. Built on LLMDet, NoInDet freezes the visual encoder and updates only a small subset of textual parameters for continual novel concept injection. At each task, it learns current concepts via region-text contrastive learning, while randomly sampled common concepts and stored historical concepts provide textual knowledge for stability regularization. **Representation Space Stability Distillation (RSSD)** stabilizes representations, and **Knowledge-aware Parameter Decoupling (KPD)** guides parameter selection, enabling continual concept learning without extra parameters or historical image replay.

Lightweight textual old knowledge. At task t, we construct an old-knowledge concept set $\mathcal{C}^t_{\text{old}} = \mathcal{C}^t_{\text{com}} \cup \mathcal{C}^t_{\text{his}}$, where $\mathcal{C}^t_{\text{com}}$ denotes a randomly sampled set of common concepts, and $\mathcal{C}^t_{\text{his}} = \bigcup_{k=1}^{t-1} \mathcal{C}^k$ denotes the set of all concepts learned before task t. Here, $\mathcal{C}^k$ is the set of novel concepts introduced at task k. We further denote the current novel concept set by $\mathcal{C}^t_{\text{novel}} = \mathcal{C}^t$. Accordingly, the student text branch receives $\mathcal{C}^t_{\text{stu}} = \mathcal{C}^t_{\text{old}} \cup \mathcal{C}^t_{\text{novel}}$, while the teacher text branch only takes the old-knowledge texts: $\mathcal{C}^t_{\text{tea}} = \mathcal{C}^t_{\text{old}}$. Since common concepts are easy to obtain and historical concept are cheap to store, this design provides effective old-knowledge supervision without any historical image replay.

Teacher-student textual representations. We use the text branch (frozen) from the previous task as the teacher and the current text branch (trainable) as the student. Let $E_{\text{tea}}(\cdot)$ and $P_{\text{tea}}(\cdot)$ denote the teacher textual encoder and projector, respectively, and let $E_{\text{stu}}(\cdot)$ and $P_{\text{stu}}(\cdot)$ denote the corresponding student modules. For the teacher and student branches, old-knowledge texts are encoded as

$$\mathbf{E}^t_{\text{tea}} = E_{\text{tea}}(\mathcal{C}^t_{\text{old}}), \quad \mathbf{P}^t_{\text{tea}} = P_{\text{tea}}(\mathbf{E}^t_{\text{tea}}), \quad \mathbf{E}^t_{\text{stu,old}} = E_{\text{stu}}(\mathcal{C}^t_{\text{old}}), \quad \mathbf{P}^t_{\text{stu,old}} = P_{\text{stu}}(\mathbf{E}^t_{\text{stu,old}}), \tag{1}$$

where $\mathbf{E}^t_{\text{tea}}$ and $\mathbf{E}^t_{\text{stu,old}}$ denote the teacher and student encoder-level text embeddings, respectively, and $\mathbf{P}^t_{\text{tea}}$ and $\mathbf{P}^t_{\text{stu,old}}$ denote their projected embeddings for old-knowledge concepts.

Dual distillation for representation space stability. RSSD preserves textual representation stability by imposing two complementary constraints on old-knowledge texts: absolute semantic distillation and relative relation distillation. The absolute semantic distillation aligns student and teacher representations of the same old concepts at both encoder and projector levels:

$$\mathcal{L}^{E,t}_{\text{abs}} = 1 - \text{Cos}\left(\mathbf{E}^t_{\text{stu,old}}, \mathbf{E}^t_{\text{tea}}\right), \qquad \mathcal{L}^{P,t}_{\text{abs}} = 1 - \text{Cos}\left(\mathbf{P}^t_{\text{stu,old}}, \mathbf{P}^t_{\text{tea}}\right), \tag{2}$$

where $\text{Cos}(\cdot,\cdot)$ denotes average cosine similarity between corresponding student and teacher features.

To preserve old textual-space structure, we impose relative relation distillation. For a feature set $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N]^\top$, its pairwise semantic relation matrix is defined as $\mathbf{R}(\mathbf{X}) = \left[\frac{\mathbf{x}_i^\top \mathbf{x}_j}{\|\mathbf{x}_i\|_2 \, \|\mathbf{x}_j\|_2}\right]_{i,j=1}^{N}$, where N is the number of concept embeddings in $\mathbf{X}$, and i, j index corresponding embeddings.

Accordingly, the teacher and student relation matrices are computed as

$$\mathbf{R}^{E,t}_{\text{tea}} = \mathbf{R}(\mathbf{E}^t_{\text{tea}}), \quad \mathbf{R}^{E,t}_{\text{stu}} = \mathbf{R}(\mathbf{E}^t_{\text{stu,old}}), \quad \mathbf{R}^{P,t}_{\text{tea}} = \mathbf{R}(\mathbf{P}^t_{\text{tea}}), \quad \mathbf{R}^{P,t}_{\text{stu}} = \mathbf{R}(\mathbf{P}^t_{\text{stu,old}}). \tag{3}$$

Then, the relation distillation losses are defined as

$$\mathcal{L}^{E,t}_{\text{rel}} = \text{SmoothL1}\left(\mathbf{R}^{E,t}_{\text{stu}}, \mathbf{R}^{E,t}_{\text{tea}}\right), \qquad \mathcal{L}^{P,t}_{\text{rel}} = \text{SmoothL1}\left(\mathbf{R}^{P,t}_{\text{stu}}, \mathbf{R}^{P,t}_{\text{tea}}\right). \tag{4}$$

Combining the above terms, the overall RSSD loss is written as $\mathcal{L}^t_{\text{RSSD}} = \mathcal{L}^{E,t}_{\text{abs}} + \mathcal{L}^{P,t}_{\text{abs}} + \mathcal{L}^{E,t}_{\text{rel}} + \mathcal{L}^{P,t}_{\text{rel}}$.

Joint optimization with current-task learning. At task t, the visual branch only receives images of current novel concepts, denoted as $\mathcal{I}^t$, while the student text branch contains both old and new

concept texts. Thus, the student old-knowledge text embeddings are not only regularized by RSSD, but also serve as negatives in current cross-modal contrastive learning with image features from $\mathcal{I}^t$. Let $\mathcal{L}^t_{\text{reg}}$ and $\mathcal{L}^t_{\text{con}}$ denote the detection loss and cross-modal contrastive loss at task t, respectively. The overall objective is $\mathcal{L}^t_{\text{task}} = \mathcal{L}^t_{\text{reg}} + \mathcal{L}^t_{\text{con}} + \mathcal{L}^t_{\text{RSSD}}$. In this way, RSSD preserves both individual old-concept semantics and textual-space structure, while anchoring newly learned concepts in a stable semantic space without storing historical images.

4.4 Knowledge-aware Parameter Decoupling (KPD)

Although RSSD preserves old knowledge at the representation level, overly strong constraints may hinder novel concept injection. We argue that not all trainable parameters should equally serve old-knowledge retention. To balance retention and injection, we further propose **KPD**, which uses old-knowledge distillation signals to identify old-knowledge-sensitive parameters, constrain their updates, and free the remaining parameters to learn novel concepts.

Restricted trainable parameter subspace. To maintain parameter efficiency, we optimize only a small text-side parameter subset, including the last several layers of the textual encoder and the projector, while freezing the rest of the detector. Formally, let Θ be the full detector parameter set and $\Theta_{\text{tr}} \subset \Theta$ the trainable subset for continual learning. This confines novel concept incorporation to a compact parameter subspace, thereby reducing the risk of overwriting old knowledge.

Old-knowledge-sensitive probe gradients. At task t, we use the preserved old-knowledge concept set $\mathcal{C}^t_{\text{old}}$ to estimate the sensitivity of trainable parameters to old knowledge via a lightweight probe loss defined only on old-knowledge texts. Using the same teacher-student textual representations as in RSSD, we define $\mathcal{L}^t_{\text{probe}} = \mathcal{L}^{E,t}_{\text{abs}} + \mathcal{L}^{P,t}_{\text{abs}}$, this probe loss is used only to estimate parameter sensitivity to old knowledge, rather than serving as the final training objective. For each trainable parameter element $\theta_i \in \Theta_{\text{tr}}$, we compute its probe gradient as $g_i^{\text{probe}} = \frac{\partial \mathcal{L}^t_{\text{probe}}}{\partial \theta_i}$, and the magnitude $\left|g_i^{\text{probe}}\right|$ indicates how sensitive θ_i is to the preserved old knowledge: a larger magnitude suggests that changing this element is more likely to disturb previously acquired semantics.

Element-wise gradient masking. Based on the probe gradients, we construct an element-wise gradient mask to suppress updates on old-knowledge-sensitive parameters. Specifically, we collect the absolute probe gradients of all trainable parameter elements in Θ_{tr} and compute a quantile threshold

$$\tau^t = Q_r\left(\left\{\left|g_i^{\text{probe}}\right| \mid \theta_i \in \Theta_{\text{tr}}\right\}\right), \tag{5}$$

where $Q_r(\cdot)$ denotes the r-th quantile operator, and $r \in (0, 1]$ controls the proportion of retained trainable parameter elements. We then define the binary mask for each trainable parameter element as $m_i^t = \mathbf{1}\left(\left|g_i^{\text{probe}}\right| \le \tau^t\right)$, where $\mathbf{1}(\cdot)$ is the indicator function. Thus, parameter elements with relatively small probe gradients are retained for updating, while those with large probe gradients are suppressed. Let g_i^{train} denote the training-loss gradient w.r.t. θ_i. KPD modifies it by $\tilde{g}_i^{\text{train}} = m_i^t \odot g_i^{\text{train}}$, where $\odot$ denotes element-wise multiplication. The effective parameter update is then performed using $\tilde{g}_i^{\text{train}}$ instead of g_i^{train}. In this way, KPD filters out high-risk updates that are likely to harm old knowledge, while leaving a safer parameter subspace for continual novel concept learning.

