Learning Expandable and Adaptable Representations for Continual Learning

Ruilong Yu[1], Mingyan Liu[2], Fei Ye[1]*, Adrian G. Bors[3], Rongyao Hu[1], Jingling Sun[1], and Shijie Zhou[1]
[1]University of Electronic Science and Technology of China, Chengdu, China
[2]Harbin Institute of Technology, Shenzhen, China
[3]University of York, York, U.K.
yrl666@outlook.com, 2023312903@stu.hit.edu.cn, feiye@uestc.edu.cn,
adrian.bors@york.ac.uk, ryhu@uestc.edu.cn,
jingling.sun910@gmail.com, sjzhou@uestc.edu.cn

Abstract

Extant studies predominantly address catastrophic forgetting within a simplified continual learning paradigm, typically confined to a singular data domain. Conversely, real-world applications frequently encompass multiple, evolving data domains, wherein models often struggle to retain many critical past information, thereby leading to performance degradation. This paper addresses this complex scenario by introducing a novel dynamic expansion approach called Learning Expandable and Adaptable Representations (LEAR). This framework orchestrates a collaborative backbone structure, comprising global and local backbones, designed to capture both general and task-specific representations. Leveraging this collaborative backbone, the proposed framework dynamically creates a lightweight expert to delineate decision boundaries for each novel task, thereby facilitating the prediction process. To enhance new task learning, we introduce a novel Mutual Information-Based Prediction Alignment approach, which incrementally optimizes the global backbone via a mutual information metric, ensuring consistency in the prediction patterns of historical experts throughout the optimization phase. To mitigate network forgetting, we propose a Kullback–Leibler (KL) Divergence-Based Feature Alignment approach, which employs a probabilistic distance measure to prevent significant shifts in critical local representations. Furthermore, we introduce a novel Hilbert-Schmidt Independence Criterion (HSIC)-Based Collaborative Optimization approach, which encourages the local and global backbones to capture distinct semantic information in a collaborative manner, thereby mitigating information redundancy and enhancing model performance. Moreover, to accelerate new task learning, we propose a novel Expert Selection Mechanism that automatically identifies the most relevant expert based on data characteristics. This selected expert is then utilized to initialize a new expert, thereby fostering positive knowledge transfer. This approach also enables expert selection during the testing phase without requiring any task information. Empirical results demonstrate that the proposed framework achieves state-of-the-art performance. Code is available at https://github.com/yrluestc/NeurIPS2025-LEAR.

*Corresponding author: feiye@uestc.edu.cn.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

1 Introduction

Modern deep learning frameworks have shown exceptional effectiveness across various visual tasks [15, 19]. However, the high performance of these approaches largely depends on large datasets, which are frequently unavailable in settings marked by constant change. This approach to learning is known as Continual/Lifelong Learning (CL), which aims to create a model that can continuously integrate new information while preserving all previously learned knowledge. Catastrophic forgetting is a significant issue that hinders the model's performance on earlier tasks [32], arising when the model tries to adjust its parameters to learn new tasks.

Among the various approaches to mitigate catastrophic forgetting in continual learning [41], Expansion-Based Methods (EBMs) have emerged as leading and highly effective strategies. The core idea is to dynamically expand the model's internal structure by adding task-specific modules to allocate dedicated capacity for each new task. However, current EBMs primarily focus on Class-Incremental Learning (CIL) [37] within a single domain, neglecting the scenario of learning across multiple domains, known as Domain-Incremental Learning (DIL). Although studies [44, 51, 21] have investigated DIL, their evaluated domains (e.g., Aircraft [29], MNIST [27]) have achieved near-perfect accuracy with pre-trained ViTs [10], making these benchmarks inadequate for assessing genuine continual learning capabilities. Therefore, we establish a more challenging and more realistic Multi-domain Continual learning (MDCL) scenario, where the task sequence comprises not only complex domains with large discrepancies but also a mixture of domains with underlying similarities. In this study, we aim to improve the model's performance in MDCL by considering three aspects including plasticity, stability and efficiency. To implement this goal, we propose a novel approach called LEAR and its core idea is to fully explore the stable and dynamic representations extracted by the pre-trained ViT backbones to achieve fast adaptation while adaptively optimizing the backbones to maintain all previously learned information.

(1) Plasticity. Existing EBMs improve downstream task performance by integrating task-specific prompts [46, 36] or adapters [30, 49] into a fixed pretrained backbone. However, these methods focus on exploring representations from a single pre-trained backbone, which fails to address more challenging data domains such as CUB-200 [39] and ImageNet-R [17]. Thus, to improve plasticity in a challenging MDCL scenario, we introduce a novel collaborative backbone architecture for LEAR, comprising a global and a local backbone, designed to capture general and task-specific information across all tasks. Leveraging this collaborative backbone structure, the proposed LEAR framework dynamically generates a lightweight expert to learn the decision boundary for each new task, thereby achieving commendable performance. The results presented in Tab. 1 and 2 demonstrate that our method achieves superior performance on most individual datasets in the MDCL scenario, which also validates that EBMs with frozen pretrained backbones cannot provide sufficient plasticity in MDCL.

(2) Stability. Many EBMs have been shown to achieve excellent stability in CIL. However, the excellent stability is usually achieved by freezing all parameters of the pre-trained models during the training, which may lead to forgetting of historical tasks in MDCL, especially when facing the severe domain shifts (e.g. ChestX [43] $\rightarrow$ ImageNet-R) in long task sequences. To address this limitation, we propose a unified optimization function to regulate the optimization behaviour of the collaborative backbone structure. This function consists of a Mutual Information-Based Prediction Alignment (MIBPA) loss and a Kullback–Leibler Divergence-Based Feature Alignment (KLDBFA) loss. The former dynamically optimizes the global backbone while preventing negative knowledge transfer at the prediction level, and the latter aligns historical and current representation distributions at the feature level. Such a design enables LEAR to achieve rehearsal-free continual learning by actively consolidating historical knowledge at both the prediction and feature levels when fine-tuning the collaborative backbones with new task data, rather than freezing parameters passively. Such a design has not been explored in the existing CL field. Furthermore, to mitigate optimization interference and information redundancy between the collaborative backbones, we propose a novel Hilbert-Schmidt Independence Criterion-Based Collaborative Optimization (HSICBCO) strategy to encourage two backbones to capture different semantic information, thus promoting effective complementary learning of MDCL tasks. The experimental results demonstrate that LEAR significantly outperforms all baseline methods in terms of overall average accuracy in three MDCL scenarios.

(3) Efficiency. Many existing EBMs usually ignore the task relevance and do not explore the previously learned parameter information to accelerate the new task learning. As a result, these methods optimize each new expert from scratch, which may result in considerable computational

costs and parameter redundancy when dealing with MDCL that contains analogous data domains. To address this issue, we aim to promote the efficient learning process of LEAR by proposing a novel Expert Selection Mechanism (ESM) that selectively transfers the parameter information learned by a selected expert into the new expert construction process. Specifically, the proposed ESM models each expert's knowledge as a Gaussian memory distribution and only preserve its critical statistical information. For each new task, the proposed ESM selects the most relevant expert by minimizing the Mahalanobis distance between stored distributions and incoming data, and reuses its parameters to facilitate new task learning. During the testing phase, ESM autonomously routes testing samples to the most suitable expert in a task-agnostic manner.

