UNDERSTANDING CATASTROPHIC INTERFERENCE ON THE IDENTIFIBILITY OF LATENT REPRESENTATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Catastrophic interference, also known as catastrophic forgetting, is a fundamental challenge in machine learning, where a trained learning model progressively loses performance on previously learned tasks when adapting to new ones. In this paper, we aim to better understand and model the catastrophic interference problem from a latent representation learning point of view, and propose a novel theoretical framework that formulates catastrophic interference as an identification problem. Our analysis demonstrates that the forgetting phenomenon can be quantified by the distance between partial-task aware (PTA) and all-task aware (ATA) setups. Building upon recent advances in identifiability theory, we prove that this distance can be minimized through identification of shared latent variables between these setups. When learning, we propose our method **`ICON`** with two-stage training strategy: First, we employ maximum likelihood estimation to learn the latent representations from both PTA and ATA configurations. Subsequently, we optimize the KL divergence to identify and learn the shared latent variables. Through theoretical guarantee and empirical validations, we establish that identifying and learning these shared representations can effectively mitigate catastrophic interference in machine learning systems. Our approach provides both theoretical guarantees and practical performance improvements across both synthetic and benchmark datasets.

1 INTRODUCTION

catastrophic interference represents a fundamental challenge in machine learning (Cha et al., 2021; Liang & Li, 2023; Xiao et al., 2024), where a model trained sequentially on multiple tasks experiences significant performance degradation on previously learned tasks when adapting to new ones. This phenomenon manifests as a direct consequence of the distributional shift between tasks, coupled with the model's capability to preserve previously learned knowledge when optimizing for new data. Therefore, the model must adapt to new tasks while preserving critical knowledge from earlier experiences, mirroring human cognitive abilities to accumulate knowledge progressively.

Handling catastrophic interference presents unique theoretical challenges, as the model involves a dynamic evolution of itself as it learns new tasks. This creates an inherent instability in the data representation process —learning from new tasks fundamentally alters the model's parameters, potentially disrupting the representations learned for previous tasks. In other words, the learned data generating process changes from task to task as the models evolve. To better understand and model this challenge, we take inspiration from recent advances in Causal Representation Learning (Schölkopf et al., 2021; Kong et al., 2022; Li et al., 2023; Kong et al., 2024), and approach this problem via modeling the data generating process through the mixing functions that map low-dimensional latent variables to high-dimensional observations. More specifically, we distinguish between two configurations in catastrophic interference handling: the partial-task aware (PTA) setting, which represents a model trained on a subset of tasks, uses a mixing function that has only seen data up to the current task, and the all-task aware (ATA) setting, which represents an ideal model trained on all tasks, leverages a single task-invariant mixing function g.

In this work, we propose a new theoretical framework that formulates catastrophic interference as a latent-variable identification problem. Our key insight is that catastrophic interference can be quantified by measuring the distance between latent representations in PTA and ATA settings. By identifying the shared latent variables between these setups, we can establish a principled approach to preserving knowledge across distributional shifts. Builds upon our theoretical findings, we introduce an a two-stage learning methodology, Identifiable CatastrOphic iNterference (**ICON**). First, we employ maximum likelihood estimation to learn the latent representations from both PTA and ATA configurations independently. Subsequently, we optimize the KL divergence between these representations to identify and learn their shared components. Through evaluating on both synthetic data and real-world benchmarks, **ICON** effectively mitigate the catastrophic interference.

Our contributions are threefold: (1) We formulate catastrophic interference as a latent-variable identification problem, providing a novel theoretical perspective that quantifies forgetting through distributional distances; (2) We establish identifiability conditions for the shared latent variables between PTA and ATA setups, proving when and how knowledge can be preserved across distributional shifts under certain assumptions; (3) Based on our theoretical findings, we develop a practical approach that demonstrates superior performance on both synthetic data and standard benchmarks on handling catastrophic interference, outperforming current state-of-the-art methods. By bridging theory and practice, our work provides the first work delving into the nature of catastrophic interference through identifications, and offers a principled framework.

2 Related Work

2.1 Handling Catastrophic Forgetting

Existing learning methods can be categorized into five primary approaches when handling catastrophic interference: (1) *Regularization-based methods* introduce constraints on model parameters or outputs within the loss function to mitigate catastrophic forgetting when learning new tasks. Representative works include Chaudhry et al. (2018); Aljundi et al. (2018); Hou et al. (2019); Cha et al. (2021). (2) *Memory replay-based methods* explicitly store and revisit past experiences by maintaining a subset of previous task samples, thereby reducing forgetting. Notable examples include Arani et al. (2022); Caccia et al. (2022); Bonicelli et al. (2022); Sarfraz et al. (2023); Wang et al. (2023b); Liang & Li (2023). (3) *Gradient-projection-based methods* mitigate forgetting by constraining gradient updates to subspaces that minimize interference with prior knowledge. Relevant studies include Chaudhry et al. (2020); Farajtabar et al. (2020); Saha et al. (2021); Wang et al. (2021); Lin et al. (2022); Qiao et al. (2024); Xiao et al. (2024). (4) *Architecture-based methods* dynamically adjust the neural network structure to integrate new tasks while preserving performance on previous ones. Key contributions in this category include Mallya & Lazebnik (2018); Serra et al. (2018); Li et al. (2019); Hung et al. (2019). (5) *Bayesian-based methods* leverage Bayesian inference principles to model uncertainty and facilitate new task while maintaining prior knowledge. Representative works include Kao et al. (2021); Henning et al. (2021); Pan et al. (2020); Titsias et al. (2020); Rudner et al. (2022).

