TOWARDS EFFICIENT CONTINUAL LEARNING: TASK DISCRIMINATION VIA INCREMENTAL LDA FOR VLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Continual learning (CL), which enables vision-language models (VLMs) to incrementally acquire new knowledge from non-stationary data streams while avoiding catastrophic forgetting (CF), is commonly addressed by parameter-efficient fine-tuning (PEFT) with task-specific low-rank adapters. However, in the task-agnostic setting, selecting the appropriate adapter remains challenging, and auxiliary trainable task-discrimination modules can introduce additional computational overhead. In this work, we empirically observe that the intermediate-layer representations of CLIP exhibit strong between-task separability on the evaluated benchmarks. Building on this observation, we propose ILDA-CL, a lightweight framework that decouples CL into two independent pathways: (i) a distribution discriminator based on Incremental Linear Discriminant Analysis (ILDA) that incrementally projects task prototypes into a discriminative subspace without gradient-based optimization, thereby enabling task-identity prediction; and (ii) a set of parameter-isolated task-specific adapters that, guided by the predicted task identity, acquire new knowledge without updating previously learned adapters. Extensive experiments demonstrate that ILDA-CL achieves state-of-the-art accuracy while incurring substantially lower computational cost compared with existing approaches.

1 INTRODUCTION

Vision-Language Model (VLM) is a core backbone of multimodal learning systems, demonstrating remarkable zero-shot generalization capabilities with widespread real-world applications, such as visual question answering for autonomous driving (Zhou et al., 2024b), industrial monitoring (Gu et al., 2024), and personal AI assistant (Nguyen et al., 2024). Such applications require VLMs adjust to evolving tasks and non-stationary data distributions, making continual fine-tuning essential for effective VLM optimization. However, real-world limitations related to data privacy and storage often make it impractical to fully retrain VLMs on all previously learned tasks; instead, they must revise and expand their knowledge incrementally as new data continuously arrives (Yang et al., 2024). Prior work (Kirkpatrick et al., 2017) has demonstrated that sequential parameter updates in such incremental learning regimes will induce *Catastrophic Forgetting* (CF). It is therefore critical to develop mechanisms that enables the model to incorporate new supervision signals continually, flexibly, and sample-efficiently, which naturally falls into the scope of *Continual Learning* (CL).

Parameter-efficient fine-tuning (PEFT), such as low-rank adaptation (LoRA) (Hu et al., 2022), enables the adaptation of pre-trained vision-language models by optimizing a small number of parameters. With task-specific adapters, new tasks can be learned while previously trained adapters remain fixed. However, this parameter-isolation strategy requires reliable task discrimination to select the appropriate adapter when task identities are unavailable at inference. Recent VLM-based continual learning methods have explored different mechanisms for task adaptation and prediction. DDAS (Yu et al., 2024) learns reconstruction-based distribution selectors, whereas RAIL (Xu et al., 2024) combines feature expansion with recursive ridge regression for continual classification. Additionally, He et al. (2026) introduce SECA, which uses textual semantic priors to guide knowledge transfer and refine visual prototypes. For task-specific adapter frameworks, a central design objective remains to achieve reliable task discrimination while limiting the computational and storage overhead of adapter selection.

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.

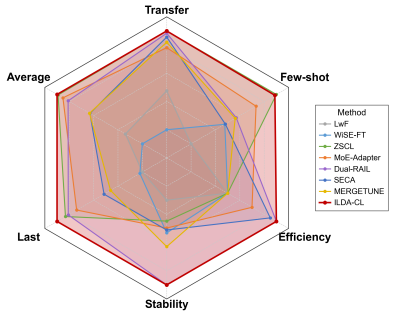


Figure 1: Performance comparison between **ILDA-CL** and several baselines on mainstream CL metrics, along with training stability and efficiency.

In this work, we identify an overlooked property of CLIP for CL: intermediate-layer representations naturally exhibit strong task-discriminative geometry because of constractive pre-training. Concretely, preliminary experiments reveal that features extracted from intermediate layers of CLIP exhibits high Fisher criterion values (see Appendix A) and inherent task-level clustering structures (see Figure 6). To interpret these observations, we analyze task discrimination under a shared-covariance Gaussian model. We derive an upper bound on the Bayes task-classification error in terms of pairwise Mahalanobis separation, together with a corresponding lower bound on mutual information. The Fisher trace summarizes aggregate separation, but does not by itself guarantee accurate task routing. Moreover, we notice that the closed-form solutions of linear models are inherently robust to CF owing to their deterministic nature. Building upon these observations, we propose **ILDA-CL**, a robust and efficient CL framework that achieves distribution discrimination via incremental linear discriminant analysis (ILDA) (Kim et al., 2007).

Specifically, during training, we take the `[CLS]` tokens from an intermediate layer of CLIP to construct the initial feature space for ILDA projection, and adopt the global mean of these tokens as the representative prototype for each distribution. Subsequently, the incremental data are employed to iteratively update the global between-class and within-class scatter matrices, ultimately yielding the optimal ILDA projection in closed form. During the inference stage, the task identity is inferred by computing the Mahalanobis distance between the test sample and each distribution prototypes. In addition, we implement PEFT with LoRA in the subsequent deeper transformer blocks to enhance plasticity for CL while maintaining parameter efficiency. To verify the effectiveness of the proposed method, We conduct extensive experiments comparing our method against recent strong baselines, achieving SOTA performance. The performance overview is shown in Figure 1.

To summarize, the contributions of our work are tri-folds. First, we proposed a novel CL framework **ILDA-CL** for task-agnostic scenarios, that combines ILDA with task-specific LoRA to effectively mitigate CF in VLMs. Second, we provide a population-level statistical interpretation of task discrimination through bounds relating pairwise Mahalanobis separation, Bayes error, and mutual information. Finally, extensive experiments under both the CIL and task-agnostic settings demonstrate that **ILDA-CL** achieves SOTA performance and exhibits strong robustness to CF and external noise, while maintaining competitive efficiency.

2 Related Works

Continual Learning. The central challenge in *Continual Learning* is to balance avoiding forgetting and enabling fast adaptation. The earliest approaches are based on regularization techniques, with Elastic Weight Consolidation (EWC) being a representative example. (Kirkpatrick et al., 2017). However, such approaches typically rely heavily on inter-class similarity and are therefore prone to negative transfer in complex scenarios. To overcome this limitation, memory-based methods (Rebuffi et al., 2017; Chaudhry et al., 2019) and optimization-based methods (Saha et al., 2021; Chen et al., 2022; Deng et al., 2021) achieve a coordinated improvement in performance across new and old tasks. Furthermore, parameter isolation-based methods (Lee et al., 2017; Li et al., 2024; Wortsman et al., 2020) assign independent or partially shared parameters to each task to achieve physical isolation, thereby achieving excellent performance.

Task Agnostic Continual Learning. To reduce dependence on task-identity information inherent in parameter-isolation-based methods, class-incremental learning scenarios have been extensively studied (Zhou et al., 2024a). In this context, Jha et al. (2024) achieves distribution discrimination in the probability space through uncertainty modeling. Zhou et al. (2025); He et al. (2026); Gao et al. (2026) leverage textual semantic information to refine visual task prototypes, thereby facilitating feature sharing across related tasks. Meanwhile, with the advancement of VLMs, task-agnostic CL scenarios have been investigated by leveraging the zero-shot generalization capabilities of VLMs to handle unseen data (Radford et al., 2021). ZSCL (Zheng et al., 2023) employs knowledge distillation

to preserve the cross-modal alignment capability of VLMs during fine-tuning. However, its reliance on the ImageNet reference dataset introduces a potential risk of information leakage. Yu et al. (2024) proposes a reconstruction-based approach to construct a Distribution Discriminative Auto-Selector (DDAS), which serves as an index for the task-specific MoE-Adapter. Wang et al. (2026) constructs low-loss pathways in the parameter space to simultaneously accommodate both new and existing tasks. Xu et al. (2024) proposes to map the intermediate features of CLIP into a higher-dimensional space, after which ridge regression is employed to achieve fine-grained classification.

3 PRELIMINARIES

Continual Learning. We consider a sequence of tasks $\{\mathcal{T}_t\}_{t=1}^{T}$ learned sequentially. Each task $\mathcal{T}_t = (\mathcal{D}_t, \mathcal{C}_t)$ consists of a training set $\mathcal{D}_t$ with $N_t = |\mathcal{D}_t|$ samples and a class set $\mathcal{C}_t$. Task boundaries are known during training, whereas task identities are unavailable at inference. After stage s, the seen class set is $\mathcal{C}_{\leq s} = \bigcup_{t=1}^{s} \mathcal{C}_t$. We consider two evaluation settings. In class-incremental learning (CIL), test samples come exclusively from seen tasks, and predictions are restricted to $\mathcal{C}_{\leq s}$. In task-agnostic evaluation, test samples may additionally come from not-yet-trained tasks (Xu et al., 2024). In this setting, the candidate class vocabulary $\mathcal{C}_{\text{all}} = \bigcup_{t=1}^{T} \mathcal{C}_t$, including the class names of unseen tasks, is available at inference for zero-shot scoring. No training images or instance-level labels from these unseen tasks are used for adaptation. Section 4.4 details the gating mechanism and subsequent routing procedure.