The principal contributions of this research are enumerated as follows : (1) This paper explores the challenging MDCL scenarios by proposing a novel approach called Learning Expandable and Adaptable Representations (LEAR) that optimizes and manages a collaborative backbone structure, comprising a global backbone and a local backbone, respectively. This design can help capture general and task-specific representations, which achieve excellent performance in MDCL. (2) A novel MIBPA approach is proposed to optimize the global backbone via a mutual information measure that ensures the consistency of the prediction pattern of each history expert when adjusting the parameters of the global backbone. (3) A novel KLDBFA approach is proposed to regulate the optimization behaviour of the local backbone by preventing significant changes in many critical local representations. Such a design can preserve task-specific representation information and prevent significant negative knowledge transfer effects. (4) A novel HSIBCO strategy is proposed to enforce the disentanglement between global and local representations, which avoids information redundancy and improves the model's performance. (5) A novel ESM is proposed to select the most relevant expert according to the data's characteristics, which is used in the training phase to promote the positive knowledge transfer process and in the testing phase to implement the expert selection procedure. The results from an extensive suite of experiments demonstrate that our proposed approach significantly outperforms existing baselines across all experimental configurations.

2 Related Work

Rehearsal-based techniques represent a widely adopted strategy for mitigating forgetting by dynamically incorporating a limited number of historical examples into the memory buffer [5, 6]. These memory samples are leveraged alongside new training instances to enhance model performance during the new task learning. Thus, the quality of the memorized samples is paramount within the rehearsal-based optimization framework [14]. Moreover, rehearsal-based approaches can be augmented through the integration of regularization techniques, with the objective of further elevating the overall efficacy of the model [2, 9, 20, 45]. In addition, memory studies have proposed to train the generative models to implement the memory system, which can provide infinite generative replay samples [1, 33, 35, 50, 23].

Prompt-based techniques leverage frozen pre-trained models like Vision Transformers (ViT) [10] as feature extractors, adapting them to sequential tasks through task-specific learnable prompt parameters. Current approaches employ diverse prompt management strategies including L2P [47]'s shared prompt pool with query-key retrieval mechanism, DualPrompt [46]'s separation of task-agnostic (G-Prompt) and task-specific (E-Prompt) components, and CODA-Prompt [36]'s attention-weighted cross-task prompt expansion. HiDe-Prompt [40] further advances performance by hierarchically decomposing class-incremental learning objectives for optimized task adaptation.

Expansion-based methods represent a robust approach to mitigating network forgetting in continual learning [8]. Such an approach dynamically expands the network architecture to enhance the learning ability of the new task [22, 38]. Beyond convolutional neural networks, expansion-based techniques have also been explored to leverage the capabilities of ViTs as the foundational network. These methods usually create self-attention blocks with the task-specific classifier to adapt to the new task learning [11, 48, 30, 49]. However, these methodologies typically involve freezing the pre-trained model, which limits their adaptability to complex and unknown data domains. We provide additional information on the related work in **Appendix-A** from Supplementary Materials (SM).

3 Methodology

3.1 Problem Definition

CL seeks to develop a model capable of acquiring knowledge across multiple sequences of tasks while retaining previously acquired information. This paper addresses a more pragmatic learning context in

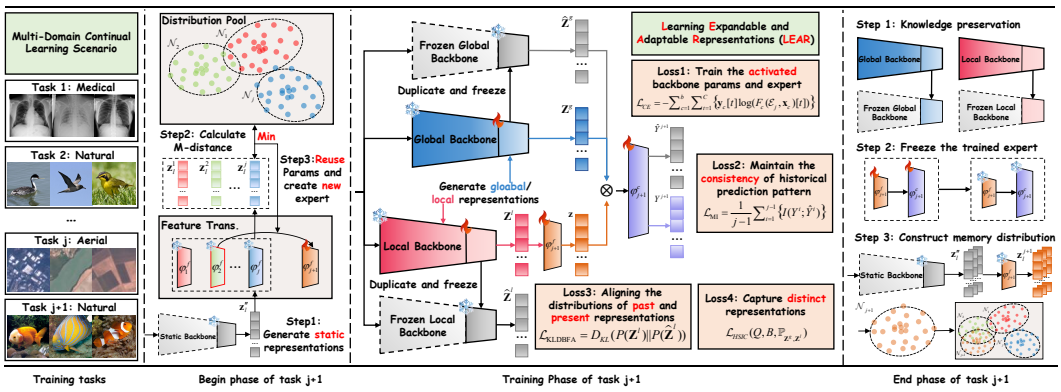


Figure 1: The training framework of the proposed LEAR. The data samples from new tasks are processed through a collaborative backbone structure to learn task-shared, task-specific and backbone-distinct representations via the proposed MIBPA, KLDBFA and HSICBCO, respectively. ESM constructs memory distributions and selects relevant experts for network expansion and test evaluation.

which each task encompasses previously unseen challenging data domains. Let $\mathcal{D}_i^S = \{\mathbf{x}_j, \mathbf{y}_j\}_{j=1}^{n_i^S}$ and $\mathcal{D}_i^T = \{\mathbf{x}_j, \mathbf{y}_j\}_{j=1}^{n_i^T}$ denote the i-th training and testing datasets, respectively. In a class-incremental learning scenario [3], a data split procedure is performed to divide the training dataset $\mathcal{D}_i^S$ into C_i subsets $\{\mathcal{D}_{i,1}^S, \cdots, \mathcal{D}_{i,C_i}^S\}$ according to the category, where each task $\mathcal{T}_j$ is associated with a training dataset $\mathcal{D}_{i,j}^S$ formed by samples from several adjacent classes. In the context of a specific task learning $\mathcal{T}_j$, the model is restricted to utilizing only the training dataset $\mathcal{D}_{i,j}^S$, with all preceding datasets $\{\mathcal{D}_{i,1}^S, \cdots, \mathcal{D}_{i,j-1}^S\}$ being inaccessible. In DIL, each task is conceptualized as a distinct data domain, denoted as $\{\mathcal{D}_1^S, \cdots, \mathcal{D}_n^S\}$, with n representing the total number of tasks. In contrast to these two CL scenarios which posit that each task comprises non-overlapping and heterogeneous data samples, a new task within our MDCL encompasses data samples that exhibit not only similar semantic characteristics with previously seen tasks but also significant domain shifts. Consequently, it becomes imperative to leverage existing parameters to facilitate learning these analogous tasks, thereby accelerating the training process and minimizing resource consumption. Once the final task learning is finished, the model's efficacy is assessed across all testing datasets $\{\mathcal{D}_1^T, \cdots, \mathcal{D}_n^T\}$ through the lens of classification accuracy.

3.2 Collaborative Backbone Structure

Recent investigations in CL have assessed the efficacy of leveraging a pre-trained ViT [10] to enhance model performance. These methodologies typically incorporate the pre-trained ViT as the primary backbone, facilitating the expert construction process while concurrently freezing its parameters to mitigate catastrophic forgetting. Nevertheless, this architectural design constrains the model's capacity for learning in the context of novel tasks, particularly when the incoming data exhibits divergent domain characteristics. This paper addresses this limitation by introducing a novel collaborative backbone architecture, comprising global and local backbones, each instantiated via a pre-trained ViT to facilitate rapid adaptation. Specifically, the global backbone incrementally updates its parameters throughout the optimization phase, with the objective of generating a task-shared representation applicable across tasks. Conversely, the local backbone is engineered to dynamically adapt to new tasks through parameter adjustments. We propose to optimize only the final three layers of the global and local backbone to mitigate computational expenses.