2.2 Identifiability of Latent Variables with Distribution Shifts

Identifying latent variables in causal representation learning has emerged as a foundational paradigm for understanding representation learning in deep neural networks (Schölkopf et al., 2021; Khemakhem et al., 2020). This approach typically assumes latent variables $\mathbf{z}$ generate observed data $\mathbf{x}$ through a generative function. However, when this function exhibits nonlinearity—as is common in deep learning models—recovering the original latent variables becomes technically challenging (Khemakhem et al., 2020).

To address this challenge, several recent works (Li et al., 2023; Song et al., 2023; Chen et al., 2024; Zheng & Zhang, 2024; Morioka & Hyvarinen, 2024) have introduced auxiliary labels $\mathbf{u}$ that induce distributional shifts in the latent components across different conditions. While effective in certain scenarios, these approaches depend critically on assuming the access to multiple disparate distributions with overlapping supports, including the target distribution. Furthermore, recent advances by (Yao et al., 2024; Kong et al., 2024) either require labeled grouping data generating process or assume identical mixing functions across different data generating processes. Our theoretical framework overcomes these limitations by eliminating the need for either labeled diverse generating processes or identical mixing functions. This broadened scope encompasses earlier work such as (Yao et al., 2024; Kong et al., 2024) as a special case or ours.

3 Problem Setup

Given T tasks in total, we aim to learn a task-invariant model to adapt to all tasks. However, the possible data distributions shift across tasks raises the challenge of catastrophic forgetting, where a model's performance on previously learned task $\mathbf{t}$ could degrade after training on all T tasks.

In this section, to formally characterize the nature of catastrophic forgetting, we present two data generating processes.

We term the first one by partial-task aware (PTA) approach, leveraging the mixing functions $g^{:\mathbf{t}}$ from task 1 to task $t-1$ ($t > 1$), resulting $T-1$ mixing functions of PTA setup in total:

$$\mathbf{x}^{\mathbf{t}} = g^{:\mathbf{t}}(\bar{\mathbf{z}}^{\mathbf{t}}) \tag{1}$$

where $\mathbf{x}^{\mathbf{t}} \in \mathbb{R}^K$ denotes the observations of task $\mathbf{t}$, the nonlinear mixing function $g^{:\mathbf{t}} : \mathbb{R}^N$ is a diffeomorphism onto $\mathbb{R}^K$, and $\bar{\mathbf{z}}^{\mathbf{t}} \in \mathbb{R}^N$ denotes the task-specific continuous latent variable.

Unlike the PTA approach, the second data generating process represents the all-task aware (ATA) paradigm, which aims to learn a mixing function g that can handle all T tasks. ATA meets the goal of continual learning in the sense that it works on all domains:

$$\mathbf{x}^{\mathbf{t}} = g(\tilde{\mathbf{z}}^{\mathbf{t}}) \tag{2}$$

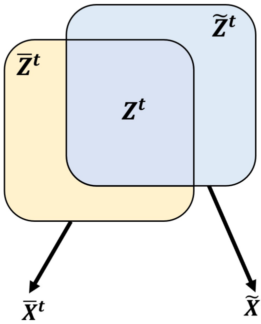


Figure 1: The illustration of our definition of subspace identification in Def. 3. Given two space of latent variables of PTA setup $\overline{\mathcal{Z}}^{\mathbf{t}} \ni \bar{\mathbf{z}}^{\mathbf{t}}$ and ATA approach $\tilde{\mathcal{Z}}^{\mathbf{t}} \ni \tilde{\mathbf{z}}^{\mathbf{t}}$, we aim to identify their intersection $\mathbf{z}^{\mathbf{t}} \in \mathcal{Z}^{\mathbf{t}} = \overline{\mathcal{Z}}^{\mathbf{t}} \cap \tilde{\mathcal{Z}}^{\mathbf{t}}$.

Simialrly, $\tilde{\mathbf{z}}^{\mathbf{t}} \in \mathbb{R}^N$ denotes the continous latent variable for the task $\mathbf{t}$, g is an nonlinear mixing function and diffeomorphism onto $\mathbb{R}^K$. For both Eqs. 1 and 2, we focus on the undercomplete case, i.e., $N \leq K$.

