Select and Distill:
Selective Dual-Teacher Knowledge Transfer for Continual Learning on Vision-Language Models

Yu-Chu Yu[1,†], Chi-Pin Huang[1], Jr-Jen Chen[1], Kai-Po Chang[1], Yung-Hsuan Lai[1], Fu-En Yang[2], and Yu-Chiang Frank Wang[1,2,‡]

[1] National Taiwan University
[2] NVIDIA
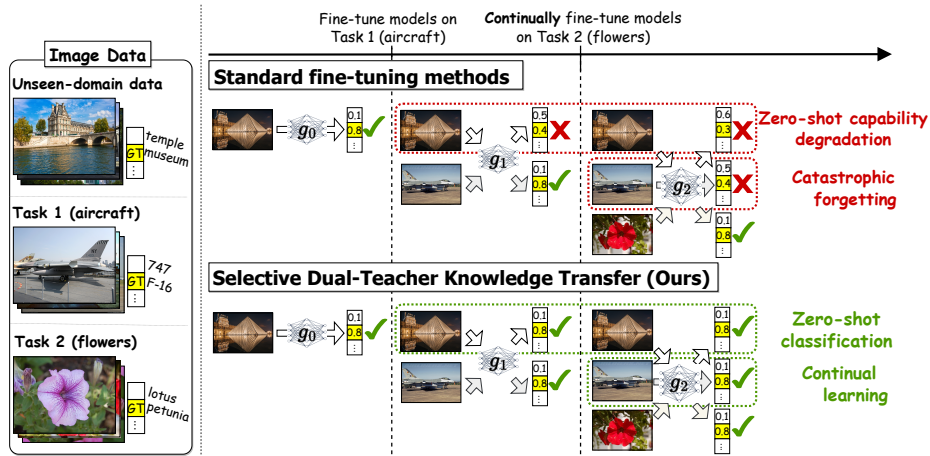[†] r09922104@ntu.edu.tw, [‡] frankwang@nvidia.com

Abstract. Large-scale vision-language models (VLMs) have shown a strong zero-shot generalization capability on unseen-domain data. However, adapting pre-trained VLMs to a sequence of downstream tasks often leads to the forgetting of previously learned knowledge and a reduction in zero-shot classification performance. To tackle this problem, we propose a unique Selective Dual-Teacher Knowledge Transfer framework that leverages the most recent fine-tuned and the original pre-trained VLMs as dual teachers to preserve the previously learned knowledge and zero-shot capabilities, respectively. With only access to an unlabeled reference dataset, our proposed framework performs a selective knowledge distillation mechanism by measuring the feature discrepancy from the dual-teacher VLMs. Consequently, our selective dual-teacher knowledge distillation mitigates catastrophic forgetting of previously learned knowledge while preserving the zero-shot capabilities of pre-trained VLMs. Extensive experiments on benchmark datasets demonstrate that our framework is favorable against state-of-the-art continual learning approaches for preventing catastrophic forgetting and zero-shot degradation. Project page: https://chuyu.org/research/snd.

Keywords: Continual Learning · Vision-Language Models · Knowledge Distillation

1 Introduction

With the access to large-scale data available for training, vision-language models (VLMs) have demonstrated unprecedented progress in visual and linguistic applications [1, 30, 46, 49]. Despite the significant achievement in static benchmark datasets, it is not easy to have VLMs incrementally accumulate the knowledge learned from previous tasks, while maintaining sufficient generalization ability. The former is known as the *catastrophic forgetting* problem [28], while the latter is the zero-shot transfer capability of VLMs.

Continual learning (CL) has emerged as a potential approach, which aims to gradually adapt the trained model to a new task without forgetting the previously learned ability. With the goal of preventing severe overfitting on currently

Fig. 1: Compared with standard fine-tuning models, our *Selective Dual-Teacher Knowledge Transfer* advances continual learning to mitigate catastrophic forgetting on previously fine-tuned tasks, while preserving the model's zero-shot capability.

available data and the consequent performance degradation on previous tasks, previous CL works [6, 7, 16, 26, 35, 36] are proposed to store previous data in a memory buffer. While effective for mitigating the catastrophic forgetting issue of past tasks, the scalability is hampered due to the limited memory size, restricting the deployment in the scenarios of growing fast new data. Instead of storing previous datasets in the memory buffer, recent methods [8,14,25,33,39,47] adopt a *data-free* manner, which synthesizes the data of the previously trained tasks from the corresponding semantic labels. However, these methods are primarily designed for *close-set* image recognition tasks that the label space is manually pre-determined. It remains challenging to handle the *open-vocabulary* nature in vision-language models (e.g., CLIP [34]) for zero-shot classification capability.

To address the degradation of zero-shot capabilities during sequential model fine-tuning, very recent work ZSCL [50] is proposed to regularize the optimization on the current task through guidance from the original pre-trained VLMs (e.g., CLIP [34]). More specifically, ZSCL [50] distills the knowledge from a teacher VLM, which remains frozen without fine-tuning, to constrain the fine-tuned student VLM using an *unlabeled* reference dataset (e.g., ImageNet [11]). This approach allows the student VLM to preserve the intrinsic zero-shot transfer capability of VLMs during fine-tuning without requiring the assessment of the pre-trained dataset. However, such a manner solely considers the zero-shot capability of VLMs. The knowledge learned from previous tasks cannot be readily preserved since the pre-trained model has never fine-tuned on previous tasks, resulting in limited performance improvement against catastrophic forgetting. Therefore, incrementally expanding the learned capability from previous tasks remains a challenging and unsolved problem.

In this paper, we propose a *Selective Dual-Teacher Knowledge Transfer* framework, as depicted in Fig. 1. Aiming at simultaneously enabling continual adaptation for sequentially arrived tasks while retaining the robust zero-shot transfer capability inherent in pre-trained VLMs, we follow the setting of [50] and leverage both the most recent fine-tuned and the original pre-trained VLMs. Without accessing the information from previous tasks, we propose a teacher selection mechanism from dual-teacher discrepancy to identify which teacher network is favored with a given image sampled from an unlabeled reference dataset. More specifically, if the reference image aligns with prior data distribution, the most recent fine-tuned VLM would be preferable to retain the knowledge learned from past tasks. On the other hand, for other reference images that are out of the previous distribution, the original pre-trained VLM is selected to prevent the degradation of zero-shot capabilities. As a result, a selective knowledge distillation from the dual-teacher VLMs could be properly performed to enable continual learning on vision-language models.

We now summarize our contributions as below:

– We propose a *Selective Dual-Teacher Knowledge Transfer* framework that simultaneously alleviates catastrophic forgetting problems and preserves the zero-shot capabilities from the pre-trained VLM.
– By observing an unlabeled reference dataset, our framework views pre-trained and the most recently finetuned models as dual-teachers, and selects the proper one for knowledge distillation based on the introduced discrepancy.
– Extensive evaluations on several benchmark datasets in various incremental learning settings confirm that our approach performs favorably against existing continual learning methods, alleviating both catastrophic forgetting and zero-shot degradation.

2 Related Work

Rehearsal-Based Continual Learning. Rehearsal-based continual learning [6, 7,16,26,35,36] mitigates catastrophic forgetting by maintaining a subset of previous training data in a memory buffer, and the stored data can then be combined with current data for regularizing model fine-tuning. For example, iCaRL [35] efficiently selects representative samples from previous tasks to maintain an evenly distributed memory buffer. LUCIR [16] addresses the issue where the model incorrectly favors newer classes caused by potential data imbalance that exists between previous and new tasks. However, retaining data from previous tasks in a memory buffer poses risks of privacy leakage and a costly storage burden, restricting the scalability in real-world deployments.