Let $F_{\theta^g}: \mathcal{X} \to \mathcal{Z}'$ denote a global backbone, implemented using a pre-trained ViT, which receives a data sample $\mathbf{x}$ over the data space and returns a feature vector $\mathbf{z}'$ over the feature space $\mathcal{Z}'$. Similarly, let $F_{\theta^l}: \mathcal{X} \to \mathcal{Z}'$ denote a local backbone, which has the same input-output pattern as the global backbone. For a given data sample $\mathbf{x}$, we can obtain its feature representations extracted by the global and local backbones, expressed as :

$$\mathbf{z}_g = F_{\theta^g}(\mathbf{x}), \mathbf{z}_l = F_{\theta^l}(\mathbf{x}), \tag{1}$$

By using Eq. (1), the proposed framework dynamically creates a lightweight expert ($\mathcal{E}_j$) consisting of a simple feature transformation module $F_{\varphi_j^f}: \mathcal{Z}' \to \mathcal{Z}$ and a linear classifier $F_{\varphi_j^c}: \mathcal{Z} \to \mathcal{Y}$, aiming to learn a decision boundary for a specific task. $F_{\varphi_j^f}$ receives the local representation $\mathbf{z}_l$ and returns a feature vector $\mathbf{z}$ over the feature space $\mathcal{Z}$, which is concatenate with global representation $\mathbf{z}_g$ and fed into the linear classifier $F_{\varphi_j^c}$ to make the prediction over the space $\mathcal{Y}$. The subscript j denotes the expert index, and $\oplus$ denotes the concatenation operation that combines two representations into a single feature vector. The prediction process of the j-th expert is expressed as :

$$F_c(\mathcal{E}_j, \mathbf{x}) = F_{\varphi_j^c}(F_{\theta^g}(\mathbf{x}) \oplus F_{\varphi_j^f}(F_{\theta^l}(\mathbf{x}))) \,. \tag{2}$$

By integrating representations derived by global and local backbones, the expert $\mathcal{E}_j$ in Eq. (2) can improve its generalization performance for a given data sample $\mathbf{x}$.

3.3 Mutual Information-Based Prediction Alignment

The global backbone's objective is to furnish a unified representation across all observed tasks. Consequently, optimization of the global backbone is susceptible to catastrophic forgetting, impacting all historical experts. To mitigate this, we introduce a novel Mutual Information-Based Prediction Alignment (MIBPA) methodology, designed to maintain the consistency of predictions of all historical experts when changing the parameters of the global backbone during the acquisition of new tasks. Specifically, we construct a parameter-shared auxiliary model $F_{\hat{\theta}^g}$ by replicating and freezing the global backbone's final three layers, then connecting them in parallel with intermediate features from the backbone's preceding layers. This auxiliary model subsequently guides the global backbone's optimization, producing two distinct prediction sets through the i-th expert, formulated as: :

$$\begin{aligned} \mathbf{Y}^i &= \left\{ \mathbf{y}_c \,|\, \mathbf{y}_c = F_{\varphi_i^c}(F_{\theta^g}(\mathbf{x}_c) \oplus F_{\varphi_i^f}(F_{\theta^l}(\mathbf{x}_c))), c = 1, \cdots, b \right\}, \\ \hat{\mathbf{Y}}^i &= \left\{ \mathbf{y}_c \,|\, \mathbf{y}_c = F_{\varphi_i^c}(F_{\hat{\theta}^g}(\mathbf{x}_c) \oplus F_{\varphi_i^f}(F_{\theta^l}(\mathbf{x}_c))), c = 1, \cdots, b \right\}, \end{aligned} \tag{3}$$

where b denotes the size of the data batch $\mathbf{X} = \{\mathbf{x}_1, \cdots, \mathbf{x}_b\}$ and $\mathbf{x}_c$ denotes the c-th data sample of $\mathbf{X}$. Let $P(Y^i, \hat{Y}^i)$ denote a joint distribution, where $P(Y^i)$ and $P(\hat{Y}^i)$ represent the marginal distributions of $\mathbf{Y}^i$ and $\hat{\mathbf{Y}}^i$, respectively. Let Y^i and $\hat{Y}^i$ denote two random variables over the joint distribution $P(Y^i, \hat{Y}^i)$. The proposed MIBPA approach minimizes the mutual information between Y^i and $\hat{Y}^i$, expressed as :

$$I(Y^i; \hat{Y}^i) = \sum\nolimits_{\hat{\mathbf{y}}^i \in \hat{Y}^i} \left\{ \sum\nolimits_{\mathbf{y}^i \in Y^i} \left\{ P(Y^i, \hat{Y}^i)(\mathbf{y}^i, \hat{\mathbf{y}}^i) \log \frac{P(Y^i, \hat{Y}^i)(\mathbf{y}^i, \hat{\mathbf{y}}^i)}{p(Y^i)(\mathbf{y}^i) p(\hat{Y})(\hat{\mathbf{y}}^i)} \right\} \right\}, \tag{4}$$

where $P(Y^i, \hat{Y}^i)(\mathbf{y}^i, \hat{\mathbf{y}}^i)$ signifies the probability density function of $P(Y^i, \hat{Y}^i)$. The mutual information term $I(Y^i; \hat{Y}^i)$, as defined in Eq. (4), evaluates the distance of the prediction made by the i-th expert built on the previously and currently learned global backbones. A small mutual information term $I(Y^i; \hat{Y}^i)$ indicates that updating the global backbone can still maintain the prediction pattern of the i-th expert. Finally, the final MIBPA regularization loss function at the j-th task learning is defined as :

$$\mathcal{L}_{\text{MI}} = \frac{1}{j-1} \sum\nolimits_{i=1}^{j-1} \left\{ I(Y^i; \hat{Y}^i) \right\}. \tag{5}$$

3.4 Kullback–Leibler (KL) Divergence-Based Feature Alignment

The iterative updating of the pre-trained backbones facilitates the temporal capture of local representations, thereby potentially enhancing the acquisition of novel tasks. However, this process risks inducing adverse knowledge transfer and performance degradation across historical experts. Regularization methods like EWC [24] and MAS [4], which typically impose constraints on parameter updates, are not desirable to capture the complex distributional shifts across domains, while knowledge distillation methods like LWF [28] and iCaRL [34] require maintaining additional teacher networks that become computationally prohibitive as the number of tasks grows. To address these limitations, we propose Kullback-Leibler Divergence-Based Feature Alignment (KLDBFA), designed to preserve crucial historical parameters during the optimization of the local backbone.

Our design rationale for selecting KL divergence stems from two key considerations: Firstly, modern generative evaluation metrics (e.g., FID [18], Kernel MMD [13]) operate on the Gaussian distribution assumption in high-dimensional feature spaces. This motivates us to model backbone features as Gaussian distributions. Fig. 1 in Appendix-C from the SM also provides additional empirical validation for the Gaussian distribution assumption. Secondly, KL divergence offers unique advantages over alternative distributional metrics: (1) Its directional property enables targeted constraint of current features toward historical distributions, unlike symmetric metrics (e.g., Jensen-Shannon divergence); (2) It maintains computational efficiency compared to expensive metrics like MMD or Wasserstein distance. These characteristics make KL divergence ideally suited for continual learning scenarios requiring efficient knowledge preservation.