KPD in training. KPD does not replace the original objective, but acts as an update constraint complementary to RSSD and current-task supervision. During back-propagation, KPD operates on updates of Θ_{tr} to *identify and suppress old-knowledge-sensitive parameters*, releasing more parameter space for novel concept learning. This enables better adaptation to current novel concepts while avoiding destructive changes to existing semantic knowledge. Overall, RSSD stabilizes the textual semantic space at the representation level, while KPD decouples retention and learning at the parameter level, enabling more stable continual novel concept incorporation.

Implementation Details. All experiments are conducted on 8 NVIDIA 3090 GPUs with a total batch size of 16. The model is trained for 24 epochs using AdamW with an initial learning rate of 1e-3 and weight decay of 1e-4, decayed by 0.1 at epochs 11 and 20. We randomly sample 100 common concepts from ImageNet [41] and Objects365 [42]. The r in Eq. (5) is set to 0.5. *At the beginning of task t, the model is initialized from task $t-1$; the student and teacher textual branches share this initialization, after which the teacher is frozen and the student is updated during training.*

5 Experiments

5.1 Setup

CoOVD settings. We build the continual novel concept incorporation benchmark on Novel-114. As described in Sec. 3, Novel-114 is divided into seven sequential tasks: Animals-25, Clothing-6, Cold Weapons-4, Food-28, Instruments-8, Plants-25, and Vessels-18. Unless specified, all experiments follow this order. The model incorporates concepts task by task. At task t, it can access only current-task training images and annotations, without historical images. This setting evaluates novel concept learning and old knowledge retention without image replay. After each task, we evaluate the model on the current and all previous tasks to measure acquisition and retention, and on MS COCO to examine whether original open-vocabulary generalization is preserved.

Comparison methods. Since this paper studies a new task, continual concept injection for OVD, and introduces a new benchmark, there is no directly applicable standard baseline in existing literature. We therefore select baselines from the perspectives of task relevance and method representativeness, including ZiRa [1] and SGVF [4], which are more related to detection-style VLMs, and MoE-Adapters [6], LADA [5], C-CLIP [8], and KeepLoRA [7], which represent adapter, MoE, reparameterization, and LoRA paradigms in continual learning for classification VLMs. For fair comparison, we adopt LLMDet-Tiny as the unified base model, adapt each method to the LLMDet [30] framework while preserving its core design as much as possible, and compare them under the same training configuration and evaluation protocol.

Evaluation protocol and metrics. To evaluate the stability–plasticity trade-off in continual novel concept injection, we report results from three perspectives: ① ***Original knowledge retention***, ② ***Novel concept acquisition***, and ③ ***Historical injected concept retention***. The gray row in Table 2 reports the zero-shot performance of LLMDet before injection. Its low performance on most Novel-114 tasks shows that these categories cannot be well detected using existing open-vocabulary knowledge alone, validating the need for Novel-114. For original knowledge retention, we evaluate MS COCO [43] after each stage, with red cells indicating preserved OVD capability. For novel concept acquisition, we test the current task after training, with yellow diagonal cells showing adaptation to novel concepts. For historical concept retention, we evaluate all previous tasks after each subsequent stage, with white cells reporting retention performance. Finally, the blue cells summarize the final average performance after all seven tasks, jointly reflecting acquisition, retention, and original OVD preservation. All experiments use COCO-style detection metrics, $AP_{50:95}$ and AP_{50}, with each entry reported as $AP_{50:95}/AP_{50}$.

5.2 Main Results

Table 2 compares our method with existing approaches on MS COCO and Novel-114 under continual novel concept injection. Overall, our method achieves superior performance across most tasks and protocols. Compared with prior SOTA methods such as SGVF and MoE-Adapters, our method obtains the best final average performance (44.3%/53.5%), demonstrating strong capability in balancing stability and plasticity during continual learning. ① ***Original knowledge retention.*** From the red-highlighted regions, which reflect the performance on MS COCO after each stage of novel concept injection, our method maintains significantly higher accuracy compared to other methods. This indicates that the proposed approach effectively mitigates catastrophic forgetting and preserves the original open-vocabulary detection (OVD) capability. ② ***Novel concept acquisition.*** From the yellow diagonal entries, which correspond to the performance on the current task after each training stage, our method achieves competitive or superior results across all tasks. This demonstrates that the model retains strong plasticity and is able to rapidly adapt to newly introduced concepts without sacrificing learning efficiency. ③ ***Historical injected concept retention.*** From the off-diagonal white regions, our method shows more stable performance on previously learned tasks compared to baseline methods. This suggests that the proposed design enables better knowledge consolidation, reducing the degradation of earlier learned concepts during subsequent training stages. ***Overall performance.*** Finally, the blue-highlighted results show that our method achieves the highest average performance among all approaches, confirming that it effectively balances preserving original knowledge, learning new concepts, and retaining historical knowledge. In contrast, KeepLoRA and C-CLIP suffer from severe forgetting or insufficient adaptation, while SGVF shows strong plasticity but weaker stability. Our method achieves a better trade-off, leading to consistent improvements across evaluation metrics.

Table 2: The main results on MS COCO and Novel-114 (%).