Data-Free Continual Learning. Data-Free Continual Learning (DF-CL) [8, 14, 24, 25, 33, 39, 47] aims to preserve knowledge learned from past tasks without accessing their data. Several DF-CL methods [8, 14, 25, 33, 39, 47] are learned to synthesize prior data given its corresponding semantic label. Then, they could

regularize the fine-tuning of the current task by distilling the knowledge from previous models using the synthetic data of prior tasks. In addition, another line of methods [41–43] focuses on learning lightweight prompts to encode task-specific information on top of a frozen pre-trained Vision Transformer (ViT). In this way, they are able to guide the frozen pre-trained ViT to perform the current task without forgetting the previous knowledge captured in the task-specific prompts. Although these methods have shown remarkable abilities in recalling previously learned data, they mainly focus on *close-set* image classification tasks, so they still cannot readily be applied to *open-vocabulary* VLMs, which require simultaneously retaining the knowledge learned from prior tasks and zero-shot capability inherent in large-scale pre-trained VLMs.
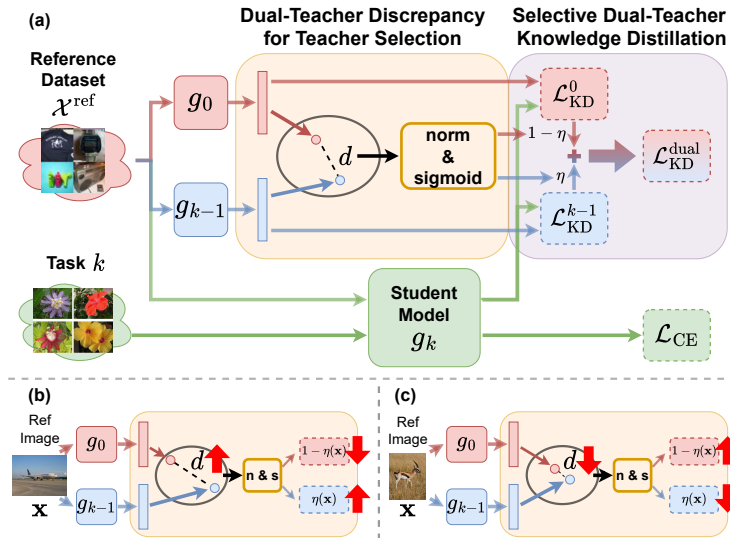
Continual Learning on Vision-Language Models. Recently, VLMs [19, 32, 34] pre-trained on large-scale datasets have demonstrated robust zero-shot transferability for open-vocabulary downstream tasks. However, recent studies [22, 44] have shown that the zero-shot capability is prone to deteriorate when fine-tuning the pre-trained VLMs to specific domains. With the aim of preserving the zero-shot capability during model fine-tuning, ZSCL [50] is proposed to regularize the model via the guidance from original pre-trained VLMs. Without the need to access the pre-trained dataset, ZSCL [50] claims that performing knowledge distillation from pre-trained VLMs on an *unlabeled* reference dataset (e.g., ImageNet [11]) can effectively preserve the zero-shot capabilities during model fine-tuning. While promising, ZSCL [50] primarily considers preventing zero-shot transfer degradation. It cannot easily expand the knowledge derived from sequentially arrived downstream tasks where only an unlabeled reference dataset is accessible. In this work, we propose a unique *Selective Dual-Teacher Knowledge Transfer* framework, which aims at simultaneously preserving zero-shot transferability while mitigating catastrophic forgetting for previous tasks.

3   Method

3.1   Problem Formulation

For the sake of completeness, we first define the problem setting in this paper. In the context of continual learning, we assume that the model has been trained on K sequentially arrived tasks $\{\mathcal{T}^1, \mathcal{T}^2, \cdots, \mathcal{T}^K\}$, where the k-th task $\mathcal{T}^k = (\mathcal{X}^k, \mathcal{Y}^k)$ contains N^k images $\mathcal{X}^k = \{\mathbf{x}_i^k\}_{i=1}^{N^k}$ with L^k class labels $\mathcal{Y}^k \subseteq \{1, 2, \cdots, L^k\}$. Following [50], we only have access to the most recent fine-tuned VLM (g_{k-1}) and the original pre-trained VLM (g_0), but *not* the data from previous tasks (i.e., $\{\mathcal{T}^j\}_{j=1}^{k-1}$). On the other hand, an unlabeled reference dataset $\mathcal{X}^{\text{ref}}$ (e.g., ImageNet [11]) can be utilized to during continual learning (as [50] did). For continual leaning on VLMs, the model g_k is expected to preserve not only the knowledge learned from previous tasks $\{\mathcal{T}^j\}_{j=1}^{k-1}$), but also the zero-shot transfer capability of large-scale pre-trained VLMs like CLIP [34] during model fine-tuning on $\mathcal{T}^k$.

Fig. 2: (a) The overall architecture of our proposed *Selective Dual-Teacher Knowledge Transfer* framework. **(b)** Selective knowledge transfer from g_{k-1} due to larger discrepancy d between dual teachers g_0 and g_{k-1}, alleviating catastrophic forgetting on Task $k-1$. **(c)** Selective knowledge transfer from g_0 due to smaller discrepancy d between dual teachers g_0 and g_{k-1}, preserving the zero-shot capability of g_0.

3.2 Selective Dual-Teacher Knowledge Transfer on VLMs

Given the most recent fine-tuned VLM g_{k-1} and the original pre-trained VLM g_0, together with an unlabeled reference dataset $\mathcal{X}^{\mathrm{ref}}$, our goal is to tackle catastrophic forgetting and to preserve zero-shot transfer capability for continual learning of VLMs. Since the alignment between $\mathcal{X}^{\mathrm{ref}}$ and data from tasks $\{\mathcal{T}^1, \mathcal{T}^2, \cdots, \mathcal{T}^{(k-1)}\}$ is not known, direct knowledge distillation from g_{k-1} on $\mathcal{X}^{\mathrm{ref}}$ might not be desirable.

As shown in Fig. 2, we propose a novel framework, *Selective Dual-Teacher Knowledge Transfer*, to perform VLM continual learning, aiming to alleviate catastrophic forgetting while preserving zero-shot transferability. We view g_{k-1} and (g_0) as *dual teachers* to perform selective knowledge transfer. That is, for each image extracted from $\mathcal{X}^{\mathrm{ref}}$, we need to identify the proper teacher to perform knowledge distillation. In the following subsections, we will detail how we utilize such irrelevant/unlabeled data and present our selection process. We will explain how our selective dual-teacher knowledge transfer would jointly alleviate catastrophic forgetting on data from previous tasks while retaining the zero-shot capabilities on unseen image data.

Dual-Teacher Discrepancy for Teacher Selection. In our work, we distill the knowledge from the dual teacher networks of the most recent fine-tuned

VLM g_{k-1} and the pre-trained VLM g_0, as depicted in Fig. 2. The problem is that, one cannot easily determine the knowledge from which teacher VLM to be distilled when observing an image $\mathbf{x}^{\text{ref}}$ from $\mathcal{X}^{\text{ref}}$. If $\mathbf{x}^{\text{ref}}$ does not match the distribution of data observed by g_{k-1}, performing knowledge distillation from g_{k-1} would not preserve the zero-shot classification ability. On the other hand, if $\mathbf{x}^{\text{ref}}$ is visually similar to the previously fine-tuned data of g_{k-1}, it is not desirable to distill knowledge from g_0, as g_0 lacks specific knowledge to alleviate catastrophic forgetting on $\mathcal{T}^{k-1}$.

To tackle the above challenge, we propose a teacher selection mechanism based on the *dual-teacher discrepancy*. To be more precise, if a sampled reference image $\mathbf{x}^{\text{ref}}$ aligns with the distribution of previous datasets, the feature derived by the g_{k-1} would differ from that obtained by the pre-trained VLM g_0, inducing large dual teacher discrepancy d. On the other hand, as a reference image is out of previous data distribution, a smaller discrepancy d would be expected due to this reference image being unfamiliar to both teacher models, so that such unseen-domain data can be leveraged to facilitate zero-shot preservation. Thus, we can denote the relation is formulated as follows:

$$\mathbb{E}_{\mathbf{x} \in \mathcal{X}^{1:k-1}} \left[d(g_{k-1}(\mathbf{x}), g_0(\mathbf{x})) \right] \geq \mathbb{E}_{\mathbf{x}' \notin \mathcal{X}^{1:k-1}} \left[d(g_{k-1}(\mathbf{x}'), g_0(\mathbf{x}')) \right], \tag{1}$$

where $d : \mathcal{F} \times \mathcal{F} \mapsto [0, \infty)$ denotes dual-teacher discrepancy measurement, which is realized by an Euclidean distance in the feature space $\mathcal{F}$. $\mathcal{X}^{1:k-1} = \bigcup_{i=1}^{k-1} \mathcal{X}^i$ collects all data fine-tuned before, and $\mathbf{x}' \notin \mathcal{X}^{1:k-1}$ represents data that has not seen by g_{k-1} before. The above observation is also empirically evident in Tab. 3 and further analyzed in Sec. 4.5.

With the dual-teacher discrepancy d derived from g_{k-1} and g_0, we are able to select the favored teacher VLM for knowledge transfer given the sampled reference image $\mathbf{x}^{\text{ref}}$. To be more specific, we define a *selection scoring function* $\eta(\cdot) : \mathcal{X} \mapsto [0, 1]$ at task k that transforms the discrepancy d into a selection score, as computed as follows,

$$\eta(\mathbf{x}) = \sigma \left(\frac{d(g_{k-1}(\mathbf{x}), g_0(\mathbf{x})) - \delta}{\gamma} \right), \tag{2}$$

where $\delta, \gamma \in \mathbb{R}$ are hyper-parameters to normalize the feature discrepancy, and $\sigma : \mathbb{R} \mapsto [0, 1]$ is a sigmoid function mapping the normalized discrepancy to a scalar score between 0 and 1.

We note that, a *larger* selection score $\eta(\mathbf{x})$ (e.g., greater than 0.5) indicates the most recent fine-tuned VLM g_{k-1} would be preferable to mitigate catastrophic forgetting on prior tasks, as depicted in Fig. 2(b). Conversely, when the selection score $\eta(\mathbf{x})$ is *small*, the pre-trained VLM g_0 is expected to transfer the zero-shot capabilities, as illustrated in Fig. 2(c).

