CaCLIP: Class-Aware CLIP for Replay-Free Multi-Label Class-Incremental Learning

Anonymous authors
Paper under double-blind review

Abstract

Multi-label class-incremental learning (MLCIL) requires a model to learn new classes over a sequence of tasks while preserving previously acquired knowledge. Adapting CLIP to MLCIL poses two challenges. First, a class-shared visual representation and class-specific text embeddings have mismatched granularities, breaking class-wise visual–textual semantic symmetry. Second, missing negative supervision under task-level partial labels leads to a high false-positive rate (FPR). To address these challenges while mitigating catastrophic forgetting, we propose CaCLIP, a replay-free Class-Aware CLIP framework comprising Semantic Evidence Allocation (SEA) and FPR-Adaptive Loss (FAL). SEA uses text-initialized semantic anchors to decouple shared patch-level evidence into dedicated class-specific visual representations, thereby promoting class-wise visual–textual semantic symmetry. Rethinking recent FPR-suppression strategies from a class-aware perspective, FAL estimates class-specific false-positive risk from confidence and adaptively strengthens negative supervision through its classification and distillation components, reducing FPR across current and old label spaces while mitigating catastrophic forgetting. Extensive experiments on PASCAL VOC, MS-COCO, and the real-world NUS-WIDEseq benchmark demonstrate that CaCLIP achieves state-of-the-art replay-free MLCIL performance with a few trainable parameters and minimal inference overhead.

1 Introduction

Class-incremental learning (CIL) aims to learn new classes over a sequence of tasks while preserving the recognition ability for learned classes. However, it is often challenged by catastrophic forgetting (Rebuffi et al., 2017). Most existing CIL studies focus on the single-label setting, where each image is assigned to one category. In many real-world recognition tasks, however, an image usually contains multiple semantic labels, leading to the more challenging multi-label class-incremental learning (MLCIL) (Dong et al., 2023; Du et al., 2024b). In MLCIL, each task provides annotations only for newly introduced classes, although concepts belonging to previous or future tasks may also appear in the image with their labels unobserved. Under this task-level partial-label (TPL) problem (Dong et al., 2023; Du et al., 2024b; 2026), the model must learn new classes from incomplete label spaces while preserving previously acquired knowledge, making it vulnerable to catastrophic forgetting as the label space expands. Meanwhile, many absent classes lack explicit negative supervision under the TPL setting, leading to a high false-positive rate (FPR) (Du et al., 2024b; 2025).

Recently, CLIP (Radford et al., 2021) has shown strong stability and generalization, making it an appealing foundation model for CIL. Existing methods (Huang et al., 2024; Zhao et al., 2024; Huang et al., 2025; He et al., 2026; Dong et al., 2026) adapt CLIP through lightweight prompts or adapters, mainly under the single-label setting. In single-label CIL, a global visual representation and a class-specific text representation form a class-wise semantic match. MLCIL breaks this visual–textual semantic symmetry: one visual representation mixes evidence from multiple classes, whereas each text representation remains class-specific. TPL further leaves much of this entangled evidence unsupervised, causing incorrect cross-class matching and class-dependent false positives.

To address these challenges while mitigating catastrophic forgetting, we propose **CaCLIP**, a replay-free Class-Aware CLIP framework that performs class-aware adaptation at the representation and optimization levels through Semantic Evidence Allocation (SEA) and FPR-Adaptive Loss (FAL),

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

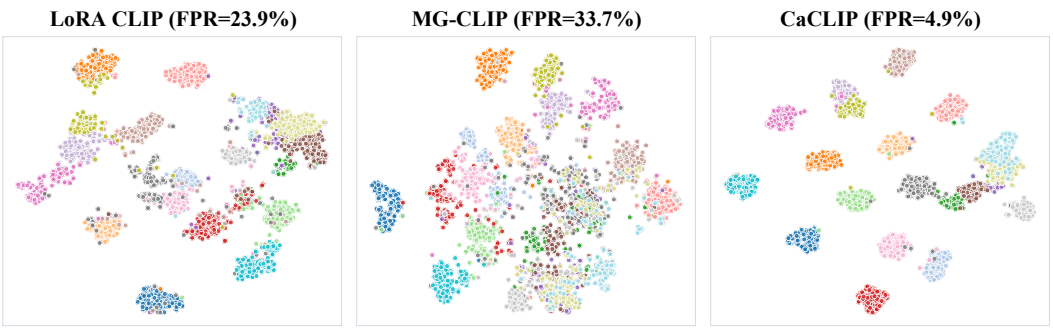

Figure 1: MLCIL t-SNE visualization on VOC B4-C2 after the final task. Each point represents an image–class pair. LoRA CLIP and MG-CLIP use the shared visual feature for each class, whereas CaCLIP uses the corresponding class-aware visual feature. Compared with LoRA CLIP and MG-CLIP, CaCLIP shows clearer class separation and a lower false-positive rate (FPR).

respectively. As shown in Fig. 1, CaCLIP yields clearer class-wise separation than LoRA CLIP (Hu et al., 2022) and MG-CLIP (Huang et al., 2025). The resulting visual representations enable class-wise matching with the inherently class-specific text embeddings, thereby promoting visual–textual semantic symmetry as in the single-label setting. CaCLIP also reduces the FPR under TPL from 23.9% to 4.9%. At the representation level, SEA decouples entangled visual evidence from shared CLIP patch tokens. Lightweight semantic anchors allocate class-relevant patches and aggregate them into dedicated visual representations for individual classes. New classes are introduced with text-initialized anchors, while previous anchors are jointly updated to mitigate catastrophic forgetting. At the optimization level, FAL addresses class-dependent false-positive bias under TPL. It uses class-wise prediction confidence as a proxy for false-positive risk and adaptively strengthens negative supervision. Unlike global confidence regularization (Du et al., 2024b) or positive-label attenuation (Du et al., 2025), FAL provides class-specific FPR control by strengthening negative supervision while leaving the positive loss terms unchanged. Its classification and distillation terms suppress FPR in the current and old label spaces, while also helping mitigate catastrophic forgetting. Extensive experiments demonstrate that CaCLIP outperforms existing methods across diverse MLCIL settings. Moreover, CaCLIP introduces only a few trainable parameters and minimal inference overhead, demonstrating favorable scalability as the label space expands. Our main contributions are summarized as follows:

- We introduce **SEA** to decouple entangled class-shared patch tokens into class-aware visual representations. SEA uses text-initialized semantic anchors to allocate visual evidence among all seen classes, promoting visual–textual semantic symmetry while mitigating catastrophic forgetting with minimal inference overhead.
- We introduce **FAL**, which rethinks FPR suppression as class-adaptive negative learning. FAL uses confidence to estimate class-specific false-positive risk and adaptively strengthens negative supervision, enabling fine-grained FPR control while mitigating forgetting.
- Extensive experiments on PASCAL VOC, MS-COCO, and the real-world NUS-WIDEseq demonstrate that CaCLIP achieves state-of-the-art (SOTA) replay-free MLCIL performance with a few trainable parameters.

2 Related Work

Single-label class-incremental learning (SLCIL). SLCIL has been extensively studied in recent years. Regularization-based methods constrain model updates to reduce forgetting (Li & Hoiem, 2017; Mohamed et al., 2023), while replay-based methods store or synthesize previous samples for later training (Rebuffi et al., 2017; Buzzega et al., 2020; Douillard et al., 2020). With the emergence of pre-trained models, ViT-based (Dosovitskiy et al., 2021) methods have become a strong direction for parameter-efficient CIL. L2P (Wang et al., 2022b) and DualPrompt (Wang et al., 2022a) learn prompt pools and select prompts for inputs. More recently, CLIP-based methods have further improved SLCIL by exploiting vision-language priors. CLAP (Jha et al., 2024) introduces probabilistic

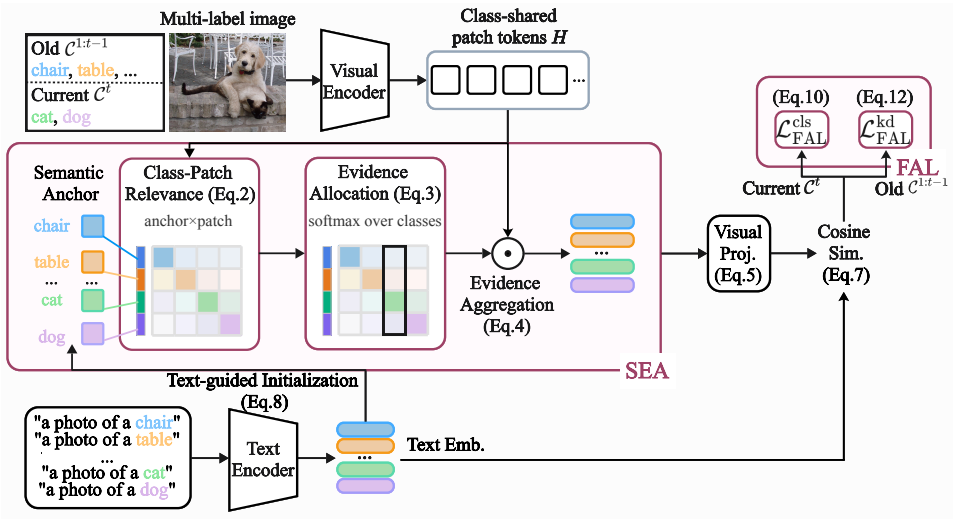

Figure 2: Training framework of CaCLIP. SEA uses text-initialized semantic anchors to allocate class-shared patch tokens among all seen classes and aggregates the allocated evidence into class-aware visual representations before visual projection. FAL optimizes current and old classes through classification and distillation. Only the LoRA parameters and semantic anchors are trainable.

modeling, RAPF (Huang et al., 2024) adds a lightweight projection layer, MG-CLIP (Huang et al., 2025) mitigates the modality gap between visual and textual representations, and SECA (He et al., 2026) further leverages visual prototype refinement for CIL.