Method	Train \ Test	MS COCO	Animals-25	Clothing-6	Cold Weapons-4	Food-28	Instruments-8	Plants-25	Vessels-18	Avg.
LLMDet	–	54.4/70.8	5.9/7.3	10.7/22.4	27.1/32.0	9.5/13.5	31.3/32.7	3.5/5.1	4.1/4.8	18.3/23.6
ZiRa	Animals-25	46.4/60.7	86.6/94.5	–	–	–	–	–	–	–
	Clothing-6	43.7/57.2	85.2/93.2	43.9/49.1	–	–	–	–	–	–
	Cold Weapons-4	43.0/56.4	69.1/75.5	37.0/40.5	30.0/36.0	–	–	–	–	–
	Food-28	34.5/45.5	58.9/66.3	14.8/22.7	20.2/24.8	70.5/80.7	–	–	–	–
	Instruments-8	35.8/47.2	61.6/69.1	13.8/20.0	18.6/23.3	68.7/78.8	44.0/51.0	–	–	–
	Plants-25	19.7/26.4	27.0/30.4	11.3/17.6	29.2/34.0	1.3/1.5	13.5/15.0	66.3/85.8	–	–
	Vessels-18	17.7/23.2	31.3/35.1	11.8/15.8	20.8/24.1	38.2/45.8	21.5/24.3	61.0/78.8	55.9/59.4	32.3/38.3
MoE-Adapters	Animals-25	39.4/52.3	84.5/91.6	–	–	–	–	–	–	–
	Clothing-6	36.4/48.7	64.9/70.7	33.5/37.4	–	–	–	–	–	–
	Cold Weapons-4	31.3/42.5	58.6/64.6	32.0/34.0	59.0/64.8	–	–	–	–	–
	Food-28	31.3/42.5	65.4/72.3	25.7/29.0	34.7/40.4	64.8/74.2	–	–	–	–
	Instruments-8	38.1/52.0	49.5/55.1	28.0/31.2	30.6/35.4	62.5/72.0	31.1/32.2	–	–	–
	Plants-25	33.5/45.7	56.0/63.2	27.2/33.0	20.7/23.9	48.7/57.5	12.4/17.6	42.7/55.0	–	–
	Vessels-18	30.7/42.0	58.4/65.1	28.8/36.8	23.2/26.8	47.1/55.4	16.2/19.8	38.5/50.0	60.3/61.3	37.9/44.7
LADA	Animals-25	35.7/47.2	87.9/95.0	–	–	–	–	–	–	–
	Clothing-6	17.6/22.6	69.7/80.9	64.2/68.2	–	–	–	–	–	–
	Cold Weapons-4	20.2/26.7	63.6/71.1	39.7/44.9	75.3/81.8	–	–	–	–	–
	Food-28	32.4/43.8	34.8/41.1	13.8/19.3	26.1/28.7	74.3/84.9	–	–	–	–
	Instruments-8	18.1/24.4	15.4/20.6	17.9/23.5	23.9/28.1	56.5/66.7	57.9/61.9	–	–	–
	Plants-25	23.4/31.8	20.2/25.0	5.6/10.6	20.2/25.0	27.5/35.0	34.2/38.9	70.1/89.3	–	–
	Vessels-18	20.8/27.9	16.0/20.5	6.0/10.6	20.6/23.6	9.3/13.1	38.4/52.6	12.3/14.1	70.9/71.9	24.3/29.3
C-CLIP	Animals-25	48.3/64.0	64.4/71.7	–	–	–	–	–	–	–
	Clothing-6	43.3/58.0	59.0/66.0	42.0/45.7	–	–	–	–	–	–
	Cold Weapons-4	45.2/61.5	27.6/32.5	30.4/32.8	51.4/60.6	–	–	–	–	–
	Food-28	45.4/63.1	17.6/20.9	11.3/19.3	34.2/40.4	46.8/55.9	–	–	–	–
	Instruments-8	46.6/62.6	14.8/17.1	17.9/26.3	30.7/35.3	31.5/37.4	41.9/45.5	–	–	–
	Plants-25	32.7/45.4	9.0/11.1	17.4/25.4	26.8/31.1	12.0/14.7	12.5/17.5	32.6/43.0	–	–
	Vessels-18	32.7/45.6	4.4/5.9	16.3/24.2	25.7/31.6	7.1/10.0	10.8/16.5	18.6/24.3	23.7/26.3	17.4/23.1
SGVF	Animals-25	26.6/35.9	83.7/92.0	–	–	–	–	–	–	–
	Clothing-6	29.5/39.1	82.2/91.6	61.1/67.4	–	–	–	–	–	–
	Cold Weapons-4	26.3/35.1	78.8/88.4	39.1/45.6	73.4/79.3	–	–	–	–	–
	Food-28	24.7/33.7	60.1/76.5	24.3/31.9	44.8/51.0	71.2/82.0	–	–	–	–
	Instruments-8	24.0/32.6	52.0/70.5	16.9/20.4	40.0/46.8	64.3/74.1	64.2/68.6	–	–	–
	Plants-25	24.7/33.7	59.8/70.9	16.5/20.4	41.1/46.8	52.3/62.7	47.2/50.4	65.9/85.9	–	–
	Vessels-18	19.3/26.4	52.1/63.1	11.9/15.1	39.4/47.0	42.7/52.0	36.3/39.1	39.3/54.5	73.4/74.7	39.3/46.5
KeepLoRa	Animals-25	50.2/67.6	14.9/19.8	–	–	–	–	–	–	–
	Clothing-6	29.5/40.6	2.6/3.8	8.7/17.7	–	–	–	–	–	–
	Cold Weapons-4	34.5/47.1	5.4/7.0	9.8/22.5	26.7/32.0	–	–	–	–	–
	Food-28	45.1/60.3	7.1/8.8	6.7/18.1	29.6/36.8	17.3/23.6	–	–	–	–
	Instruments-8	31.9/46.2	4.9/6.2	9.4/17.9	21.0/25.9	6.9/10.4	23.8/26.8	–	–	–
	Plants-25	44.7/59.9	7.0/9.2	13.2/29.0	26.5/32.7	9.3/14.0	26.7/30.4	5.5/8.0	–	–
	Vessels-18	43.8/59.7	6.1/7.6	10.1/18.8	25.2/31.0	8.4/13.3	28.2/32.1	3.7/5.2	6.8/8.9	16.5/22.1
NoInDet	Animals-25	49.8/65.1	87.9/95.2	–	–	–	–	–	–	–
	Clothing-6	49.5/64.3	80.7/88.6	47.1/54.6	–	–	–	–	–	–
	Cold Weapons-4	49.1/63.7	81.9/89.0	40.5/53.5	71.0/79.6	–	–	–	–	–
	Food-28	47.0/62.0	73.3/81.0	24.3/32.7	55.8/62.7	72.8/82.9	–	–	–	–
	Instruments-8	47.1/61.9	75.9/83.4	14.9/25.0	51.0/61.3	68.8/78.7	62.5/67.1	–	–	–
	Plants-25	45.7/60.2	48.7/54.5	20.6/29.6	44.7/50.1	55.3/64.8	63.3/69.5	69.6/89.1	–	–
	Vessels-18	41.7/54.8	50.1/57.3	18.2/26.7	31.1/38.8	43.2/54.1	35.7/39.2	63.7/83.3	70.4/73.7	44.3/53.5

Table 3: Component ablation of RSSD and KPD on MS COCO and Novel-114 (%).

Method	Protocol	MS COCO	Animals-25	Clothing-6	Cold Weapons-4	Food-28	Instruments-8	Plants-25	Vessels-18	Avg.
LLMDet	–	54.4/70.8	5.9/7.3	10.7/22.4	27.1/32.0	9.5/13.5	31.3/32.7	3.5/5.1	4.1/4.8	18.3/23.6
+ RSSD	(i) OVD retention	–	39.0/54.4	38.2/50.5	45.2/60.3	43.9/58.8	45.3/59.8	31.3/42.9	30.4/40.4	39.0/52.4
	(ii) Current-task acquisition	–	87.4/94.8	45.6/56.3	73.8/81.9	72.5/82.7	62.4/66.9	69.2/88.6	67.3/70.3	68.3/77.4
	(iii) Final performance	30.4/40.4	46.7/55.0	20.9/28.9	29.7/37.8	43.7/52.0	24.9/30.2	59.8/80.1	67.3/70.3	40.4/49.3
+ RSSD + KPD (NoInDet)	(i) OVD retention	–	49.8/65.1	49.5/64.3	49.1/63.7	47.0/62.0	47.1/61.9	45.7/60.2	41.7/54.8	47.1/61.7
	(ii) Current-task acquisition	–	87.9/95.2	47.1/54.6	71.0/79.6	72.8/82.9	62.5/67.1	69.6/89.1	70.4/73.7	68.8/77.5
	(iii) Final performance	41.7/54.8	50.1/57.3	18.2/26.7	31.1/38.8	43.2/54.1	35.7/39.2	63.7/83.3	70.4/73.7	44.3/53.5

5.3 Ablation Study

Component ablation. As shown in Table 3, we conduct ablation studies on MS COCO and Novel-114 to evaluate the contributions of RSSD and KPD in NoInDet. Since KPD relies on the stable gradient signals induced by RSSD for parameter-level constraints, it cannot be applied independently. We therefore adopt a progressive ablation strategy: adding RSSD first to stabilize the representation space, and then adding KPD to further constrain parameter updates.

Effect of RSSD. Compared with LLMDet, LLMDet+RSSD improves original OVD knowledge retention. As shown by the red-highlighted results, the model retains higher COCO performance after continual learning, indicating that RSSD mitigates forgetting by stabilizing the textual representation space. Meanwhile, strong current-task performance shows that RSSD does not hinder plasticity.

Effect of KPD. Adding KPD brings further gains. COCO retention improves from 39.0%/52.4% to 47.1%/61.7%, and final average performance increases from 40.4%/49.3% to 44.3%/53.5%. This shows that KPD strengthens knowledge consolidation while preserving novel concept learning.

Overall analysis. RSSD and KPD are complementary: RSSD stabilizes the semantic representation space, while KPD regularizes parameter updates at a finer granularity. Together, they achieve a better stability–plasticity trade-off and superior overall performance.

Effect of fine-tuned layers in the textual encoder. We further study how the number of fine-tuned top layers in the textual encoder affects RSSD, as shown in Table 4. We vary the number of updated

Table 4: Effect of the number of fine-tuned textual encoder layers in RSSD (%).