We are now ready to connect catastrophic forgetting with Eq. 1 and 2. Let us define a $\mathcal{G} \subset g^{:t} : \mathbb{R}^{\bar{\mathbf{z}}} \to \mathbb{R}^{\mathbf{x}}$, where G denotes a hypothesis class. $l : \mathcal{G} \times \mathbb{R}^{\mathbf{x}} \to [0, B_l]$ denotes the loss function, where $B_l > 0$ is a constant. In this work, we leverage the negative log-likelihood for l, i.e., $l(\hat{g}^{:t}, \mathbf{x}^{\mathbf{t}}) = -\log p_{\hat{g}^{:t}}(\mathbf{x}^{\mathbf{t}})$. Similar, we define a lost function $l(\hat{g}, \mathbf{x}^{\mathbf{t}}) = -\log p_{\hat{g}}(\mathbf{x}^{\mathbf{t}})$ Following the defintion 1.1 in Wang et al. (2024), we reinterpret the catastrophfic forgetting $\mathcal{F}$ by:

$$\mathcal{F} = \mathbb{E}_{\mathbf{t}} \frac{1}{|\mathbf{x}^{\mathbf{t}}|} \sum_{\mathbf{x}^{\mathbf{t}}} (-l(\hat{g}^{:t}, \mathbf{x}^{\mathbf{t}}) + l(g, \mathbf{x}^{\mathbf{t}})) \tag{3}$$

where $|\mathbf{x}^{\mathbf{t}}|$ denotes the sample numbers of $\mathbf{x}^{\mathbf{t}}$. This formulation aligns with Definition 1.1 in Wang et al. (2024) by utilizing negative log-likelihood as the performance measurement metric, i.e., Eq. 3 quantifies the difference between two data generating process from Eq. 1 and Eq. 2, respectively.

Given the invertibility of both mapping functions $g^{:\mathbf{t}}$ and g, and both $\bar{\mathbf{z}}^{\mathbf{t}}$ and $\tilde{\mathbf{z}}^{\mathbf{t}}$ live in $\mathbb{R}^N$, we establish the observed differences under Eq. 2 and Eq. 1 uniquely determined by the underlying differences between their respective latent representations $\bar{\mathbf{z}}^{\mathbf{t}}$ and $\tilde{\mathbf{z}}^{\mathbf{t}}$. This allows us to decompose $\bar{\mathbf{z}}^{\mathbf{t}}$ and $\tilde{\mathbf{z}}^{\mathbf{t}}$ into two parts: their difference and their overlap. In this work, to minimize $\mathcal{F}$ in Eq. 3, we focus on identifying the overlap between $\bar{\mathbf{z}}^{\mathbf{t}}$ and $\tilde{\mathbf{z}}^{\mathbf{t}}$ across the PTA and ATA settings. Formally, for the latent variable manifolds $\overline{\mathcal{Z}}^{\mathbf{t}} \ni \bar{\mathbf{z}}^{\mathbf{t}}$ and $\tilde{\mathcal{Z}}^{\mathbf{t}} \ni \tilde{\mathbf{z}}^{\mathbf{t}}$, we denote their intersection by $\mathcal{Z}^{\mathbf{t}} = \overline{\mathcal{Z}}^{\mathbf{t}} \cap \tilde{\mathcal{Z}}^{\mathbf{t}}$. $\forall \mathbf{z}^{\mathbf{t}} \in \mathcal{Z}^{\mathbf{t}}, \mathbf{z}^{\mathbf{t}} = \tilde{\mathbf{z}}^{\mathbf{t}} = \bar{\mathbf{z}}^{\mathbf{t}}$. We introduce the definition of the identifiability in the following (subspace identifiability suffices):

(Identifiability of latent variables shared by Eqs. 1 and 2): *For any pair of mixing functions* (g^t, g) *defined in Equations 1 and 2 respectively, there exists a shared latent space region* $\mathcal{Z}^{\mathbf{t}}$ *where the latent variables from both the PTA and ATA settings coincide.* $\forall \mathbf{z}^{\mathbf{t}} \in \mathcal{Z}^{\mathbf{t}}$*, there exists an invertible transformation* t *such that:* $\hat{\mathbf{z}}^{\mathbf{t}} = t(\mathbf{z}^{\mathbf{t}})$.

4 Identifiability Theory

To establish our identifiability results, we begin with formalizing the distance between two manifolds $\mathcal{Z}_1^{\mathbf{t}}$ and $\mathcal{Z}_2\mathbf{t}$ as:

$$\mathcal{D}(\mathcal{Z}_1^{\mathbf{t}}, \mathcal{Z}_2^{\mathbf{t}}) = inf_{\mathbf{z}'_{\mathit{1}} \in \mathcal{Z}_{\mathit{1}}^{\mathbf{t}}, \mathbf{z}'_{\mathit{2}} \in \mathcal{Z}_{\mathit{2}}^{\mathbf{t}}} \|\mathbf{z}'_{\mathit{1}} - \mathbf{z}'_{\mathit{2}}\|_{\mathit{2}} \tag{4}$$

where $\overline{\mathbf{z}}'$ denotes points on the boundary support of $\overline{\mathcal{Z}}^{\mathbf{t}}$, and $\tilde{\mathbf{z}}'$ denotes points on the boundary support of $\tilde{\mathcal{Z}}^{\mathbf{t}}$. Here, $\|\cdot\|_2$ represents the Euclidean distance.