CLIP Model. Our method builds on CLIP (Radford et al., 2021) that learns a joint vision-language embedding space via contrastive pre-training. It comprises a text encoder f_{T} and an image encoder f_{I}. For task $\mathcal{T}_t$ with label set $\mathcal{C}_t = \{c_j\}_{j=1}^{C}$, each class name c_j is embedded as $\mathbf{e}_{\text{T}}^{j} = f_{\text{T}}(c_j) \in \mathbb{R}^{d_{\text{emb}}}$ using a fixed prompt (*e.g.*, "`a photo of a [class]`"). The visual input $\mathbf{x}$ is encoded as $\mathbf{e}_{\text{I}} = f_{\text{I}}(\mathbf{x}) \in \mathbb{R}^{d_{\text{emb}}}$. Denoting $s_j^i = \langle \mathbf{e}_{\text{I}}^{i}, \mathbf{e}_{\text{T}}^{j} \rangle_{\cos}$, the fine-tuning objective minimizes:

$$\mathcal{L}_{\text{CLIP}} = -\frac{1}{N} \sum_{i=1}^{N} \log \frac{\exp(\tau \cdot s_{y_i}^{i})}{\sum_{j=1}^{C} \exp(\tau \cdot s_j^{i})}, \tag{1}$$

where τ is the logit scale and y_i is the ground-truth label of the i-th sample.

Linear Discriminant Analysis. Given labeled samples $\mathcal{D} = \{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^{N}$ from C classes, LDA learns a projection $\mathbf{W} \in \mathbb{R}^{d \times k}$, $k \leq C-1$, that maximizes class separability. Let $\boldsymbol{\mu}_c$, N_c, and $\boldsymbol{\mu}$ denote the mean and size of class c and the global mean, respectively. The scatter matrices are

$$\mathbf{S}_{\text{W}} = \sum_{c=1}^{C} \sum_{i:y^{(i)}=c} (\mathbf{x}^{(i)} - \boldsymbol{\mu}_c)(\mathbf{x}^{(i)} - \boldsymbol{\mu}_c)^{\top}, \quad \mathbf{S}_{\text{B}} = \sum_{c=1}^{C} N_c (\boldsymbol{\mu}_c - \boldsymbol{\mu})(\boldsymbol{\mu}_c - \boldsymbol{\mu})^{\top}. \tag{2}$$

LDA optimizes

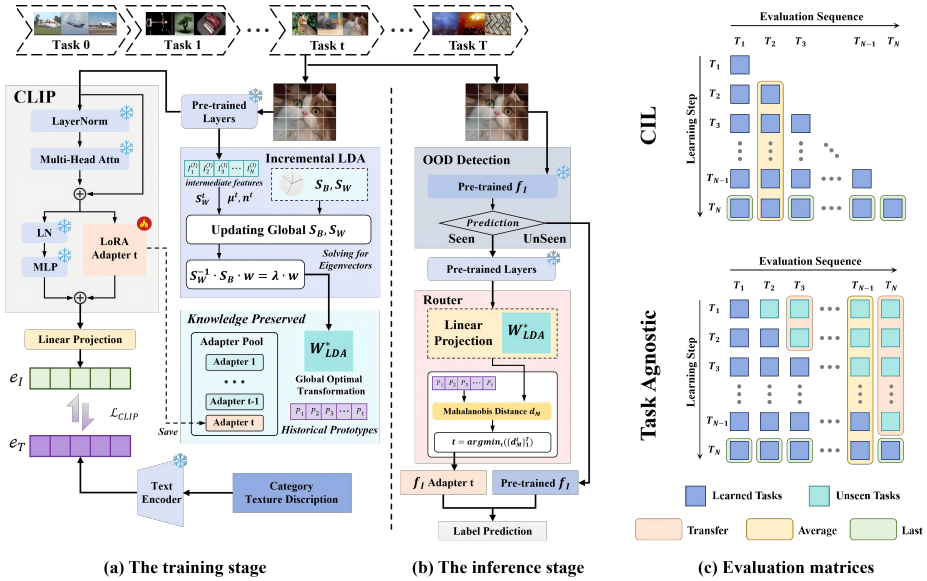
$$\mathbf{W}^{*} = \arg\max_{\mathbf{W}} \frac{\det(\mathbf{W}^{\top} \mathbf{S}_{\text{B}} \mathbf{W})}{\det(\mathbf{W}^{\top} \mathbf{S}_{\text{W}} \mathbf{W})}, \tag{3}$$

whose solution comprises the leading generalized eigenvectors satisfying $\mathbf{S}_{\text{B}}\mathbf{w} = \lambda \mathbf{S}_{\text{W}}\mathbf{w}$; see Appendix B.

4 METHODOLOGY

4.1 MOTIVATION AND OVERVIEW

ILDA-CL combines an incremental statistical router with task-specific LoRA adapters. Motivated by the task-level clustering observed in frozen CLIP features (Appendix A), we treat each task as a meta-class and summarize its feature distribution through first- and second-order statistics. These statistics define a shared LDA subspace in which task prototypes support adapter selection. The frozen representation provides a stable coordinate system for accumulating task statistics, while task-specific adapters learn new tasks without modifying previously trained adapter parameters. Figure 2 summarizes the framework. We first describe the incremental router, then interpret task separation statistically, and finally specify the training and inference procedures.



(a) The training stage (b) The inference stage (c) Evaluation matrices

Figure 2: Overview of **ILDA-CL**. Panels (a) and (b) illustrate training and inference, respectively. Panel (c) illustrates *Average* and *Last* accuracy in both evaluation settings and *Transfer* accuracy in task-agnostic evaluation

4.2 Incremental LDA for Task Discrimination

We treat each task as a meta-class using frozen CLIP features $\mathbf{z}_i^{(t)} = g^{(l^*)}(x_i^{(t)}) \in \mathbb{R}^d$. For task $\mathcal{T}_t$ with $N_t \geq 2$, its mean and scatter are

$$\hat{\boldsymbol{\mu}}_t = \frac{1}{N_t} \sum_{i=1}^{N_t} \mathbf{z}_i^{(t)}, \qquad \mathbf{S}_{\mathrm{W}}^{(t)} = \sum_{i=1}^{N_t} (\mathbf{z}_i^{(t)} - \hat{\boldsymbol{\mu}}_t)(\mathbf{z}_i^{(t)} - \hat{\boldsymbol{\mu}}_t)^\top. \tag{4}$$

At stage s, we balance task sizes through normalized weights:

$$p_{t,s} = \frac{N_t^{1-\beta}}{\sum_{j=1}^{s} N_j^{1-\beta}}, \qquad \hat{\boldsymbol{\mu}}_{\leq s} = \sum_{t=1}^{s} p_{t,s} \hat{\boldsymbol{\mu}}_t, \tag{5}$$

where fixed $\beta \in [0, 1]$ interpolates between sample-count and equal task weighting. The weighted covariance matrices are

$$\begin{aligned} \mathbf{S}_{\mathrm{W},s} &= \sum_{t=1}^{s} p_{t,s} \frac{\mathbf{S}_{\mathrm{W}}^{(t)}}{N_t - 1}, \\ \mathbf{S}_{\mathrm{B},s} &= \sum_{t=1}^{s} p_{t,s} (\hat{\boldsymbol{\mu}}_t - \hat{\boldsymbol{\mu}}_{\leq s})(\hat{\boldsymbol{\mu}}_t - \hat{\boldsymbol{\mu}}_{\leq s})^\top. \end{aligned} \tag{6}$$

These matrices are updated from accumulated task statistics without retaining past images.

For $s \geq 2$, we solve

$$\mathbf{S}_{\mathrm{B},s} \mathbf{w} = \lambda \mathbf{G}_s \mathbf{w}, \qquad \mathbf{G}_s = \mathbf{S}_{\mathrm{W},s} + \epsilon \mathbf{I}_d, \tag{7}$$

where $\epsilon > 0$. The leading $1 \leq k_s \leq \min(d, s-1)$ generalized eigenvectors form $\mathbf{W}_s$, with $\mathbf{W}_s^\top \mathbf{G}_s \mathbf{W}_s = \mathbf{I}_{k_s}$. Under this normalization, the projected regularized Mahalanobis distance reduces to Euclidean distance. With equal task priors, routing selects

$$d_{t,s}^2(x) = \left\| \mathbf{W}_s^\top \left(g^{(l^*)}(x) - \hat{\boldsymbol{\mu}}_t \right) \right\|_2^2, \qquad \hat{t}_s(x) = \arg\min_{1 \leq t \leq s} d_{t,s}^2(x). \tag{8}$$

Task means are stored in the original space and reprojected after each update; at $s = 1$, the sole adapter is selected directly.

Table 1: Performance comparison across different methods and datasets in terms of "*Transfer*", "*Average*" and "*Last*" accuracy (%). The best results are highlighted with **bold** style.