Multi-label class-incremental learning (MLCIL). MLCIL is challenging because of the TPL issue. Early replay-based methods such as PRS (Kim et al., 2020) design sampling or memory strategies for MLCIL, and KAR (Cao et al., 2026) proposes a knowledge-aware replay framework. Replay-free methods have also been explored for MLCIL. KRT (Dong et al., 2023) restores and transfers knowledge; CLIN (Ding et al., 2025) is its extension. CSC (Du et al., 2024b) calibrates label relationships and suppresses over-confidences by max-entropy, RebLL (Du et al., 2025) reveals the positive-negative imbalance and proposes asymmetric knowledge distillation (AKD). L3A (Zhang et al., 2025b) introduces analytic continual learning into MLCIL. However, it adopts a different evaluation protocol from others. N-CGCN (Du et al., 2026) models negative correlations. The ViT-based MULTI-LANE (De Min et al., 2024) introduces patch selectors and task prompts. Recent methods also attempt to adapt CLIP to MLCIL. DPA (Zhao et al., 2024) adjusts prompts with replay; we use its non-memory version in replay-free comparison. P2L-CA+ (Dong et al., 2026) uses CLIP to initialize class prompts inside a ViT.

3 Method

3.1 Preliminary

MLCIL consists of a sequence of T tasks. Each task t provides a training set $\mathcal{D}^t_{\text{trn}}$ and a test set $\mathcal{D}^t_{\text{tst}}$ associated with the current class set $\mathcal{C}^t$, where $\mathcal{C}^i \cap \mathcal{C}^j = \varnothing$ for $i \neq j$. After task t, the model predicts over all seen classes $\mathcal{C}^{1:t} = \bigcup_{i=1}^{t} \mathcal{C}^i$.

3.2 Overview

We propose CaCLIP, a replay-free Class-Aware CLIP framework for MLCIL. As illustrated in Fig. 2, CaCLIP consists of two components: **SEA** and **FAL**. The figure also depicts the task-level partial-label (TPL) setting: "cat" and "dog" form the observed current-class set $\mathcal{C}^t$, while "chair", "table", ... form the missing old-class set $\mathcal{C}^{1:t-1}$. Since most missing labels correspond to absent classes, the resulting lack of negative supervision leads to frequent false-positive predictions (Du et al., 2025). Nevertheless, during training at task t, SEA produces predictions for all seen classes $\mathcal{C}^{1:t}$ using both

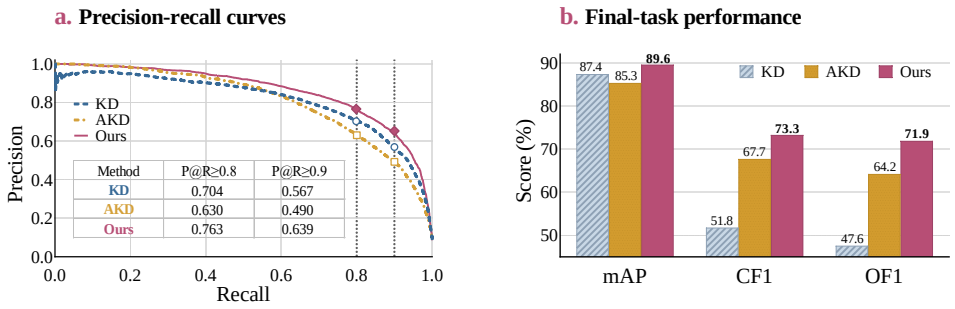

Figure 3: (a) Precision–recall curves and precision at recall ≥ 0.8 and ≥ 0.9. (b) final-task performance on VOC B4-C2. CaCLIP maintains higher precision in the high-recall regime and achieves better mAP and F1 scores than knowledge distillation (KD) and AKD.

current and previous semantic anchors. FAL supervises current classes $\mathcal{C}^t$ with observed labels and old classes $\mathcal{C}^{1:t-1}$ with predictions from the frozen previous-task model. Only the LoRA parameters and semantic anchors are trainable, while all other parameters remain frozen.

3.3 Semantic Evidence Allocation

As shown in Fig. 2, a multi-label image may contain co-occurring concepts from both current ("cat" and "dog") and old ("chair" and "table") classes. CLIP encodes their visual evidence into the same class-shared patch tokens, whereas each text embedding remains class-specific. This visual–textual asymmetry can increase interference among seen classes (shown in Fig. 1) and exacerbate catastrophic forgetting. To this end, **Semantic Evidence Allocation** (SEA) uses per-class semantic anchors to decouple shared patch evidence into class-aware visual representations, preserving old-class evidence and mitigating catastrophic forgetting.

We build CaCLIP upon LoRA CLIP (Hu et al., 2022). Let $E_v(\cdot)$ and $E_t(\cdot)$ denote the CLIP visual and text encoders, respectively. As shown in Fig. 2, given a multi-label image x, the visual encoder produces class-shared patch tokens

$$H = E_v(x) = [h_1, h_2, \ldots, h_N], \qquad h_n \in \mathbb{R}^{d_v}. \tag{1}$$

During training at task t, SEA maintains a learnable semantic anchor $q_c \in \mathbb{R}^{d_v}$ for each seen class $c \in \mathcal{C}^{1:t}$. The relevance between patch n and class c is computed as

$$a_{n,c} = h_n^\top q_c. \tag{2}$$

We normalize the class-patch relevance scores over all seen classes to obtain the evidence-allocation weight $\alpha_{n,c}$:

$$\alpha_{n,c} = \frac{\exp(a_{n,c})}{\sum_{j \in \mathcal{C}^{1:t}} \exp(a_{n,j})}. \tag{3}$$

It represents the proportion of evidence from patch n allocated to class c, allowing different classes to collect relevant evidence from the shared set of patch tokens.

As shown in Fig. 2, the class-aware visual feature is obtained by aggregating its allocated patch-level evidence.

$$\tilde{h}_c = \frac{\sum_{n=1}^{N} \alpha_{n,c} h_n}{\sum_{n=1}^{N} \alpha_{n,c} + \epsilon}, \tag{4}$$

where $\epsilon > 0$ is a small constant for stability. As shown in Fig. 2, SEA is inserted after the CLIP visual encoder and before its visual projection layer. This allows evidence allocation to be performed in the visual token space, while the original projection maps the resulting class-specific features into the pretrained vision-language embedding space, where $P_v(\cdot)$ denotes the CLIP visual projection:

$$z_c = \frac{P_v(\tilde{h}_c)}{\|P_v(\tilde{h}_c)\|_2}. \tag{5}$$

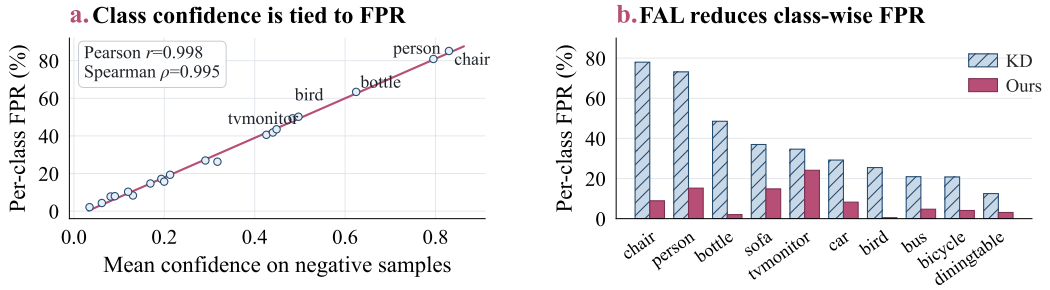

Figure 4: Class-wise FPR analysis on VOC B4-C2. All variants use SEA. (a) The mean confidence on negative samples is strongly correlated with class-wise FPR. (b) FAL substantially reduces class-wise FPR compared with standard KD.

For the textual branch, as shown in Fig. 2, we use a fixed class prompt p_c, such as "a photo of a [CLASS]", and obtain its normalized text embedding:

$$t_c = \frac{E_t(p_c)}{\|E_t(p_c)\|_2}. \tag{6}$$

The prediction for class c is produced by matching its class-specific visual embedding with the corresponding text embedding. Let $\sigma(\cdot)$ denote the sigmoid function. The scaled logit and class probability are

$$s_c = \gamma z_c^\top t_c, \qquad \hat{y}_c = \sigma(s_c), \tag{7}$$

where γ is the frozen pretrained CLIP logit scale. To provide a semantic prior for each newly introduced class, we initialize its semantic anchor using the corresponding CLIP text embedding. Since q_c lies in the pre-projection visual token space while t_c lies in the shared vision-language embedding space, we map t_c to the visual token space through the transpose of the frozen visual projection matrix:

$$q_c^{(0)} = \frac{P_v^\top t_c}{\|P_v^\top t_c\|_2}. \tag{8}$$

At the beginning of task t, only the anchors for newly introduced classes $\mathcal{C}^t$ are initialized from their text embeddings. The inherited anchors preserve previously learned semantic knowledge and class-aware evidence pathways, helping mitigate catastrophic forgetting.

At inference, SEA is retained to construct visual representations for all seen classes. Its post-encoder design requires only one visual encoding pass, yielding low inference cost, while lightweight anchors keep CaCLIP parameter-efficient.

3.4 FPR-Adaptive Loss

Existing false-positive rate (FPR) suppression methods rely on coarse-grained confidence regularization. The max-entropy method (Du et al., 2024b) globally suppresses prediction confidence and is sensitive to its regularization weight. More recently, AKD (Du et al., 2025) applies exponential attenuation to positive-label learning for stronger bias correction. Under the same SEA, Fig. 3(a) shows that AKD yields lower precision than KD in high-recall regime. Fig. 3(b) shows that, although AKD improves CF1 and OF1, it reduces mAP compared with KD. This performance degradation is associated with the rapid attenuation of positive gradients (see FAL analysis in the Appendix).