Selective Knowledge Distillation from Dual-Teachers. With the estimated teacher selection score $\eta(\mathbf{x})$ obtained, we are able to perform selective knowledge transfer from the dual teacher networks. As depicted in Fig. 2(a), at

the current task k, our framework selectively distills the knowledge from g_{k-1} and g_0 to the student VLM g_k to alleviate the forgetting of previous tasks and the degradation of zero-shot capabilities through the control by the teacher selection score $\eta(\mathbf{x})$. As a selective knowledge transfer from g_{k-1} is preferable when $\eta(\mathbf{x})$ is large while a knowledge distillation from g_0 is encouraged as relatively low $\eta(\mathbf{x})$, we compute the dual-teacher knowledge distillation objective $\mathcal{L}_{\text{dual}}$ as,

$$\mathcal{L}_{\text{KD}}^{\text{dual}} = \sum_{\mathbf{x} \sim \mathcal{X}^{\text{ref}}} \eta(\mathbf{x}) \cdot \mathcal{L}_{\text{KD}}^{k-1} + (1 - \eta(\mathbf{x})) \cdot \mathcal{L}_{\text{KD}}^0, \tag{3}$$

where $\mathcal{L}_{\text{KD}}^{k-1} = d(g_{k-1}(\mathbf{x}), g_k(\mathbf{x}))$ denotes a knowledge distillation objective which aligns the feature representations of the input $\mathbf{x}$ with that of the most recent fine-tuned model g_{k-1}, and $\mathcal{L}_{\text{KD}}^0 = d(g_0(\mathbf{x}), g_k(\mathbf{x}))$ aims to the align the feature representation with the pre-trained model g_0.

Combining with the standard cross-entropy loss function $\mathcal{L}_{\text{CE}}$ on the current task $\mathcal{T}^k$, our proposed framework is capable of retaining both previously fine-tuned knowledge from g_{k-1} and the inherent zero-shot transferability from the pre-trained VLM g_0 during fine-tuning on task k. In summary, the overall objective function of our proposed *Selective Dual-Teacher Knowledge Transfer* framework is formulated as below:

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \mathcal{L}_{\text{KD}}^{\text{dual}}. \tag{4}$$

3.3   Training and Inference

Training Phase. Following previous settings for continual learning on Vision-Language Models [50], we fine-tune the original pre-trained model CLIP [34] to the downstream tasks in a sequential manner. We summarize the training algorithm in our supplementary material. Note that at each stage k, we do not have access to data from previous tasks $\{\mathcal{T}^1, \cdots, \mathcal{T}^{k-1}\}$. After sequentially fine-tuning over all K different tasks, we derive a final model g_K that exhibits zero-shot classification capabilities with catastrophic forgetting suppressed.

Inference Phase. Once the learning of the proposed framework is complete, we deploy the derived g_K for performing image recognition tasks on each task $\{\mathcal{T}^1, \mathcal{T}^2, \cdots, \mathcal{T}^K\}$. Following the inference manner proposed in CLIP [34]. Let h denote the text encoder of the CLIP model g. Given a set of labels $\mathcal{Y}$ with L different categories, we convert the corresponding class name of each label y to a text feature vector $\mathbf{w}_y$. Then, the probability of a data $\mathbf{x}$ to the class y is calculated as below:

$$p(y|\mathbf{x}) = \frac{\exp(\cos(g(\mathbf{x}), \mathbf{w}_y)/\tau)}{\sum_{j=1}^{L} \exp(\cos(g(\mathbf{x}), \mathbf{w}_j)/\tau)}, \tag{5}$$

where τ is a temperature parameter learned by CLIP [34].

4   Experiment

4.1   Implementation Detail

In our experiments, we use CLIP [34] implemented by open_clip [17] with the ViT-B/16 [12] image encoder as our backbone. We optimize our model with AdamW, dynamically adjusting learning rates with a cosine scheduler started by 1×10^{-5} and a weight decay regularization set to 5×10^{-4}. During training, only the image encoder is updated while the text encoder is kept frozen. We standardize the text prompt to "a photo of a $<$CLASS$>$" during both the training and testing phases for classification purposes.

4.2   Datasets

We evaluate our proposed method on eight fine-grained classification datasets, including FGVC-Aircraft [27], DTD [9], EuroSAT [15], Flowers-102 [29], Food-101 [2], Oxford-Pets [31], Stanford-Cars [20], and UCF-101 [40]. Note that to avoid potential overlapping label spaces among different datasets, we alleviate coarse-grained datasets such as Caltech-101 [13], CIFAR100 [21], and SUN397 [45] used in previous works [50].
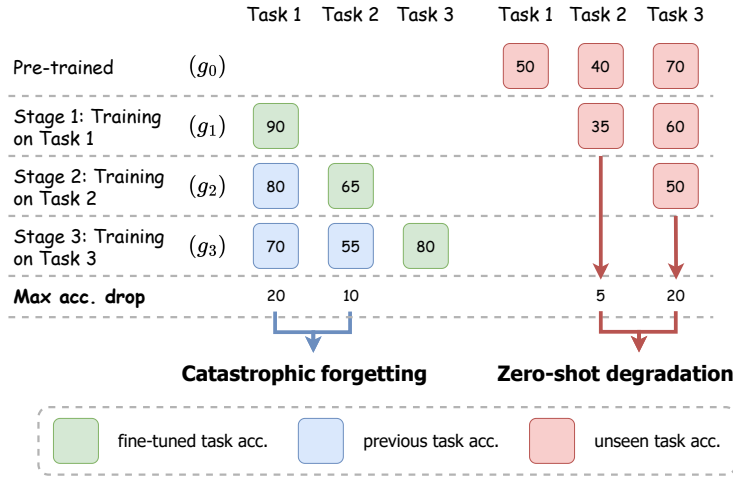
4.3   Evaluation Protocol

Multiple Training Sequences. Unlike previous benchmarks [50], which picked only one or two sequences to evaluate the performance, we construct K unique sequences to fully understand the level of forgetting after multiple training rounds for each dataset. Specifically, given K different image classification tasks, we first construct the first ordered sequence $\mathcal{S}^1 = (\mathcal{T}^1, \mathcal{T}^2, \cdots, \mathcal{T}^K)$. Based on $\mathcal{S}^1$, we shift the sequence to the left to derive the next sequence $\mathcal{S}^2 = (\mathcal{T}^2, \mathcal{T}^3, \cdots, \mathcal{T}^K, \mathcal{T}^1)$. Thus, the k-th sequence $\mathcal{S}^k$ contains ordered tasks:

$$\mathcal{S}^k = (\mathcal{T}^{k \ \% \ K}, \mathcal{T}^{(k+1) \ \% \ K}, \cdots, \mathcal{T}^{(k+K-1) \ \% \ K}), \tag{6}$$

where % indicates the mod operator. By training and testing on these K different sequences, each dataset has one chance to be the first and the last dataset during continual training progress. This allows us to thoroughly evaluate the catastrophic forgetting and the degradation of zero-shot transferability after training on K multiple rounds.

Specifically, we pick up the first training sequence $\mathcal{S}^1$ in the following order: (Aircraft, DTD, EuroSAT, Flowers, Food, Pets, Cars, UCF101), and the next 7 training sequences can be derived through Eq. (6). We leave the detailed order of tasks for each training sequence in the supplementary.

Metrics. For each sequence, we measure three metrics, *Average accuracy*, *Catastrophic forgetting*, and *Zero-shot degradation*. Following previous works [5, 7, 26], Average accuracy is calculated as the mean value of the final performance on

Fig. 3: Illustration of training and evaluation schemes for continual learning. From top to bottom rows, the pre-trained model g_0 is incrementally finetuned on different tasks (in green). For the incrementally learned model g in each row, data of unseen tasks are shown in red, while that of previously fine-tuned ones are in blue.

each task. Catastrophic forgetting measures the average of the maximum performance drop on each previous task. Also, Zero-shot degradation evaluates the average of the maximum performance drop on each unseen task. We illustrate the calculation of Catastrophic forgetting and Zero-shot degradation in Fig. 3.

4.4 Main Result

Baseline Methods. We compare our proposed framework with several baseline methods. Continual FT is the most straightforward strategy that continually fine-tunes the model to the current task. LwF [24] proposes to distill knowledge from the previous model with the current data. iCaRL [35] maintains a memory buffer that stores previous data and performs knowledge distillation to acquire knowledge from the previous models. ZSCL [50] is the most related method to our approach, which also introduces a reference dataset and distills knowledge from the pre-trained CLIP model. In addition, they further apply a weight-space ensemble for certain intervals to ensure a gradual transition of model parameters. MoE-Adapters [48] is the newest state-of-the-art that involves incremental adapters as mixture-of-experts [18, 38] upon a frozen CLIP model, and further apply an automatic selector to allocate data to the experts during test phase.