Specifically, upon each task transition, the proposed KLDBFA approach duplicates and immobilizes the local backbone F_{θ^l} as a frozen model $F_{\hat{\theta}^l}$ following a similar procedure in MIBPA. $F_{\hat{\theta}^l}$ serves to regulate the optimization dynamics of the local backbone. For a given data batch $\mathbf{X} = \{\mathbf{x}_1, \cdots, \mathbf{x}_b\}$, two distinct sets of feature vectors are derived utilizing F_{θ^l} and $F_{\hat{\theta}^l}$, respectively, as follows :

$$\mathbf{Z}^l = \{\mathbf{z}_c \,|\, \mathbf{z}_c = F_{\theta^l}(\mathbf{x}_c), c = 1, \cdots, b\}, \hat{\mathbf{Z}}^l = \{\mathbf{z}_c \,|\, \mathbf{z}_c = F_{\hat{\theta}^l}(\mathbf{x}_c), c = 1, \cdots, b\}. \tag{6}$$

Building upon the Gaussian assumption stated above, we model two Gaussian distributions $P(\mathbf{Z}^l) = \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and $P(\hat{\mathbf{Z}}^l) = \mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$, through calculating the mean vectors $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$ and covariance matrix $\boldsymbol{\Sigma}_1$, $\boldsymbol{\Sigma}_2$ of $\mathbf{Z}^l$ and $\hat{\mathbf{Z}}^l$, respectively. We propose to employ the KL divergence to evaluate the discrepancy between $P(\mathbf{Z}^l)$ and $P(\hat{\mathbf{Z}}^l)$ as a regularization loss term, expressed as :

$$\begin{aligned} D_{KL}(P(\mathbf{Z}^l)||P(\hat{\mathbf{Z}}^l)) &= \frac{1}{2}\left[\log\left(\frac{\det(\Sigma_2)}{\det(\Sigma_1)}\right) - d + \mathrm{tr}(\Sigma_2^{-1}\Sigma_1) + (\mu_2 - \mu_1)^\top \Sigma_2^{-1}(\mu_2 - \mu_1)\right] \\ \mathcal{L}_{\text{KLDBFA}} &= D_{KL}(P(\mathbf{Z}^l) \,||\, P(\hat{\mathbf{Z}}^l)), \end{aligned} \tag{7}$$

where D_{KL} denotes the KL divergence. $\det(\cdot)$, d, and $\mathrm{tr}(\cdot)$ represent the determinant, dimension, and trace of a matrix, respectively.

3.5 HSIC-Based Collaborative Optimization

The global and local backbones are designed to capture distinct feature representations, thereby potentially improving model efficacy. To further facilitate the disentanglement between these backbones, we introduce a novel Hilbert-Schmidt Independence Criterion (HSIC)-Based Collaborative Optimization (HSICBCO) methodology. This approach leverages an independence criterion to maximize the divergence of knowledge between the global and local backbones. Specifically, we employ the HSIC measure [12], given its property of ranging from 0 to infinity, with 0 signifying statistical independence. Consequently, minimizing the HSIC term allows for enhanced disentanglement between the global and local backbones, which can be easily added to the primary loss function.

Let Z^g and Z^l denote two distinct domains, and let $\mathbb{P}_{\mathbf{Z}^g,\mathbf{Z}^l}$ represent a joint distribution from which a sample pair $\{\mathbf{z}_g, \mathbf{z}_l\}$ is drawn using global and local backbones across $Z^g \times Z^l$. The primary objective of HSIC, as delineated in [12], within the framework of Reproducing Kernel Hilbert Space (RKHS), [42], is to quantify the dependency between the domains of the variables $\mathbf{z}_g$ and $\mathbf{z}_l$ by assessing the norm of the cross-covariance operator over the domain $Z^g \times Z^l$. Let Q and B be the RKHSs defined on Z^g and Z^l, respectively, and let $f_Q \colon Z^g \to Q$, and $f_B \colon Z^l \to B$ denote their respective feature mappings. The associated reproducing kernels are defined as $k(\mathbf{z}_g, \mathbf{z}'_g) = \langle f_Q(\mathbf{z}_g), f_Q(\mathbf{z}'_g)\rangle$ and $l(\mathbf{z}_l, \mathbf{z}'_l) = \langle f_B(\mathbf{z}_l), f_B(\mathbf{z}'_l)\rangle$, where $\mathbf{z}_g, \mathbf{z}'_g \in Z^g$ and $\mathbf{z}_l, \mathbf{z}'_l \in Z^l$. The cross-covariance operator between f_Q and f_B is defined as follows :

$$C_{\mathbf{z}_g\mathbf{z}_l} = \mathbb{E}_{\mathbf{z}_g\mathbf{z}_l}\Big\{ \big(f_Q(\mathbf{z}_g) - \mathbb{E}_{\mathbf{z}_g}\left[f_Q(\mathbf{z}_g)\right]\big) \otimes \big(f_B(\mathbf{z}_l) - \mathbb{E}_{\mathbf{z}_l}\left[f_B(\mathbf{z}_l)\right]\big) \Big\}, \tag{8}$$

where $\otimes$ is the tensor product. HSIC is defined as the square of the Hilbert-Schmidt norm of $C_{\mathbf{z}_g,\mathbf{z}_l}$:

$$\begin{aligned} &\mathcal{L}_{HSIC}(Q, B, \mathbb{P}_{\mathbf{Z}^g,\mathbf{Z}^l}) = \left\|C_{\mathbf{z}_g,\mathbf{z}_l}\right\|_{HS}^2 = \mathbb{E}_{\mathbf{z}_g,\mathbf{z}'_g,\mathbf{z}_l,\mathbf{z}'_l}[k(\mathbf{z}_g,\mathbf{z}'_g)l(\mathbf{z}_l,\mathbf{z}'_l)] \\ &+ \mathbb{E}_{\mathbf{z}_g,\mathbf{z}'_g}[k(\mathbf{z}_g,\mathbf{z}'_g)]\mathbb{E}_{\mathbf{z}_l,\mathbf{z}'_l}[l(\mathbf{z}_l,\mathbf{z}'_l)] - 2\mathbb{E}_{\mathbf{z}_g,\mathbf{z}_l}[\mathbb{E}_{\mathbf{z}'_g}[k(\mathbf{z}_g,\mathbf{z}'_g)]\mathbb{E}_{\mathbf{z}'_l}[l(\mathbf{z}_l,\mathbf{z}'_l)]], \end{aligned} \tag{9}$$

where $\mathbb{E}_{\mathbf{z}_g,\mathbf{z}'_g,\mathbf{z}_l,\mathbf{z}'_l}$ represents the expectation over samples $\{\mathbf{z}_g, \mathbf{z}_l\}$ and $\{\mathbf{z}'_g, \mathbf{z}'_l\}$ drawn from $\mathbb{P}_{\mathbf{Z}^g,\mathbf{Z}^l}$.

3.6 Expert Selection Mechanism

In scenarios where analogous data domains are encountered in subsequent tasks, the reuse of pertinent parameters and information becomes imperative for the efficient learning of such domains, thereby accelerating the training process of a novel task. To this end, this study introduces a novel Expert Selection Mechanism (ESM), designed to identify the most relevant expert for a given new task, facilitating the reuse of existing parameters to initialize a new expert, which, in turn, can engender positive knowledge transfer effects.