Method	Protocol	MS COCO	Animals-25	Clothing-6	Cold Weapons-4	Food-28	Instruments-8	Plants-25	Vessels-18	Avg.
LLMDet	–	54.4/70.8	5.9/7.3	10.7/22.4	27.1/32.0	9.5/13.5	31.3/32.7	3.5/5.1	4.1/4.8	18.3/23.6
NoInDet (12 layers)	(i) OVD retention	–	41.2/53.0	38.1/53.9	37.6/49.0	39.3/51.5	40.7/52.8	40.8/54.1	20.3/28.3	36.9 / 48.9
	(ii) Current-task acquisition	–	86.9/94.6	55.6/60.1	74.8/82.4	73.4/83.7	59.9/66.0	69.9/89.1	68.3/70.2	69.8/78.0
	(iii) Final performance	20.3/28.3	44.4/61.6	2.3/4.8	15.7/21.4	30.7/47.9	29.3/43.4	46.3/64.3	68.3/70.2	32.2/42.7
NoInDet (10 layers)	(i) OVD retention	–	43.9/58.1	39.3/51.6	43.5/56.6	41.5/55.0	43.4/56.4	43.0/57.8	34.2/44.8	41.3/54.3
	(ii) Current-task acquisition	–	88.8/96.0	55.4/63.0	71.1/79.8	72.2/82.4	66.2/70.4	70.4/89.6	67.9/70.9	70.3/78.9
	(iii) Final performance	34.2/44.8	71.1/79.9	7.9/13.0	19.4/29.4	50.7/64.0	27.9/38.1	63.1/82.0	67.9/70.9	42.8/52.8
NoInDet (8 layers)	(i) OVD retention	–	42.5/57.5	36.4/48.3	38.7/51.6	25.5/34.1	31.2/41.2	40.1/53.5	29.8/39.4	34.9/46.5
	(ii) Current-task acquisition	–	88.1/95.4	51.0/58.3	66.8/76.1	71.1/81.7	61.3/66.6	68.7/88.6	69.7/71.6	68.1/76.9
	(iii) Final performance	29.8/39.4	61.7/72.6	14.3/19.7	27.7/39.5	48.4/58.1	32.4/39.5	56.5/74.9	69.7/71.6	42.6/51.9
NoInDet (6 layers)	(i) OVD retention	–	48.4/63.7	42.8/56.1	50.3/65.7	40.7/55.2	49.8/65.2	40.5/55.4	41.4/55.3	44.8/59.5
	(ii) Current-task acquisition	–	87.4/94.4	52.4/55.9	72.3/80.3	72.6/82.6	62.8/66.4	69.8/89.3	66.3/70.1	69.1/77.0
	(iii) Final performance	41.4/55.3	58.1/66.4	8.1/13.2	14.5/23.8	49.4/55.9	46.9/51.5	62.0/81.2	66.3/70.1	43.3/52.2
NoInDet (4 layers)	(i) OVD retention	–	49.8/65.1	49.5/64.3	49.1/63.7	47.0/62.0	47.1/61.9	45.7/60.2	41.7/54.8	47.1/61.7
	(ii) Current-task acquisition	–	87.9/95.2	47.1/54.6	71.0/79.6	72.8/82.9	62.5/67.1	69.6/89.1	70.4/73.7	68.8/77.5
	(iii) Final performance	41.7/54.8	50.1/57.3	18.2/26.7	31.1/38.8	43.2/54.1	35.7/39.2	63.7/83.3	70.4/73.7	44.3/53.5
NoInDet (2 layers)	(i) OVD retention	–	52.7/68.5	51.6/67.2	51.6/67.2	47.0/63.0	50.0/65.5	46.3/60.8	40.4/53.3	48.5/63.6
	(ii) Current-task acquisition	–	86.6/94.2	47.9/51.9	71.4/80.2	68.6/78.8	54.9/59.0	69.0/88.4	69.0/71.7	66.8/74.9
	(iii) Final performance	40.4/53.3	19.4/27.0	5.7/8.9	15.9/20.9	40.4/48.6	31.6/37.2	63.0/81.9	69.0/71.7	35.7/43.7
NoInDet (0 layers)	(i) OVD retention	–	53.7/70.0	53.5/69.7	53.5/69.7	52.5/68.2	52.9/68.9	52.3/68.1	50.2/65.7	52.7/68.6
	(ii) Current-task acquisition	–	27.0/31.0	10.1/21.8	24.6/30.1	19.3/24.4	20.4/22.9	5.3/7.5	7.4/8.6	16.3/20.9
	(iii) Final performance	50.2/65.7	4.7/8.2	1.5/3.8	6.4/9.8	18.7/24.2	9.0/13.8	3.6/5.5	7.4/8.6	12.7/17.5

Table 5: Sensitivity analysis of the update ratio r in KPD on MS COCO and Novel-114 (%).

Method	Protocol	MS COCO	Animals-25	Clothing-6	Cold Weapons-4	Food-28	Instruments-8	Plants-25	Vessels-18	Avg.
LLMDet	–	54.4/70.8	5.9/7.3	10.7/22.4	27.1/32.0	9.5/13.5	31.3/32.7	3.5/5.1	4.1/4.8	18.3/23.6
NoInDet (r=0.1)	(i) OVD retention	–	41.5/55.5	45.4/59.2	48.4/63.1	39.8/53.8	44.4/58.6	31.2/43.4	41.2/54.5	41.7/55.4
	(ii) Current-task acquisition	–	87.9/95.1	47.0/54.5	68.7/78.3	71.5/81.7	65.8/72.7	69.5/89.2	61.6/63.9	67.4/76.5
	(iii) Final performance	41.2/54.5	51.9/60.4	1.7/5.8	18.6/28.9	52.9/62.8	29.2/33.9	67.0/86.5	61.6/63.9	40.5/49.6
NoInDet (r=0.3)	(i) OVD retention	–	38.6/52.6	35.7/47.0	43.2/56.7	40.7/55.0	41.4/55.4	37.6/51.0	39.0/52.5	39.5/52.9
	(ii) Current-task acquisition	–	86.6/93.5	53.6/60.3	71.8/80.5	69.0/79.5	55.1/60.6	69.6/88.9	64.2/67.4	67.1/75.8
	(iii) Final performance	39.0/52.5	44.4/53.9	4.0/8.5	15.5/24.6	37.0/51.6	28.2/31.9	63.5/82.4	64.2/67.4	37.0/46.6
NoInDet (r=0.5)	(i) OVD retention	–	49.8/65.1	49.5/64.3	49.1/63.7	47.0/62.0	47.1/61.9	45.7/60.2	41.7/54.8	47.1/61.7
	(ii) Current-task acquisition	–	87.9/95.2	47.1/54.6	71.0/79.6	72.8/82.9	62.5/67.1	69.6/89.1	70.4/73.7	68.8/77.5
	(iii) Final performance	41.7/54.8	50.1/57.3	18.2/26.7	31.1/38.8	43.2/54.1	35.7/39.2	63.7/83.3	70.4/73.7	44.3/53.5
NoInDet (r=0.6)	(i) OVD retention	–	48.2/63.8	46.7/61.1	47.3/62.0	47.9/63.7	48.5/64.2	41.4/55.7	40.2/53.0	45.7/60.5
	(ii) Current-task acquisition	–	86.6/94.3	54.4/63.5	71.5/80.7	71.6/81.5	62.2/65.8	69.7/89.5	69.0/71.5	69.3/78.1
	(iii) Final performance	40.2/53.0	48.8/56.8	5.4/10.7	10.4/17.5	49.9/59.7	34.8/39.1	65.3/84.4	69.0/71.5	40.5/49.1
NoInDet (r=0.8)	(i) OVD retention	–	47.5/62.1	41.1/54.1	46.7/60.7	48.9/64.0	48.7/63.6	42.0/55.5	39.9/52.9	45.0/59.0
	(ii) Current-task acquisition	–	87.5/94.9	50.6/56.2	68.3/78.0	72.1/82.3	58.9/63.9	69.4/88.4	62.4/65.0	67.0/75.5
	(iii) Final performance	39.9/52.9	59.9/67.6	4.2/7.2	21.1/29.3	39.7/49.6	24.8/28.3	66.0/85.1	62.4/65.0	39.7/48.1
NoInDet (r=1.0)	(i) OVD retention	–	39.0/54.4	38.2/50.5	45.2/60.3	43.9/58.8	45.3/59.8	31.3/42.9	30.4/40.4	39.0/52.4
	(ii) Current-task acquisition	–	87.4/94.8	45.6/56.3	73.8/81.9	72.5/82.7	62.4/66.9	69.2/88.6	67.3/70.3	68.3/77.4
	(iii) Final performance	30.4/40.4	46.7/55.0	20.9/28.9	29.7/37.8	43.7/52.0	24.9/30.2	59.8/80.1	67.3/70.3	40.4/49.3

layers from 0 to 12 to analyze the stability–plasticity trade-off. Updating no layers best preserves original knowledge (52.7%/68.6%) but provides limited adaptation. As more layers are fine-tuned, novel concept acquisition improves, especially with 4–10 updated layers, but OVD retention gradually decreases, indicating stronger forgetting. Updating too few layers limits learning, while updating too many layers disrupts the original semantic space. The best final performance is achieved by fine-tuning a moderate number of layers, *e.g.*, 4 layers, which obtains the highest average score (44.3%/53.5%). This suggests that lower textual layers encode more general and transferable semantics and should remain stable, whereas higher layers are more task-adaptive. Selectively updating top layers therefore enables effective concept injection while preserving the core representation space.

Effect of update ratio in KPD. We analyze the update ratio r, which controls the proportion of trainable parameter elements selected for updating, as shown in Table 5. Performance varies non-monotonically with r, reflecting a stability–plasticity trade-off. While $r = 0.6$ achieves the highest current-task acquisition (69.3%/78.1%), $r = 0.5$ yields the best final performance (44.3%/53.5%) and stronger COCO retention (47.1%/61.7%). In contrast, updating all trainable elements at $r = 1.0$ results in the lowest COCO retention (39.0%/52.4%) and a final score of 40.4%/49.3%. These results support selective parameter updates, with $r = 0.5$ providing the best overall balance among the tested configurations. **Additional experimental results are provided in Appendix A, and qualitative comparisons are presented in Appendix B.4.**

6 Conclusion

In this paper, we introduce CoOVD, a new task that continually injects novel concepts while preserving previously injected concepts and original OVD knowledge. To support CoOVD, we construct Novel-114, a benchmark with 114 novel concepts organized into seven sequential stages. We propose NoInDet, a parameter-efficient framework that freezes the visual encoder and selectively updates the textual branch. With Representation Space Stability Distillation (RSSD) and Knowledge-aware Parameter Decoupling (KPD), NoInDet stabilizes the semantic space and constrains old-knowledge-sensitive parameters. Experiments on Novel-114 and COCO show that NoInDet effectively balances novel concept learning, historical concept retention, and original OVD knowledge preservation, consistently outperforming existing continual learning methods.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grants U2574216, 62476196, and 62402490; by the Shenzhen Basic Research Foundation under Grant QNXMC20250701101037019; and by the Emerging Frontiers Cultivation Program of Tianjin University Interdisciplinary Center.