Multi-Domain Task-Incremental Learning. We evaluate the efficacy of our proposed method on the Multi-Domain Task-Incremental Learning (MTIL) benchmark [50]. This benchmark introduces a sequence of tasks with varying data distributions and distinct label spaces. Following the standard setting of

Table 1: Quantitative comparisons on Multi-Domain Task-Incremental Learning (MTIL) benchmark. In MTIL, inference is performed in a sequential manner on each dataset. $\mathcal{S}^i$ denotes the i-th training sequence.

Method / Sequence	$\mathcal{S}^1$	$\mathcal{S}^2$	$\mathcal{S}^3$	$\mathcal{S}^4$	$\mathcal{S}^5$	$\mathcal{S}^6$	$\mathcal{S}^7$	$\mathcal{S}^8$	**Mean**
Catastrophic forgetting ($\downarrow$)									
Continual FT	10.98	10.60	8.80	19.17	10.11	11.95	15.19	9.48	12.04
LwF [24]	10.38	6.52	6.37	10.22	7.99	7.70	10.41	8.91	8.56
iCaRL [35]	8.42	7.00	6.45	10.21	7.03	7.33	9.68	8.23	8.04
ZSCL [50]	4.67	2.35	2.13	2.97	3.15	4.28	4.89	4.70	3.64
MoE-Adapters [48]	2.74	4.71	4.28	1.15	1.50	1.60	2.94	2.77	2.71
Ours	**1.70**	**1.16**	**0.89**	**1.04**	**0.59**	**1.34**	**1.12**	**1.79**	**1.20**
Zero-shot degradation ($\downarrow$)									
Continual FT	24.81	23.58	19.54	16.46	22.22	19.02	19.54	24.02	21.15
LwF [24]	10.75	10.23	8.63	8.25	12.02	10.33	8.98	11.01	10.03
iCaRL [35]	13.77	12.68	11.28	12.14	13.20	13.20	13.09	14.01	12.92
ZSCL [50]	3.44	3.94	4.02	2.85	3.79	2.31	1.86	1.84	3.00
MoE-Adapters [48]	1.62	2.58	**1.04**	2.37	4.31	3.05	**1.77**	**0.63**	2.17
Ours	**1.55**	**2.04**	1.21	**1.92**	**2.79**	**2.18**	1.90	2.08	**1.96**
Average accuracy ($\uparrow$)									
Continual FT	76.16	76.24	78.03	68.69	76.64	75.44	72.71	77.45	75.17
LwF [24]	76.78	80.45	80.65	77.52	79.64	79.45	77.31	78.70	78.81
iCaRL [35]	77.99	79.77	79.93	76.66	79.26	79.08	77.06	78.61	78.55
ZSCL [50]	81.89	83.98	84.30	83.49	83.41	82.38	81.92	81.97	82.92
MoE-Adapters [48]	82.71	80.74	81.15	83.97	83.68	83.68	82.73	79.68	82.29
Ours	**84.48**	**84.92**	**84.97**	**84.89**	**85.50**	**85.07**	**85.02**	**84.52**	**84.92**

task-incremental learning [23], we evaluate the model on each dataset in a sequential manner during inference.

As shown in Tab. 1, we present the quantitative comparisons with different methods on the MTIL benchmark [50]. The results demonstrate that our method outperforms SOTA CL approaches, showing the effectiveness of our proposed method. By leveraging dual teachers with the proposed teacher selective mechanism, our framework is able to alleviate catastrophic forgetting on all training sequences with less than 2% of performance degradation and properly preserves the zero-shot classification capability.

Multi-Domain Class-Incremental Learning. As mentioned in [42, 43], the above *task-incremental* learning setting requires the information of task identity (*e.g.*, label space) of each test image, so that might not reflect practical scenarios. In this work, we further consider a more challenging scenario, Multi-Domain Class-Incremental Learning (MCIL), where the task (data domain) to be evaluated is not known during inference. To realize this, we conduct a unified label space by merging label spaces from all datasets at the inference stage.

As we can observe in Tab. 2, while there is a slight performance drop for all methods, our method consistently surpasses the other SOTA approaches, with about $1 \sim 3\%$ improvement for almost all metrics across different sequences. From the above results, we successfully confirm the effectiveness and robustness of our proposed method in the more challenging class-incremental setting.

Table 2: Quantitative comparisons on Multi-Domain Class-Incremental Learning (MCIL) benchmark. In MCIL, the task (data domain) to be evaluated is not known during inference and thus can be viewed as *open-set* classification.
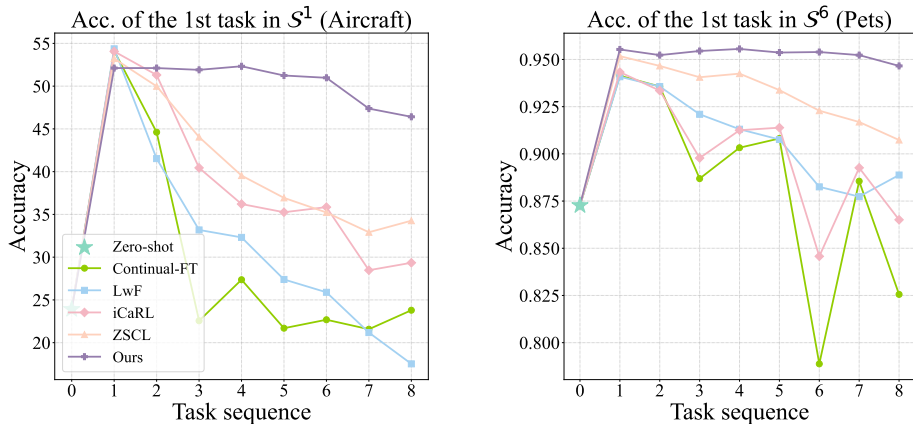
Method / Sequence	$\mathcal{S}^1$	$\mathcal{S}^2$	$\mathcal{S}^3$	$\mathcal{S}^4$	$\mathcal{S}^5$	$\mathcal{S}^6$	$\mathcal{S}^7$	$\mathcal{S}^8$	**Mean**
Catastrophic forgetting ($\downarrow$)									
Continual FT	11.17	10.89	10.16	20.12	10.57	12.14	15.62	9.80	12.56
LwF [24]	9.56	6.38	6.93	11.09	8.37	7.69	10.24	8.44	8.59
iCaRL [35]	8.43	6.90	6.83	10.69	7.09	7.37	10.17	8.66	8.27
ZSCL [50]	4.21	**1.41**	2.08	3.32	2.85	4.39	5.22	5.13	3.58
Ours	**1.92**	1.53	**0.97**	**1.14**	**0.58**	**1.55**	**1.29**	**1.81**	**1.35**
Zero-shot degradation ($\downarrow$)									
Continual FT	24.54	24.10	19.53	17.60	21.96	18.92	20.26	24.31	21.40
LwF [24]	11.94	11.82	8.27	9.99	13.36	11.47	10.95	12.63	11.30
iCaRL [35]	13.02	12.78	10.89	11.81	12.74	12.87	11.92	13.34	12.42
ZSCL [50]	3.59	4.71	4.17	2.81	3.55	1.97	**1.47**	2.30	3.07
Ours	**1.44**	**1.80**	**1.01**	**1.53**	**2.17**	**1.80**	1.65	**1.82**	**1.65**
Average accuracy ($\uparrow$)									
Continual FT	75.17	75.13	76.01	67.17	75.54	74.47	71.66	76.40	73.94
LwF [24]	74.14	77.44	77.94	74.89	77.30	77.43	75.50	76.51	76.39
iCaRL [35]	76.97	78.82	78.57	75.43	78.08	78.10	75.70	77.52	77.40
ZSCL [50]	80.49	82.54	82.99	82.08	82.17	80.99	80.30	80.09	81.46
Ours	**83.35**	**83.57**	**83.88**	**83.70**	**84.46**	**83.82**	**83.89**	**83.43**	**83.76**

4.5   Analysis

Assessment of Catastrophic Forgetting on the first Dataset. The performance on the first dataset in a training sequence inevitably suffers from the most severe catastrophic forgetting during continual learning. In Fig. 4, we plot the evaluation results on the first task through all training rounds to visualize the degree of catastrophic forgetting. The results clearly demonstrate that our method effectively maintains stable performance on the initial task even after multiple training rounds, verifying the effectiveness of our method in alleviating catastrophic forgetting on previous tasks. More experimental results are provided in the supplementary.

Assessment of Zero-Shot Degradation on the last Dataset. In addition to maintaining the knowledge learned from previous tasks, it is also crucial to preserve the zero-shot transferability for continual learning on VLMs. Fig. 5 indicates the level of zero-shot degradation for the last dataset, which experiences the most significant zero-shot degradation. The results show that we are able to keep almost the same zero-shot performance compared with the original pre-trained CLIP, demonstrating the effectiveness of our method in preserving the pre-trained zero-shot classification capability during sequential fine-tuning. More examples on different training sequences are shown in the supplementary.