The memory distribution. Specifically, we utilize a frozen pre-trained ViT as a feature extractor, denoted as $F_{\theta^f}: \mathcal{X} \rightarrow \mathcal{Z}''$, to generate static data representations, where θ^f represents the fixed parameters of the ViT backbone. To mitigate parameter redundancy, we duplicate and freeze the local backbone's final three layers as in KLDBFA in the first task. Upon the completion of a specific task learning phase ($\mathcal{T}_j$), a subset of training samples $\{\mathbf{x}_k\}_{k=1}^m$ is randomly selected from $\mathcal{D}_j^S$ and processed by F_{θ^f} to extract the class token representation $\mathbf{z}_k'' = F_{\theta^f}(\mathbf{x}_k)$. These extracted features are subsequently propagated through the fully connected layer of the current expert $\mathcal{E}_j$, yielding transformed features :

$$\mathbf{z}_k = F_{\varphi_j^f}(\mathbf{z}_k'')\,. \tag{10}$$

By using the transformed features, we obtain the empirical mean $\boldsymbol{\mu}_j$ and covariance matrix $\boldsymbol{\Sigma}_j$ by :

$$\boldsymbol{\mu}_j = \frac{1}{m}\sum\nolimits_{k=1}^{m}\{\mathbf{z}_k\}, \quad \boldsymbol{\Sigma}_j = \frac{1}{m-1}\sum\nolimits_{k=1}^{m}\{(\mathbf{z}_k-\boldsymbol{\mu}_j)(\mathbf{z}_k-\boldsymbol{\mu}_j)^\top\}\,. \tag{11}$$

Subsequently, a multivariate Gaussian distribution $\mathcal{N}_j = \mathcal{N}(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$ is constructed to preserve the statistical information about the j-th task. We call $\mathcal{N}_j$ as the memory distribution for the expert $\mathcal{E}_j$, which is always fixed during the subsequent learning.

The expert selection process. When a new task $\mathcal{T}_{j+1}$ begins, its training samples $\{\mathbf{x}_l\}_{l=1}^{m'}$ are first processed by the frozen backbone F_{θ^f} to obtain $\mathbf{z}_l'' = F_{\theta^f}(\mathbf{x}_l)$. For the l-th representation $\mathbf{z}_l''$, we can employ the feature transformation modules $\{F_{\varphi_1^f}, \cdots, F_{\varphi_j^f}\}$ of all existing experts $\{\mathcal{E}_1, \cdots, \mathcal{E}_j\}$ to generate a set of transformed features $\mathbf{z}_l^c = F_{\varphi_j^f}(\mathbf{z}_l''), \quad \forall c = 1, \cdots, j$. Based on the transformed features, the most relevant expert $\mathcal{E}_{c^*}$ with the minimum average Mahalanobis distance at the new task learning ($\mathcal{T}_{j+1}$) is selected by :

$$c^* = \underset{c=1,\cdots,j}{\operatorname{argmin}}\Big\{\frac{1}{m'}\sum\nolimits_{l=1}^{m'}\sqrt{(\mathbf{z}_l^c-\boldsymbol{\mu}_c)^\top\boldsymbol{\Sigma}_c^{-1}(\mathbf{z}_l^c-\boldsymbol{\mu}_c)}\Big\}\,, \tag{12}$$

where c^* denotes the index of the selected expert. Finally, a new expert $\mathcal{E}_{j+1}$ is initialized using the parameters of the selected expert $\mathcal{E}_{c^*}$, particularly inheriting its feature transformation module $F_{\varphi_{c^*}^f}$ and linear classifier $F_{\varphi_{c^*}^c}$. The new expert is then fine-tuned on the subset of incoming task samples to adapt to the new domain. During inference, test samples are assigned to the most suitable expert for prediction using the same selection strategy. The reason for choosing the Mahalanobis distance is that it accounts for the scale and correlation of variables, making it suitable for datasets where features are correlated or have different units. Other distance measures are analyzed in **Appendix-D** from SM. This Mahalanobis distance-based expert selection strategy enables the model to dynamically identify the most semantically related expert for knowledge transfer, thereby accelerating convergence and reducing parameter overhead during continual learning.

3.7 Algorithm Implementation

The learning procedure of the proposed LEAR (illustrated in Fig. 1) consists of three stages:

Step 1: Collaborative backbone initialization. We initialize both global and local backbones using pre-trained ViT models. These networks serve as the feature extractors for all tasks throughout the Multi-Domain Continual Learning process. For a given input $\mathbf{x}$, we obtain its feature representations extracted by both backbones using Eq. (1).

Step 2: Dynamic expert creation and selection. During the training of the j-th task ($\mathcal{T}_j$), we dynamically create a lightweight expert tailored to this domain. To ensure effective knowledge transfer, we select the most relevant historical expert based on the Mahalanobis distance computed

Table 1: The classification accuracy (%) of all testing datasets after learning the **CDM** task sequence.

Methods	C10	Disease	MNIST	RESISC45	EuroSAT	TImg	C100	ChestX	ImgR	CUB200	Avg
DER++(Re)	22.78	34.00	11.33	8.24	18.18	7.66	31.79	25.99	54.04	78.23	29.22
CLS-ER	22.58	27.22	12.50	14.32	24.17	14.14	34.23	25.99	52.46	75.78	30.34
RanPAC	87.17	96.76	87.45	84.20	92.53	71.46	51.44	39.63	43.30	56.18	71.01
MoE	92.74	32.44	91.87	54.02	41.85	6.12	78.98	19.60	**78.43**	77.24	57.33
L2P	20.66	11.15	14.08	4.50	13.65	11.67	31.76	21.52	58.26	81.30	26.85
DAP	8.83	2.78	18.94	3.36	18.19	3.06	11.29	16.34	60.68	80.37	22.38
D-Prompt	25.99	9.08	16.57	4.13	7.30	22.73	38.36	17.33	59.06	82.44	28.30
C-Prompt	13.57	2.33	9.10	1.90	14.18	0.68	3.40	14.35	4.14	60.55	12.42
Ours	**95.44**	**98.46**	**96.59**	**92.04**	**95.00**	**81.24**	**85.10**	**45.95**	70.47	**85.80**	**84.61**

Table 2: The classification accuracy (%) of all testing datasets after learning the **ETI** task sequence.

Methods	EuroSAT	TImg	ImgR	CUB200	C100	MNIST	RESISC45	ChestX	C10	Disease	Avg
DER++(Re)	51.82	36.25	2.32	6.20	23.08	65.97	40.45	27.41	82.69	97.77	43.40
CLS-ER	45.76	29.33	16.19	33.08	30.74	67.10	46.44	30.26	80.29	97.62	47.68
RanPAC	92.64	70.87	43.75	56.13	51.83	88.08	83.51	40.34	86.74	96.88	71.08
MoE	43.01	0.94	55.48	25.16	74.22	97.67	84.57	33.38	**96.88**	**99.90**	61.12
L2P	10.88	1.45	0.97	4.04	14.55	12.17	22.48	9.16	89.69	98.62	26.40
DAP	9.93	1.07	1.84	17.65	15.74	15.44	22.19	16.34	86.20	98.44	28.48
D-Prompt	10.61	1.31	1.44	3.09	17.03	23.83	25.03	14.77	93.42	99.33	28.99
C-Prompt	11.66	0.67	0.62	0.33	1.72	13.04	6.14	16.62	21.74	96.21	16.88
Ours	**95.89**	**81.25**	**69.57**	**84.12**	**85.30**	**98.56**	**92.92**	**45.45**	96.60	99.30	**84.90**

over previous feature distributions through Eq. (12). As a result, the selected expert is employed to initialize a new expert.