	Method	Air.	Cal.	CIF.	DTD	Eur.	Flo.	Foo.	MNI.	Oxf.	Sta.	*Avg.*↑	*Std.* ↓
	Full Fine-tune	60.6	96.7	89.7	78.7	95.2	93.4	93.9	99.2	94.3	85.4	88.7	-
	Zero-shot	24.8	88.4	68.4	43.0	55.0	71.0	88.5	59.4	89.1	64.0	65.2	-
Transfer	Continual-FT	-	66.8	47.2	32.5	33.7	35.0	58.2	46.1	61.6	22.5	44.8	1.27
	LwF	-	65.4	61.2	32.3	22.1	45.3	69.5	29.4	63.9	36.5	47.3	2.70
	WiSE-FT	-	79.2	56.6	34.3	36.0	23.1	48.9	46.7	49.8	14.6	43.2	2.97
	ZSCL	-	86.2	63.4	42.6	49.3	68.5	83.3	50.2	87.6	61.5	65.8	1.24
	MoE-Adapter	-	80.4	68.3	42.7	51.7	64.0	85.1	**52.0**	84.4	60.2	65.4	0.77
	Dual-RAIL	-	**88.4**	**68.4**	43.3	51.3	67.6	88.2	22.9	88.3	63.7	64.7	**0.00**
	SECA	-	77.3	66.3	42.4	51.2	63.7	85.8	45.7	83.1	60.8	64.0	0.52
	MERGETUNE	-	81.1	67.2	41.0	50.2	63.1	84.9	34.8	79.7	59.6	62.4	1.48
	ILDA-CL	-	78.6	60.8	**43.7**	**54.9**	**69.7**	**88.4**	43.6	**88.9**	**64.7**	**65.9**	**0.00**
Average	Continual-FT	8.4	21.6	26.4	17.5	49.9	19.1	53.8	61.5	59.6	28.8	34.7	2.75
	LwF	24.5	82.9	65.3	41.8	35.5	63.4	64.1	40.2	65.0	40.9	52.4	2.79
	WiSE-FT	9.1	79.1	60.8	41.3	63.4	31.7	57.8	62.6	58.6	19.5	48.4	1.45
	ZSCL	45.6	93.0	81.4	64.8	78.3	**81.4**	85.2	64.5	89.5	64.1	74.8	2.33
	MoE-Adapter	50.7	92.4	**84.3**	54.3	79.8	76.5	87.8	**65.6**	85.3	61.4	73.8	2.19
	Dual-RAIL	49.0	**94.6**	76.5	63.6	76.8	79.7	89.7	45.4	89.4	65.6	73.0	**0.00**
	SECA	45.5	86.3	68.6	50.0	65.5	70.5	86.5	57.7	84.4	62.2	67.7	1.96
	MERGETUNE	33.8	87.6	70.6	49.7	59.8	66.0	85.7	44.0	82.2	61.2	64.1	1.23
	ILDA-CL	**59.1**	90.5	77.3	**65.8**	**80.7**	**81.4**	**89.9**	60.4	**89.8**	**66.8**	**76.2**	**0.00**
Last	Continual-FT	1.2	2.6	3.6	6.0	18.7	1.5	24.1	**99.6**	49.6	85.1	29.2	3.15
	LwF	20.6	78.3	64.5	40.5	36.8	69.5	84.3	45.4	87.1	80.9	60.8	3.10
	WiSE-FT	7.5	76.4	50.1	35.2	76.4	36.2	81.8	**99.6**	90.0	63.3	61.7	0.85
	ZSCL	40.3	93.7	83.0	74.6	96.9	92.0	92.4	99.2	**95.4**	**87.6**	85.5	2.89
	MoE-Adapter	50.7	92.9	**88.3**	56.0	98.6	79.2	91.7	**99.6**	88.9	72.6	81.9	2.81
	Dual-RAIL	47.2	95.3	80.0	72.0	94.4	90.7	92.0	97.9	93.8	82.6	84.6	**0.00**
	SECA	44.2	85.7	69.0	53.1	74.9	73.6	86.9	82.1	86.8	74.5	73.1	3.08
	MERGETUNE	45.7	91.7	75.6	53.4	61.7	71.1	89.0	53.9	91.6	76.0	71.0	1.17
	ILDA-CL	**59.1**	**96.3**	87.7	**76.0**	**98.9**	**93.1**	**92.5**	**99.6**	93.5	85.4	**88.2**	**0.00**

4.3 Statistical Interpretation of Task Separation

To interpret task discrimination, we consider Gaussian distributions for the frozen features of the $s \geq 2$ seen tasks, with task-specific means and a shared positive-definite covariance. Under equal task priors, the Bayes error satisfies

$$P_e^\star \leq \min\left\{1 - \frac{1}{s},\ \frac{s-1}{2}\exp\left(-\frac{\delta_{\min}^2}{8}\right)\right\}, \tag{9}$$

where $\delta_{\min}^2$ is the minimum squared Mahalanobis distance between task means. Under this model, the Bayes classifier selects the nearest task prototype under the shared Mahalanobis metric, providing a statistical interpretation of our routing rule. The Fisher trace captures aggregate separation but does not control the least separated task pair. Appendix C presents the general-prior result, its mutual-information consequence, and the proof of Proposition C.1. These population results do not directly bound the error of the estimated, regularized router; routing performance and feature-layer selection are evaluated empirically.

4.4 Training and Inference Pipeline

For each task $\mathcal{T}_t$, we extract `[CLS]` features $g^{(l^*)}(x)$ using the frozen CLIP image encoder and update the cumulative statistics and, for $t \geq 2$, the projection $\mathbf{W}_t$ as described in Section 4.2. The empirical mean $\hat{\boldsymbol{\mu}}_t$ is stored in the original feature space as the task prototype. The choice of layer l^* is examined in Section 5.3.

Table 2: Performance comparison across different methods and datasets in 5-shot configuration.

	Method	Air.	Cal.	CIF.	DTD	Eur.	Flo.	Foo.	MNI.	Oxf.	Sta.	***Avg.*** ↑	***Std.*** ↓
	Full Fine-tune	39.8	95.8	73.9	72.0	88.6	90.2	91.1	93.8	92.7	72.6	81.1	-
	Zero-shot	24.8	88.4	68.4	43.0	55.0	71.0	88.5	59.4	89.1	64.0	65.2	-
Transfer	Continual-FT	-	85.1	51.5	36.7	34.6	39.8	55.1	24.9	72.9	36.9	48.6	1.56
	LwF	-	71.5	55.3	42.7	41.9	55.2	70.1	36.8	75.8	44.6	54.9	2.69
	WiSE-FT	-	83.8	64.0	42.1	42.3	52.9	77.8	**66.0**	77.8	50.6	61.9	4.07
	ZSCL	-	87.3	68.0	**43.9**	51.7	**71.4**	**88.5**	55.8	**91.0**	61.9	**68.8**	2.20
	MoE-Adapter	-	**87.6**	**68.2**	42.1	53.8	65.8	88.2	50.4	87.7	64.6	67.6	0.52
	Dual-RAIL	-	65.2	40.0	36.6	38.7	61.2	86.8	12.6	87.4	63.8	54.7	**0.00**
	SECA	-	78.5	67.1	42.9	51.2	64.0	86.3	40.0	82.3	60.9	63.7	0.94
	MERGETUNE	-	87.1	67.9	42.2	50.7	68.7	87.1	30.1	85.7	62.9	64.7	1.32
	ILDA-CL	-	78.6	60.8	43.7	**54.9**	69.7	88.4	43.6	88.9	**64.7**	65.9	**0.00**
Average	Continual-FT	18.0	83.5	49.7	46.2	56.2	58.0	58.0	33.5	74.3	38.2	51.6	2.15
	LwF	23.5	79.3	48.7	41.5	46.3	54.9	73.9	51.6	70.5	46.8	53.7	2.68
	WiSE-FT	21.8	85.1	63.3	45.7	62.9	58.9	78.9	**72.5**	78.9	51.7	62.0	2.52
	ZSCL	32.6	92.0	74.0	56.8	68.8	**82.7**	88.9	65.4	**91.5**	63.4	**71.6**	2.43
	MoE-Adapter	28.8	91.4	**74.1**	46.7	63.8	73.6	88.2	52.7	88.0	64.7	67.2	1.98
	Dual-RAIL	34.4	83.9	56.0	52.1	62.4	77.6	87.4	31.0	87.9	64.8	63.8	**0.00**
	SECA	25.1	85.7	66.8	42.5	45.0	65.4	86.7	44.7	83.6	61.2	60.7	1.62
	MERGETUNE	25.9	89.3	68.0	45.6	53.0	71.0	87.7	38.1	87.0	63.3	62.9	1.73
	ILDA-CL	**39.2**	**92.7**	65.6	**57.5**	**74.3**	82.4	**90.5**	59.6	89.2	**65.4**	**71.6**	0.05
Last	Continual-FT	15.9	82.1	52.5	50.3	70.0	71.4	62.3	50.7	79.2	49.9	58.4	2.95
	LwF	21.1	64.9	43.4	47.7	71.0	65.2	63.4	72.1	73.5	66.9	58.9	3.03
	WiSE-FT	16.7	74.2	57.5	43.5	68.2	60.0	77.7	87.1	82.2	61.7	62.9	1.87
	ZSCL	30.6	92.4	74.3	**61.8**	76.7	92.9	89.4	86.8	**93.4**	**76.4**	77.5	2.63
	MoE-Adapter	29.4	90.0	**75.8**	47.6	70.8	76.1	88.3	74.4	89.1	65.0	70.7	3.10
	Dual-RAIL	34.8	88.7	66.0	60.4	80.5	**93.8**	88.4	73.3	90.0	74.1	75.0	**0.00**
	SECA	25.5	85.7	67.5	43.5	44.8	62.8	86.7	53.6	85.7	64.0	62.0	2.59
	MERGETUNE	26.0	91.1	69.5	44.3	52.5	72.2	88.8	50.8	92.2	66.6	65.4	0.92
	ILDA-CL	**39.2**	**93.3**	72.7	60.3	**88.1**	89.1	**89.8**	**95.1**	90.5	71.5	**79.0**	**0.00**

Task adaptation uses a separate LoRA adapter for each task. For each adapted linear layer m, the effective weight is

$$\mathbf{M}_t^{(m)} = \mathbf{M}_0^{(m)} + \mathbf{B}_t^{(m)}\mathbf{A}_t^{(m)}, \tag{10}$$

where $\mathbf{M}_0^{(m)}$ is the pretrained weight and $\mathbf{B}_t^{(m)}\mathbf{A}_t^{(m)}$ is a task-specific update of rank at most r, with any fixed LoRA scaling absorbed into the factors. Only the new adapter is optimized using the objective in Eq. equation 1; the pretrained parameters and previously trained adapters remain fixed.