Empirical Average Dual-Teacher Discrepancy. In Sec. 3.2, we argue that the dual-teacher discrepancy is highly related to determining whether an image is visually similar to previously fine-tuned data. Here, we verify the idea with

Fig. 4: Assessment of catastrophic forgetting with Aircraft (left) and Pets (right) as the first task in the continual learning sequence (i.e., the horizontal axis). It can be seen that our method is able to maintain their accuracies at the end of learning sequence.
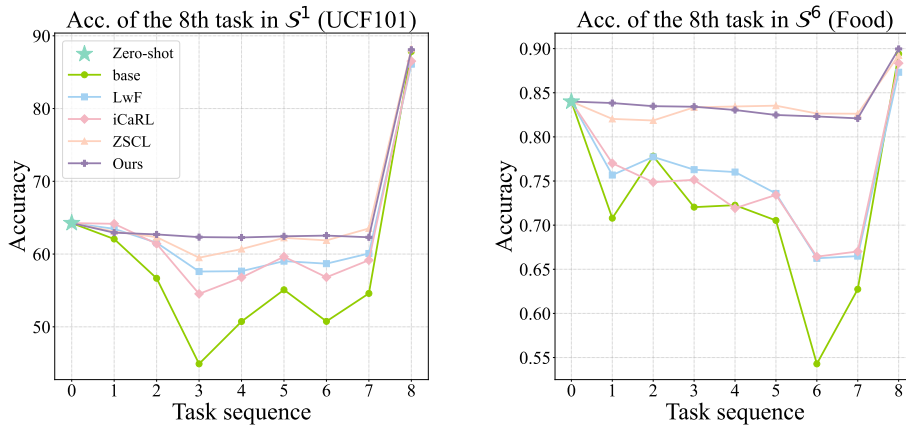
Table 3: Empirical average dual-teacher discrepancy d between the model g_1 trained on Aircraft ($\mathcal{T}^1$) and the orig1inal pre-trained g_0. Take Food for example, its discrepancy is calculated by $d(g_0(\mathbf{x}), g_1(\mathbf{x}))$ where $\mathbf{x}$ denotes images from Food. Since g_1 is finetuned on Aircraft only, a large discrepency score d of 1.059 for Aircraft is expected.

Dataset	Aircraft	DTD	EuroSAT	Flowers	Food	Pets	Cars	UCF101
Distance	1.059	0.090	0.126	0.073	0.091	0.067	0.170	0.112

empirical analysis. At stage $k = 2$, given a model g_1 fine-tuned on the Aircraft dataset ($\mathcal{T}^1$) and a pre-trained model g_0, we calculate the average dual-teacher discrepancy between g_1 and g_0 across various datasets, as shown in Tab. 3. The results show that the discrepancy d on the previously fine-tuned data (Aircraft, $\mathcal{T}^1$) is significantly greater than on other datasets not fine-tuned by g_1, while data not aligned with $\mathcal{X}^1$ shows a lower discrepancy, empirically supporting our intuition in Eq. (1).

Ablation study to the choice of different teachers. To verify the performance improvement of our proposed dual-teacher distillation mechanism, we conduct an ablation study to different choices of teacher models and present the results in Tab. 4. It can be seen that while distilling from the pre-trained model (g_0) results in satisfactory zero-shot performance with a drop of 2.51, it fails to prevent catastrophic forgetting. Conversely, distilling from the most recent model g_{k-1} preserve the continual learning performance but compromise zero-shot capability. Our method, an adaptive distillation scheme, is shown to alleviate catastrophic forgetting while preserving zero-shot performance.
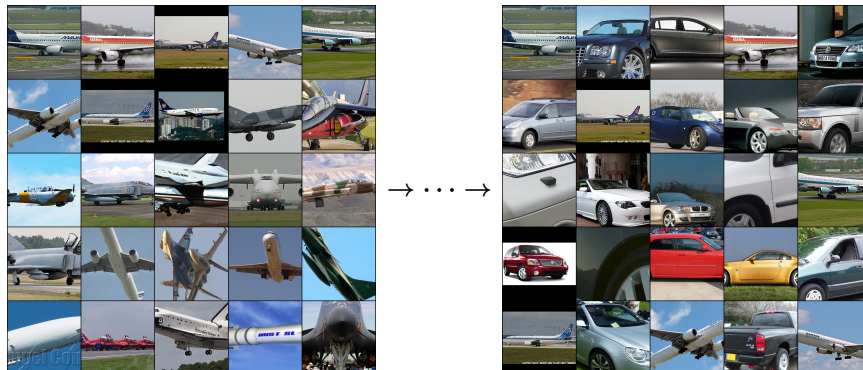
Fig. 5: Assessment of zero-shot degradation with UCF101 (left) and Food (right) as the last task in the continual learning sequence (i.e., the horizontal axis). It can be seen that our method shows satisfactory accuracies before finetuning on the last task.

Table 4: An ablation study on knowledge distillation from different teacher selections in $\mathcal{S}^1$. Distilling from either g_0 or g_{k-1} leads to unsatisfactory continual learning and zero-shot performance, respectively. Our method, distilling from both teachers, preserves the zero-shot capability and mitigates catastrophic forgetting in previous tasks.

Method	Forgetting ($\downarrow$)	Degradation ($\downarrow$)	Avg. Accuracy ($\uparrow$)
Distill from g_0	5.26	2.51	81.35
Distill from g_{k-1}	2.63	3.36	83.61
Ours	**1.70**	**1.55**	**84.48**

Visualization of Reference Images with Large Selection Scores η. Our proposed selection function η (Eq. (2)) aims to select the appropriate teacher by estimating the visual similarity between a reference image and previously fine-tuned data of g_{k-1}. We verify the effectiveness of the selection function by selecting Top-K reference images with large η scores. Fig. 6 visualize Top-25 reference images selected after training on certain datasets. We highlight two observations from the visualized results in the following,

- As shown in the left half of Fig. 6, after fine-tuning the model g_1 on the 1st task (*i.e.*, Aircraft), the Top-25 reference images with large η scores are highly similar to previously fine-tuned task without actually accessing any information related to previous data.
- After the model has been sequentially trained on multiple rounds (with the last task as Cars), our scheme is still able to select reference images closer to prior tasks (*e.g.*, Aircraft), demonstrating the ability to preserve the earliest

Fig. 6: Example images selected from the reference dataset with large η scores. **Left**: Top-25 reference images selected *after* fine-tuning on the 1st task of Aircraft. This suggests how we utilize such data to prevent possible catastrophic forgetting of Aircraft. **Right**: Top-25 reference images selected after continual learning across all tasks (with the last task as Cars). It can be seen that our scheme still selects reference images closer to prior tasks (e.g., Aircraft), explaining how zero-shot transfer ability is preserved.

fine-tuned knowledge even after multiple downstream tasks, as visualized in the right part of Fig. 6.

5   Conclusion

In this paper, we propose a Selective Dual-Teacher Knowledge Transfer framework for continual learning that tackles catastrophic forgetting and preserves zero-shot generalization ability simultaneously. By leveraging the most recent fine-tuned and original pre-trained VLMs as dual teachers, our framework selectively distills knowledge based on a dual-teacher discrepancy observed from an auxiliary reference dataset, without requiring label supervision. Comprehensive experiments, including comparisons with state-of-the-art continual learning methods and extensive analysis, quantitatively and qualitatively verify the effectiveness of our framework over existing approaches.

Limitation. Following [50], our work leverages an unlabeled reference dataset with the proposed teacher selection mechanism to identify the proper teacher network for knowledge distillation. If the reference dataset is very different from the previously fine-tuned tasks (e.g., medical images), the most recent fine-tuned VLM (g_{k-1}) would rarely be selected during training, so that the catastrophic forgetting might not be addressed well.

To further assess the impact of different reference datasets, we leave additional experimental results in the supplementary material due to page limitations. For example, subsets of larger-scale datasets such as ConceptualCaptioning 12M [3] and LAION 5B [37]) can be further considered and exploited.