These results highlight the limitations of existing FPR suppression methods. We instead introduce **FPR-Adaptive Loss** (FAL), which adaptively adjusts negative learning across classes. FAL comprises a classification component $\mathcal{L}_{\text{FAL}}^{\text{cls}}$ for $\mathcal{C}^t$ and a distillation component $\mathcal{L}_{\text{FAL}}^{\text{kd}}$ for $\mathcal{C}^{1:t-1}$:

$$\mathcal{L}_{\text{FAL}} = \lambda \mathcal{L}_{\text{FAL}}^{\text{cls}} + (1-\lambda)\mathcal{L}_{\text{FAL}}^{\text{kd}}, \tag{9}$$

where λ balances the two objectives. All expectations are computed over the current training batch.

FPR-Adaptive classification loss. As shown in Fig. 4(a), class-wise prediction confidence on negative samples is strongly correlated with class-wise FPR, with Pearson and Spearman correlation coefficients of 0.998 and 0.995, respectively. This observation motivates us to use negative prediction confidence to characterize the false-positive risk of each current class. For the current classes $\mathcal{C}^t$, we build on binary cross-entropy (BCE) and adaptively reweight its negative term according to the estimated risk:

$$\mathcal{L}_{\mathrm{FAL}}^{\mathrm{cls}} = -\sum_{c\in\mathcal{C}^t}\left[y_c^t \log \hat{y}_c^t + \alpha w_c^{\mathrm{cls}}(1-y_c^t)\log(1-\hat{y}_c^t)\right], \tag{10}$$

where $y_c^t \in \{0,1\}$ is ground-truth and α controls the negative learning strength. The class-specific weight is defined as

$$w_c^{\mathrm{cls}} = \frac{\mathbb{E}[\hat{y}_c^t \mid y_c^t = 0]}{\frac{1}{|\mathcal{C}^t|}\sum_{j\in\mathcal{C}^t}\mathbb{E}\big[\hat{y}_j^t \mid y_j^t = 0\big] + \epsilon}. \tag{11}$$

Classes with higher mean confidence on negative samples have greater estimated false-positive risk and therefore receive stronger negative supervision. The w_c^{cls} is detached.

FPR-Adaptive distillation loss. For old classes $\mathcal{C}^{1:t-1}$, ground-truth labels are unavailable during training, so we use the frozen teacher's class-wise mean confidence as a label-free proxy for ranking relative FPR risk. Its Pearson/Spearman correlations with class-wise FPR are $0.945/0.934$ on VOC B4-C2 and $0.900/0.881$ on COCO B0-C5, close to the corresponding full-label negative-only oracle correlations of $0.961/0.934$ and $0.920/0.890$. We therefore use class-wise teacher confidence as a relative signal to adapt the negative distillation strength for old classes and formulate the FPR-Adaptive distillation component as

$$\mathcal{L}_{\mathrm{FAL}}^{\mathrm{kd}} = -\sum_{c\in\mathcal{C}^{1:t-1}}\Big[\hat{y}_c^{t-1}\log\hat{y}_c^t + (1+w_c^{\mathrm{kd}})(1-\hat{y}_c^{t-1})\log(1-\hat{y}_c^t)\Big]. \tag{12}$$

The class-specific distillation weight is defined as

$$w_c^{\mathrm{kd}} = \frac{\mathbb{E}\big[\hat{y}_c^{t-1}\big]}{\frac{1}{|\mathcal{C}^{1:t-1}|}\sum_{j\in\mathcal{C}^{1:t-1}}\mathbb{E}\big[\hat{y}_j^{t-1}\big] + \epsilon}. \tag{13}$$

FAL leaves the positive loss terms unchanged while strengthening negative supervision. For soft teacher targets, this changes the effective distillation target, its pointwise optimum and gradient are derived in the FAL Analysis of Appendix. Old classes with higher teacher confidence have higher estimated false-positive risk and thus receive stronger negative distillation constraints. As shown in Fig. 4(b), FAL substantially reduces the FPR of high-risk classes. Overall, SEA constructs class-aware visual representations, while FAL optimizes their predictions according to class-specific false-positive risk. Together, they promote semantic symmetry, reduce FPR, and mitigate forgetting.

4 Experiments

4.1 Experimental Setup

Datasets. We conduct the main experiments on PASCAL VOC 2007 (Everingham et al., 2010) and MS-COCO 2014 (Lin et al., 2014). VOC contains 20 classes, with 5,011 training images and 4,952 test images. COCO contains 80 classes, with 82,081 training images and 40,504 validation images. Following (Du et al., 2024b; 2025), we use the challenging long-sequence settings VOC B4-C2, VOC B5-C3, COCO B0-C5 and COCO B20-C4, where Bx-Cy denotes x base classes and y classes introduced per incremental task. To test CaCLIP in a more realistic setting, we further use NUS-WIDEseq (Kim et al., 2020), a long-tailed benchmark built from the real-world NUS-WIDE dataset (Chua et al., 2009). NUS-WIDEseq contains 49 classes, with 48,724 training images and 2,367 test images.

Table 1: Replay-free MLCIL results on the PASCAL VOC dataset (%). Type S and M denote SLCIL and MLCIL methods, respectively. Results are grouped by backbone families: TResNet, ViT-B/16, and CLIP with ViT-B/16. CaCLIP results are reported as mean ± standard deviation over three runs. The best results are in bold, and the second-best results are underlined.

	Method	Type	VOC B4-C2				VOC B5-C3			
			Avg.	Last			Avg.	Last		
			mAP	mAP	CF1	OF1	mAP	mAP	CF1	OF1
TResNet	KRT	M	71.0	43.6	13.7	31.0	86.2	74.6	39.3	46.6
	CSC	M	83.3	74.1	52.4	42.9	88.0	82.1	62.5	56.6
	RebLL	M	84.6	73.1	57.2	<u>60.0</u>	87.7	79.4	64.6	66.9
	N-CGCN	M	<u>91.1</u>	<u>85.8</u>	<u>67.1</u>	59.9	<u>92.8</u>	<u>88.3</u>	<u>76.3</u>	<u>74.9</u>
ViT	L2P	S	85.8	77.4	54.5	53.6	88.8	81.9	64.0	57.1
	DualPrompt	S	86.5	78.2	43.7	44.7	88.8	81.7	57.3	55.2
	CODA-P	S	87.6	78.6	45.7	44.9	89.3	82.3	60.8	57.8
	MULTI-LANE	M	89.8	82.1	43.2	36.1	92.1	87.1	66.9	56.8
CLIP	CLAP	S	81.7	61.4	46.1	49.4	90.0	78.5	51.9	61.8
	RAPF	S	80.8	66.2	27.3	25.7	83.5	68.4	50.1	46.1
	DPA	M	83.5	70.9	42.8	38.0	86.1	77.7	59.7	55.8
	MG-CLIP	S	85.2	77.9	24.6	25.1	90.7	82.0	35.2	27.2
	P2L-CA+	M	88.2	79.9	-	-	91.3	85.8	-	-
	SECA	S	89.8	81.6	56.0	56.5	90.8	86.7	72.2	73.9
	CaCLIP	M	**94.8** ±0.2	**89.6** ±0.4	**73.3** ±0.4	**71.9** ±0.5	**95.2** ±0.1	**91.2** ±0.1	**82.0** ±0.2	**81.5** ±0.5

Evaluation metrics. Following prior MLCIL studies (Dong et al., 2023; Du et al., 2024b), we use mAP, class-wise F1 (CF1), and overall F1 (OF1). We report Avg. mAP over all incremental tasks and Last mAP, CF1, and OF1 over the cumulative label space after the final task.

Baselines. We compare CaCLIP with representative replay-free methods grouped by backbones. For TResNet-based (Ridnik et al., 2021) MLCIL methods, we include KRT (Dong et al., 2023), CSC (Du et al., 2024b), RebLL (Du et al., 2025), and N-CGCN (Du et al., 2026), where N-CGCN is the SOTA method in long-sequence settings. For pre-trained ViT-based methods, we compare with L2P (Wang et al., 2022b), DualPrompt (Wang et al., 2022a), CODA-P (Smith et al., 2023), and MULTI-LANE (De Min et al., 2024). MULTI-LANE is the SOTA ViT-based MLCIL method. For CLIP-based methods, we include CLAP (Jha et al., 2024), RAPF (Huang et al., 2024), DPA (Zhao et al., 2024), MG-CLIP (Huang et al., 2025), SECA (He et al., 2026), and P2L-CA+ (Dong et al., 2026).

Implementation details. We use ViT-B/16 CLIP as the backbone. Following MG-CLIP (Huang et al., 2025), we insert LoRA modules into both the visual and text encoders. We optimize only the LoRA parameters and our semantic anchors, while freezing all remaining CLIP parameters. We use the same fixed threshold as prior methods (Dong et al., 2023; Du et al., 2024b; 2026). All images are resized to 224×224. We optimize the parameters with Adam ($\beta_1 = 0.9$, $\beta_2 = 0.999$). The batch size is set to 32. Each task is trained for 10 epochs, and the learning rate is 3.0×10^{-3}. We set $\alpha = 8.5$, $\lambda = 0.1$ for COCO and NUS-WIDEseq, and $\lambda = 0.5$ for VOC without test tuning. We use CLIP's frozen pretrained logit scale. The hyperparameter analysis can be found in the Appendix.