We define the minimum distance from a point $\mathbf{z}^{\mathbf{t}}$ to any point in the manifold $\tilde{\mathcal{Z}}^{\mathbf{t}}$ for the ATA setting:

$$\mathcal{D}(\tilde{\mathbf{z}}^{\mathbf{t}}, \mathbf{z}^{\mathbf{t}}) = \arg\min_{\tilde{\mathbf{z}}^{\mathbf{t}} \in \tilde{\mathcal{Z}}^{\mathbf{t}}} \|\tilde{\mathbf{z}}^{\mathbf{t}} - \mathbf{z}^{\mathbf{t}}\|_2 \tag{5}$$

Similarly, for the PTA setting, we define:

$$\mathcal{D}(\overline{\mathbf{z}}^{\mathbf{t}}, \mathbf{z}^{\mathbf{t}}) = \arg\min_{\overline{\mathbf{z}}^{\mathbf{t}} \in \overline{\mathcal{Z}}^{\mathbf{t}}} \|\overline{\mathbf{z}}^{\mathbf{t}} - \mathbf{z}^{\mathbf{t}}\|_2 \tag{6}$$

Having established the necessary distance metrics, we are now ready to present our identifiability results.

Theorem 1. *Given the data generating process in Eq. 1 and Eq. 2, if the following assumptions are satisfied:*

1. *(Smoothness and invertibility) The mixing function $g^{\mathbf{t}}$ in Eq. 1, and g in Eq. 2 are smooth functions and invertible everywhere;*
2. *(The existence of intersection) For the latent variable manifolds $\overline{\mathcal{Z}}^{\mathbf{t}} \ni \overline{\mathbf{z}}^{\mathbf{t}}$ and $\tilde{\mathcal{Z}}^{\mathbf{t}} \ni \tilde{\mathbf{z}}^{\mathbf{t}}$, their intersection $\mathcal{Z}^{\mathbf{t}} = \overline{\mathcal{Z}}^{\mathbf{t}} \cap \tilde{\mathcal{Z}}^{\mathbf{t}} \neq \emptyset$.*
3. *(Path-connected) both $\overline{\mathcal{Z}}^{\mathbf{t}}$ and $\tilde{\mathcal{Z}}^{\mathbf{t}}$ are path-connected;*
4. *(Compactness) The spaces of observed variables $\overline{\mathbf{x}}^{\mathbf{t}}$ and $\tilde{\mathbf{x}}^{\mathbf{t}}$ are closed and bounded;*
5. *(Constrained out-of-intersection distance) $\forall \tilde{\mathbf{z}}^{\mathbf{t}} \notin \mathcal{Z}^{\mathbf{t}}$, if $\exists \tilde{\mathcal{Z}}_1^{\mathbf{t}}, \tilde{\mathcal{Z}}_2^{\mathbf{t}}$, the distance $\mathcal{D}(\tilde{\mathbf{z}}^{\mathbf{t}}, \mathbf{z}^{\mathbf{t}})$ between the out-of-intersection $\tilde{\mathbf{z}}^{\mathbf{t}}$ and $\mathbf{z}^{\mathbf{t}}$ is constrained by $\mathcal{D}(\tilde{\mathbf{z}}^{\mathbf{t}}, \mathbf{z}^{\mathbf{t}}) \leq \frac{\mathcal{D}(\tilde{\mathcal{Z}}_1^{\mathbf{t}}, \tilde{\mathcal{Z}}_2^{\mathbf{t}})}{2J_{\tilde{\mathbf{u}}}}$. $J_{\tilde{\mathbf{u}}}$ denotes the spectrum norm of $\{J_g(\tilde{\mathbf{z}}_1^{\mathbf{t}}), J_g(\tilde{\mathbf{z}}_2^{\mathbf{t}})\}$.*

If we can learn the optimal estimation $\hat{\mathbf{z}}^{\mathbf{t}}$ of $\mathbf{z}^{\mathbf{t}} \in \mathcal{Z}^{\mathbf{t}}$ such that:

$$sup(p(\hat{\mathbf{z}})),\ s.t.\ \ sup(p_{g^{\hat{:}t}}(\mathbf{x}^{\mathbf{t}})) \quad \& \quad sup(p_{\hat{g}}(\mathbf{x}^{\mathbf{t}})) \tag{7}$$

then the identifiability results stated in Def. 3 are obtained.

Proof Sketch: Our proof establishes identifiability via contradiction by assuming a latent variable $\mathbf{z}^t$ simultaneously resides on two distinct latent manifolds. Differences between latent points are expressed as integrals involving the Jacobian of the generative function g, bounded by the spectral norm of the Jacobian multiplied by their latent-space distance. Since both latent points generate the same observed data, their output difference must be zero, implying a zero distance between distinct latent points. This result contradicts the theorem's assumption of a strictly positive minimal separation between distinct manifold points. Therefore, each observed data point must originate from a unique latent point, ensuring identifiability.

Remarks: Assumption 1 & Assumption 2 guarantee the the existence of an intersection between $\bar{\mathcal{Z}}$ and $\tilde{\mathcal{Z}}$. The path-connectedness specified in Assumption 3 and The compactness property in Assumption 4 impose the geometric boundedness of the overlap between $\bar{\mathcal{Z}}$ and $\tilde{\mathcal{Z}}$. Assumption 5 establishes critical upper bounds for the distances $\mathcal{D}(\tilde{\mathbf{z}}^{\mathbf{t}}, \mathbf{z}^{\mathbf{t}})$.

5 ICON APPROACH

Building upon our identifiability results, we now introduce **ICON** to estimate the latent causal variables. Our approach aims to achieve the observational equivalence in Eq. 7 by modeling the data generating processes in Eqs. 1 and 2. In what follows, we introduce each part of our network individually.

5.1 NETWORK DESIGN

Our network architecture is designed to uncover the latent variables for both PTA and ATA setup through carefully constructed flow-based models.