Step 3: Interactive optimization with alignment constraints. Once the new expert $\mathcal{E}_j$ is created, we perform joint optimization incorporating mutual information-based prediction alignment via Eq. (5) and KL-divergence feature alignment using Eq. (7). Furthermore, we calculate the HSIC regularization losses to encourage disentanglement between global and local backbones using Eq. (9). All regularization terms are incorporated into the main objective function for expert optimization :

$$\begin{aligned}\mathcal{L}_{\text{final}}(\mathbf{X}) = -\sum\nolimits_{c=1}^{b}\sum\nolimits_{t=1}^{C}\left\{\mathbf{y}_c[t]\log(F_c(\mathcal{E}_j,\mathbf{x}_c)[t])\right\} + \lambda_1\mathcal{L}_{\text{MI}} \\ + \lambda_2\mathcal{L}_{\text{KLDBFA}} + \lambda_3\mathcal{L}_{HSIC}(Q,B,\mathbb{P}_{\mathbf{Z}^g,\mathbf{Z}^l})\,,\end{aligned} \tag{13}$$

where $F_c(\mathcal{E}_j,\mathbf{x}_c)[t]$ denotes the predicted probability of class t for sample $\mathbf{x}_c$. λ_1, λ_2, λ_3 are trade-off hyperparameters balancing different loss components. The model parameters $\{\theta^g,\theta^l,\varphi_j^f,\varphi_j^c\}$ is updated using Eq. (13). The detailed algorithm is summarized in **Appendix-B** from SM.

4 Experiment

4.1 Experimental Setup

Metrics. In the context of the MDCL scenarios, we assess and compare model efficacy at the final task through two key metrics: the classification accuracy of a single domain (e.g., **C10** or **CUB200**) and the overall performance across all domains (**Avg**).

Datasets. The datasets used in our experiment can be logically categorized into three primary fields according to [21]. **Natural Domains** include CIFAR-10 [25] (C10), TinyImageNet [26] (TImg), CUB-200 [39], MNIST [27] and ImageNet-R [17] (ImgR), covering a range of tasks from basic image classification to fine-grained recognition and robustness testing across various visual styles. **Aerial Domains** comprise EuroSAT [16] and RESISC45 [7], focusing on satellite imagery for land cover classification and environmental monitoring. **Medical Domains** consist of CropDiseases [31] (Disease) and ChestX [43], specialized for identifying plant diseases and diagnosing medical conditions through radiographic images, respectively. Then, we randomly shuffle these datasets to construct three highly challenging MDCL scenarios (**CDM**, **ETI** and **TRC** which are derived from the initial letters of the first three domains). Detailed experimental configurations are provided in **Appendix-C** from SM.

Table 3: Comparison of Baselines in terms of parameter and computational efficiency (**in CDM**).

Methods	Train Params↓	Iter/s↑	GPU Avg↓	GPU Max↓	CPU Avg↓	CPU Max↓
DER++(Re)	42.84M	1.04	12919.37MB	12919.37MB	9756.08MB	16805.70MB
CLS-ER	42.84M	2.13	6199.39MB	6199.39MB	9658.21MB	16700.56MB
RanPAC	1.19M	4.27	3065.27MB	3087.97MB	9823.56MB	17465.18MB
MoE	4.03M	1.77	11098.88MB	11098.88MB	14203.61MB	17262.77MB
L2P	**0.20M**	**5.08**	3420.06MB	3420.06MB	9654.75MB	16798.61MB
DAP	0.51M	2.76	3686.95MB	3686.95MB	9911.54MB	16958.32MB
D-Prompt	0.41M	3.11	3556.64MB	3556.64MB	9866.40MB	17040.50MB
C-Prompt	3.99M	4.91	4775.65MB	4775.65MB	9869.61MB	16941.20MB
Ours	42.54M	3.06	**2926.20MB**	**2926.20MB**	**9525.63MB**	**16681.54MB**

Table 4: Impact of individual and combined components on model performance in ETI.

Methods	EuroSAT	TImg	ImgR	CUB200	C100	MNIST	RESISC45	ChestX	C10	Disease	Avg
CB	19.59	8.37	11.81	12.29	36.92	49.88	61.14	30.04	92.79	99.00	42.18
SBE	85.38	70.49	55.42	73.45	77.05	90.05	85.93	37.73	93.65	99.13	76.83
CBE	92.38	76.49	62.42	78.45	82.05	95.05	90.93	38.73	94.17	99.04	80.97
CBE+MI	95.08	80.93	68.19	82.49	84.68	97.15	91.29	39.56	96.18	98.28	83.38
CBE+KL	92.18	78.02	66.13	78.97	82.50	92.24	90.54	44.58	95.81	99.13	82.01
CBE+HSIC	93.65	76.97	64.37	78.56	82.35	91.88	91.29	45.53	95.49	99.06	81.92
CBE+MI+KL	95.35	**81.61**	68.30	83.56	**85.54**	97.43	92.28	**46.44**	96.22	99.01	84.57
CBE+MI+HSIC	95.87	80.85	68.74	83.15	85.52	98.03	92.08	42.43	96.33	98.91	84.19
CBE+KL+HSIC	93.24	78.22	66.45	80.49	82.55	92.21	91.39	45.26	95.59	99.08	82.45
LEAR	**95.89**	81.25	**69.57**	**84.12**	85.30	**98.56**	**92.92**	45.45	**96.60**	**99.30**	**84.90**
LEAR w/o ESM	3.72	1.25	2.36	83.95	8.24	4.91	14.63	1.05	17.85	99.15	23.71

4.2 Experimental Results

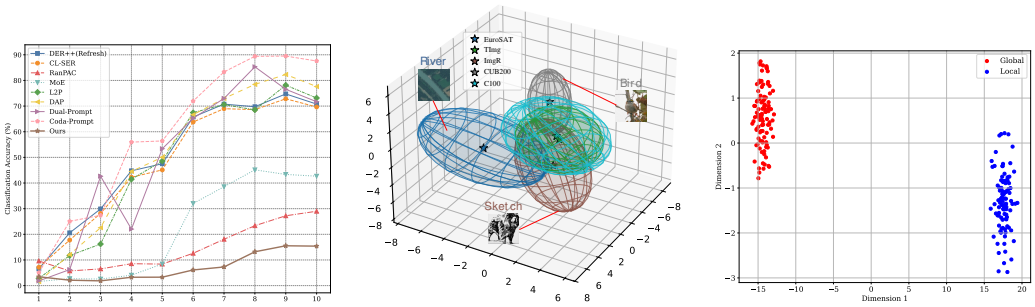
Results in Multi-Domain Continual Learning. Our comprehensive evaluation compares LEAR against state-of-the-art approaches across three distinct domain sequences (Tables 1 and 2). As shown in Table 1, LEAR achieves an outstanding average accuracy of 84.61% in the CDM scenario. Specifically, LEAR outperforms the replay-based DER++ (Refresh) by 55.39%, highlighting its superior ability to mitigate catastrophic forgetting without requiring memory buffers. When compared to expansion-based methods, LEAR shows substantial advantages over both RanPAC (71.01%) and MoE-adapters (57.33%), particularly in challenging domains like TinyImageNet and ChestX, while maintaining consistent performance across all evaluated domains.