References

[1] Jieren Deng, Haojian Zhang, Kun Ding, Jianhua Hu, Xingxuan Zhang, and Yunkuan Wang. Zero-shot generalizable incremental learning for vision-language object detection. *Advances in Neural Information Processing Systems*, 37:136679–136700, 2024. 2, 3, 4, 8

[2] Chunyuan Li, Haotian Liu, Liunian Li, Pengchuan Zhang, Jyoti Aneja, Jianwei Yang, Ping Jin, Houdong Hu, Zicheng Liu, Yong Jae Lee, et al. Elevater: A benchmark and toolkit for evaluating language-augmented visual models. *Advances in Neural Information Processing Systems*, 35:9287–9301, 2022. 2, 4

[3] Longxiang Tang, Zhuotao Tian, Kai Li, Chunming He, Hantao Zhou, Hengshuang Zhao, Xiu Li, and Jiaya Jia. Mind the interference: Retaining pre-trained knowledge in parameter efficient continual learning of vision-language models. In *European conference on computer vision*, pages 346–365. Springer, 2024. 2, 3

[4] Ximeng Cheng, Qiaonan Zhu, Shukun Jia, Yichao Cao, and Xiaobo Lu. Incremental mixture of experts: Continual learning for object detection in forestry scenarios. *Pattern Recognition*, page 113001, 2026. 2, 3, 4, 8

[5] Mao-Lin Luo, Zi-Hao Zhou, Tong Wei, and Min-Ling Zhang. Lada: Scalable label-specific clip adapter for continual learning. *arXiv preprint arXiv:2505.23271*, 2025. 2, 3, 8

[6] Jiazuo Yu, Yunzhi Zhuge, Lu Zhang, Ping Hu, Dong Wang, Huchuan Lu, and You He. Boosting continual learning of vision-language models via mixture-of-experts adapters. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23219–23230, 2024. 2, 3, 8

[7] Mao-Lin Luo, Zi-Hao Zhou, Yi-Lin Zhang, Yuanyu Wan, Tong Wei, and Min-Ling Zhang. Keeplora: Continual learning with residual gradient adaptation. *arXiv preprint arXiv:2601.19659*, 2026. 2, 3, 4, 8

[8] Wenzhuo Liu, Fei Zhu, Longhui Wei, and Qi Tian. C-clip: Multimodal continual learning for vision-language model. In *International Conference on Learning Representations*, volume 2025, pages 46461–46477, 2025. 3, 8

[9] Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. Open-vocabulary object detection using captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14393–14402, 2021. 3

[10] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763, 2021. 3, 4

[11] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916, 2021. 3

[12] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921*, 2021. 3

[13] Luting Wang, Yi Liu, Penghui Du, Zihan Ding, Yue Liao, Qiaosong Qi, Biaolong Chen, and Si Liu. Object-aware distillation pyramid for open-vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11186–11196, 2023. 3

[14] Liangqi Li, Jiaxu Miao, Dahu Shi, Wenming Tan, Ye Ren, Yi Yang, and Shiliang Pu. Distilling detr with visual-linguistic knowledge for open-vocabulary object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6501–6510, 2023. 3

[15] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In *European Conference on Computer Vision*, pages 350–368, 2022. 3

[16] Hanoona Bangalath, Muhammad Maaz, Muhammad Uzair Khattak, Salman H Khan, and Fahad Shahbaz Khan. Bridging the gap between object and image-level representations for open-vocabulary detection. *Advances in Neural Information Processing Systems*, 35:33781–33794, 2022. 3

[17] Shiyu Zhao, Samuel Schulter, Long Zhao, Zhixing Zhang, Yumin Suh, Manmohan Chandraker, Dimitris N Metaxas, et al. Taming self-training for open-vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13938–13947, 2024. 3

[18] Shilin Xu, Xiangtai Li, Size Wu, Wenwei Zhang, Yunhai Tong, and Chen Change Loy. Dst-det: Open-vocabulary object detection via dynamic self-training. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024. 3

[19] Junjie Wang, Bin Chen, Bin Kang, Yulin Li, Weizhi Xian, Yichi Chen, and Yong Xu. Ov-dquo: Open-vocabulary detr with denoising text query training and open-world unknown objects supervision. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 7762–7770, 2025. 3

[20] Joonhyun Jeong, Geondo Park, Jayeon Yoo, Hyungsik Jung, and Heesu Kim. Proxydet: Synthesizing proxy novel classes via classwise mixup for open-vocabulary object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 2462–2470, 2024. 3

[21] Yu Du, Fangyun Wei, Zihe Zhang, Miaojing Shi, Yue Gao, and Guoqi Li. Learning to prompt for open-vocabulary object detection with vision-language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14084–14093, 2022. 3

[22] Jiaming Li, Jiacheng Zhang, Jichang Li, Ge Li, Si Liu, Liang Lin, and Guanbin Li. Learning background prompts to discover implicit knowledge for open vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16678–16687, 2024. 3

[23] Xiaoshi Wu, Feng Zhu, Rui Zhao, and Hongsheng Li. Cora: Adapting clip for open-vocabulary detection with region prompting and anchor pre-matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7031–7040, 2023. 3

[24] Weicheng Kuo, Yin Cui, Xiuye Gu, AJ Piergiovanni, and Anelia Angelova. F-vlm: Open-vocabulary object detection upon frozen vision and language models. *arXiv preprint arXiv:2209.15639*, 2022. 3

[25] 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. *arXiv preprint arXiv:2310.01403*, 2023. 3

[26] Junjie Wang, Bin Chen, Yulin Li, Bin Kang, Yichi Chen, and Zhuotao Tian. Declip: Decoupled learning for open-vocabulary dense perception. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 14824–14834, 2025. 3

[27] Dahun Kim, Anelia Angelova, and Weicheng Kuo. Region-aware pretraining for open-vocabulary object detection with vision transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11144–11154, 2023. 3

[28] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10965–10975, 2022. 3

[29] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, pages 38–55. Springer, 2024. 3

[30] Shenghao Fu, Qize Yang, Qijie Mo, Junkai Yan, Xihan Wei, Jingke Meng, Xiaohua Xie, and Wei-Shi Zheng. Llmdet: Learning strong open-vocabulary object detectors under the supervision of large language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 14987–14997, 2025. 3, 4, 8

[31] Tianheng Cheng, Lin Song, Yixiao Ge, Wenyu Liu, Xinggang Wang, and Ying Shan. Yolo-world: Real-time open-vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16901–16911, 2024. 3

[32] Ao Wang, Lihao Liu, Hui Chen, Zijia Lin, Jungong Han, and Guiguang Ding. Yoloe: Real-time seeing anything. *arXiv preprint arXiv:2503.07465*, 2025. 3

[33] Wuyang Li, Xinyu Liu, Jiayi Ma, and Yixuan Yuan. Cliff: Continual latent diffusion for open-vocabulary object detection. In *European Conference on Computer Vision*, pages 255–273. Springer, 2024. 3

[34] Zangwei Zheng, Mingyuan Ma, Kai Wang, Ziheng Qin, Xiangyu Yue, and Yang You. Preventing zero-shot transfer degradation in continual learning of vision-language models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 19125–19136, 2023. 3

[35] Yaoyao Liu, Bernt Schiele, Andrea Vedaldi, and Christian Rupprecht. Continual detection transformer for incremental object detection. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 23799–23808. IEEE, 2023. 4

[36] Matthias Neuwirth-Trapp, Maarten Bieshaar, Danda Pani Paudel, and Luc Van Gool. Rico: Two realistic benchmarks and an in-depth analysis for incremental learning in object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 5212–5223, October 2025. 4

[37] Chiara Cappellino, Gianluca Mancusi, Matteo Mosconi, Angelo Porrello, Simone Calderara, and Rita Cucchiara. Dithub: A modular framework for incremental open-vocabulary object detection. *Advances in Neural Information Processing Systems*, 38:22898–22926, 2026. 4

[38] Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2935–2947, 2017. 4

[39] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017. 4

[40] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5356–5364, 2019. 4

[41] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 248–255, 2009. 7

[42] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8430–8439, 2019. 7

[43] 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 *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755, 2014. 8

Appendix

A More Results

A.1 Second-round Concept Injection

To further evaluate the scalability of our method, we conduct a second round of concept injection by initializing from the final model obtained after the first round and repeating the same training protocol on Novel-114. As shown in Table 6, our method continues to achieve strong performance under extended continual learning. In particular, it obtains the best final average performance of 55.1%/64.7%, clearly outperforming SGVF (49.2%/56.3%), ZiRa (44.8%/52.0%), and MoE-Adapters (39.5%/45.3%).

Knowledge accumulation. Compared with the first round, our method improves from 44.3%/53.5% to 55.1%/64.2% in final average performance, indicating that the model can accumulate knowledge across multiple injection rounds rather than merely relearning or overwriting previous concepts.

Retention and stability. Despite undergoing an additional round of concept injection, our method maintains stronger original OVD retention than competing methods. After the final second-round task, our method still achieves 39.5%/52.5% on MS COCO, while SGVF and ZiRa drop to 16.9%/23.3% and 15.0%/19.5%, respectively. This further confirms that RSSD and KPD effectively mitigate catastrophic forgetting in longer continual learning processes.