Network Structure for PTA approach For the PTA setup described in Eq. 1, **ICON** formalizes the probabilistic joint distribution as:

$$p(\mathbf{x^t}, \bar{\mathbf{z}}^\mathbf{t}) = p_{g^{:\mathbf{t}}}(\mathbf{x^t}|\bar{\mathbf{z}}^\mathbf{t})p(\bar{\mathbf{z}}^\mathbf{t}) \tag{8}$$

To implement the invertible mapping function $g^\mathbf{t}$ in Eq. 1, we employ General Incompressible-flow Network (GIN) Sorrenson et al. (2020), which provide a highly expressive class of normalizing flows with the following inverse mapping:

$$\hat{\bar{\mathbf{z}}}^\mathbf{t} \sim \mathcal{N}(\hat{\bar{\mu}}^\mathbf{t}, \hat{\bar{\sigma}}^\mathbf{t}),\ \hat{\bar{\mu}}^\mathbf{t}, \hat{\bar{\sigma}}^\mathbf{t} = \hat{g}^{:t}_{-1}(\mathbf{x^t}) \tag{9}$$

where $\hat{g}^{:\mathbf{t}}_{-1}$ denotes the estimation of inverse of the mixing function $g^\mathbf{t}$ for PTA settings.

Network Structure for ATA framework In contrast to the PTA setting, the ATA framework in Eq. 2 requires to learn a task-invariant mixing function g, The joint distribution in this case is modeled as:

$$p(\mathbf{x^t}, \tilde{\mathbf{z}}^\mathbf{t}) = p_g(\mathbf{x^t}|\tilde{\mathbf{z}}^\mathbf{t})p(\tilde{\mathbf{z}}^\mathbf{t})d \tag{10}$$

We implement this task-invariant mapping using another GIN model that processes data from all tasks:

$$\hat{\tilde{\mathbf{z}}}^\mathbf{t} \sim \mathcal{N}(\hat{\tilde{\mu}}^\mathbf{t}, \hat{\tilde{\sigma}}^\mathbf{t}),\ \hat{\tilde{\mu}}^\mathbf{t}, \hat{\tilde{\sigma}}^\mathbf{t} = \hat{g}_{-1}(\mathbf{x^t}) \tag{11}$$

5.2 Learning Objective

In this work, we learn $\hat{\bar{\mathbf{z}}}^\mathbf{t}$ through maximum likelihood estimation (MLE). Specifically, we estimate $p(\hat{\mathbf{x}}^\mathbf{t})$ of Eq. 8 by optimizing the following objective:

$$\mathcal{L}^\mathbf{t} = \frac{1}{\mathbf{t}} \sum_{i=1}^{\mathbf{t}} \Big(\frac{1}{|\mathbf{x^t}|} \sum_{\mathbf{x^t}} \big(\log p(\hat{g}^{:\mathbf{t}}_{-1}(\mathbf{x^t})) \big) \Big) \tag{12}$$

where $|\mathbf{x^t}|$ denotes the number of $\mathbf{x^t}$ of the task of $\mathbf{t}$. Eq. 12 leverages the volume-preservation from GIN (Sorrenson et al., 2020).

Similarly, the learning objective for $p(\hat{\mathbf{x}}^\mathbf{t})$ of Eq. 10 from the true observations $\mathbf{x^t}$ also employs MLE:

$$\mathcal{L} = \frac{1}{T} \sum_{t=1}^{T} \Big(\frac{1}{|\mathbf{x^t}|} \sum_{\mathbf{x^t}} \big(\log p(\hat{g}_{-1}(\mathbf{x^t})) \big) \Big) \tag{13}$$

In this equation, following Sorrenson et al. (2020), the volume-preservation is used. When comparing with Eq. 12, we observe that Eq. 13 indicates $\hat{g}$ being trained across all T tasks, highlighting the fundamental distinction between our PTA and ATA approaches.

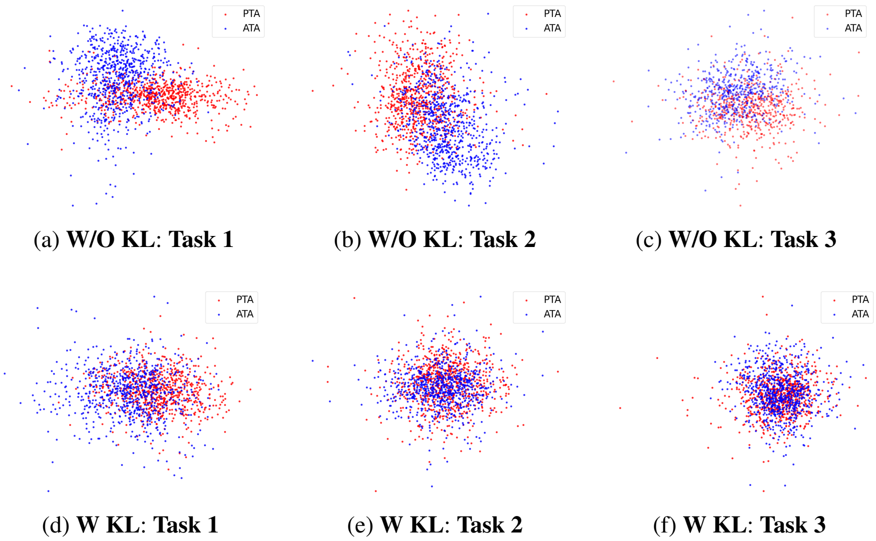
In addition to Eq. 12 and Eq. 13, our work focuses on identifying and learning the sharing of latent variable $\mathbf{z^t}$. Based on our reinterpretation of Eq. 3, maximizing the distribution similarity of $\mathbf{z^t}$ directly contributes to minimizing the catastrophic forgetting term $\mathcal{F}$. To achieve this objective, for each task $\mathbf{t}$, we employ the KL divergence to further tune both $\hat{g}^{:\mathbf{t}}$ and g:

$$\mathcal{KL} = \frac{1}{|\mathbf{x^t}|} \sum_{\mathbf{x^t}} q(\hat{\tilde{\mathbf{z}}}^\mathbf{t}) \log \big(\frac{q(\hat{\tilde{\mathbf{z}}}^\mathbf{t})}{q(\hat{\bar{\mathbf{z}}}^\mathbf{t})} \big) = \frac{1}{\mathbf{t}} \sum_{i=1}^{\mathbf{t}} \Big(\frac{1}{|\mathbf{x^t}|} \sum_{\mathbf{x^t}} q(\hat{g}_{-1}(\mathbf{x^t})) \log \big(\frac{q(\hat{g}_{-1}(\mathbf{x^t}))}{q(\hat{g}^{:\mathbf{t}}_{-1}(\mathbf{x^t}))} \big) \Big) \tag{14}$$

where $q(\hat{\tilde{\mathbf{z}}}^\mathbf{t})$ and $q(\hat{\bar{\mathbf{z}}}^\mathbf{t})$ denote the posterior of $\hat{\tilde{\mathbf{z}}}^\mathbf{t}$ and $\hat{\bar{\mathbf{z}}}^\mathbf{t}$, which are learned using Eq. 9 and Eq. 11, respectively.

5.3 Training and Inference

Two-stage Training ICON takes inspiration from Li et al. (2024) to implement a two-stage training mechanism. The first stage focuses on independently optimizing the objectives in Eq. 12 and Eq. 13. In the second stage, we jointly train both $\hat{g}^{:\mathbf{t}}$ and $\hat{g}$ by minimizing the KL divergence defined in



(a) **W/O KL**: **Task 1** (b) **W/O KL**: **Task 2** (c) **W/O KL**: **Task 3**

(d) **W KL**: **Task 1** (e) **W KL**: **Task 2** (f) **W KL**: **Task 3**

Figure 2: Visualization of latent space distributions across three tasks under PTA (blue) and ATA (red) setups from our simulations. The top row shows representations without optimizing the KL divergence in Eq. 14, displaying significant disparity between PTA and ATA. The bottom row demonstrates improved alignment through our proposed KL-based identification approach, illustrating effective mitigation of catastrophic forgetting across sequential tasks. For clear visualizations, each figure displays 1,000 uniformly sampled points $\hat{\bar{\mathbf{z}}}^{\mathbf{t}}$ and $\hat{\tilde{\mathbf{z}}}^{\mathbf{t}}$, respectively.

Eq. 14. This process effectively reduces the discrepancies between the latent representations $\hat{\bar{\mathbf{z}}}^{\mathbf{t}}$ and $\hat{\tilde{\mathbf{z}}}^{\mathbf{t}}$. Notably, during training, we process tasks sequentially, maintaining access to the task identity $\mathbf{t}$ to calculate the objective functions in Section 5.2.

Inference During inference stage, only $\hat{g}$ is used as the goal of continual learning is to learn a task-invariant mixing function. Thus, **ICON** does not require $\mathbf{t}$ during inference since $\hat{g}$ is task invariant.

6 Experiments

6.1 Synthetic Experiments

Experimental Setup To evaluate **ICON** 's ability to mitigate catastrophic forgetting ($\mathcal{F}$), we first conducted simulation experiments. We generated synthetic datasets satisfying the identifiability assumptions outlined in Theorem 4.

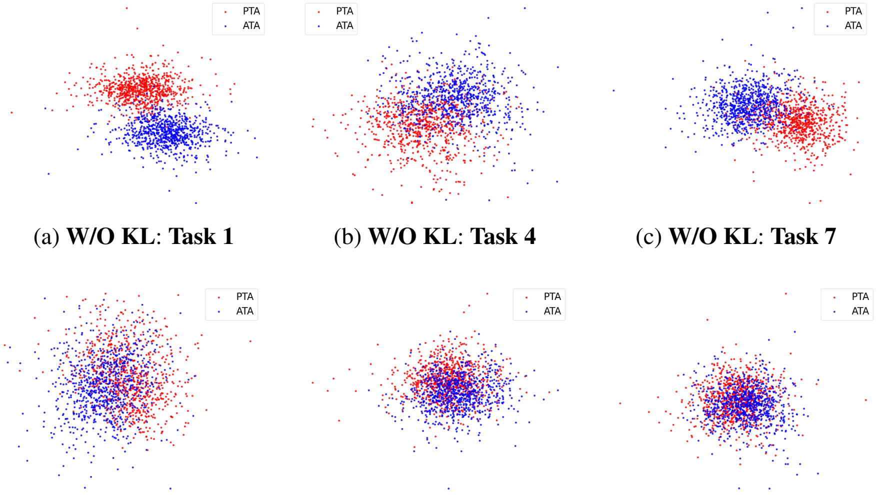
Specifically, our approach generated four distinct scenarios of observations, with each scenario corresponding to a particular task. The latent variables for each scenario comprised 16 dimensions, which we partitioned into two parts: (1) a 8-dimensional task-invariant component drawn from $\mathcal{N}(0, I)$ that remained constant across all tasks, and (2) a 8-dimensional task-specific component drawn from $\mathbf{z_v} \sim \mathcal{N}(\mu, \sigma^2 I)$, which varied between tasks. For each task, the data generation process begins with $10,000$ latent data points, where $\mu \sim \text{Uniform}(-4, 4)$ and $\sigma^2 \sim \text{Uniform}(0.1, 1)$. Following the practices established in Kong et al. (2022); Li et al. (2023), we used a two-layer MLP to generate the observations, which compriese 16 dimensions.