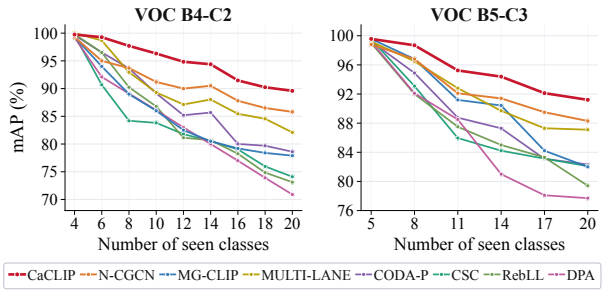

Figure 5: Performance curves during MLCIL.

4.2 Comparison Results

Tables 1 and 2 report the results on PASCAL VOC and MS-COCO. CaCLIP achieves better performance across all four scenarios. On VOC B4-C2, CaCLIP improves over the strong previous method N-CGCN by **3.7%** in Avg. mAP (91.1%→94.8%) and **3.8%** in Last mAP (85.8%→89.6%). Com-

Table 2: Replay-free MLCIL results on the MS-COCO dataset (%).

	Method	Type	MS-COCO B0-C5				MS-COCO B20-C4			
			Avg.	Last			Avg.	Last		
			mAP	mAP	CF1	OF1	mAP	mAP	CF1	OF1
TResNet	KRT	M	63.1	44.5	22.6	37.5	64.0	45.2	17.6	33.0
	CSC	M	75.0	67.3	50.5	47.2	72.6	65.5	48.4	44.2
	RebLL	M	71.7	63.5	53.5	51.9	69.2	60.1	51.3	49.2
	N-CGCN	M	76.4	69.4	56.2	54.2	73.9	68.5	55.0	51.8
ViT	L2P	S	69.8	64.0	44.8	37.0	69.2	65.0	43.9	36.3
	DualPrompt	S	70.7	64.6	45.0	40.7	69.6	65.0	42.0	38.8
	CODA-P	S	71.4	64.3	43.9	38.5	69.5	64.6	43.7	38.5
	MULTI-LANE	M	76.3	70.5	52.5	43.3	76.3	71.0	50.8	40.6
CLIP	CLAP	S	48.8	32.3	40.0	34.2	55.6	38.1	38.6	31.0
	RAPF	S	63.4	53.7	48.1	39.1	62.2	53.3	41.7	35.4
	DPA	M	67.7	59.7	51.3	45.7	67.8	59.2	46.3	43.5
	MG-CLIP	S	69.4	59.2	22.3	22.0	70.6	65.8	46.9	42.6
	SECA	S	72.0	68.0	51.8	54.6	70.8	67.7	51.6	54.8
	CaCLIP	M	**78.7** $\pm$0.1	**71.8** $\pm$0.2	**60.3** $\pm$0.4	**59.1** $\pm$0.5	**78.3** $\pm$0.4	**72.1** $\pm$0.5	**60.1** $\pm$0.3	**58.7** $\pm$0.4

Table 3: Class-incremental results on the real-world NUS-WIDEseq benchmark under two task schedules: a shuffled task order (3, 1, 0, 5, 4, 2) and a sequential task order (0, 1, 2, 3, 4, 5). #P denotes the number of trainable parameters.

Method	#P	Schedule: 3, 1, 0, 5, 4, 2				Schedule: 0, 1, 2, 3, 4, 5			
		Avg.	Last			Avg.	Last		
		mAP	mAP	CF1	OF1	mAP	mAP	CF1	OF1
RebLL	29.4M	51.7	38.4	26.5	25.7	45.2	32.8	19.4	20.6
CSC	34.9M	55.7	44.4	27.3	26.6	50.5	40.4	30.7	26.4
N-CGCN	34.9M	58.3	46.2	29.5	27.6	51.4	42.9	33.7	29.5
MULTI-LANE	0.8M	59.6	52.9	46.3	50.3	63.3	53.8	46.6	54.9
CLAP	9.5M	34.9	23.5	31.7	36.7	36.1	25.4	35.4	38.9
RAPF	0.3M	55.4	48.0	45.9	44.1	59.9	49.7	46.0	39.5
MG-CLIP	0.5M	58.9	49.5	44.6	48.6	61.1	50.0	47.2	48.2
SECA	4.3M	58.0	53.7	45.2	35.1	57.7	54.0	34.6	24.4
CaCLIP	0.5M	**65.4**	**56.9**	**49.7**	**54.0**	**65.1**	**55.8**	**50.4**	**59.3**

pared with the strong CLIP-based baseline P2L-CA+, CaCLIP improves Avg. mAP and Last mAP by **6.6%** and **9.7%**, respectively. On VOC B5-C3, CaCLIP improves over N-CGCN by **2.4%** in Avg. mAP and **2.9%** in Last mAP, while also increasing CF1 and OF1 by **5.7%** and **6.6%**.

On MS-COCO B0-C5, CaCLIP improves over N-CGCN by **2.3%** in Avg. mAP, **2.4%** in Last mAP, **4.1%** in CF1, and **4.9%** in OF1. On MS-COCO B20-C4, CaCLIP surpasses MULTI-LANE by **2.0%** in Avg. mAP, **1.1%** in Last mAP, **9.3%** in CF1, and **18.1%** in OF1.

Fig. 5 shows the performance curves along MLCIL process. CaCLIP maintains higher performance across tasks and exhibits less degradation after new classes are introduced.

4.3 Real-World Results

To examine CLIP-based adaptation in a more realistic MLCIL setting, we further evaluate CaCLIP on NUS-WIDEseq, which contains diverse web images and a naturally long-tailed class distribution. We consider both a shuffled task order $(3, 1, 0, 5, 4, 2)$ and a sequential order $(0, 1, 2, 3, 4, 5)$. As shown in Table 3, CaCLIP achieves better results under both task schedules, demonstrating robustness to varying task orders in real-world MLCIL.

4.4 Ablation and Efficiency

Effect of semantic evidence allocation. Table 4 studies each component of SEA on VOC B4-C2. Starting from the LoRA CLIP baseline (row (a)), adding SEA brings a large improvement (row (b)),

Table 4: SEA ablation on VOC B4-C2 (%). All variants use $\mathcal{L}_{\text{FAL}}^{\text{cls}} + \mathcal{L}_{\text{FAL}}^{\text{kd}}$.

ID Variant	Avg.	Last		
	mAP	mAP	CF1	OF1
(a) LoRA baseline	86.0	76.2	48.2	46.3
(b) + SEA (Post-proj.)	92.6	87.2	67.6	65.8
(c) + SEA (Pre-proj.)	94.1	88.5	70.4	66.9
(d) ++ Text init.	**94.8**	**89.6**	**73.3**	**71.9**

Table 5: FAL ablation on VOC B4-C2 (%). All variants use complete SEA.

ID Variant	Avg.	Last		
	mAP	mAP	CF1	OF1
(a) $\mathcal{L}_{\text{BCE}}^{\text{cls}}$	92.0	86.8	49.5	44.8
(b) $\mathcal{L}_{\text{FAL}}^{\text{cls}}$	93.7	88.3	62.2	57.1
(c) $\mathcal{L}_{\text{FAL}}^{\text{cls}} + \mathcal{L}_{\text{BCE}}^{\text{kd}}$	94.0	88.8	65.0	60.0
(d) $\mathcal{L}_{\text{FAL}}^{\text{cls}} + \mathcal{L}_{\text{FAL}}^{\text{kd}}$	**94.8**	**89.6**	**73.3**	**71.9**

demonstrating the benefit of allocating class-specific visual evidence. Placing SEA before rather than after the CLIP visual projection further improves performance (row (c)), indicating that evidence allocation in the visual token space better preserves the pretrained vision-language alignment. Text-guided initialization of the semantic anchors provides an additional gain (row (d)), confirming that CLIP text semantics offer useful priors. Fig. 6 further shows that SEA allocates stronger patch evidence to regions relevant to each queried class across both VOC and COCO.

Effect of FPR-Adaptive Loss. Table 5 evaluates the classification and distillation components of FAL. Compared with standard BCE classification (row (a)), the FPR-Adaptive classification component consistently improves mAP and F1 scores (row (b)). Adding standard BCE distillation provides further gains (row (c)), while replacing it with FPR-Adaptive distillation substantially improves CF1 and OF1 (row (d)). These results demonstrate that the two components jointly suppress false positives and catastrophic forgetting.

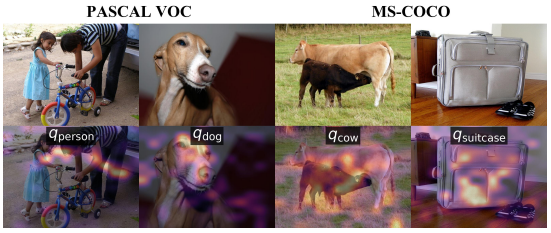

Figure 6: Anchor-guided patch evidence produced by SEA on PASCAL VOC B4-C2 and MS-COCO B0-C5. Each heatmap shows evidence allocated by semantic anchor q_c for the queried class.

Table 6: Computational cost on VOC B4-C2. ΔP/task denotes per-task parameter growth; GFLOPs are measured at inference.

Method	#P↓	ΔP/task↓	GFLOPs↓	mAP↑
CLAP	9.5M	0.5253M	17.8	61.4
DPA	0.3M	0.0338M	17.7	70.9
MG-CLIP	0.5M	0.0010M	17.7	77.9
P2L-CA+	1.1M	0.1144M	18.6	79.9
SECA	4.3M	0.3277M	17.7	81.6
CaCLIP	**0.5M**	**0.0016M**	**17.7**	**89.6**

Computational cost. Table 6 compares trainable parameters, per-task parameter growth, inference GFLOPs, and final mAP. CaCLIP introduces only 0.0016M parameters per new task and keeps per-image inference GFLOPs close to the vanilla CLIP-based method, while achieving better mAP.

Further algorithmic details, gradient analyses, ablation and sensitivity studies, forgetting analysis, comparisons with replay-based methods, results under additional MLCIL scenarios, and FAL theoretical analysis are provided in the **Appendix**.