References

1. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C.L., Parikh, D.: Vqa: Visual question answering. In: Proceedings of the IEEE international conference on computer vision. pp. 2425–2433 (2015)
2. Bossard, L., Guillaumin, M., Van Gool, L.: Food-101–mining discriminative components with random forests. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VI 13. pp. 446–461. Springer (2014)
3. Changpinyo, S., Sharma, P., Ding, N., Soricut, R.: Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts supplementary material. training **36**, 13
4. Changpinyo, S., Sharma, P., Ding, N., Soricut, R.: Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3558–3568 (2021)
5. Chaudhry, A., Dokania, P.K., Ajanthan, T., Torr, P.H.: Riemannian walk for incremental learning: Understanding forgetting and intransigence. In: Proceedings of the European conference on computer vision (ECCV). pp. 532–547 (2018)
6. Chaudhry, A., Ranzato, M., Rohrbach, M., Elhoseiny, M.: Efficient lifelong learning with a-gem. arXiv preprint arXiv:1812.00420 (2018)
7. Chaudhry, A., Rohrbach, M., Elhoseiny, M., Ajanthan, T., Dokania, P.K., Torr, P.H., Ranzato, M.: On tiny episodic memories in continual learning. arXiv preprint arXiv:1902.10486 (2019)
8. Choi, Y., El-Khamy, M., Lee, J.: Dual-teacher class-incremental learning with data-free generative replay. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3543–3552 (2021)
9. Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S., Vedaldi, A.: Describing textures in the wild. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3606–3613 (2014)
10. Cui, Y., Jia, M., Lin, T.Y., Song, Y., Belongie, S.: Class-balanced loss based on effective number of samples. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9268–9277 (2019)
11. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
12. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)

13. Fei-Fei, L., Fergus, R., Perona, P.: 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. pp. 178–178. IEEE (2004)

14. Gao, Q., Zhao, C., Ghanem, B., Zhang, J.: R-dfcil: Relation-guided representation learning for data-free class incremental learning. In: European Conference on Computer Vision. pp. 423–439. Springer (2022)

15. Helber, P., Bischke, B., Dengel, A., Borth, D.: 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)

16. Hou, S., Pan, X., Loy, C.C., Wang, Z., Lin, D.: Learning a unified classifier incrementally via rebalancing. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 831–839 (2019)

17. Ilharco, G., Wortsman, M., Wightman, R., Gordon, C., Carlini, N., Taori, R., Dave, A., Shankar, V., Namkoong, H., Miller, J., Hajishirzi, H., Farhadi, A., Schmidt, L.: Openclip (Jul 2021), if you use this software, please cite it as below.

18. Jacobs, R.A., Jordan, M.I., Nowlan, S.J., Hinton, G.E.: Adaptive mixtures of local experts. Neural computation **3**(1), 79–87 (1991)

19. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: International conference on machine learning. pp. 4904–4916. PMLR (2021)

20. Krause, J., Stark, M., Deng, J., Fei-Fei, L.: 3d object representations for fine-grained categorization. In: Proceedings of the IEEE international conference on computer vision workshops. pp. 554–561 (2013)

21. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images (2009)

22. Kumar, A., Raghunathan, A., Jones, R., Ma, T., Liang, P.: Fine-tuning can distort pretrained features and underperform out-of-distribution. arXiv preprint arXiv:2202.10054 (2022)

23. Li, X., Zhou, Y., Wu, T., Socher, R., Xiong, C.: Learn to grow: A continual structure learning framework for overcoming catastrophic forgetting. In: International Conference on Machine Learning. pp. 3925–3934. PMLR (2019)

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

25. Liu, H., Gu, L., Chi, Z., Wang, Y., Yu, Y., Chen, J., Tang, J.: Few-shot class-incremental learning via entropy-regularized data-free replay. In: European Conference on Computer Vision. pp. 146–162. Springer (2022)

26. Lopez-Paz, D., Ranzato, M.: Gradient episodic memory for continual learning. Advances in neural information processing systems **30** (2017)

27. Maji, S., Rahtu, E., Kannala, J., Blaschko, M., Vedaldi, A.: Fine-grained visual classification of aircraft. arXiv preprint arXiv:1306.5151 (2013)

28. McCloskey, M., Cohen, N.J.: Catastrophic interference in connectionist networks: The sequential learning problem. In: Psychology of learning and motivation, vol. 24, pp. 109–165. Elsevier (1989)

29. Nilsback, M.E., Zisserman, A.: Automated flower classification over a large number of classes. In: 2008 Sixth Indian conference on computer vision, graphics & image processing. pp. 722–729. IEEE (2008)

30. Parelli, M., Delitzas, A., Hars, N., Vlassis, G., Anagnostidis, S., Bachmann, G., Hofmann, T.: Clip-guided vision-language pre-training for question answering in 3d scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5606–5611 (2023)

31. Parkhi, O.M., Vedaldi, A., Zisserman, A., Jawahar, C.: Cats and dogs. In: 2012 IEEE conference on computer vision and pattern recognition. pp. 3498–3505. IEEE (2012)

32. Pham, H., Dai, Z., Ghiasi, G., Kawaguchi, K., Liu, H., Yu, A.W., Yu, J., Chen, Y.T., Luong, M.T., Wu, Y., et al.: Combined scaling for zero-shot transfer learning. Neurocomputing **555**, 126658 (2023)

33. PourKeshavarzi, M., Zhao, G., Sabokrou, M.: Looking back on learned experiences for class/task incremental learning. In: International Conference on Learning Representations (2021)

34. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)

35. Rebuffi, S.A., Kolesnikov, A., Sperl, G., Lampert, C.H.: icarl: Incremental classifier and representation learning. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 2001–2010 (2017)

36. Riemer, M., Cases, I., Ajemian, R., Liu, M., Rish, I., Tu, Y., Tesauro, G.: Learning to learn without forgetting by maximizing transfer and minimizing interference. arXiv preprint arXiv:1810.11910 (2018)

37. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al.: Laion-5b: An open large-scale dataset for training next generation image-text models. Advances in Neural Information Processing Systems **35**, 25278–25294 (2022)

38. Shazeer, N., Mirhoseini, A., Maziarz, K., Davis, A., Le, Q., Hinton, G., Dean, J.: Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. arXiv preprint arXiv:1701.06538 (2017)

39. Smith, J., Hsu, Y.C., Balloch, J., Shen, Y., Jin, H., Kira, Z.: Always be dreaming: A new approach for data-free class-incremental learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9374–9384 (2021)

40. Soomro, K., Zamir, A.R., Shah, M.: Ucf101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402 (2012)

41. Wang, R., Duan, X., Kang, G., Liu, J., Lin, S., Xu, S., Lü, J., Zhang, B.: Attriclip: A non-incremental learner for incremental knowledge learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3654–3663 (2023)

42. Wang, Z., Zhang, Z., Ebrahimi, S., Sun, R., Zhang, H., Lee, C.Y., Ren, X., Su, G., Perot, V., Dy, J., et al.: Dualprompt: Complementary prompting for rehearsal-free continual learning. In: European Conference on Computer Vision. pp. 631–648. Springer (2022)

43. Wang, Z., Zhang, Z., Lee, C.Y., Zhang, H., Sun, R., Ren, X., Su, G., Perot, V., Dy, J., Pfister, T.: Learning to prompt for continual learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 139–149 (2022)

44. Wortsman, M., Ilharco, G., Kim, J.W., Li, M., Kornblith, S., Roelofs, R., Lopes, R.G., Hajishirzi, H., Farhadi, A., Namkoong, H., et al.: Robust fine-tuning of zero-shot models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7959–7971 (2022)

45. Xiao, J., Hays, J., Ehinger, K.A., Oliva, A., Torralba, A.: Sun database: Large-scale scene recognition from abbey to zoo. In: 2010 IEEE computer society conference on computer vision and pattern recognition. pp. 3485–3492. IEEE (2010)
46. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention. In: International conference on machine learning. pp. 2048–2057. PMLR (2015)
47. Yin, H., Molchanov, P., Alvarez, J.M., Li, Z., Mallya, A., Hoiem, D., Jha, N.K., Kautz, J.: Dreaming to distill: Data-free knowledge transfer via deepinversion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8715–8724 (2020)
48. Yu, J., Zhuge, Y., Zhang, L., Hu, P., Wang, D., Lu, H., He, Y.: 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. pp. 23219–23230 (2024)
49. Zellers, R., Bisk, Y., Farhadi, A., Choi, Y.: From recognition to cognition: Visual commonsense reasoning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6720–6731 (2019)
50. Zheng, Z., Ma, M., Wang, K., Qin, Z., Yue, X., You, Y.: Preventing zero-shot transfer degradation in continual learning of vision-language models. arXiv preprint arXiv:2303.06628 (2023)
51. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. International Journal of Computer Vision **130**(9), 2337–2348 (2022)

Appendix

A   Evaluation Details

Datasets Statistics. We provide the detailed statistics of 8 fine-grained datasets and the reference dataset (*i.e.*, ImageNet [11]) in Tab. 5. The splits for training, validation and test of each dataset basically follow the setting provided by Zhou *et al.* [51]. Following the setting proposed in ZSCL [50], we sample 100,000 unlabeled images from ImageNet as the reference dataset.

Details of Multiple Training Sequences. We introduce *Multip Training Sequences* evaluation protocol to thoroughly evaluate every method over different training sequences in Sec. 4.3. Here we provide the detailed order of tasks for each sequence in Tab. 6.