In our synthetic experiments, we aim to determine whether $\mathcal{F}$ could be minimized using our training objectives presented in Section 5.2. Therefore, we compare the average Root Mean Squared Error (RMSE) of the reconstructions across all 4 tasks obtained from the $\hat{g}$ trained with our objective in Eq. 13, and the mixing function of ATA setup without Eq. 14.

Results and Discussions Table 1 summarizes our main findings on our simulations. We evaluate both PTA and ATA setup of **ICON** against the baseline without KL divergence of Eq. 14 in identifying the shared latent varaible $\mathbf{z}^{\mathbf{t}}$.

Table 1: Average Root Mean Squared Error (RMSE) comparison between PTA and ATA frameworks, with and without optimizing KL divergence (Eq. 14). Lower values indicate better performance.

		Average RMSE $\times 10^{-1}$
PTA setup	w/o KL	0.12
ATA setup		0.20
PTA setup	w KL	0.12
ATA setup		0.13



(a) **W/O KL**: **Task 1** (b) **W/O KL**: **Task 4** (c) **W/O KL**: **Task 7**

(d) **W KL**: **Task 1** (e) **W KL**: **Task 4** (f) **W KL**: **Task 7**

Figure 3: Visualization of latent space distributions across Tasks 1, 4, and 7 on ImageNet-100 dataset, comparing representations from PTA (blue) and ATA (red) frameworks. The top row shows results without using KL divergence optimization in Eq. 14, where significant distribution misalignment indicates catastrophic forgetting as training progresses through tasks. The bottom row demonstrates our **ICON** with KL divergence optimization, exhibiting substantially improved alignment. For visualization clarity, each subfigure displays 1,000 uniformly sampled points from the estimated latent representations $\hat{\bar{\mathbf{z}}}^t$ (PTA framework) and $\hat{\tilde{\mathbf{z}}}^t$ (ATA framework).

We can observe that **ICON** incorporating the KL divergence term substantially improves performance in the ATA setup, reducing the average MSE from 0.20 to 0.13 (a 35% improvement) across all tasks. Notably, this evidently indicates that our approach effectively handles the catastrophic forgetting by identifying the shared latent variables.

Figure 2 provides visual evidence of this improvement compared to the baseline without KL divergence. The latent space distributions across three tasks reveal that in the top row (without KL), there exists significant disparity between the PTA (blue) and ATA (red) representations, particularly pronounced in Tasks 1 and 2. This disparity directly corresponds to catastrophic forgetting in the model. In contrast, the bottom row (with KL) exhibits markedly reduced differences between the PTA and ATA setups. This improved alignment of latent representations confirms that our approach demonstrates its effectiveness in addressing catastrophic forgetting.

6.2 Real-world Experiments

To verify the efficacy of our theory in complex real-world scenarios, we further conduct real-work experiments.

Experimental Setup We evaluate our approach on two standard benchmarks for handling catastrophic interference: CIFAR-100 (Krizhevsky et al., 2009) and ImageNet-100 (Deng et al., 2009).

Table 2: Average classification accuracy (%) comparison on ImageNet-100 and CIFAR-100 datasets. The best results are highlighted in **bold**, and the second best are in **underline**.

	ImageNet-100	CIFAR-100
CoOp (Zhou et al., 2022)	79.14	81.17
MaPLe (Khattak et al., 2023)	79.23	82.74
AttriCLIP (Wang et al., 2023a)	82.39	79.31
Continual-CLIP (Thengane et al., 2022)	83.99	78.65
CLIP-Adapter (Gao et al., 2024)	84.13	78.75
CLAP (Jha et al., 2024)	<u>87.76</u>	<u>86.13</u>
ICON (ours)	**88.91**	**87.07**

CIFAR-100 comprises 60,000 RGB images (32×32 pixels) distributed across 100 classes. Following established protocols, we partition this dataset into 10 sequential tasks, each containing 10 distinct classes. Each class contains 500 training and 100 testing samples, ensuring a balanced evaluation framework. For ImageNet-100, a carefully curated subset of the full ImageNet dataset, we utilize higher-resolution images (224×224 pixels) from 100 classes. The dataset provides approximately 1,300 training and 50 testing samples per class. Consistent with recent state-of-the-art approach, such as CLAP (Zhou et al., 2022; Thengane et al., 2022; Gao et al., 2024; Wang et al., 2023a; Khattak et al., 2023; Derakhshani et al., 2023; Jha et al., 2024), ImageNet-100 is divided into 10 tasks with 10 classes per task. Across both benchmarks, we report the average classification accuracy on test data across all tasks.