Comparison with existing methods. Most baseline methods show limited robustness under repeated knowledge injection. For example, KeepLoRA obtains only 16.1%/21.9% final average performance, while C-CLIP reaches 27.6%/34.0%. In contrast, our method maintains both strong plasticity and stability, leading to consistent performance improvements.

Overall analysis. These results demonstrate that our method is not only effective for single-round continual learning, but also scalable to multi-round concept injection. This highlights its potential for real-world scenarios where new concepts are continuously introduced over time.

A.2 Robustness to Common Concept Sampling

Since RSSD uses randomly sampled common concepts as lightweight textual old knowledge, we further evaluate whether the performance is sensitive to the sampled common set. We randomly construct five Common-100 sets from ImageNet and Objects365, while keeping all other settings unchanged. As shown in Table 7, NoInDet achieves stable performance across different common sets, with only minor variance in both COCO retention and Novel-114 performance. This indicates that RSSD does not rely on a specific common concept set, and the sampled common texts mainly serve as general semantic anchors for stabilizing the textual representation space.

A.3 Effect of the Adaptation Branch

To examine the choice of text-side adaptation, we compare three variants under the same Novel-114 data, training budget, and no historical image replay setting. *Text Only* is the default NoInDet configuration, which updates selected text-branch parameters with RSSD and KPD. *Vision Only* updates visual-side parameters with SGVF-style visual cosine regularization. *Text+Vision* updates both branches with their corresponding text-side and visual-side regularization. The comparison therefore evaluates branch-adaptation strategies, including their regularization schemes.

For this study and the task-order study in Section A.4, OVD retention is the mean MS COCO performance over the seven stages, current-task acquisition is the mean performance of each concept group immediately after it is learned, and final performance is the equal-weight average over MS COCO and the seven concept groups after the last stage.

Table 8 reveals a stability–plasticity trade-off. Vision Only and Text+Vision obtain higher current-task acquisition than Text Only, reaching 70.6/78.7 and 72.3/79.9, respectively. However, their OVD retention decreases to 27.4/37.6 and 25.5/34.9, compared with 47.1/61.7 for Text Only. Consequently, Text Only achieves the highest final performance, exceeding Vision Only by 8.3/9.6 points and Text+Vision by 4.1/4.8 points. Under the evaluated no-image-replay setting, the text-side strategy

Table 6: Second-round concept injection on MS COCO and Novel-114 (%).

Method	Train \ Test	MS COCO	Animals-25	Clothing-6	Cold Weapons-4	Food-28	Instruments-8	Plants-25	Vessels-18	Avg.
LLMDet	–	54.4/70.8	5.9/7.3	10.7/22.4	27.1/32.0	9.5/13.5	31.3/32.7	3.5/5.1	4.1/4.8	18.3/23.6
ZiRa	Animals-25	18.9/24.7	87.7/95.3	16.3/24.8	22.6/27.1	2.6/4.0	18.4/21.7	47.6/62.0	37.2/40.8	31.4/37.6
	Clothing-6	17.2/22.5	85.3/93.2	46.4/50.3	20.4/26.1	43.3/50.7	15.2/19.6	46.3/60.0	41.2/44.6	39.4/45.9
	Cold Weapons-4	18.8/24.5	86.1/94.2	46.0/51.6	66.0/73.6	42.3/49.4	12.3/15.7	45.5/58.9	30.2/35.0	43.4/50.4
	Food-28	20.4/26.5	83.7/91.7	27.5/30.3	43.5/51.9	73.2/83.7	24.1/28.6	43.9/57.0	32.8/35.3	43.6/50.6
	Instruments-8	21.7/28.0	84.2/92.2	21.5/24.7	43.6/51.1	71.2/81.4	48.1/53.2	45.3/59.1	33.1/35.4	46.1/53.1
	Plants-25	17.3/22.6	65.8/72.5	13.7/18.7	43.5/51.0	64.7/74.3	44.9/50.5	68.7/87.9	33.8/36.1	44.1/51.7
	Vessels-18	15.0/19.5	65.5/72.1	13.5/16.6	43.6/51.0	57.2/67.4	36.3/40.0	65.1/83.7	62.1/65.3	44.8/52.0
MoE-Adapters	Animals-25	33.4/44.9	85.5/92.7	12.8/23.2	22.7/26.5	8.0/10.0	11.3/12.0	2.3/3.7	6.5/7.8	22.8/27.6
	Clothing-6	30.3/41.0	84.5/92.1	34.2/41.2	23.1/27.0	6.2/8.3	8.3/9.6	1.8/3.0	7.5/9.2	24.5/28.9
	Cold Weapons-4	28.4/38.4	82.0/89.9	33.3/43.5	43.6/49.6	5.2/6.9	6.3/7.8	2.1/3.3	8.4/9.1	26.2/31.1
	Food-28	31.2/42.8	78.9/86.9	32.7/40.4	46.5/53.7	62.4/71.8	14.2/17.0	3.2/4.6	6.7/7.3	34.5/40.6
	Instruments-8	32.6/44.8	80.1/88.0	33.4/40.5	42.0/49.2	60.5/69.8	23.4/25.5	3.8/5.2	6.4/7.1	35.3/41.3
	Plants-25	29.0/40.1	77.9/85.7	27.0/33.9	41.9/48.4	50.8/59.8	20.1/22.6	35.6/44.9	3.4/4.2	35.7/42.5
	Vessels-18	28.2/38.8	66.1/73.0	25.6/28.6	33.5/40.0	48.0/56.1	22.1/23.7	34.1/43.3	58.1/59.1	39.5/45.3
LADA	Animals-25	31.9/43.2	87.3/95.6	13.0/23.0	28.8/32.5	27.7/33.9	22.8/24.9	24.3/32.1	21.7/23.1	32.2/38.5
	Clothing-6	23.2/30.7	78.3/89.0	64.0/67.9	18.7/22.2	16.9/24.3	24.1/26.8	17.3/23.1	26.7/28.1	33.7/39.0
	Cold Weapons-4	22.5/30.2	72.7/81.8	38.7/44.7	75.1/82.6	15.9/20.6	11.2/15.1	16.2/21.8	18.0/19.0	33.8/39.5
	Food-28	27.6/37.4	61.3/70.6	19.2/24.8	41.0/47.8	74.5/85.3	28.2/31.3	6.9/9.9	6.0/6.7	33.1/39.2
	Instruments-8	18.0/24.3	43.4/51.4	25.8/31.7	40.5/46.9	58.3/69.2	70.5/75.1	4.9/6.7	11.8/13.1	34.2/39.8
	Plants-25	20.8/28.7	43.3/52.3	19.3/25.9	35.6/43.7	45.4/54.7	33.9/38.5	70.4/89.1	10.4/12.0	34.9/43.1
	Vessels-18	17.6/23.9	35.8/41.0	14.1/17.1	26.7/28.6	20.2/26.0	21.9/24.2	49.6/66.3	75.6/76.4	32.7/37.9
C-CLIP	Animals-25	42.7/57.2	75.2/82.6	25.8/39.0	30.6/36.2	21.8/27.1	12.4/14.2	10.2/12.9	7.8/8.7	28.3/34.7
	Clothing-6	41.9/55.9	56.5/63.1	43.7/47.2	56.5/63.1	17.5/21.8	8.1/10.7	8.2/10.8	8.9/9.8	30.2/35.3
	Cold Weapons-4	45.5/61.6	48.4/54.3	36.5/38.5	50.3/59.1	21.3/26.6	20.3/24.0	7.6/10.0	12.7/14.5	30.3/36.1
	Food-28	45.6/61.0	28.0/32.2	19.6/25.8	29.5/34.4	56.0/65.4	31.1/33.3	6.7/8.9	6.4/7.4	27.9/33.6
	Instruments-8	46.2/62.0	21.7/25.3	21.4/29.3	27.8/33.5	42.2/49.9	44.1/46.9	6.8/8.9	6.9/9.1	27.1/33.1
	Plants-25	38.3/52.7	19.1/22.4	22.1/29.0	34.7/43.0	33.3/39.9	40.7/53.0	39.5/52.1	7.0/10.1	29.3/37.8
	Vessels-18	32.1/44.9	19.9/23.2	25.6/33.1	43.6/50.0	18.7/23.4	16.1/20.1	33.0/44.2	31.6/32.9	27.6/34.0
SGVF	Animals-25	26.1/34.8	85.2/93.8	17.6/21.7	39.5/45.1	45.9/53.7	41.3/44.8	29.0/39.4	42.1/43.1	40.8/47.1
	Clothing-6	24.6/32.9	83.3/92.3	67.3/72.5	38.3/44.2	41.0/48.5	46.3/50.5	28.6/39.9	32.8/34.0	45.3/51.9
	Cold Weapons-4	23.6/31.7	81.9/90.4	48.6/53.4	74.1/81.4	39.6/46.1	31.8/35.1	25.5/35.4	23.2/25.8	43.5/49.9
	Food-28	25.3/34.3	76.9/87.7	35.5/40.1	60.2/67.2	73.6/84.5	28.0/32.4	12.3/17.9	17.1/18.1	41.1/47.8
	Instruments-8	21.8/29.8	71.2/82.6	34.2/37.6	50.1/57.3	71.4/81.9	63.6/69.3	12.2/17.6	15.1/16.3	42.5/49.1
	Plants-25	22.5/30.7	72.5/83.3	35.1/39.1	48.8/54.6	64.2/75.2	54.1/58.5	68.4/88.3	15.8/16.7	47.7/55.8
	Vessels-18	16.9/23.3	69.3/78.8	27.8/29.7	46.1/51.3	60.0/70.2	43.6/48.6	50.5/68.4	79.4/80.3	49.2/56.3
KeepLoRa	Animals-25	43.6/58.9	7.7/11.2	15.8/32.0	28.5/34.5	10.2/15.5	27.9/31.3	3.3/5.0	6.0/7.5	17.9/24.5
	Clothing-6	40.6/56.9	6.1/7.4	14.2/24.6	27.1/34.3	9.3/14.0	24.6/29.5	3.2/4.7	4.4/5.2	16.2/22.1
	Cold Weapons-4	38.5/54.1	5.8/7.1	10.2/20.6	25.9/32.4	8.7/12.8	26.5/32.2	3.0/4.3	3.9/5.0	15.3/21.1
	Food-28	46.0/62.0	7.3/9.1	9.4/21.2	29.1/36.4	16.0/22.2	27.7/32.1	3.6/5.3	4.5/5.6	18.0/24.2
	Instruments-8	40.0/56.0	6.0/7.5	8.4/18.2	23.0/28.7	8.6/13.2	26.0/28.6	3.1/4.6	4.2/5.3	14.9/20.3
	Plants-25	32.2/45.0	7.2/9.5	11.2/24.7	24.8/29.8	8.3/12.9	25.3/29.2	5.3/7.3	4.8/6.2	14.9/20.6
	Vessels-18	40.8/55.9	6.1/7.6	11.9/22.8	22.5/28.6	8.2/13.3	28.0/32.2	3.6/5.2	7.9/9.7	16.1/21.9
NoInDet	Animals-25	46.6/62.0	88.4/95.9	22.4/29.4	36.5/44.8	41.8/49.0	32.3/35.7	20.5/28.2	35.1/37.4	40.5/47.8
	Clothing-6	42.5/56.3	75.9/90.8	59.8/65.4	31.6/36.6	35.9/46.4	30.1/34.1	17.1/25.9	33.5/35.8	40.8/48.9
	Cold Weapons-4	45.2/59.7	76.9/91.5	55.6/63.7	74.4/82.7	44.3/54.5	31.6/34.5	19.2/28.0	37.0/39.3	48.0/56.7
	Food-28	44.8/58.7	81.7/90.4	25.3/30.7	57.2/64.1	72.6/83.0	32.8/37.5	28.9/38.8	38.3/39.6	47.7/55.4
	Instruments-8	44.9/59.1	74.8/88.4	28.0/34.0	62.9/69.4	70.0/81.4	58.5/64.7	26.3/36.8	38.1/39.9	50.4/59.2
	Plants-25	38.4/52.2	80.9/89.2	20.4/27.9	51.9/58.8	66.5/77.0	64.6/73.1	71.2/90.1	41.0/43.4	54.4/64.0
	Vessels-18	39.5/52.5	77.0/88.0	27.9/34.5	50.8/57.1	61.5/73.6	47.4/53.6	66.2/86.0	70.1/72.1	55.1/64.7