5 Conclusion

We proposed CaCLIP for replay-free MLCIL under task-level partial labels, where old classes may appear without annotations. SEA uses text-initialized anchors to turn shared CLIP patch evidence into class-specific visual representations. FAL adapts negative learning for current and old classes while retaining positive signals from the frozen teacher, helping limit false positives and forgetting. Experiments on PASCAL VOC, MS-COCO, and NUS-WIDEseq show gains over the evaluated baselines with small parameter growth and a single visual-encoder pass. A limitation is that teacher confidence remains an indirect proxy for old-class false-positive risk. Future work will investigate adaptive LoRA designs tailored to MLCIL.

AI USE STATEMENT

In this work, we used generative AI tools solely to assist with manuscript writing and language polishing, including improving grammar, clarity, readability, and presentation. We did not use generative AI tools for generating research ideas, designing the proposed method, conducting experiments, analyzing experimental results, or drawing scientific conclusions.

ETHICS STATEMENT

This work uses existing benchmark datasets, including PASCAL VOC, MS-COCO, and NUS-WIDEseq, and involves no new data collection, human participants, or user studies. The method targets general-purpose multi-label image recognition rather than consequential decisions about individuals. Potential biases or privacy limitations in the source images may be inherited by our evaluation. We use the datasets according to their stated access terms.

REPRODUCIBILITY STATEMENT

The Method section specifies SEA, FAL, and the training objective. The Experimental Setup describes the datasets, task protocols, backbone, optimization settings, and evaluation procedure. The Appendix provides the training algorithm, additional implementation details, sensitivity analyses, and metric definitions. We will submit an anonymized code package with the paper, including training configurations, data-split and evaluation scripts, and instructions for reproducing the reported results. We commit to releasing the source code publicly after review.

REFERENCES

Pietro Buzzega, Matteo Boschini, Angelo Porrello, Davide Abati, and Simone Calderara. Dark experience for general continual learning: a strong, simple baseline. In *Proceedings of the Advances in Neural Information Processing Systems*, pp. 15920–15930, 2020.

Chengtai Cao, Xinhong Chen, Qun Song, Rui Tan, Yung-Hui Li, and Jianping Wang. Knowledge-aware replay for multi-label class-incremental learning. *Expert Systems with Applications*, 331: 133167, 2026. ISSN 0957-4174.

Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yantao Zheng. Nus-wide: a real-world web image database from national university of singapore. In *Proceedings of the ACM International Conference on Image and Video Retrieval*, pp. 1–9, 2009.

Thomas De Min, Massimiliano Mancini, Stéphane Lathuilière, Subhankar Roy, and Elisa Ricci. Less is more: Summarizing patch tokens for efficient multi-label class-incremental learning. In *Proceedings of the third Conference on Lifelong Learning Agents*, 2024.

Chenhao Ding, Songlin Dong, Zhengdong Zhou, Jizhou Han, Qiang Wang, Yuhang He, and Yihong Gong. Class-independent increment: An efficient approach for multi-label class-incremental learning. *arXiv preprint arXiv:2503.00515*, 2025.

Songlin Dong, Haoyu Luo, Yuhang He, Xing Wei, Jie Cheng, and Yihong Gong. Knowledge restore and transfer for multi-label class-incremental learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 18711–18720, 2023.

Songlin Dong, Jiangyang Li, Chenhao Ding, Zhiheng Ma, Haoyu Luo, Yuhang He, and Yihong Gong. P2l-ca: An effective parameter tuning framework for rehearsal-free multi-label class-incremental learning. *arXiv preprint arXiv:2601.12714*, 2026.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of the International Conference on Learning Representations*, 2021.

Arthur Douillard, Matthieu Cord, Charles Ollion, Thomas Robert, and Eduardo Valle. Podnet: Pooled outputs distillation for small-tasks incremental learning. In *Proceedings of the European Conference on Computer Vision*, pp. 86–102, 2020.

Kaile Du, Fan Lyu, Linyan Li, Fuyuan Hu, Wei Feng, Fenglei Xu, Xuefeng Xi, and Hanjing Cheng. Multi-label continual learning using augmented graph convolutional network. *IEEE Transactions on Multimedia*, 26:2978–2992, 2024a.

Kaile Du, Yifan Zhou, Fan Lyu, Yuyang Li, Chen Lu, and Guangcan Liu. Confidence self-calibration for multi-label class-incremental learning. In *European Conference on Computer Vision*, pp. 234–252, 2024b.

Kaile Du, Yifan Zhou, Fan Lyu, Yuyang Li, Junzhou Xie, Yixi Shen, Fuyuan Hu, and Guangcan Liu. Rebalancing multi-label class-incremental learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 39, pp. 16372–16380, 2025.

Kaile Du, Junzhou Xie, Fan Lyu, Yifan Zhou, Zihan Ye, Wei Li, Yuyang Li, and Guangcan Liu. Negative-weighted knowledge distillation regularized graph convolutional network for multi-label class-incremental learning. *Pattern Recognition*, pp. 113657, 2026.

Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2): 303–338, 2010.

Lingfeng He, De Cheng, Di Xu, Huaijie Wang, and Nannan Wang. Harnessing textual semantic priors for knowledge transfer and refinement in clip-driven continual learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 21645–21653, 2026.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2):3, 2022.

Ping Hu, Ximeng Sun, Stan Sclaroff, and Kate Saenko. Dualcoop++: Fast and effective adaptation to multi-label recognition with limited annotations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(5):3450–3462, 2023.

Linlan Huang, Xusheng Cao, Haori Lu, and Xialei Liu. Class-incremental learning with clip: Adaptive representation adjustment and parameter fusion. In *European Conference on Computer Vision*, pp. 214–231. Springer, 2024.

Linlan Huang, Xusheng Cao, Haori Lu, Yifan Meng, Fei Yang, and Xialei Liu. Mind the gap: Preserving and compensating for the modality gap in clip-based continual learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3777–3786, 2025.

Saurav Jha, Dong Gong, and Lina Yao. Clap4clip: Continual learning with probabilistic finetuning for vision-language models. *Advances in neural information processing systems*, 37:129146–129186, 2024.

Chris Dongjoo Kim, Jinseo Jeong, and Gunhee Kim. Imbalanced continual learning with partitioning reservoir sampling. In *Proceedings of the European Conference on Computer Vision*, pp. 411–428, 2020.

Z Li and D Hoiem. Learning without forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(12):2935–2947, 2017.

Yan-Shuo Liang and Wu-Jun Li. Optimizing class distribution in memory for multi-label online continual learning. *arXiv preprint arXiv:2209.11469*, 2022.

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 *Proceedings of the European Conference on Computer Vision*, pp. 740–755, 2014.

Abdelrahman Mohamed, Rushali Grandhe, KJ Joseph, Salman Khan, and Fahad Khan. D3former: Debiased dual distilled transformer for incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2420–2429, 2023.

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 *Proceedings of the International Conference on Machine Learning*, pp. 8748–8763, 2021.

Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2001–2010, 2017.

Tal Ridnik, Hussam Lawen, Asaf Noy, Emanuel Ben Baruch, Gilad Sharir, and Itamar Friedman. Tresnet: High performance gpu-dedicated architecture. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1400–1409, 2021.

David Rolnick, Arun Ahuja, Jonathan Schwarz, Timothy Lillicrap, and Gregory Wayne. Experience replay for continual learning. In *Proceedings of the Advances in Neural Information Processing Systems*, pp. 350–360, 2019.

James Seale Smith, Leonid Karlinsky, Vyshnavi Gutta, Paola Cascante-Bonilla, Donghyun Kim, Assaf Arbelle, Rameswar Panda, Rogerio Feris, and Zsolt Kira. Coda-prompt: Continual decomposed attention-based prompting for rehearsal-free continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11909–11919, 2023.

Ximeng Sun, Ping Hu, and Kate Saenko. Dualcoop: Fast adaptation to multi-label recognition with limited annotations. *Advances in Neural Information Processing Systems*, 35:30569–30582, 2022.

Xiaoyu Tao, Xinyuan Chang, Xiaopeng Hong, Xing Wei, and Yihong Gong. Topology-preserving class-incremental learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIX 16*, pp. 254–270. Springer, 2020.

Xinrui Wang, Shao-yuan Li, Jiaqiang Zhang, and Songcan Chen. Cut out and replay: A simple yet versatile strategy for multi-label online continual learning. In *Proceedings of the International Conference on Machine Learning*, 2025.

Zifeng Wang, Zizhao Zhang, Sayna Ebrahimi, Ruoxi Sun, Han Zhang, Chen-Yu Lee, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, et al. Dualprompt: Complementary prompting for rehearsal-free continual learning. In *European Conference on Computer Vision*, pp. 631–648. Springer, 2022a.

Zifeng Wang, Zizhao Zhang, Chen-Yu Lee, Han Zhang, Ruoxi Sun, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, and Tomas Pfister. Learning to prompt for continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 139–149, 2022b.

Yue Wu, Yinpeng Chen, Lijuan Wang, Yuancheng Ye, Zicheng Liu, Yandong Guo, and Yun Fu. Large scale incremental learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 374–382, 2019.

Aoting Zhang, Dongbao Yang, Chang Liu, Xiaopeng Hong, and Yu Zhou. Specifying what you know or not for multi-label class-incremental learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 22345–22353, 2025a.

Xiang Zhang, Run He, Chen Jiao, Di Fang, Ming Li, Ziqian Zeng, Cen Chen, and Huiping Zhuang. L3a: Label-augmented analytic adaptation for multi-label class incremental learning. In *Proceedings of the International Conference on Machine Learning*, pp. 74938–74949, 2025b.