B   More Implementation Details

Re-Weighted Dual-Teacher Knowledge Distillation Loss. Our proposed Dual-Teacher Knowledge Distillation loss shows the way to select the appropriate teacher model for a reference image according to the dual-teacher discrepancy and selection score η. In practice, there are few reference images with higher dual-teacher discrepancy. To address this potential imbalance problem, we apply a loss re-weighting strategy [10] as a post-processing technique. Specifically, the re-weighted dual-teacher knowledge distillation loss is shown below:

$$\tilde{\mathcal{L}}_{\mathrm{KD}}^{\mathrm{dual}} = \lambda \cdot \sum_{\mathbf{x} \sim \mathcal{X}^{\mathrm{ref}}} \eta(\mathbf{x}) \cdot \mathcal{L}_{\mathrm{KD}}^{k-1} + \sum_{\mathbf{x} \sim \mathcal{X}^{\mathrm{ref}}} (1 - \eta(\mathbf{x})) \cdot \mathcal{L}_{\mathrm{KD}}^{0}, \tag{7}$$

where λ is a hyper-parameter to control the imbalance ratio between the KD loss to the most recent fine-tuned model g_{k-1} and the KD loss to the pre-trained model g_0. Emprically we set $\lambda = 9$ to properly deal with the imbalance issue for every experiment in this work.

Hyper-Parameters to the η Selection Function. Our proposed η selection function:

$$\eta(\mathbf{x}) = \sigma\left(\frac{d(g_{k-1}(\mathbf{x}), g_0(\mathbf{x})) - \delta}{\gamma}\right), \tag{8}$$

involves two hyper-parameters: δ and γ. At a high-level, δ serves as a threshold that determining whether to select more from g_{k-1} or g_0. As the threshold δ increases, more reference data points are likely to be assigned values lower than 0.5, *i.e.*, select KD Loss more from g_0. On the other hand, γ works as a scaling factor to scale the value before applying the sigmoid function. As $\gamma \to 0$, the

Table 5: Detailed statistics for each dataset.

Dataset	Classes	Train	Val	Test
ImageNet [11]	1,000	1.28M	N/A	50,000
Aircraft [27]	100	3,334	3,333	3,333
DTD [9]	47	2,820	1,128	1,692
EuroSAT [15]	10	13,500	5,400	8,100
Flowers-102 [29]	102	4,093	1,633	2,463
Food-101 [2]	101	50,500	20,200	30,300
Oxford-Pets [31]	37	2,944	736	3,669
Stanford-Cars [20]	196	6,509	1,635	8,041
UCF-101 [40]	101	7,639	1,898	3,783

Table 6: The order of tasks for each training sequence.

Sequence	1st Task	2nd Task	3rd Task	4th Task	5th Task	6th Task	7th Task	8th Task
$\mathcal{S}^1$	Aircraft	DTD	EuroSAT	Flowers	Food	Pets	Cars	UCF101
$\mathcal{S}^2$	DTD	EuroSAT	Flowers	Food	Pets	Cars	UCF101	Aircraft
$\mathcal{S}^3$	EuroSAT	Flowers	Food	Pets	Cars	UCF101	Aircraft	DTD
$\mathcal{S}^4$	Flowers	Food	Pets	Cars	UCF101	Aircraft	DTD	EuroSAT
$\mathcal{S}^5$	Food	Pets	Cars	UCF101	Aircraft	DTD	EuroSAT	Flowers
$\mathcal{S}^6$	Pets	Cars	UCF101	Aircraft	DTD	EuroSAT	Flowers	Food
$\mathcal{S}^7$	Cars	UCF101	Aircraft	DTD	EuroSAT	Flowers	Food	Pets
$\mathcal{S}^8$	UCF101	Aircraft	DTD	EuroSAT	Flowers	Food	Pets	Cars

selection function move towards a *hard selection* mechanism, where the η scores tend to output either 1 or 0, depending on the discrepancy $d(g_{k-1}(\mathbf{x}), g_0(\mathbf{x}))$.

Tab. 7 provides a sensitivity analysis for hyper-parameters δ and γ. In general, the performance shows no significant difference when $\delta = 0.1$ or 0.2, hinting that it is stable enough for a proper range. By default, we select $\delta = 0.2$ and $\gamma = 1/6$ across all experiments in this work.

C   Different Choices of Reference Datasets

Our *Selective Dual-Teacher Knowledge Transfer* framework leverages an unlabeled reference dataset, following the settings in [50]. As mentioned in the limitation, the composition and the diversity of the images in the reference dataset might greatly affect the final performance. To examine the effect, we conduct ablation studies using different reference datasets (e.g., ConceptualCaptioning 12M [4]) and exploring the impact of varying the size of the reference dataset. Tab. 8 shows the performance of different reference datasets with varying size. While increasing the size of the reference dataset typically enhances performance, empirically there are no significant differences when the size exceeds 100k. By default, we use ImageNet with 100k images as our reference dataset, which also aligns with the same settings in [50].

Table 7: Sensitivity analysis on $\mathcal{S}^1$ to the hyper-parameters δ and γ in the η selection function. We highlight the results of our default setting across all experiments in the main paper in light red.

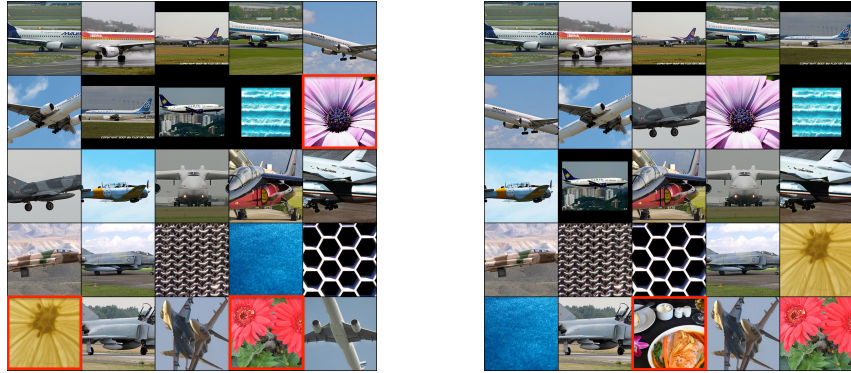| δ	γ	Forgetting ($\downarrow$)	Degradation ($\downarrow$)	Avg. Accuracy ($\uparrow$)
	1/3	1.72	1.58	84.42
0.1	1/6	1.68	1.57	84.43
	1/9	**1.65**	1.58	84.47
	1/3	1.67	1.60	84.46
0.2	1/6	1.70	**1.55**	**84.48**
	1/9	1.82	1.86	84.31
	1/3	1.81	1.52	84.23
0.3	1/6	2.13	1.99	84.03
	1/9	2.45	1.99	83.93

Table 8: The performance of different reference datasets with varying size. The default setting for all experiments is marked in light red.

Ref. Dataset	Size	Forgetting ($\downarrow$)	Degradation ($\downarrow$)	Avg. Accuracy ($\uparrow$)
	10k	1.92	2.12	84.18
ImageNet	100k	1.70	1.55	84.48
	200k	1.65	**1.11**	84.80
	10k	2.28	2.17	83.84
Conceptual Captions 12M	100k	**1.50**	1.88	84.48
	200k	1.60	1.25	**84.99**

D   Experiments Details

Detailed Explanation to the Visualization of Reference Images with Large η Scores. To illustrate the reference images with the highest η scores, we train our model on the first sequence $\mathcal{S}^1$ (the detailed task orders are shown in Tab. 6). For each stage $k \geq 2$, we calculate the η scores for each reference image using **only** the original pre-trained model g_0 and the most recent fine-tuned model g_{k-1} according to Eq. (2). Then, we select the Top-25 images with the highest η scores. Given that the visual concepts in some datasets are challenging to depict (*e.g.*, EuroSAT, UCF101), we focus our visualizations on datasets with more concrete concepts, such as Flowers and Food, as visualized in Fig. 7.

Detailed results for Catastrophic Forgetting and Zero-Shot Degradation. In Fig. 4 and Fig. 5, we present examples of the assessment of catastrophic forgetting for the first task and evaluation of zero-shot degradation for the last task, respectively. Here we plot the impact of catastrophic forgetting on the first task and the impact of zero-shot degradation on the last task across each sequence in Fig. 8 and Fig. 9. For catastrophic forgetting, our method clearly outperform other methods by stably preserving the performance on the previously fine-tuned task (1st task in this case). Regarding the issue of zero-shot

Fig. 7: Example images selected from the reference dataset with large η scores. **Left**: Top-25 reference images selected *after* fine-tuning on the Flowers Dataset. **Right**: Top-25 reference images selected *after* fine-tuning on the Food Dataset.

degradation, our method effectively maintains the original zero-shot capabilities in most scenarios, highlighting our success in preserving both pre-trained and previously fine-tuned knowledge across diverse datasets and various sequences.