Table 7: Robustness to common concept sampling (%).

Method	Protocol	Set 1	Set 2	Set 3	Set 4	Set 5
NoInDet	(i) OVD retention	47.3/60.9	46.8/60.6	47.1/62.0	46.6/60.9	47.0/61.1
	(ii) Current-task acquisition	68.6/77.1	68.7/78.1	68.3/77.6	67.9/76.9	69.1/77.8
	(iii) Final performance	44.1/53.0	43.8/52.8	43.9/53.2	44.5/53.2	44.1/52.7

Table 8: Effect of the adaptation branch on MS COCO and Novel-114 (%).

Metric	Text Only	Vision Only	Text+Vision
OVD retention	**47.1/61.7**	27.4/37.6	25.5/34.9
Current-task acquisition	68.8/77.5	70.6/78.7	**72.3/79.9**
Final performance	**44.3/53.5**	36.0/43.9	40.2/48.7

with RSSD and KPD provides a better observed balance between current-concept acquisition and preservation of the original OVD capability than these visual-side adaptation variants.

A.4 Sensitivity to Task Ordering

We further examine task-order sensitivity by evaluating NoInDet with the original, reverse, and random orders of the same seven concept groups. The original order is shared by all methods in the main experiments and was not selected based on NoInDet's performance. We summarize the results using the same three metrics as in the branch-adaptation study.

Table 9: Sensitivity to task ordering on MS COCO and Novel-114.

Metric	Original	Reverse	Random
OVD retention	47.1/61.7	45.5/59.8	47.8/62.5
Current-task acquisition	68.8/77.5	69.1/77.6	67.9/75.6
Final performance	44.3/53.5	43.0/51.6	45.1/53.8

As shown in Table 9, final performance ranges from 43.0/51.6 to 45.1/53.8 across the three orders. Relative to the original order, the reverse order reduces final performance by 1.3/1.9 points, whereas

Table 10: Task-wise category composition of Novel-114.

Task	Concepts
Animals-25	Rupicola Peruvianus, Alpheus, Okapia Johnstoni, Andrias Japonicus, Dryococelus Australis, Nipponia Nippon, Megadyptes Antipodes, Muntiacus Muntjak, Nephrurus Amyae, Dipodidae, Pacu Fish, Penelopides Panini, Hyalinobatrachium Pellucidum, Varecia Variegata, Moloch Horridus Gray, Uroplatus Phantasticus, Chrysiridia Rhipheus, Tapiridae, Odontodactylus Scyllarus, Litocranius, Lipotes, Opisthocomus Hoazin, Leontopithecus Rosalia, Cebidae, Setonix Brachyurus
Clothing-6	CoatSkirt, RuQun, ZhuZiShenYi, QuJu, ZhiDuo, ZhiJu
Cold Weapons-4	Mao, Ge, Yue, Zu
Food-28	Stew, Dried Fish, Haupia, Three Cup Chicken, Zoni, Chip Butty, Cold Tofu, Brownie, Mung Bean Cake, Zha Jiang Mian, Dry Curry, Mozuku, Namero, Lightly Roasted Fish, Ganmodoki, Kushikatu, Chop Suey, Goya Chanpuru, Kamameshi, Oshiruko, Zeng Cake, Nanbanzuke, Trunip Pudding, Samul, BianZhong, Ruan, KongHou, HuLuSi
Instruments-8	BianQing, Nao, Sheng, Se, Xun, Tao, Duo, Bo
Plants-25	Abutilon Pictum, Conophytum Caculus, Dionaea Muscipula, Saussurea Involucrata, Thysanocarpus Radians, Ochna KirkiiOliv, Conophytum Burgeri, Diplocyclos Palmatus, Ecballium Elaterium, Peristeria Elata Hook, Hedysarum, Cookeina Speciosa, Adenium Obesum, Gentianella Hirculus, Rafflesia Arnoldi, Saussurea Gossipiphora, Davidia Involucrata, Darlingtonia Californica, Centropogon Cornutus, Nepenthes Distillatoria, Sauropus Androgynus, Euphorbia Obesa, Dolichandrone Stipulata, Rhodomyrtus Tomentosa, Lasiosphaera
Vessels-18	Xu, Gui, Li, Ding, Jue, Jia, Yi, Zun, Gu, Gong, You, Bu, Fu, He, Lei, Zhi, Yan, Fou

the random order increases it by 0.8/0.3 points. The reverse order yields the highest current-task acquisition but the lowest OVD retention and final performance, indicating an order-dependent stability–plasticity trade-off. The original order does not yield the highest final score among the evaluated orders, suggesting that the main result is not tied to the most favorable tested sequence.

B More Details of Novel-114

B.1 Task-wise Concept Composition

Table 10 presents the task-wise concept composition of Novel-114. The seven tasks cover diverse concept domains with clear scenario and semantic differences, including animals, clothing, cold weapons, food, instruments, plants, and vessels. These groups span natural objects, man-made objects, specialized concepts, and daily-life concepts, with task sizes ranging from small concept groups to larger ones. Importantly, the task order is not organized by category number, visual difficulty, or semantic similarity, but follows a heterogeneous cross-scenario injection sequence. This design better reflects realistic deployment scenarios, where novel concepts may arrive from different domains and in non-uniform batches. Therefore, Novel-114 not only supports continual concept injection, but also provides a challenging benchmark for evaluating cross-scenario generalization, new concept acquisition, and old knowledge retention.

B.2 Sunburst Visualization

Fig. 4 provides an intuitive visualization of the hierarchical composition and data distribution of Novel-114. The inner ring denotes the seven continual learning tasks, and the outer ring shows their corresponding concepts. The sector area reflects the relative data scale of each task/concept, showing that Novel-114 contains both small and large concept groups. Together with the distinct semantic domains, this non-uniform structure better simulates realistic continual learning scenarios, where novel concepts arrive from different domains and with varying data scales.

B.3 Visualization Examples

We visualize Novel-114 in Fig. 5. The benchmark covers diverse novel concepts across animals, clothing, cold weapons, food, instruments, plants, and vessels, spanning natural objects, man-made objects, specialized concepts, and daily-life concepts.