The reshuffled domain sequence in Table 2 further validates LEAR's adaptability, where it achieves an even higher average accuracy of 84.90% in the ETI scenario, outperforms dual-branch method CLS-ER (47.68%) by 37.22%, with especially large gap on ImageNet-R. Moreover, LEAR demonstrates over 55% higher average accuracy than the domain-incremental method DAP (28.48%) and other listed prompt-based approaches. These results highlight LEAR's ability to effectively learn and retain knowledge regardless of the domain order.

Moreover, as shown in Figure 2 (a), the proposed LEAR achieves the lowest forgetting rate in CDM scenario compared to alternative methods. As the number of tasks increases, LEAR consistently maintains stable and superior performance across domains with various fields and different complexities, effectively addressing the catastrophic forgetting prevalent in existing approaches. Detailed results for the TRC scenario are provided in the **Appendix-D** from SM.

4.3 Ablation Studies

The impact of components in LEAR. Table 4 provides empirical validation for the theoretical contributions of each proposed module. Where "CB" denotes using only the collaborative backbone with a single shared expert network across all data domains; "CBE" extends CB with task-specific expert network expansion and ESM expert selection; "SBE" denotes the configuration where "CBE"'s dual backbone architecture is replaced with a single backbone; and "CBE+MI/KL/HSIC" represents CBE augmented with individual components or combination of components. "LEAR w/o ESM" denotes randomly selecting experts during the beginning and testing phase of each task. This table



(a) The forgetting curve of CDM. (b) Visualization of distributions. (c) Clustering results of features.

Figure 2: (a): The comparison of the forgetting curves between LEAR and other baseline methods after learning the CDM sequence. (b): Visualization of memory distributions from the ETI sequence after PCA dimensionality reduction, showing mean positions and covariance ellipsoids. (c): Feature visualization of global backbone and local backbone with HSICBCO regularization.

demonstrates that the collaborative backbone design ("SBE"->"CBE") and the dynamically expanded expert network ("CB"->"CBE") significantly enhance model performance in ETI. In addition, each regularization term and its combinations contribute to varying degrees of performance improvement over "CBE" on these two sequences. Furthermore, the model's performance will significantly drop when inappropriate experts are chosen ("LEAR"->"LEAR w/o ESM"), thereby demonstrating the necessity of ESM. These evaluation results align well with our methodological design.

The analysis of parameter and computational efficiency. As shown in Table 3, LEAR achieves the lowest GPU and CPU utilization among all baseline methods. While baselines including prompt-based approaches (e.g., DualPrompt) and adapter-based variants (e.g., MoE-Adapters) indeed contain fewer trainable parameters (0.5%-5% per ViT block), they inevitably require complete backpropagation through all ViT blocks, necessitating the storage of intermediate activations throughout the entire backbone due to chain rule dependencies, which maintain extensive computation graphs. Conversely, LEAR's innovative design strategically fine-tunes only the last three ViT layers and terminates backpropagation after the third-last layer, thereby significantly reducing the computation graph and GPU usage.

Visualization of the Expert Selection Mechanism. Figure 2b shows the first five ESM-preserved distributions from the ETI scenario, visualized in 3D space after PCA reduction. ESM computes Mahalanobis distances between these distributions and incoming task samples to select experts for either network expansion or test evaluation.

Visualization of the HSICBCO approach. Both the global and local backbones are initialized with identical pretrained weights. Under the guidance of MIBPA and KLDBFA, they learn task-general and task-specific representations, respectively. However, their feature representations still exhibit strong correlations. As illustrated in Figure 2c, the proposed HSICBCO module effectively decouples these representations, demonstrating its capability to promote distinct feature learning. Additional ablation results are provided in the **Appendix-D** from SM.

5 Conclusion and Limitation

In this paper, we propose LEAR, a novel framework for Multi-Domain Continual Learning that simultaneously addresses stability plasticity and efficiency. Specifically, built on a collaborative backbone structure, we introduce MIBPA and KLDBFA to maintain historical prediction consistency and task-specific feature alignment during model updates, while HSICBCO ensures disentangled and complementary representations. Additionally, ESM dynamically selects relevant experts for efficient network expansion and task-agnostic prediction. The empirical results demonstrate the effectiveness of the proposed approach. The primary limitation of this paper is that the proposed approach would contain a considerable number of parameters. To address this issue, we will propose a novel expert merging technology with self-distillation for effective model compression.

Acknowledgements

This work was supported by the Sichuan Provincial Natural Science Foundation Project (No.2025ZNSFSC0510) and National Natural Science Foundation of China (Grant No: 62506067)

References

[1] A. Achille, T. Eccles, L. Matthey, C. Burgess, N. Watters, A. Lerchner, and I. Higgins. Life-long disentangled representation learning with cross-domain latent homologies. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 9873–9883, 2018.

[2] Hongjoon Ahn, Sungmin Cha, Donggyu Lee, and Taesup Moon. Uncertainty-based continual learning with adaptive regularization. In *Advances in Neural Information Processing Systems*, pages 4394–4404, 2019.

[3] Rahaf Aljundi, Klaas Kelchtermans, and Tinne Tuytelaars. Task-free continual learning. In *Proc. of IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, pages 11254–11263, 2019.

[4] Rahaf Aljundi, Klaas Kelchtermans, and Tinne Tuytelaars. Task-free continual learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11254–11263, 2019.

[5] Jihwan Bang, Hyunseo Koh, Seulki Park, Hwanjun Song, Jung-Woo Ha, and Jonghyun Choi. Online continual learning on a contaminated data stream with blurry task boundaries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9275–9284, June 2022.

[6] Hyuntak Cha, Jaeho Lee, and Jinwoo Shin. Co2l: Contrastive continual learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9516–9525, 2021.

[7] Gong Cheng, Junwei Han, and Xiaoqiang Lu. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 105(10):1865–1883, 2017.

[8] C. Cortes, X. Gonzalvo, V. Kuznetsov, M. Mohri, and S. Yang. Adanet: Adaptive structural learning of artificial neural networks. In *Proc. of Int. Conf. on Machine Learning (ICML), vol. PMLR 70*, pages 874–883, 2017.

[9] 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.

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[11] Arthur Douillard, Alexandre Ramé, Guillaume Couairon, and Matthieu Cord. Dytox: Transformers for continual learning with dynamic token expansion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9285–9295, 2022.

[12] A. Gretton, O. Bousquet, A. Smola, and B. Schölkopf. Measuring statistical dependence with Hilbert-Schmidt norms. In *Proc. Int. Conf. on Algorithmic Learning Theory, vol. Lecture Notes in Artif. Intell. (LNAI) 3734*, pages 63–77, 2005.

[13] Arthur Gretton, Karsten Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alex Smola. A kernel method for the two-sample-problem. *Advances in neural information processing systems*, 19, 2006.