At inference stage s, CIL samples are routed according to Eq. equation 8 to select a task-specific adapter, with predictions restricted to the seen class set $\mathcal{C}_{\leq s}$. In task-agnostic evaluation, the frozen CLIP model first performs zero-shot classification over the candidate vocabulary $\mathcal{C}_{\text{all}}$, including the class names of unseen tasks. If the predicted class belongs to $\mathcal{C}_{\leq s}$, the sample is assigned to the seen-task branch, where ILDA routing selects the corresponding adapter for classification. Otherwise, the sample is assigned to the unseen-task branch and retains its zero-shot prediction.

5 Experimental Results

5.1 Experimental Setup

Datasets. Following Xu et al. (2024); Yu et al. (2024), we construct the non-stationary learning sequence using common benchmarks to evaluate model performance under the task-agnostic and CIL settings. Consistent with Xu et al. (2024); Yu et al. (2024), the learning orders of the different datasets are arranged alphabetically: Aircraft (Maji et al., 2013), Caltech101 (Fei-Fei et al., 2004), CIFAR-100 (Krizhevsky et al., 2009), DTD (Aljundi et al., 2018), EuroSAT (Helber et al., 2019), Flowers (Nilsback and Zisserman, 2008), Food (Bossard et al., 2014), MNIST (Deng, 2012), Oxford-Pet (Parkhi et al., 2012), and StanfordCars (Krause et al., 2013). We refer to these datasets using their

respective three-letter abbreviations, for example, "Air." for Aircrafts.

Metrics. We follow (Zheng et al., 2023) by employing the metrics "*Transfer*", "*Average*", and "*Last*" to evaluate the resistance to CF and zero-shot capability. Specifically, as shown in Figure 2, "*Transfer*" accuracy is employed to assess model performance across unseen distributions. "*Average*" accuracy is defined as the global mean accuracy computed across the dataset throughout the training stage. "*Last*" accuracy denotes the final performance attained upon completion of the entire training stage. Under task-agnostic scenarios, all three metrics are employed. In contrast, only "*Average*' and "*Last*" accuracy are reported under the CIL setting, as unseen data are excluded and the *Transfer* accuracy is therefore not applicable.

Baselines. We select several strong CL methods (Li and Hoiem, 2017; Wortsman et al., 2022; Zheng et al., 2023; Yu et al., 2024; Xu et al., 2024; He et al., 2026) designed for the task-agnostic setting as baselines, and additionally include several recent approaches (Gao et al., 2026; Li et al., 2025; Lu et al., 2024; Wang et al., 2024) developed for the CIL setting. Due to space limitations, a detailed description of all baselines is provided in Appendix E.

5.2 Comparison with Baselines

Task Agnostic Scenario. Table 1 presents the performance comparison between the proposed **ILDA-CL** and several strong baselines under the task-agnostic CL setting. As demonstrated in Table 1, **ILDA-CL** achieves SOTA *Transfer* accuracy of 65.9%, *Average* accuracy of 76.2% and *Last* accuracy of 88.2%, significantly outperforming all existing task-agnostic CL baselines. Furthermore, we also conduct a 5-shot experiment in accordance with standard practice in this field (Yu et al., 2024; Xu et al., 2024), as shown in Table 2. **ILDA-CL** exhibits robust performance even under scenarios with scarce training data. Moreover, as demonstrated in Figure 3, upon exposure to a dataset, **ILDA-CL**'s performance on the corresponding dataset shows no further degradation.

Table 1 compares **ILDA-CL** with representative task-agnostic continual learning methods. Under the full-data setting, **ILDA-CL** achieves the highest aggregate *Transfer*, *Average*, and *Last* accuracies among the compared methods, reaching 65.9%, 76.2%, and 88.2%, respectively. In particular, the improvement is more pronounced for *Average* and *Last* accuracy, whereas the *Transfer* result is close to ZSCL (65.9% vs. 65.8%). These results indicate that **ILDA-CL** maintains strong performance on previously learned tasks while preserving competitive zero-shot performance on unseen tasks. We further evaluate all methods in the 5-shot setting following prior work (Yu et al., 2024; Xu et al., 2024). As shown in Table 2, **ILDA-CL** achieves the highest *Last* accuracy of 79.0% and an *Average* accuracy of 71.6%, matching ZSCL at the reported precision. Figure 3 further illustrates the evolution of accuracy throughout the learning sequence. Notably, ZSCL retains access to the full ImageNet reference set even under the 5-shot setting, which provides additional information beyond the limited task-specific training samples. This may contribute to its stronger *Transfer* performance.

CIL Scenario. In addition, we conduct experiments under the CIL setting using the same benchmarks. Upon excluding OOD detection, **ILDA-CL** is demonstrated to exhibit near-zero CF. As illustrated in Table 3, **ILDA-CL** achieves best performance on "*Average*" and "*Last*" accuracy, outperforming all rencent works. Moreover, **ILDA-CL** maintains its excellent generalization capability under 5-shot CIL scenarios. As illustrated in Figure 4a, the "*Last*" accuracy of **ILDA-CL** is observed to closely approximate that of low-rank fine-tuning and full fine-tune with known task identities. The performance degradation induced by the absence of task boundary information is constrained to within 1%.

Table 3: Experimental results under CIL scenarios. Here, "*Average** " and "*Last** " denote the global mean of each metric across all datasets.

Method	CIL		5-shot CIL	
	Average*	Last*	Average*	Last*
LwF	69.1	61.5	62.1	58.5
WISE-FT	63.9	62.1	67.2	62.9
ZSCL	86.6	85.8	78.4	77.5
MoE-Adapters	83.8	80.7	72.6	71.1
Dual-RAIL	83.5	83.3	75.5	75.2
SECA	83.2	82.7	69.0	67.9
MERGETUNE	71.0	70.3	67.0	65.1
DPeCLIP	82.6	81.4	72.3	70.1
Coleclip	80.6	80.1	71.0	69.2
ConDU	86.9	85.7	74.1	73.3
ILDA-CL	**88.3**	**88.2**	**79.0**	**78.8**

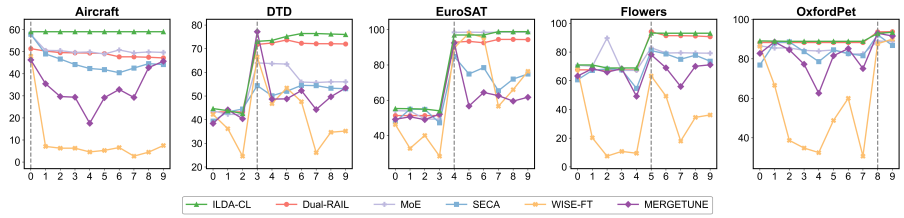
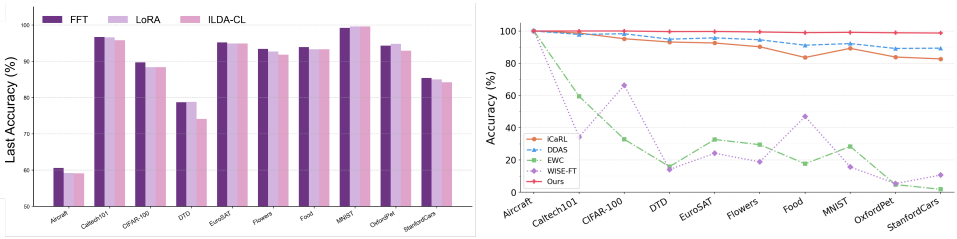


Figure 3: The accuracy curves on five datasets throughout the learning sequence. The gray dashed line denotes the training step at which the corresponding dataset is introduced.



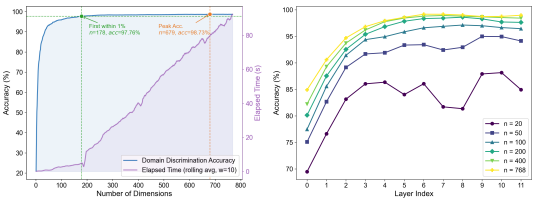
(a) (b)

Figure 4: (a) "*Last*" accuracy of **ILDA-CL**, LoRA and FFT across all datasets following the whole training sequence. (b) Comparison of domain discrimination accuracy between the proposed **ILDA-CL** and several baseline methods. Both experiments are under the CIL configuration.

5.3 Ablation Study

ILDA distribution Discrimination. As aforementioned, the ILDA-based task discrimination plays a crucial role in **ILDA-CL**. In this section, we compare its performance against several mainstream distribution discrimination approaches, including DDAS, EWC, WISE-FT, and iCaRL. Because these methods do not possess built-in OOD detection mechanisms, we conduct experiments within the CIL framework for a fair comparison, and the resulting average accuracy curves are shown in Figure 4b. **ILDA-CL** demonstrates superior performance across all datasets, with the corresponding confusion matrices provided in Appendix F. Furthermore, **ILDA-CL** does not rely on data replay, whereas iCaRL employs an experience replay buffer with a capacity of $2k$ samples. Additionally, ILDA-CL contains merely $7.68k$ parameters, significantly fewer than the $8.7M$ parameters of DDAS and the 13.6M parameters of iCaRL.

Intermediate Feature Dimension. Given the strong distribution-based discriminative ability of **ILDA-CL**, we further investigate how much redundancy is preserved across all dimensions of the `[CLS]` token. As shown in Figure 5a, the discrimination accuracy of ILDA using only the first 168 dimensions of the intermediate features falls within 1% of the peak value. This observation yields two insights. First, for the benchmarks used in this study, **ILDA-CL** can reduce the number of retained dimensions to further improve efficiency, with detailed experimental results presented in Appendix G. Second, **ILDA-CL** preserves a performance margin that allows it to handle more complex tasks.