Haifeng Zhao, Yuguang Jin, and Leilei Ma. Dynamic prompt adjustment for multi-label class-incremental learning. In *International Conference on Brain Inspired Cognitive Systems*, pp. 52–61. Springer, 2024.

APPENDIX

This appendix provides additional details and analyses for CaCLIP. We first present the complete training algorithm and explain the incremental optimization procedure. We then report results under additional MLCIL scenarios and analyze forgetting, parameter sensitivity, and the class-wise designs in SEA and FAL. Finally, we analyze the classification gradients and distillation objective of FAL, provide formal definitions of the evaluation metrics.

A TRAINING PROCEDURE

Algorithm 1 summarizes the replay-free training procedure of CaCLIP. Unless otherwise specified, the LoRA configurations, including the target layers, scaling factor, and dropout, follow MG-CLIP (Huang et al., 2025). The LoRA parameters are initialized and the semantic anchor set is created in line 1. At task t, the seen label space is expanded in line 3. For $t > 1$, the previous model is copied and frozen as the teacher (lines 4–6). Each newly introduced class is assigned a semantic anchor initialized using Eq. 8 (lines 7–9).

Algorithm 1 Training CaCLIP

Require: $\{\mathcal{D}^t_{\text{trn}}, \mathcal{C}^t\}_{t=1}^{T}$, pretrained CLIP, α, and λ
Ensure: Final model f^T

1: Initialize LoRA parameters Θ and semantic anchors $\mathcal{Q}$
2: **for** $t = 1, \ldots, T$ **do**
3: $\quad\mathcal{C}^{1:t} \leftarrow \mathcal{C}^{1:t-1} \cup \mathcal{C}^t$
4: $\quad$**if** $t > 1$ **then**
5: $\quad\quad\bar{f}^{t-1} \leftarrow f^{t-1}$
6: $\quad$**end if**
7: $\quad$**for** $c \in \mathcal{C}^t$ **do**
8: $\quad\quad$Initialize semantic anchor q_c using Eq. (8)
9: $\quad$**end for**
10: $\quad$**for** each mini-batch $(x, \mathbf{y}^t) \sim \mathcal{D}^t_{\text{trn}}$ **do**
11: $\quad\quad H \leftarrow E_v(x)$
12: $\quad\quad\hat{\mathbf{y}}^t \leftarrow \text{SEA}(H, \mathcal{Q})$
13: $\quad\quad\mathcal{L}_{\text{cls}} \leftarrow \mathcal{L}^{\text{cls}}_{\text{FAL}}(\hat{\mathbf{y}}^t, \mathbf{y}^t)$
14: $\quad\quad$**if** $t > 1$ **then**
15: $\quad\quad\quad\hat{\mathbf{y}}^{t-1} \leftarrow \bar{f}^{t-1}(x)$
16: $\quad\quad\quad\mathcal{L}_{\text{kd}} \leftarrow \mathcal{L}^{\text{kd}}_{\text{FAL}}(\hat{\mathbf{y}}^t, \hat{\mathbf{y}}^{t-1})$
17: $\quad\quad\quad\mathcal{L} \leftarrow \lambda\mathcal{L}_{\text{cls}} + (1-\lambda)\mathcal{L}_{\text{kd}}$
18: $\quad\quad$**else**
19: $\quad\quad\quad\mathcal{L} \leftarrow \mathcal{L}_{\text{cls}}$
20: $\quad\quad$**end if**
21: $\quad\quad$Update $(\Theta, \mathcal{Q})$ using $\nabla\mathcal{L}$
22: $\quad$**end for**
23: **end for**

For each mini-batch, the LoRA-adapted visual encoder produces patch tokens H (line 11), from which SEA allocates class-aware visual evidence and predicts all seen classes according to Eqs. 1–7 (line 12). The FPR-Adaptive classification loss is computed for current classes using Eqs. 10–11 (line 13). The case where a current class has no negative instance in a mini-batch is negligibly rare in practice. When old classes exist, the frozen teacher provides their soft predictions (line 15) for computing the FPR-Adaptive distillation loss in Eqs. 12–13 (line 16). The two components are combined following Eq. 9 (line 17); for the first task, only the classification component is used (line 19). Here, α controls the negative-learning strength, while λ balances classification and distillation. The FAL implementation uses mean reduction. Finally, only the LoRA parameters and semantic anchors are updated (line 21), with the remaining pretrained CLIP parameters frozen.

Validation of the label-free risk proxy. Our goal is not to estimate the absolute FPR without bias, but to rank old classes by their relative false-positive risk, because FAL uses normalized class-wise weights. On VOC B4-C2 and COCO B0-C5, the label-free teacher mean confidence achieves Pearson/Spearman correlations of $0.945/0.934$ and $0.900/0.881$ with class-wise FPR, respectively. These results closely match the corresponding full-label negative-only oracles of $0.961/0.934$ and $0.920/0.890$. This result indicates that including unlabeled positives has little effect on the relative risk ranking. The oracle is used only for post-hoc analysis and is never accessed during training. Post-hoc per-class threshold calibration is not directly applicable to replay-free MLCIL because old-class validation labels are unavailable. We therefore compare FAL with feasible training-time confidence-control approaches, including CSC/Max-entropy (Du et al., 2024b), RebLL/AKD (Du et al., 2025), and standard KD.

Relation to CLIP-based multi-label recognition. DualCoOp (Sun et al., 2022) and Dual-CoOp++ (Hu et al., 2023) adapt CLIP to multi-label recognition with learnable text prompts

Table 7: Multi-label class-incremental results on MS-COCO dataset (%). A memory size of 0 indicates replay-free. Type S and M denote SLCIL and MLCIL methods. For single-label baselines, we replace softmax with BCE for multi-label training, while following their original settings otherwise. Results are grouped by backbone families: TResNet, ViT-B/16, and CLIP with ViT-B/16. All results of our method are averaged over three runs.

	Method	Type	Memory	MS-COCO B40-C10				MS-COCO B0-C10			
				Avg.	Last			Avg.	Last		
				mAP	mAP	CF1	OF1	mAP	mAP	CF1	OF1
TResNet	TPCIL (Tao et al., 2020)	S		63.1	53.1	25.3	25.1	63.8	50.8	20.1	21.6
	PODNet (Douillard et al., 2020)	S	5/class	65.4	57.8	24.2	23.4	65.7	53.4	13.6	17.3
	DER++ (Buzzega et al., 2020)	S		69.6	59.0	41.9	43.7	68.1	54.6	33.3	36.7
	iCaRL (Rebuffi et al., 2017)	S		65.6	55.7	22.1	25.5	59.7	43.8	19.3	22.8
	BiC (Wu et al., 2019)	S		65.5	55.9	38.1	40.7	65.0	51.1	31.0	38.1
	ER (Rolnick et al., 2019)	S	20/class	68.9	61.6	58.6	61.1	60.3	47.2	40.6	43.6
	PODNet (Douillard et al., 2020)	S		71.0	64.2	46.6	42.1	70.0	58.8	45.2	48.7
	DER++ (Buzzega et al., 2020)	S		73.6	66.3	51.5	53.5	72.7	63.1	45.2	48.7
	PRS (Kim et al., 2020)	M		50.8	33.2	9.3	15.1	48.8	27.9	8.5	14.7
	OCDM (Liang & Li, 2022)	M	1000	51.3	34.0	9.5	15.5	49.5	28.5	8.6	14.9
	CUTER (Wang et al., 2025)	M		-	-	-	-	60.1	47.8	35.9	39.2
	AGCN (Du et al., 2024a)	M		73.9	69.1	58.7	59.9	72.4	61.4	53.9	56.6
	KRT (Dong et al., 2023)	M		77.8	74.0	64.4	63.4	74.6	65.9	55.6	56.5
	CSC (Du et al., 2024b)	M	0	78.2	75.0	65.7	67.0	78.0	72.8	64.9	66.8
	HCP (Zhang et al., 2025a)	M		78.9	75.3	64.9	68.6	77.9	71.2	60.4	65.3
	RebLL (Du et al., 2025)	M		76.4	72.5	60.7	64.9	77.2	70.4	60.9	63.8
	N-CGCN (Du et al., 2026)	M		79.2	76.3	68.1	70.4	78.3	73.5	65.2	68.1
ViT	L2P (Wang et al., 2022b)	S		73.7	71.1	62.8	62.3	73.4	68.0	58.8	56.0
	DualPrompt (Wang et al., 2022a)	S	0	74.3	71.4	63.0	64.3	74.1	68.6	55.3	56.3
	CODA-P (Smith et al., 2023)	S		73.0	73.0	65.4	66.6	75.8	70.5	59.0	59.4
	MULTI-LANE (De Min et al., 2024)	M		78.9	76.6	65.9	66.5	79.4	74.3	65.7	64.8
CLIP	CLAP (Jha et al., 2024)	S		55.8	56.2	62.9	62.5	65.2	55.9	62.3	66.9
	RAPF (Huang et al., 2024)	S		70.8	64.1	58.0	59.7	69.3	59.3	54.4	54.0
	MG-CLIP (Huang et al., 2025)	S		74.8	71.0	61.3	65.2	77.8	70.4	60.8	63.7
	SECA (He et al., 2026)	S		73.3	69.4	57.4	59.4	71.8	70.1	59.5	61.2
	DPA (Zhao et al., 2024)	M		81.1	77.2	68.2	68.5	80.7	73.3	65.1	65.1
	CaCLIP	M		**82.7**	**79.9**	**71.0**	**72.3**	**82.9**	**78.0**	**68.7**	**71.0**

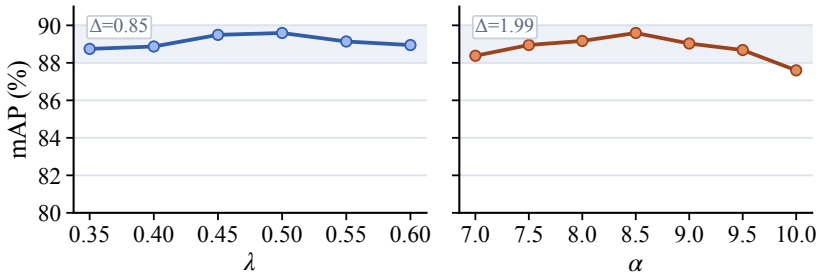

Figure 7: Sensitivity of λ and α on VOC B4-C2. CaCLIP remains stable across different settings.

and class-conditioned aggregation of spatial evidence. DualCoOp++ additionally uses evidential prompts and Winner-Take-All regularization to model inter-class interactions. SEA instead constructs class-aware visual features with learnable anchors in the pre-projection visual token space. Text-initialized anchors allocate patch evidence through a softmax over seen classes; as new classes arrive, new anchors are introduced while existing anchors are carried forward and jointly updated. Thus, SEA applies class-aware aggregation within the task-level partial-label, replay-free incremental setting.