As shown in the figure, Novel-114 exhibits rich variation in object appearance, scale, pose, and context. Some categories require fine-grained discrimination among visually similar concepts, while others appear in cluttered scenes, unusual viewpoints, or challenging lighting conditions. Small, sparse, and partially occluded instances further increase the difficulty of localization and recognition.

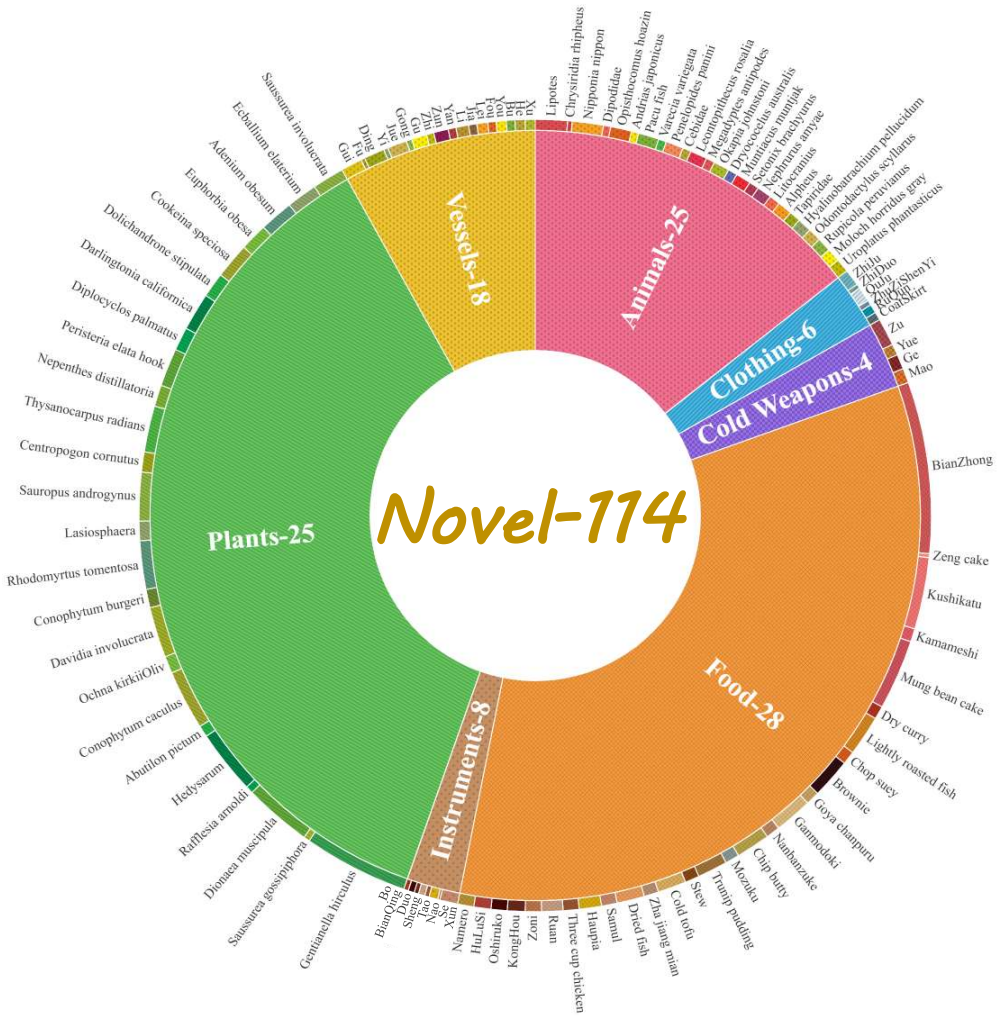


Figure 4: Sunburst visualization of the task-wise composition and data distribution of Novel-114. The inner ring denotes the seven continual learning tasks, the outer ring shows their concepts, and sector areas reflect relative data scales.

These properties make Novel-114 a challenging benchmark for continual novel concept injection: models must acquire genuinely novel concepts, distinguish them from related known semantics, and remain robust across diverse visual scenarios.

B.4 Qualitative Results

We further provide qualitative comparisons in Fig. 6 to validate the effectiveness of our method. Existing methods exhibit different failure patterns under continual novel concept injection. LLMDet and KeepLoRA tend to miss novel objects or produce incomplete detections, indicating insufficient adaptation to newly introduced concepts and unstable retention of historically injected concepts. SGVF shows stronger adaptability to current-task concepts, but often produces redundant detections and false positives, especially on background regions or semantically related objects, suggesting limited stability and knowledge retention.

In contrast, our method produces more accurate and consistent predictions across diverse scenarios. It not only localizes novel objects more precisely and covers object regions more completely, but also avoids spurious or redundant bounding boxes. Its stable behavior across object scales, complex backgrounds, and visually similar categories shows that NoInDet learns more reliable concept-region alignment rather than relying on unstable task-specific cues. These qualitative results demonstrate that our method better balances plasticity and stability, accurately detecting newly injected concepts while preserving historical knowledge and original open-vocabulary capability.

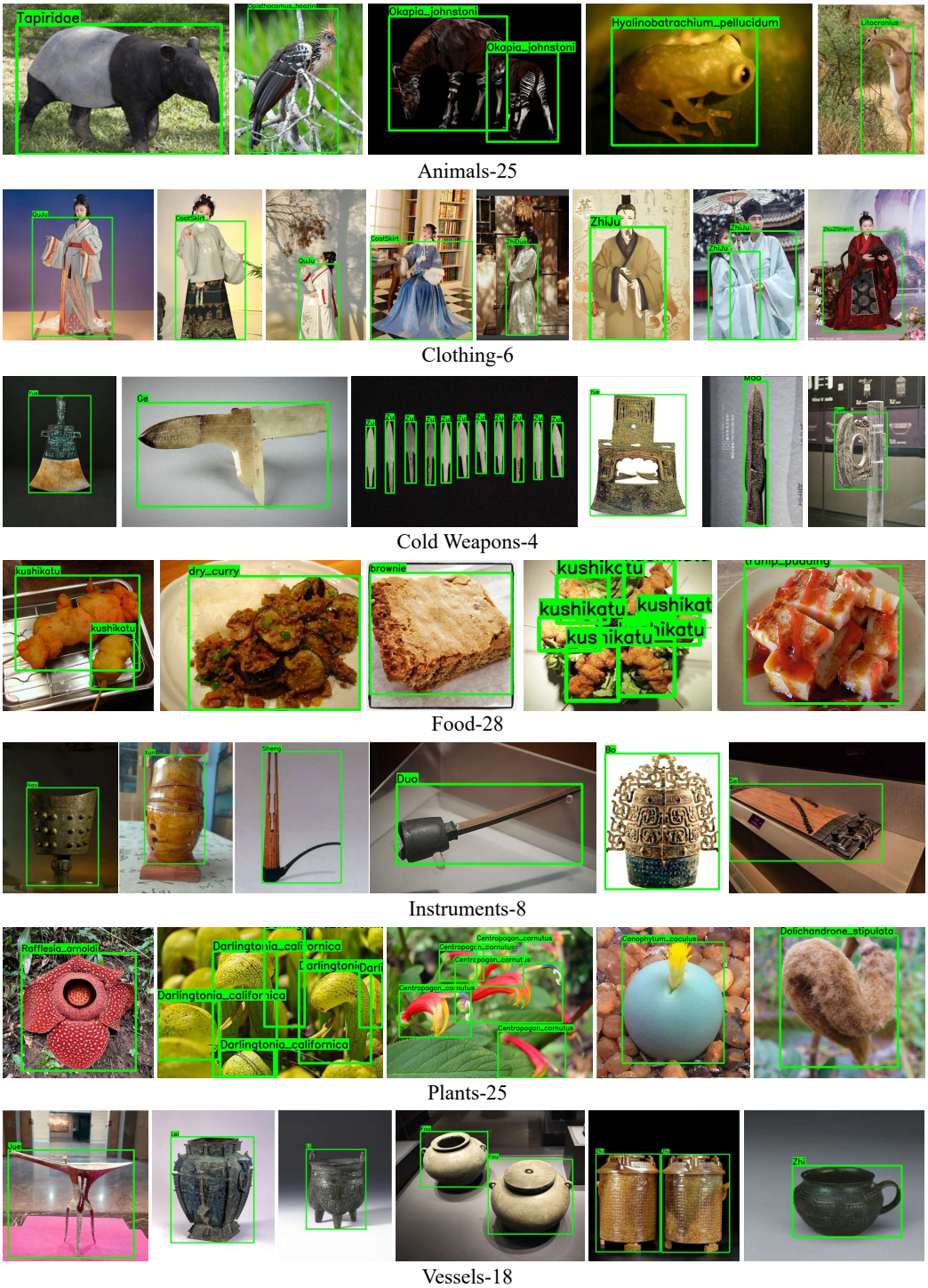


Figure 5: Visualization of Novel-114. The benchmark contains diverse novel, rare, and specialized concepts that are difficult to cover with existing open-vocabulary knowledge, while complex backgrounds and scale variations further increase the detection challenge.



Figure 6: Qualitative comparison of different methods on Novel-114. Our method produces more accurate and stable detections across various scenarios.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [No]

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A].

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.

- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.

- Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.