(a) (b)

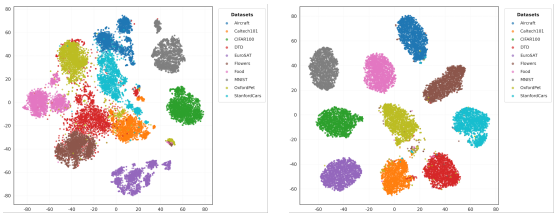
Figure 5: (a) Discrimination accuracy and training time versus the number of retained dimensions. (b) Discrimination accuracy versus the layer selection.

Layer Selection. As shown in Figure 5b, we perform ablation experiments to identify the optimal choice of intermediate layers. Specifically, we evaluate how effectively the ILDA-based distribution discriminator classifies `[CLS]` tokens derived from various intermediate layers under different

settings for the number of preserved dimensions. From the fifth layer onward, the accuracy stabilizes, with variations remaining within 1%. The experimental results demonstrate that **ILDA-CL** is fairly robust with respect to this hyperparameter setting.

5.4 DISCUSSION

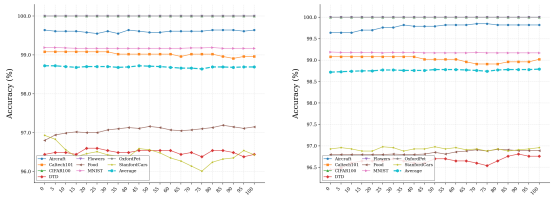
Visualization of t-SNE. Figure 6 presents a comparison of the t-SNE projections between the raw CLIP intermediate feature space and the ILDA space. The raw CLIP intermediate features inherently exhibit task-level discriminative geometry. Following ILDA projection, the samples corresponding to different distributions formed more compact clustering structures. These findings provide compelling evidence for the effectiveness of the ILDA-based distribution discriminator in CL scenarios. In Appendix H, we further present detailed visualizations of the ILDA projection space.



(a) Raw Space (b) LDA Space

Figure 6: t-SNE Visualization.

Robustness to Noise. Considering that linear classifiers are inherently sensitive to noise in the data, we further examine the robustness of **ILDA-CL**. We first incorporate images randomly sampled from a Gaussian distribution into each dataset and then evaluate whether the noisy images affect discrimination accuracy. As illustrated in Figure 7a, **ILDA-CL** exhibits strong robustness against externally introduced noise. Furthermore, we extract contamination samples from the Aircraft dataset to corrupt the remaining datasets and then evaluate their impact on the discrimination accuracy in a comparable manner. As shown in Figure 7b, this contamination exerts no significant impact on the performance of **ILDA-CL**. These verified the robustness of our method.



(a) (b)

Figure 7: The domain discrimination accuracy of ILDA-CL across various datasets (y-axis) versus the number of corrupted images (x-axis).

Computational Efficiency. By leveraging a low-rank adapter, **ILDA-CL** requires fewer parameters. Thus, we further compare the efficiency of the proposed method in terms of trainable parameters, GPU memory costs and inference time efficiency. As shown in Table 4, **ILDA-CL** substantially outperforms baselines in terms of both the number of trainable parameters and GPU memory utilization. Furthermore, the differences in inference time are relatively minor, indicating that our method can achieve similar time efficiency (marginally better) while requiring less memory.

Table 4: The effieciency analysis.

Method	Train Params. ↓	GPU ↓	Time ↓
LWF	149.6M	22172MiB	0.39s/it
WISE-FT	149.6M	20453MiB	0.37s/it
ZSCL	149.6M	22290MiB	0.38s/it
MoE-Adapters	59.8M	22358MiB	0.58s/it
SECA	1.6M	8694MiB	0.32s/it
MERGETUNE	149.6M	25746MiB	0.37s/it
ILDA-CL	**1.4M**	**7848MiB**	0.37s/it

6 CONCLUSION

In this work, we present **ILDA-CL**, a robust CL framework that effectively mitigates CF and resists external noise during sequential learning in the task-agnostic setting. Motivated by task separation in frozen CLIP features, we develop an incremental router and relate pairwise separation to Bayes error under a shared-covariance Gaussian model. We validate the estimated, regularized router empirically. Furthermore, we employ ILDA to achieve task routing, thereby resolving the dependence of parameter-isolation-based CL methods on task-identity information during inference. Extensive experiments have demonstrated **ILDA-CL**'s SOTA performance and computational efficiency. Moreover, owing to its considerable flexibility, the proposed framework can be readily extended to a wide range of mainstream vision tasks.

AI USAGE DISCLOSURE

Generative AI tools were used to assist with writing and language polishing, including improving grammar, clarity, and readability. The authors are responsible for verifying the accuracy of the text and take full responsibility for the final content of the paper.

REFERENCES

Rahaf Aljundi, Francesca Babiloni, Mohamed Elhoseiny, Marcus Rohrbach, and Tinne Tuytelaars. Memory aware synapses: Learning what (not) to forget. In *Proceedings of the European conference on computer vision (ECCV)*, pages 139–154, 2018.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *European conference on computer vision*, pages 446–461. Springer, 2014.

Arslan Chaudhry, Marcus Rohrbach, Mohamed Elhoseiny, Thalaiyasingam Ajanthan, Puneet K Dokania, Philip HS Torr, and Marc'Aurelio Ranzato. On tiny episodic memories in continual learning. *arXiv preprint arXiv:1902.10486*, 2019.

Cheng Chen, Ji Zhang, Jingkuan Song, and Lianli Gao. Class gradient projection for continual learning. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 5575–5583, 2022.

Danruo Deng, Guangyong Chen, Jianye Hao, Qiong Wang, and Pheng-Ann Heng. Flattening sharpness for dynamic gradient projection memory benefits continual learning. *Advances in Neural Information Processing Systems*, 34:18710–18721, 2021.

Li Deng. The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE signal processing magazine*, 29(6):141–142, 2012.

Luc Devroye. *A course in density estimation*. Birkhauser Boston Inc., 1987.

Ky Fan. On a theorem of weyl concerning eigenvalues of linear transformations i. *Proceedings of the National Academy of Sciences*, 35(11):652–655, 1949.

Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *2004 conference on computer vision and pattern recognition workshop*, pages 178–178. IEEE, 2004.

Keinosuke Fukunaga. *Introduction to statistical pattern recognition*. Elsevier, 2013.

Haoyuan Gao, Zicong Zhang, Yuqi Wei, Linglan Zhao, Guilin Li, Yexin Li, Bo Wang, Linghe Kong, and Weiran Huang. Enhanced continual learning of vision-language models with model fusion. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=ptFP9yT9DK.

Zhaopeng Gu, Bingke Zhu, Guibo Zhu, Yingying Chen, Ming Tang, and Jinqiao Wang. Anomalygpt: Detecting industrial anomalies using large vision-language models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 1932–1940, 2024.

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*, 2026.

Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019.

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

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.

Tae-Kyun Kim, Shu-Fai Wong, Bjorn Stenger, Josef Kittler, and Roberto Cipolla. Incremental linear discriminant analysis using sufficient spanning set approximations. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2007.

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.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 554–561, 2013.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Jeongtae Lee, Jaehong Yoon, Eunho Yang, and Sung Ju Hwang. Lifelong learning with dynamically expandable networks. *CoRR*, 2017.

Jiaqi Li, Yuanhao Lai, Rui Wang, Changjian Shui, Sabyasachi Sahoo, Charles X Ling, Shichun Yang, Boyu Wang, Christian Gagné, and Fan Zhou. Hessian aware low-rank perturbation for order-robust continual learning. *IEEE Transactions on Knowledge and Data Engineering*, 36(11):6385–6396, 2024.

Yukun Li, Guansong Pang, Wei Suo, Chenchen Jing, Yuling Xi, Lingqiao Liu, Hao Chen, Guoqiang Liang, and Peng Wang. Coleclip: Open-domain continual learning via joint task prompt and vocabulary learning. *IEEE Transactions on Neural Networks and Learning Systems*, 2025.

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

Yadong Lu, Shitian Zhao, Boxiang Yun, Dongsheng Jiang, Yin Li, Qingli Li, and Yan Wang. Boosting open-domain continual learning via leveraging intra-domain category-aware prototype. *arXiv preprint arXiv:2408.09984*, 2024.

Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013.

Thao Nguyen, Haotian Liu, Yuheng Li, Mu Cai, Utkarsh Ojha, and Yong Jae Lee. Yo'llava: Your personalized language and vision assistant. *Advances in Neural Information Processing Systems*, 37:40913–40951, 2024.

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian conference on computer vision, graphics & image processing*, pages 722–729. IEEE, 2008.

Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3498–3505. IEEE, 2012.

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. PmLR, 2021.

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

Gobinda Saha, Isha Garg, and Kaushik Roy. Gradient projection memory for continual learning. *arXiv preprint arXiv:2103.09762*, 2021.

Liyuan Wang, Xingxing Zhang, Hang Su, and Jun Zhu. A comprehensive survey of continual learning: Theory, method and application. *IEEE transactions on pattern analysis and machine intelligence*, 46(8):5362–5383, 2024.

Wenqing Wang, Da Li, Xiatian Zhu, and Josef Kittler. Mergetune: Continued fine-tuning of vision-language models. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=MAApSY32Z6.