B Additional Experimental Results

B.1 Results under Additional MLCIL Scenarios

Table 8: Multi-label class-incremental results on the PASCAL VOC dataset (%).

	Method	Type	Memory	VOC B0-C4				VOC B10-C2			
				Avg.	Last			Avg.	Last		
				mAP	mAP	CF1	OF1	mAP	mAP	CF1	OF1
TResNet	ER (Rolnick et al., 2019)	S	2/class	82.6	67.7	47.2	47.2	78.8	64.3	40.8	37.6
	PODNet (Douillard et al., 2020)	S		84.1	72.3	52.0	50.3	80.2	66.2	54.3	40.8
	DER++ (Buzzega et al., 2020)	S		84.9	74.4	59.8	56.4	81.6	67.1	55.5	45.9
	CUTER (Wang et al., 2025)	M	1000	82.1	67.9	51.4	60.0	-	-	-	-
	OCDM (Liang & Li, 2022)	M		76.1	45.4	36.4	40.5	-	-	-	-
	CSC (Du et al., 2024b)	M	0	90.4	85.1	67.7	62.2	89.0	83.8	62.0	47.2
	HCP (Zhang et al., 2025a)	M		92.9	87.9	-	-	90.1	81.9	-	-
	RebLL (Du et al., 2025)	M		91.4	85.1	72.6	75.8	88.8	80.9	64.1	66.1
	N-CGCN (Du et al., 2026)	M		94.1	90.0	80.7	80.6	92.5	87.3	63.2	60.9
ViT	L2P (Wang et al., 2022b)	S	0	89.1	82.7	55.2	51.2	88.3	83.4	51.2	40.5
	DualPrompt (Wang et al., 2022a)	S		89.9	83.4	60.6	58.9	88.5	83.0	57.5	52.2
	CODA-P (Smith et al., 2023)	S		90.3	84.0	65.5	62.1	89.1	83.4	59.8	54.0
	MULTI-LANE (De Min et al., 2024)	M		93.5	88.8	74.8	69.9	93.0	88.0	63.8	49.4
CLIP	CLAP (Jha et al., 2024)	S		77.9	68.3	75.6	72.8	66.6	60.3	68.6	62.5
	RAPF (Huang et al., 2024)	S		84.6	71.2	67.7	66.6	80.3	68.7	39.2	38.5
	MG-CLIP (Huang et al., 2025)	S		92.5	85.4	72.7	64.4	90.8	86.4	52.8	53.5
	SECA (He et al., 2026)	S		92.9	88.0	75.1	77.9	91.7	87.0	52.5	54.7
	DPA (Zhao et al., 2024)	M		93.5	88.1	-	-	88.5	79.0	-	-
	CaCLIP	M		**95.6**	**91.8**	**84.1**	**83.9**	**94.0**	**90.9**	**75.7**	**71.5**

Table 9: Comparison of forgetting on VOC B4-C2 (%).

Method	Avg. mAP ↑	Last mAP ↑	Forgetting ↓	FPR ↓
MG-CLIP	85.2	77.9	1.0	33.7
SECA	89.8	81.6	1.6	10.8
CaCLIP	**94.8**	**89.6**	**0.2**	**4.9**

To complement the long-sequence settings reported in the main paper, we evaluate CaCLIP under four additional MLCIL scenarios. Tables 7 and 8 present results on MS-COCO B40-C10 and B0-C10, and PASCAL VOC B0-C4 and B10-C2, respectively.

The compared baseline results are either taken from prior work (Du et al., 2024b; Zhao et al., 2024; Du et al., 2026), or reproduced using the officially released implementations. For SLCIL baselines, we replace softmax with BCE while otherwise following their official settings.

On COCO B40-C10, CaCLIP exceeds the strongest competing results by 1.6, 2.7, 2.8, and 1.9 points in Avg. mAP, Last mAP, CF1, and OF1, respectively. On COCO B0-C10, the corresponding improvements are 2.2, 3.7, 3.0, and 2.9 points. CaCLIP achieves these gains without storing any previous samples, and also outperforms several replay-based approaches.

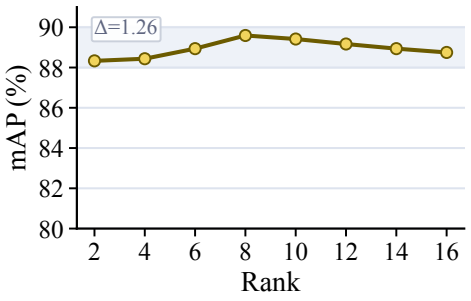

Figure 8: Sensitivity of LoRA rank on VOC B4-C2. CaCLIP shows stable performance across ranks.

CaCLIP similarly achieves the best results across all metrics on both additional PASCAL VOC settings. It improves over the strongest competing results by 1.5, 1.8, 3.4, and 3.3 points on VOC B0-C4, and by 1.0, 2.9, 7.1, and 5.4 points on VOC B10-C2. These results demonstrate that CaCLIP generalizes consistently across different base-class sizes, incremental step sizes, datasets, and task lengths.

B.2 Forgetting

As shown in Table 9, compared with advanced CLIP-based methods, including MG-CLIP (Huang et al., 2025) and SECA (He et al., 2026), CaCLIP achieves less forgetting and a lower FPR. This

demonstrates its stronger ability to preserve previously learned knowledge while suppressing false-positive predictions.

B.3 Comparison with Replay-Based Methods

For replay-based comparison, we include PODNet (Douillard et al., 2020), DER++ (Buzzega et al., 2020), CSC (Du et al., 2024b), RebLL (Du et al., 2025), CLIP-based DPA (Zhao et al., 2024) and KAR (Cao et al., 2026), where KAR and DPA represent the current SOTA replay-based MLCIL methods. Table 10 reports the comparison on COCO B40-C10 and B0-C10 where replay-based methods are commonly evaluated. Although CaCLIP does not use any replay memory, it consistently outperforms all replay-based competitors.

Table 10: mAP (%) comparison with replay-based methods. Replay-based methods store 20 samples per class, whereas CaCLIP is replay-free.

		B40-C10		B0-C10	
Method	**Memory**	**Avg.**	**Last**	**Avg.**	**Last**
PODNet	20/class	71.0	64.2	70.0	58.8
DER++		73.6	66.3	72.7	63.1
CSC		78.7	76.7	79.6	74.8
RebLL		78.3	76.0	78.6	74.0
DPA		<u>81.5</u>	78.4	<u>81.7</u>	76.1
KAR		81.4	<u>78.9</u>	81.2	<u>76.8</u>
CaCLIP	0	**82.7**	**79.9**	**82.9**	**78.0**

B.4 Parameter Sensitivity

We investigate the sensitivity of CaCLIP to the loss-balancing coefficient λ, the negative-learning coefficient α, and the LoRA rank. As shown in Fig. 7, the mAP varies by 0.85 points across the evaluated values of λ and by 1.99 points across those of α. We select $\lambda = 0.5$ and $\alpha = 8.5$. The sensitivity plots are diagnostic analyses and were not used for test-set model selection. Overall, CaCLIP is not highly sensitive to either coefficient within the evaluated ranges.

Fig. 8 further evaluates different LoRA ranks from 2 to 16. The performance varies by only 1.26 points. CaCLIP therefore remains effective with low-rank adaptation and is robust to the model capacity introduced by LoRA. We use a LoRA rank of 8.

B.5 Effect of Class-Wise Modeling

Table 13 examines the class-wise designs in SEA and FAL. For SEA, we replace class-softmax with independent sigmoid gates while keeping FAL and all other components unchanged. Class-softmax jointly normalizes the assignment scores over all seen classes and consistently improves performance, demonstrating its benefit for fine-grained visual evidence allocation. For FAL, we replace the class-specific weights with a shared global weight while retaining the complete SEA. Class-wise weighting notably improves the performances, confirming the benefit of adapting negative learning to class-specific false-positive risks.

B.6 Effect of SEA normalization

We compare SEA with a class-query aggregation variant that normalizes class-patch relevance over patches instead of seen classes. Both variants use the same text-initialized anchors, projection position, FAL, and training configuration; only the softmax dimension differs. As shown in Table 14, class-wise normalization improves final-task mAP and F1 scores. This indicates that SEA's evidence-allocation rule contributes beyond class-aware aggregation alone.

B.7 Ablations on MS-COCO

We further evaluate SEA and FAL on COCO B0-C5. Table 11 compares alternative visual representation designs under the same FAL objective. SEA consistently outperforms the LoRA baseline under the same FAL objective, with pre-projection aggregation and text initialization providing further improvements. With SEA fixed, FAL classification improves upon BCE, while FAL distillation further enhances mAP and F1 scores over standard distillation. These trends agree with the VOC

Table 11: SEA ablation on COCO B0-C5 (%). All variants use $\mathcal{L}_{\text{FAL}}^{\text{cls}} + \mathcal{L}_{\text{FAL}}^{\text{kd}}$.