[14] Yiduo Guo, Bing Liu, and Dongyan Zhao. Online continual learning through mutual information maximization. In *International Conference on Machine Learning*, pages 8109–8126. PMLR, 2022.

[15] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[16] 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.

[17] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. *ICCV*, 2021.

[18] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

[19] G. E. Hinton, S. Osindero, and Y. W. Teh. A fast learning algorithm for deep belief nets. *Neural Computation*, 18(7):1527–1554, 2006.

[20] Saurav Jha, Dong Gong, He Zhao, and Lina Yao. Npcl: Neural processes for uncertainty-aware continual learning. *Advances in Neural Information Processing Systems*, 36, 2024.

[21] Dahuin Jung, Dongyoon Han, Jihwan Bang, and Hwanjun Song. Generating instance-level prompts for rehearsal-free continual learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11847–11857, 2023.

[22] Haeyong Kang, Rusty John Lloyd Mina, Sultan Rizky Hikmawan Madjid, Jaehong Yoon, Mark Hasegawa-Johnson, Sung Ju Hwang, and Chang D Yoo. Forget-free continual learning with winning subnetworks. In *International Conference on Machine Learning*, pages 10734–10750. PMLR, 2022.

[23] Junsu Kim, Hoseong Cho, Jihyeon Kim, Yihalem Yimolal Tiruneh, and Seungryul Baek. Sddgr: Stable diffusion-based deep generative replay for class incremental object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28772–28781, 2024.

[24] J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska, D. Hassabis, C. Clopath, D. Kumaran, and R. Hadsell. Overcoming catastrophic forgetting in neural networks. *Proc. of the National Academy of Sciences (PNAS)*, 114(13):3521–3526, 2017.

[25] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, Univ. of Toronto, 2009.

[26] Ya Le and Xuan Yang. Tiny imageNet visual recognition challenge. Technical report, Univ. of Stanford, 2015.

[27] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proc. of the IEEE*, 86(11):2278–2324, 1998.

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

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

[30] Mark D McDonnell, Dong Gong, Amin Parvaneh, Ehsan Abbasnejad, and Anton van den Hengel. Ranpac: Random projections and pre-trained models for continual learning. *Advances in Neural Information Processing Systems*, 36, 2024.

[31] Sharada P Mohanty, David P Hughes, and Marcel Salathé. Using deep learning for image-based plant disease detection. *Frontiers in plant science*, 7:215232, 2016.

[32] G. I. Parisi, R. Kemker, J. L. Part, C. Kanan, and S. Wermter. Continual lifelong learning with neural networks: A review. *Neural Networks*, 113:54–71, 2019.

[33] J. Ramapuram, M. Gregorova, and A. Kalousis. Lifelong generative modeling. In *Proc. Int. Conf. on Learning Representations (ICLR), arXiv preprint arXiv:1705.09847*, 2017.

[34] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. iCaRL: Incremental classifier and representation learning. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 2001–2010, 2017.

[35] H. Shin, J. K. Lee, J. Kim, and J. Kim. Continual learning with deep generative replay. In *Advances in Neural Inf. Proc. Systems (NIPS)*, pages 2990–2999, 2017.

[36] James Seale Smith, Leonid Karlinsky, Vyshnavi Gutta, Paola Cascante-Bonilla, Donghyun Kim, Assaf Arbelle, Rameswar Panda, Rogerio Feris, and Zsolt Kira. Coda-prompt: Continual decomposed attention-based prompting for rehearsal-free continual learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11909–11919, 2023.

[37] Gido M van de Ven and Andreas S Tolias. Three continual learning scenarios. In *NeurIPS Continual Learning Workshop*, volume 1, 2018.

[38] Vinay Kumar Verma, Kevin J Liang, Nikhil Mehta, Piyush Rai, and Lawrence Carin. Efficient feature transformations for discriminative and generative continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13865–13875, 2021.

[39] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

[40] Liyuan Wang, Jingyi Xie, Xingxing Zhang, Mingyi Huang, Hang Su, and Jun Zhu. Hierarchical decomposition of prompt-based continual learning: Rethinking obscured sub-optimality. *Advances in Neural Information Processing Systems*, 36:69054–69076, 2023.

[41] 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.

[42] T. Wang and W. Li. Kernel learning and optimization with Hilbert–Schmidt independence criterion. *Int. Jour. of Machine Learning and Cyber.*, 9(10):1707–1717, 2018.

[43] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2097–2106, 2017.

[44] Yabin Wang, Zhiwu Huang, and Xiaopeng Hong. S-prompts learning with pre-trained transformers: An occam's razor for domain incremental learning. *Advances in Neural Information Processing Systems*, 35:5682–5695, 2022.

[45] Zhenyi Wang, Yan Li, Li Shen, and Heng Huang. A unified and general framework for continual learning. *arXiv preprint arXiv:2403.13249*, 2024.

[46] Zifeng Wang, Zizhao Zhang, Sayna Ebrahimi, Ruoxi Sun, Han Zhang, Chen-Yu Lee, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, et al. Dualprompt: Complementary prompting for rehearsal-free continual learning. In *European conference on computer vision*, pages 631–648. Springer, 2022.

[47] Zifeng Wang, Zizhao Zhang, Chen-Yu Lee, Han Zhang, Ruoxi Sun, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, and Tomas Pfister. Learning to prompt for continual learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 139–149, 2022.

[48] Mengqi Xue, Haofei Zhang, Jie Song, and Mingli Song. Meta-attention for vit-backed continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 150–159, 2022.

[49] 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.

[50] M. Zhai, L. Chen, F. Tung, J He, M. Nawhal, and G. Mori. Lifelong GAN: Continual learning for conditional image generation. In *Proc. of the IEEE/CVF Int. Conf. on Computer Vision (ICCV)*, pages 2759–2768, 2019.

[51] 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.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The abstract and introduction accurately reflect the paper's contribution and scope.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: We discuss the limitations of our paper in the conclusion and Appendix-E from SM.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

Answer: [NA]

Justification: The paper does not include theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: We provide the algorithm implementation in Section 3.7 and **Appendix-B** from SM, along with the source code. The detailed experimental setup is documented in **Appendix-C**.

   Guidelines:

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

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

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

Answer: [Yes]

Justification: We provide the algorithm implementation in Section 3.7 and **Appendix-B** from SM, along with the source code. The detailed experimental setup is documented in **Appendix-C**.

Guidelines:

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

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: We provide the algorithm implementation in Section 3.7 and **Appendix-B** from SM, along with the source code. The detailed experimental setup is documented in **Appendix-C**.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [Yes]

   Justification: Detailed experimental results are provided in **Appendix-D** from SM.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

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

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: We provide the detailed experimental setup (include compute resources) in **Appendix-C**.

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: The research conducted in our paper conform, in every respect, with the NeurIPS Code of Ethics.

   Guidelines:

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

10. **Broader impacts**

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

    Answer: [Yes]

    Justification: We provide the societal impacts of the work in **Appendix-E**.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.

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

11. **Safeguards**

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

    Answer: [NA]

    Justification: The paper poses no such risks.

    Guidelines:

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

12. **Licenses for existing assets**

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

    Answer: [NA]

    Justification: The paper does not use existing assets.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

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

    Answer: [NA]

    Justification: The paper does not release new assets.

    Guidelines:

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

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

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

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

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

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

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

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

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

    Answer: [NA]

    Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.