Mitchell Wortsman, Vivek Ramanujan, Rosanne Liu, Aniruddha Kembhavi, Mohammad Rastegari, Jason Yosinski, and Ali Farhadi. Supermasks in superposition. *Advances in neural information processing systems*, 33:15173–15184, 2020.

Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, et al. Robust fine-tuning of zero-shot models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7959–7971, 2022.

Yicheng Xu, Yuxin Chen, Jiahao Nie, Yusong Wang, Huiping Zhuang, and Manabu Okumura. Advancing cross-domain discriminability in continual learning of vision-language models. *Advances in Neural Information Processing Systems*, 37:51552–51576, 2024.

Xin Yang, Hao Yu, Xin Gao, Hao Wang, Junbo Zhang, and Tianrui Li. Federated continual learning via knowledge fusion: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 36(8): 3832–3850, 2024.

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.

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.

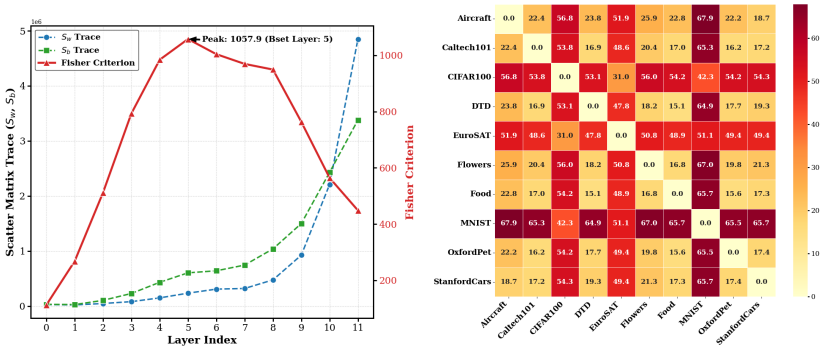
Da-Wei Zhou, Qi-Wei Wang, Zhi-Hong Qi, Han-Jia Ye, De-Chuan Zhan, and Ziwei Liu. Class-incremental learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):9851–9873, 2024a.

Da-Wei Zhou, Yuanhan Zhang, Yan Wang, Jingyi Ning, Han-Jia Ye, De-Chuan Zhan, and Ziwei Liu. Learning without forgetting for vision-language models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Xingcheng Zhou, Mingyu Liu, Ekim Yurtsever, Bare Luka Zagar, Walter Zimmer, Hu Cao, and Alois C. Knoll. Vision language models in autonomous driving: A survey and outlook. *IEEE Transactions on Intelligent Vehicles*, pages 1–20, 2024b. doi: 10.1109/TIV.2024.3402136.

A Results of Preliminary Experiments

Our preliminary experiments reveal that features drawn from distinct distributions already exhibit notably high Fisher criterion values within the CLIP embedding space itself, as shown in Figure 8a. It is demonstrated in Figure 8b that the Mahalanobis distances between these prototypes are notably substantial within the LDA space.



(a) (b)

Figure 8: Preliminary Experiments Results. (a) The Fisher criterion across different layers of the CLIP backbone. The trace of the between-class scatter $\mathbf{S}_\mathrm{B}$ and the within-class scatter $\mathbf{S}_\mathrm{W}$ are also reported. (b) The pairwise Mahalanobis distances between all distinct task prototypes in the LDA space.

B Solution for Equation 3

In this section, we provide a detailed derivation showing that the Fisher criterion maximization in Eq. equation 3 is equivalent to solving the generalized eigenvalue problem $\mathbf{S}_\mathrm{B}\mathbf{w} = \lambda\mathbf{S}_\mathrm{W}\mathbf{w}$.

B.1 The One-Dimensional Case

We first consider the case where $k = 1$, in which the projection matrix $\mathbf{W}$ reduces to a single vector $\mathbf{w} \in \mathbb{R}^d$. The Fisher criterion in Eq. equation 3 simplifies to the generalized Rayleigh quotient:

$$J(\mathbf{w}) = \frac{\mathbf{w}^\top \mathbf{S}_\mathrm{B}\mathbf{w}}{\mathbf{w}^\top \mathbf{S}_\mathrm{W}\mathbf{w}}. \tag{11}$$

Since $J(\mathbf{w})$ is invariant to the scaling of $\mathbf{w}$ (*i.e.*, $J(\alpha\mathbf{w}) = J(\mathbf{w})$ for any $\alpha \neq 0$), we can impose the normalization constraint $\mathbf{w}^\top \mathbf{S}_\mathrm{W}\mathbf{w} = 1$ without loss of generality. The optimization problem then becomes:

$$\max_{\mathbf{w}} \ \mathbf{w}^\top \mathbf{S}_\mathrm{B}\mathbf{w} \quad \text{s.t.} \quad \mathbf{w}^\top \mathbf{S}_\mathrm{W}\mathbf{w} = 1. \tag{12}$$

Introducing the Lagrange multiplier λ, the Lagrangian is given by:

$$\mathcal{L}(\mathbf{w}, \lambda) = \mathbf{w}^\top \mathbf{S}_\mathrm{B}\mathbf{w} - \lambda(\mathbf{w}^\top \mathbf{S}_\mathrm{W}\mathbf{w} - 1). \tag{13}$$

Setting the gradient with respect to $\mathbf{w}$ to zero yields:

$$\frac{\partial \mathcal{L}}{\partial \mathbf{w}} = 2\mathbf{S}_\mathrm{B}\mathbf{w} - 2\lambda\mathbf{S}_\mathrm{W}\mathbf{w} = \mathbf{0}, \tag{14}$$

which directly gives the generalized eigenvalue problem:

$$\mathbf{S}_\mathrm{B}\mathbf{w} = \lambda\mathbf{S}_\mathrm{W}\mathbf{w}. \tag{15}$$

Left-multiplying both sides of Eq. equation 15 by $\mathbf{w}^\top$ and applying the constraint $\mathbf{w}^\top \mathbf{S}_\mathrm{W}\mathbf{w} = 1$, we obtain $\mathbf{w}^\top \mathbf{S}_\mathrm{B}\mathbf{w} = \lambda$. Hence, the value of the Fisher criterion at any stationary point equals its corresponding eigenvalue, and the maximum is attained by the eigenvector associated with the largest eigenvalue $\lambda_{\max}$.

B.2 Generalization to the Multi-Dimensional Case

For $k > 1$, the determinant-based Fisher criterion in Eq. equation 3 can be reformulated by exploiting its scale invariance. Specifically, under any non-singular transformation $\mathbf{W} \to \mathbf{WA}$ with $\mathbf{A} \in \mathbb{R}^{k\times k}$, the criterion remains unchanged:

$$\frac{\det(\mathbf{A}^\top \mathbf{W}^\top \mathbf{S}_\mathrm{B} \mathbf{W}\mathbf{A})}{\det(\mathbf{A}^\top \mathbf{W}^\top \mathbf{S}_\mathrm{W} \mathbf{W}\mathbf{A})} = \frac{\det(\mathbf{A})^2 \det(\mathbf{W}^\top \mathbf{S}_\mathrm{B} \mathbf{W})}{\det(\mathbf{A})^2 \det(\mathbf{W}^\top \mathbf{S}_\mathrm{W} \mathbf{W})} = \frac{\det(\mathbf{W}^\top \mathbf{S}_\mathrm{B} \mathbf{W})}{\det(\mathbf{W}^\top \mathbf{S}_\mathrm{W} \mathbf{W})}. \tag{16}$$

This invariance allows us to impose the orthonormality constraint $\mathbf{W}^\top \mathbf{S}_\mathrm{W} \mathbf{W} = \mathbf{I}_k$, reducing Eq. equation 3 to:

$$\max_{\mathbf{W}} \det(\mathbf{W}^\top \mathbf{S}_\mathrm{B} \mathbf{W}) \quad \text{s.t.} \quad \mathbf{W}^\top \mathbf{S}_\mathrm{W} \mathbf{W} = \mathbf{I}_k. \tag{17}$$

Assuming $\mathbf{S}_\mathrm{W}$ is positive definite (otherwise, regularization $\mathbf{S}_\mathrm{W} + \epsilon \mathbf{I}$ can be applied), it admits a Cholesky factorization $\mathbf{S}_\mathrm{W} = \mathbf{L}\mathbf{L}^\top$. Defining $\mathbf{V} = \mathbf{L}^\top \mathbf{W}$, the constraint becomes $\mathbf{V}^\top \mathbf{V} = \mathbf{I}_k$ and the objective is reformulated as:

$$\max_{\mathbf{V}} \det\left(\mathbf{V}^\top \mathbf{L}^{-1} \mathbf{S}_\mathrm{B} \mathbf{L}^{-\top} \mathbf{V}\right) \quad \text{s.t.} \quad \mathbf{V}^\top \mathbf{V} = \mathbf{I}_k. \tag{18}$$

Let $\tilde{\mathbf{S}} = \mathbf{L}^{-1} \mathbf{S}_\mathrm{B} \mathbf{L}^{-\top}$, which is symmetric positive semi-definite. By the Ky Fan theorem (Fan, 1949), the maximum of $\det(\mathbf{V}^\top \tilde{\mathbf{S}} \mathbf{V})$ subject to $\mathbf{V}^\top \mathbf{V} = \mathbf{I}_k$ is attained when the columns of $\mathbf{V}$ are the k eigenvectors of $\tilde{\mathbf{S}}$ corresponding to its k largest eigenvalues, *i.e.*, $\tilde{\mathbf{S}}\mathbf{v} = \lambda \mathbf{v}$.