ID Variant	Avg.	Last		
	mAP	mAP	CF1	OF1
(a) LoRA baseline	71.7	62.8	43.3	44.0
(b) + SEA (Post-proj.)	76.9	69.1	50.0	46.9
(c) + SEA (Pre-proj.)	77.7	70.0	56.9	50.6
(d) ++ Text init.	**78.7**	**71.8**	**60.3**	**59.1**

Table 12: FAL ablation on COCO B0-C5 (%). All variants use complete SEA.

ID Variant	Avg.	Last		
	mAP	mAP	CF1	OF1
(a) $\mathcal{L}_{\text{BCE}}^{\text{cls}}$	73.0	62.2	48.3	43.6
(b) $\mathcal{L}_{\text{FAL}}^{\text{cls}}$	73.9	63.7	50.5	49.1
(c) $\mathcal{L}_{\text{FAL}}^{\text{cls}} + \mathcal{L}_{\text{BCE}}^{\text{kd}}$	77.6	70.9	56.6	54.9
(d) $\mathcal{L}_{\text{FAL}}^{\text{cls}} + \mathcal{L}_{\text{FAL}}^{\text{kd}}$	**78.7**	**71.8**	**60.3**	**59.1**

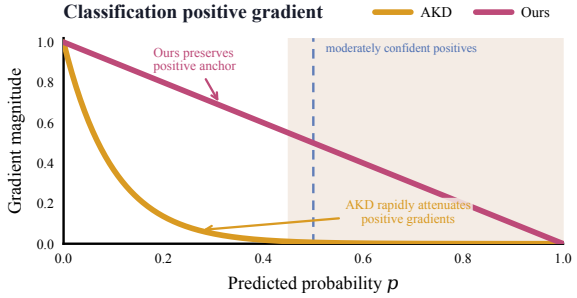

Figure 9: Positive-gradient comparison between AKD and FAL for hard-positive classification targets ($y_c^t = 1$). FAL retains the standard BCE positive gradient, whereas AKD attenuates the positive gradient.

results, supporting the effectiveness of the proposed components across datasets and label-space sizes.

C Analysis of FPR-Adaptive Loss

We analyze the classification and distillation losses in Eqs. 10 and 12. Both losses are sums over the corresponding classes for a single example, before batch reduction and the outer loss-balancing coefficients. At stage t, the student probability is $\hat{y}_c^t = \sigma(s_c)$.

C.1 FPR-Adaptive Classification

For a current class $c \in \mathcal{C}^t$ with $y_c^t = 1$, its negative term in Eq. 10 vanishes. Treating the detached weights w_c^{cls} as constants, the partial derivative with respect to the corresponding student logit is

$$\left.\frac{\partial \mathcal{L}_{\text{FAL}}^{\text{cls}}}{\partial s_c}\right|_{y_c^t=1} = \hat{y}_c^t - 1. \tag{14}$$

Thus, FAL leaves the hard-positive BCE gradient unchanged while adapting negative supervision. To further explain the performance difference between AKD and FAL, we analyze the gradient magnitude associated with positive-label learning. As shown in Fig. 9, the positive gradient of AKD (Du et al., 2025) decreases rapidly as the predicted probability increases and becomes nearly vanishing for moderately confident positive predictions. Consequently, these predictions may receive insufficient supervision before being reliably learned. In contrast, FAL retains the positive-label term and performs FPR suppression by adaptively strengthening negative learning. It therefore preserves meaningful positive gradients over a broader confidence range. This analysis supports the results, where AKD exhibits lower high-recall precision and mAP, whereas CaCLIP achieves a better balance between false-positive suppression and positive-class learning.

Table 13: Ablation of the class-wise designs in SEA and FAL on VOC B4-C2 (%).

Component	Variant	Avg. mAP	Last mAP	CF1	OF1
SEA	Sigmoid gates	92.9	86.2	70.3	67.5
SEA	Class-softmax (Ours)	**94.8**	**89.6**	**73.3**	**71.9**
FAL	Class-agnostic weighting	93.5	87.6	64.0	65.8
FAL	Class-wise weighting (Ours)	**94.8**	**89.6**	**73.3**	**71.9**

Table 14: Effect of the softmax dimension in SEA on VOC B4-C2 (%).

Normalization	Avg. mAP	Last mAP	CF1	OF1
Over patches	94.0	88.3	71.4	67.7
Over seen classes (SEA)	**94.8**	**89.6**	**73.3**	**71.9**

C.2 FPR-Adaptive Distillation

For an old class $c \in \mathcal{C}^{1:t-1}$ and a fixed training batch, the teacher probability $\hat{y}_c^{t-1} \in (0,1)$ and the class-specific weight $w_c^{\mathrm{kd}} \geq 0$ are fixed with respect to the student, since both are computed from the frozen teacher. Define the auxiliary coefficient and effective distillation target as

$$A_c = 1 + w_c^{\mathrm{kd}}(1 - \hat{y}_c^{t-1}), \qquad \widetilde{y}_c^t = \frac{\hat{y}_c^{t-1}}{A_c}. \tag{15}$$

Since $A_c \widetilde{y}_c^t = \hat{y}_c^{t-1}$ and $A_c(1-\widetilde{y}_c^t) = (1+w_c^{\mathrm{kd}})(1-\hat{y}_c^{t-1})$, Eq. 12 can be rewritten exactly as

$$\mathcal{L}_{\mathrm{FAL}}^{\mathrm{kd}} = -\sum_{c \in \mathcal{C}^{1:t-1}} A_c \left[\widetilde{y}_c^t \log \hat{y}_c^t + (1-\widetilde{y}_c^t)\log(1-\hat{y}_c^t)\right]. \tag{16}$$

Thus, the distillation loss is a sum of weighted BCE terms with effective targets $\widetilde{y}_c^t$. Its partial derivative with respect to the student logit is

$$\frac{\partial \mathcal{L}_{\mathrm{FAL}}^{\mathrm{kd}}}{\partial s_c} = A_c \hat{y}_c^t - \hat{y}_c^{t-1}. \tag{17}$$

As $A_c > 0$, the unique pointwise minimizer for each independently optimized student probability is

$$(\hat{y}_c^t)^\star = \widetilde{y}_c^t = \frac{\hat{y}_c^{t-1}}{1 + w_c^{\mathrm{kd}}(1-\hat{y}_c^{t-1})}. \tag{18}$$

Equivalently, the effective target satisfies

$$\log \frac{\widetilde{y}_c^t}{1-\widetilde{y}_c^t} = \log \frac{\hat{y}_c^{t-1}}{1-\hat{y}_c^{t-1}} - \log(1+w_c^{\mathrm{kd}}). \tag{19}$$

When $w_c^{\mathrm{kd}} = 0$, the corresponding term reduces to standard BCE distillation with $\widetilde{y}_c^t = \hat{y}_c^{t-1}$. For $w_c^{\mathrm{kd}} > 0$, the effective target satisfies $\widetilde{y}_c^t < \hat{y}_c^{t-1}$. Thus, FAL retains the coefficient of the teacher's positive loss term while introducing a class-dependent downward adjustment of the effective soft target and a corresponding loss weight A_c. This adjustment is intended to suppress overconfident prediction and reduce FPR.

D Evaluation Metrics

Multi-label evaluation. Following prior MLCIL studies (Du et al., 2024a; Dong et al., 2023), we evaluate performance using mean average precision (mAP), class-level F1 (CF1), and overall F1 (OF1). Let C denote the number of evaluated classes. For class c, TP_c is the number of true positives, P_c is the number of predicted positives, and G_c is the number of ground-truth positives. We first compute the class-averaged and overall precision and recall as

$$\begin{aligned} \mathrm{CP} &= \frac{1}{C}\sum_{c=1}^{C}\frac{TP_c}{P_c}, \quad \mathrm{CR} = \frac{1}{C}\sum_{c=1}^{C}\frac{TP_c}{G_c}, \\ \mathrm{OP} &= \frac{\sum_{c=1}^{C} TP_c}{\sum_{c=1}^{C} P_c}, \quad \mathrm{OR} = \frac{\sum_{c=1}^{C} TP_c}{\sum_{c=1}^{C} G_c}. \end{aligned} \tag{20}$$

CF1 and OF1 are then defined as

$$\mathrm{CF1} = \frac{2\,\mathrm{CP}\,\mathrm{CR}}{\mathrm{CP}+\mathrm{CR}}, \qquad \mathrm{OF1} = \frac{2\,\mathrm{OP}\,\mathrm{OR}}{\mathrm{OP}+\mathrm{OR}}. \tag{21}$$

For the false-positive analysis, we additionally report the macro-averaged false positive rate:

$$\mathrm{FPR} = \frac{1}{C}\sum_{c=1}^{C}\frac{FP_c}{FP_c + TN_c}, \tag{22}$$

where FP_c and TN_c are the false-positive and true-negative counts for class c, respectively.

Following prior MLCIL studies (Dong et al., 2023; Du et al., 2024b; 2026), we use the same fixed decision threshold of 0.8 for all methods and variants without method-specific threshold tuning, ensuring a consistent protocol and fair comparison.

Forgetting metric. Let $M_{t,i}$ denote the mAP on the classes introduced in task i, evaluated using the model after learning task t, where $t \geq i$. The forgetting after the final task is defined as

$$\mathcal{F} = \frac{1}{T-1}\sum_{i=1}^{T-1}\left(\max_{k\in\{i,\ldots,T-1\}} M_{k,i} - M_{T,i}\right). \tag{23}$$

It measures the average performance degradation on previous tasks from their best historical performance to the final model. A lower $\mathcal{F}$ indicates less catastrophic forgetting.