Algorithm 1 Selective Dual-Teacher Knowledge Transfer

Input: A pre-trained VLM g_0, hyper-parameters $\delta, \gamma, \lambda_{\mathrm{dual}}$.
Data: A sequence of training tasks $\mathcal{S} = (\mathcal{T}^1, \cdots, \mathcal{T}^K)$ and a reference dataset $\mathcal{X}^{\mathrm{ref}}$.
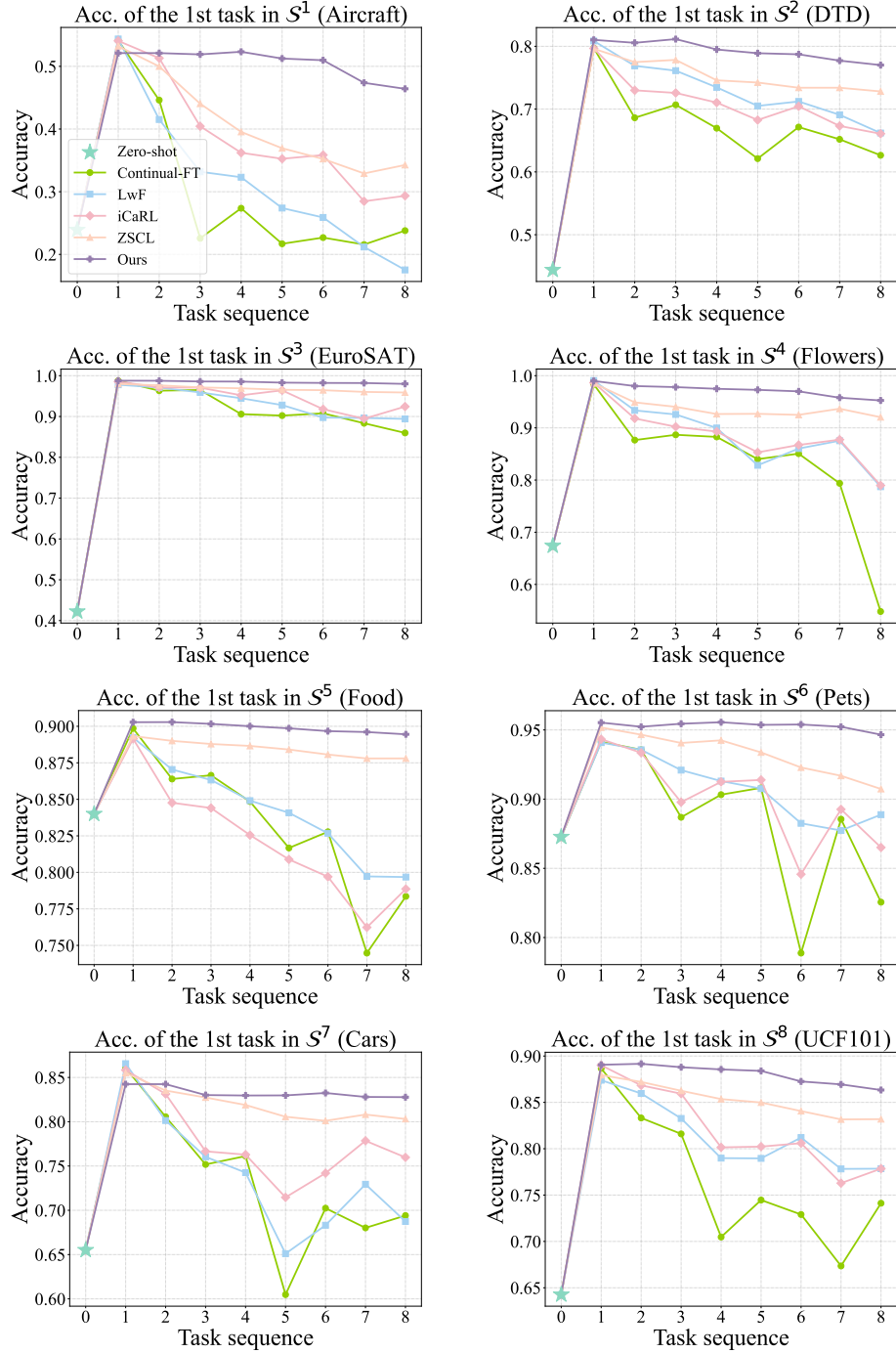**Output**: The final fine-tuned model g_K.

1: **for** k in $1 : K$ **do**
2:     Freeze g_0 as the pre-trained knowledge teacher.
3:     Freeze g_{k-1} as the previously fine-tuned knowledge teacher.
4:     Initialize the current model g_k by g_{k-1}.
5:     **for** e in E **do**
6:         **while** not traverse over all current data $\mathcal{T}^k$ **do**
7:             Sample a batch of current data B^k.
8:             Sample a batch of ref data B^{ref}.
9:             Calculate $\mathcal{L}_{\mathrm{CE}}$ with the current data B^k.
10:            Calculate Eq. (3) with g_0, g_{k-1}, and B^{ref}.
11:            Update g_k with loss function Eq. (4).
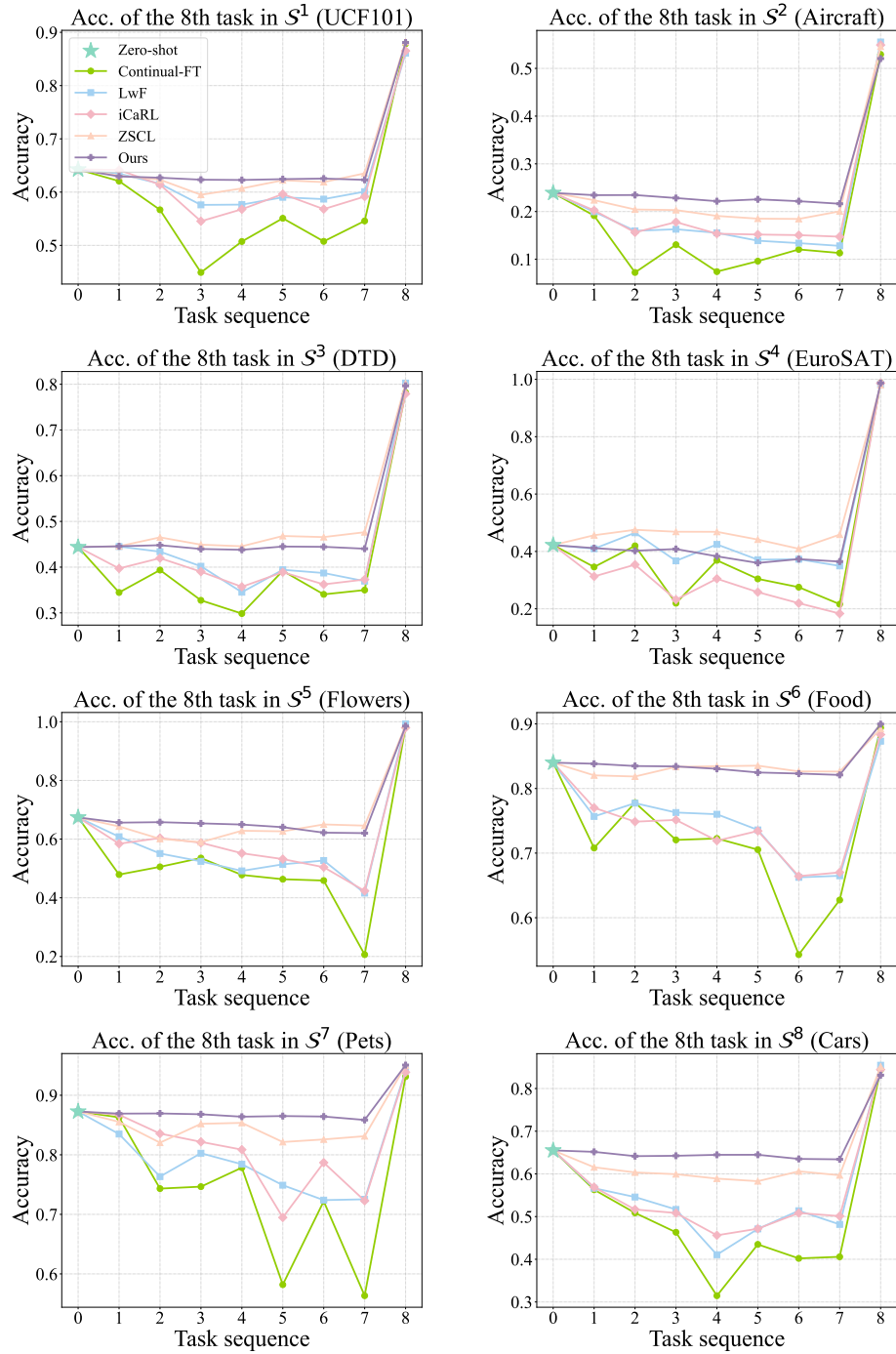12:   **end while**
13:    **end for**
14: **end for**

E   The Training Algorithm of Our Proposed Framework

As discussed in Sec. 3.3, we provide the detailed training algorithm of our *Selective Dual-Teacher Knowledge Transfer* framework in Algorithm 1.

Fig. 8: Assessment of catastrophic forgetting with the first task in the continual learning sequence (i.e., the horizontal axis). It can be seen that our method is able to maintain their accuracies at the end of learning sequence.

Fig. 9: Assessment of zero-shot degradation with the last task in the continual learning sequence (i.e., the horizontal axis). It can be seen that our method shows satisfactory accuracies before finetuning on the last task