Substituting $\tilde{\mathbf{S}} = \mathbf{L}^{-1} \mathbf{S}_\mathrm{B} \mathbf{L}^{-\top}$ and $\mathbf{v} = \mathbf{L}^\top \mathbf{w}$, we obtain:

$$\mathbf{L}^{-1} \mathbf{S}_\mathrm{B} \mathbf{L}^{-\top} \mathbf{L}^\top \mathbf{w} = \lambda \mathbf{L}^\top \mathbf{w} \implies \mathbf{L}^{-1} \mathbf{S}_\mathrm{B} \mathbf{w} = \lambda \mathbf{L}^\top \mathbf{w}. \tag{19}$$

Left-multiplying both sides by $\mathbf{L}$ and using $\mathbf{L}\mathbf{L}^\top = \mathbf{S}_\mathrm{W}$, we recover the generalized eigenvalue problem:

$$\mathbf{S}_\mathrm{B} \mathbf{w} = \lambda \mathbf{S}_\mathrm{W} \mathbf{w}. \tag{20}$$

C Statistical Analysis of Task Separation

This section establishes the statistical results used in Section 4.3. We first derive a pairwise separation bound on the Bayes error and its mutual-information consequence, then discuss the interpretation of Fisher separation and its connection to prototype routing.

Setting. At a fixed training stage, let $Y \in \{1, \ldots, s\}$ denote the identity of a seen task, where $s \geq 2$, and let $Z \in \mathbb{R}^d$ denote its frozen feature representation. We assume

$$Z \mid Y = t \sim \mathcal{N}(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}), \qquad \boldsymbol{\Sigma} \succ 0, \tag{21}$$

with priors $\pi_t = \Pr(Y = t) > 0$ satisfying $\sum_{t=1}^{s} \pi_t = 1$. Here, $\boldsymbol{\mu}_t$ is the population task mean, distinguished from its empirical estimate $\hat{\boldsymbol{\mu}}_t$. Define the squared pairwise Mahalanobis distance as

$$\delta_{ij}^2 = (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j), \qquad \delta_{\min}^2 = \min_{i<j} \delta_{ij}^2. \tag{22}$$

All logarithms are natural, and $h_2(u) = -u \log u - (1-u)\log(1-u)$ denotes binary entropy, with $0 \log 0 = 0$.

The following proposition combines classical Gaussian overlap bounds with Fano's inequality (Fukunaga, 2013; Devroye, 1987).

Proposition C.1 (Pairwise separation and Bayes error)**.** *Under the model above, let* $P_e^\star = \inf_r \Pr(r(Z) \neq Y)$ *denote the Bayes error over measurable task classifiers* r*. Define*

$$q = \min\left\{ 1 - \max_{1\leq t\leq s} \pi_t, \sum_{1\leq i<j\leq s} \sqrt{\pi_i \pi_j} \exp\left(-\frac{\delta_{ij}^2}{8}\right) \right\}. \tag{23}$$

Then $P_e^\star \leq q$*, and*

$$I(Y; Z) \geq \max\left\{0,\ H(Y) - h_2(q) - q \log(s-1)\right\}. \tag{24}$$

Proof. Let $f_t(z)$ denote the density of $Z \mid Y = t$, and write $a_t(z) = \pi_t f_t(z)$. The Bayes error can be expressed as

$$\begin{aligned} P_e^\star &= \int_{\mathbb{R}^d} \left[\sum_{t=1}^{s} a_t(z) - \max_{1 \le t \le s} a_t(z) \right] dz \\ &\le \sum_{i<j} \int_{\mathbb{R}^d} \min\{a_i(z), a_j(z)\}\, dz \\ &\le \sum_{i<j} \sqrt{\pi_i \pi_j} \int_{\mathbb{R}^d} \sqrt{f_i(z) f_j(z)}\, dz. \end{aligned} \tag{25}$$

To see the first inequality, fix z and choose an index m attaining the largest $a_t(z)$. The sum of the remaining terms equals $\sum_{t \neq m} \min\{a_t(z), a_m(z)\}$, which is included in the sum over all pairs. The second inequality follows from $\min\{a, b\} \le \sqrt{ab}$ for nonnegative a, b.

For Gaussians with a common covariance, completing the square gives

$$\sqrt{f_i(z) f_j(z)} = \exp\left(-\frac{\delta_{ij}^2}{8}\right) \varphi\left(z; \frac{\boldsymbol{\mu}_i + \boldsymbol{\mu}_j}{2}, \boldsymbol{\Sigma}\right), \tag{26}$$

where $\varphi(z; \boldsymbol{m}, \boldsymbol{\Sigma})$ is the Gaussian density with mean $\boldsymbol{m}$ and covariance $\boldsymbol{\Sigma}$. Integrating this identity and substituting into Eq. equation 25 yields

$$P_e^\star \le \sum_{i<j} \sqrt{\pi_i \pi_j} \exp\left(-\frac{\delta_{ij}^2}{8}\right). \tag{27}$$

Always predicting the most probable task gives error $1 - \max_t \pi_t$. Since the Bayes classifier minimizes the error, it also satisfies $P_e^\star \le 1 - \max_t \pi_t$. Combining these two bounds gives $P_e^\star \le q \le 1 - 1/s$.

Fano's inequality gives

$$H(Y \mid Z) \le h_2(P_e^\star) + P_e^\star \log(s-1). \tag{28}$$

Define $F_s(u) = h_2(u) + u \log(s-1)$. For $0 < u \le 1 - 1/s$,

$$F_s'(u) = \log\left(\frac{(s-1)(1-u)}{u}\right) \ge 0. \tag{29}$$

By continuity at zero, F_s is nondecreasing on $[0, 1 - 1/s]$. Therefore,

$$H(Y \mid Z) \le F_s(P_e^\star) \le F_s(q). \tag{30}$$

Using $I(Y; Z) = H(Y) - H(Y \mid Z)$ and the nonnegativity of mutual information establishes Eq. equation 24. □

Equal task priors. For $\pi_t = 1/s$, the pairwise overlap bound satisfies

$$\sum_{i<j} \sqrt{\pi_i \pi_j} \exp\left(-\frac{\delta_{ij}^2}{8}\right) = \frac{1}{s} \sum_{i<j} \exp\left(-\frac{\delta_{ij}^2}{8}\right) \le \frac{s-1}{2} \exp\left(-\frac{\delta_{\min}^2}{8}\right). \tag{31}$$

Combining this inequality with $P_e^\star \le 1 - 1/s$ gives Eq. equation 9 in the main text. The mutual-information bound also holds with $H(Y) = \log s$ and q replaced by the clipped upper bound in Eq. equation 9.

Relation to Fisher separation. Let $\bar{\boldsymbol{\mu}} = \sum_t \pi_t \boldsymbol{\mu}_t$ and define the population between-task covariance as

$$\mathbf{B}_{\text{pop}} = \sum_{t=1}^{s} \pi_t (\boldsymbol{\mu}_t - \bar{\boldsymbol{\mu}})(\boldsymbol{\mu}_t - \bar{\boldsymbol{\mu}})^\top. \tag{32}$$

Expanding the centered outer products yields $\mathbf{B}_{\text{pop}} = \frac{1}{2} \sum_{i,j} \pi_i \pi_j (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^\top$. Consequently,

$$J_{\text{pop}} = \operatorname{tr}\left(\boldsymbol{\Sigma}^{-1} \mathbf{B}_{\text{pop}}\right) = \sum_{i<j} \pi_i \pi_j \delta_{ij}^2. \tag{33}$$

Thus, the Fisher trace measures a weighted sum of pairwise separations. This aggregate quantity does not provide a positive lower bound on the minimum pairwise separation.

For example, consider three equally likely one-dimensional tasks with unit variance and means 0, 0, and M. Then

$$J_{\text{pop}} = \frac{2M^2}{9}, \qquad \delta^2_{\min} = 0, \qquad P_e^\star \geq \frac{1}{3}. \tag{34}$$

As M increases, the Fisher trace becomes arbitrarily large, while the first two tasks remain indistinguishable. Therefore, the shared-covariance Gaussian assumption alone does not justify replacing pairwise separation by a positive multiple of the Fisher trace in the error bound. A large Fisher score describes aggregate separation but does not by itself guarantee accurate task routing.

Connection to prototype routing. Under the assumed Gaussian model, maximizing the posterior task probability is equivalent to

$$\hat{t}_{\text{Bayes}}(z) = \arg\min_{1\leq t\leq s} \left\{ (z-\boldsymbol{\mu}_t)^\top \boldsymbol{\Sigma}^{-1}(z-\boldsymbol{\mu}_t) - 2\log \pi_t \right\}. \tag{35}$$

For equal priors, the prior term is constant across tasks, giving nearest-prototype classification under the shared Mahalanobis metric.

For any fixed full-column-rank projection $\mathbf{W}$,

$$\mathbf{W}^\top Z \mid Y = t \sim \mathcal{N}\left(\mathbf{W}^\top \boldsymbol{\mu}_t,\, \mathbf{W}^\top \boldsymbol{\Sigma} \mathbf{W}\right). \tag{36}$$

Proposition C.1 therefore applies in the projected space after recomputing the pairwise distances using these projected means and covariance. This provides a population-level interpretation of covariance-aware routing. The practical ILDA router uses finite-sample estimates and regularization, so the proposition does not directly bound its implemented routing error. The analysis concerns discrimination among seen tasks; the separate gating procedure for task-agnostic inference is described in Section 4.4.

D Implementation Details

Architecture. We build our method on top of the CLIP model (Radford et al., 2021) with a ViT-B/16 vision encoder, initialized from the official OpenAI checkpoint. The text encoder is kept completely frozen throughout training, and we use the basic prompt template "`a photo of a {}.`" for all categorical embeddings. The intermediate rank of LoRA is set to 128.

Training configuration. All models are trained with the AdamW optimizer, using an initial learning rate of 5×10^{-3}, weight decay of 0.01, and a cosine learning rate schedule with warmup. We train for 30 epochs per task with a batch size of 128. Standard data augmentations in CLIP are applied during training; images are resized to 224×224 for training and evaluation.

Hardware. All experiments are conducted on a single NVIDIA RTX 4090 GPU. Training sequentially across the full suite of ten datasets takes approximately 8 hours of wall-clock time. The ILDA-based distribution discriminator is training-free and incurs no additional GPU cost.

E Baselines

We compare ILDA-CL against the following three groups of baselines.

E.1 Task-agnostic CL Methods

LwF (Li and Hoiem, 2017). Learning without Forgetting is a classical regularization-based approach that mitigates catastrophic forgetting via knowledge distillation: the predictions of the previous model on current-task data are used as soft targets to constrain parameter drift. We adapt LwF to CLIP by distilling logits computed with frozen text embeddings.

WiSE-FT (Wortsman et al., 2022). WiSE-FT linearly interpolates the weights of the fine-tuned model and the original pre-trained CLIP model, achieving a trade-off between downstream adaptation and zero-shot robustness without additional training cost.

ZSCL (Zheng et al., 2023). ZSCL prevents zero-shot transfer degradation during continual fine-tuning by distilling from the frozen CLIP model on a large reference image set, while also performing weight ensembling along the optimization trajectory to preserve generalization.

MoE-Adapter (Yu et al., 2024). MoE-Adapter introduces a Mixture-of-Experts adapter on top of frozen CLIP, where a learnable router dynamically activates task-relevant experts. It enables parameter-efficient continual adaptation while preserving CLIP's pre-trained knowledge in the frozen backbone.

Dual-RAIL (Xu et al., 2024). Dual-RAIL employs a dual-branch architecture combining a primary adaptation branch with a regularization branch grounded in CLIP's pre-trained representations, enabling continual learning while retaining zero-shot capability on unseen distributions.

SECA (He et al., 2026). SECA leverages the semantic structure of CLIP's text embeddings to guide continual adaptation, harnessing the alignment between visual and textual modalities to reduce interference between sequentially arriving tasks.

MERGETUNE (Wang et al., 2026). MERGETUNE is a post-hoc, model-agnostic method that restores CLIP's pretrained knowledge by searching for a continued model linearly mode-connected to both the zero-shot and fine-tuned checkpoints, with a second-order Taylor surrogate replacing the data-replay term on the zero-shot side.

E.2 CIL-oriented CL Methods

ConDU (Gao et al., 2026). ConDU enhances CLIP-based class-incremental learning by [explicitly modeling class-conditional distributions / decoupling representation and classifier updates / etc.], enabling improved separation between previously learned and newly introduced classes.

CoLeCLIP (Li et al., 2025). CoLeCLIP performs continual learning of CLIP by [introducing task-specific prompts / lightweight adapters] while maintaining a unified classification head across tasks, supporting class-incremental evaluation without task identity at inference.

DPeCLIP (Lu et al., 2024). DPeCLIP boosts CLIP's class-incremental learning performance through [dual-prompt / parameter-efficient] tuning, balancing stability on previously seen classes and plasticity for newly introduced ones.

E.3 Reference Points

Continual-FT. Unrestricted continual fine-tuning of all CLIP parameters across the task sequence without any forgetting-mitigation mechanism. This serves as a lower-bound reference, illustrating the severity of catastrophic forgetting in the absence of dedicated continual learning techniques.

Full Fine-tune. Full-parameter joint training on the union of all ten datasets, with all data available simultaneously. This violates the continual learning assumption but provides an approximate upper bound on achievable accuracy.

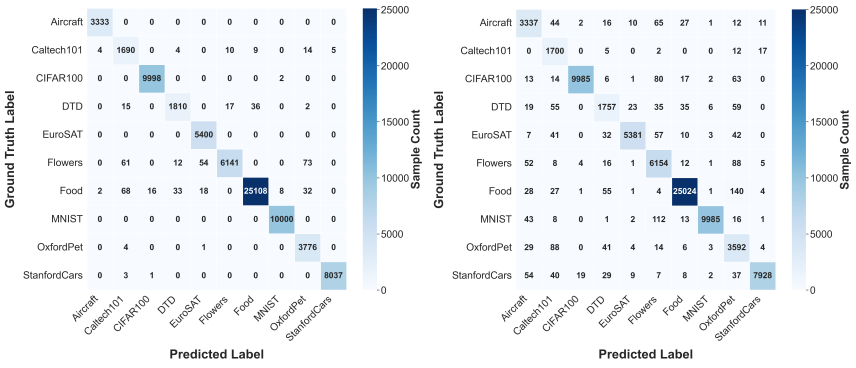
Zero-shot CLIP. The original CLIP model evaluated without any continual updates, using only the prompt template "`a photo of a {}.`". This reflects the pre-trained model's intrinsic generalization to the benchmark datasets and serves as a reference for the value added by continual adaptation.

F The Confusion Matrix of ILDA-based Discriminator

The confusion matrix of ILDA serving as a domain discriminator upon completion of the entire training procedure is presented in Figure 9, further demonstrating its domain discrimination accuracy.

G Detailed Results with Reduced Dimensions

Table 5 presents an ablation study on the dimensionality of the intermediate CLIP features, comparing two reduced variants (ILDA-CL-178 and ILDA-CL-697, corresponding to the results in Section 5)



(a) Complete datasets. (b) 5-shot datasets.

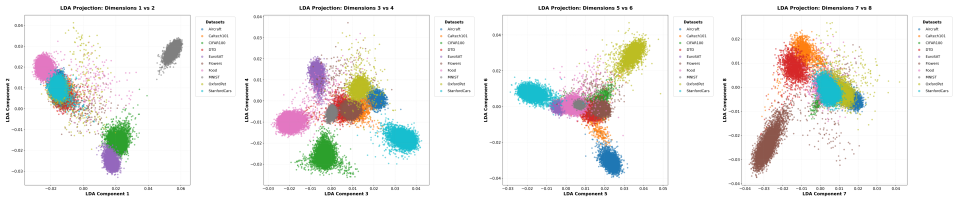
Figure 9: Confusion matrices of domain discrimination achieved by the proposed **ILDA-CL** under the CIL configuration.

against the default **ILDA-CL** under the task-agnostic CL setting. The results indicate that dimensionality reduction has minimal impact on **ILDA-CL**. When 178 dimensions are retained, both the *Average* and *Last* accuracies exhibit slight declines; the results with 697 dimensions, however, are virtually identical to those obtained under the default **ILDA-CL** configuration.

Table 5: Experimental results with reduced intermediate CLIP dimensions.

	Method	Air.	Cal.	CIF.	DTD	Eur.	Flo.	Foo.	MNI.	Oxf.	Sta.	***Avg.***
Transfer	ILDA-CL-178	-	78.6	60.8	43.7	54.9	69.7	88.4	43.6	88.9	64.7	65.9
	ILDA-CL-679	-	78.6	60.8	43.7	54.9	69.7	88.4	43.6	88.9	64.7	65.9
	ILDA-CL	-	78.6	60.8	43.7	54.9	69.7	88.4	43.6	88.9	64.7	65.9
Average	ILDA-CL-178	59.2	89.7	77.3	63.1	80.7	81.2	89.7	60.4	89.1	66.5	76.1
	ILDA-CL-679	59.1	90.5	77.3	65.8	80.7	81.4	89.9	60.4	89.8	66.8	76.2
	ILDA-CL	59.1	90.5	77.3	65.8	80.7	81.4	89.9	60.4	89.8	66.8	76.2
Last	ILDA-CL-178	59.2	95.5	87.8	73.9	98.9	92.9	92.0	99.6	91.0	83.5	87.4
	ILDA-CL-679	59.1	96.3	87.7	76.0	98.9	93.1	92.5	99.6	93.5	85.4	88.2
	ILDA-CL	59.1	96.3	87.7	76.0	98.9	93.1	92.5	99.6	93.5	85.4	88.2

H VISUALIZATION OF LDA SPACE



(a) (b) (c) (d)

Figure 10: Visualization of the first eight dimensions of the intermediate features mapped to the ILDA space.

To further elucidate the representation power of the proposed incremental learning framework, we visualize the intermediate features projected into the LDA space. The first eight discriminant dimensions is shown in Figure 10. Superior task-to-task decoupling capability is exhibited within the LDA space. Global features exhibiting significant statistical differences are effectively captured within the low-dimensional space, as exemplified by the outlier MNIST dataset depicted in Figure 10a.

With the increase in dimensionality (as illustrated in Figure 10c and Figure 10d), distinct manifolds are gradually formed along specific discriminative directions by categories that previously exhibited potential semantic overlap within low-dimensional spaces. These visualizations demonstrate that the eigenvectors corresponding to different dimensions of the optimal projection $\mathbf{W}^*$ effectively separate samples across distinct classes.

I LIMITATION

While this work provides both theoretical analysis and empirical validation, several limitations warrant discussion. First, the theoretical analysis in Section 4.3 is derived under a Gaussian assumption on feature distributions, which may not hold exactly in practice. As a result, the maximum of the Fisher criterion does not always coincide with the truly optimal intermediate layer. Developing more refined evaluation metrics that relax this assumption constitutes a promising direction for future work. Second, the benchmarks adopted in our experiments are characterized by pronounced semantic discrepancies across domains, which provide a favorable setting for distribution-level discrimination.