Since both datasets focus on image classification tasks, we adapt **ICON** using noise contrastive estimation (NCE) (Gutmann & Hyvärinen, 2010). First, we train our model using the objectives defined in Equations 12, 13, and 14. Subsequently, we define our NCE loss $\mathcal{L}^{\mathrm{c}}$ as: $\mathcal{L}^{\mathrm{c}} = -\sum_k \log \frac{\exp(\mathrm{sim}(\hat{\tilde{\mathbf{z}}}_k^{\mathbf{t}}, \hat{e}_k)/\tau)}{\sum_m \exp(\mathrm{sim}(\hat{\tilde{\mathbf{z}}}_k^{\mathbf{t}}, \hat{e}_m)/\tau)}$, where $\mathrm{sim}(\cdot, \cdot)$ represents the cosine similarity between the text embeddings $\hat{e}_*$ of the class labels and the learned latent variables $\hat{\tilde{\mathbf{z}}}_k^{\mathbf{t}}$, and τ is a temperature parameter controlling the sharpness of the distribution. Notably, the text embeddings $\hat{e}_*$ are from all task. Prior to computing $\mathcal{L}^{\mathrm{c}}$, we project $\hat{\tilde{\mathbf{z}}}^t$ to a 512-dimensional embedding space using an MLP layer, aligning with the dimensionality of the text embeddings. For both datasets, we set the dimensionality of the latent representations $\hat{\bar{\mathbf{z}}}^{\mathrm{t}}$ and $\hat{\tilde{\mathbf{z}}}^{\mathrm{t}}$ to 24.

For feature extraction across all tasks and experiments, we employ the Vision Transformer (ViT-B/16) backbone from Radford et al. (2021) to obtain image features $\mathbf{x}^{\mathbf{t}}$, by following Zhou et al. (2022); Thengane et al. (2022); Gao et al. (2024); Wang et al. (2023a); Khattak et al. (2023); Derakhshani et al. (2023); Jha et al. (2024). To ensure fair comparison with prior work, we maintain consistent replay memory configurations. Specifically, we randomly sample 2,000 exemplars for the CIFAR-100 dataset and 1,000 exemplars for the ImageNet-100 dataset. We optimize our network using AdamW Loshchilov & Hutter (2019) with an initial learning rate of 0.002 and weight decay of 10^{-2}, employing a cosine annealing schedule for learning rate decay. For the contrastive learning objective, we set the temperature parameter $\tau = 0.07$ in the NCE loss $\mathcal{L}^{\mathbf{c}}$. We implement our framework in PyTorch and conduct all experiments on a single NVIDIA GeForce RTX 3090 GPU with 24GB memory.

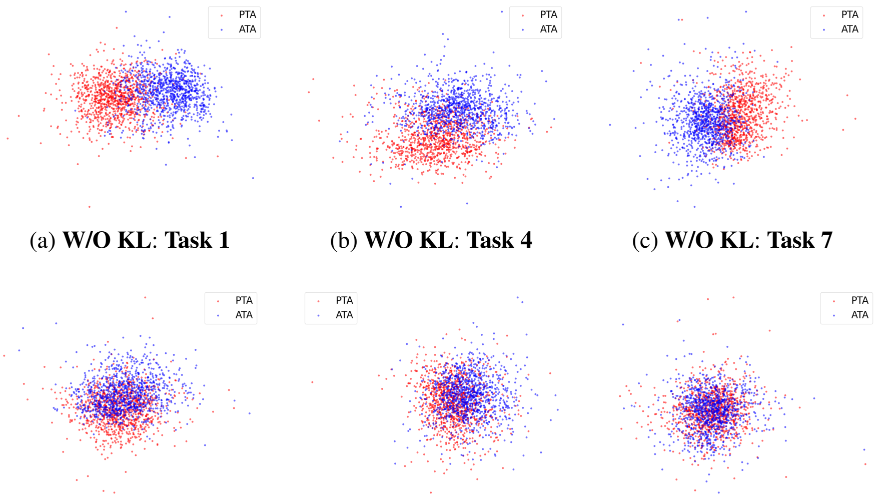
Results and Discussions Table 2 presents our comparative evaluation against state-of-the-art approaches for handling catastrophic interference on ImageNet-100 and CIFAR-100 benchmarks. Our **ICON** demonstrates superior performance across both datasets, achieving 88.91% and 87.07% average accuracy on CIFAR-100 and ImageNet-100, respectively.

On ImageNet-100, **ICON** surpasses the current state-of-the-art method (CLAP) by 1.15%, demonstrating the effectiveness of our identification-based framework on the real-world scenarios. Similarly, **ICON** outperforms the previous best method (CLAP) from 86.13% to 87.07%, representing a substantial margin improvement. The consistent performance improvements across both datasets validate the capability of **ICON** to handle catastrophic forgetting with respect to previous approaches. We trace this capability back to the advantage of identifying shared latent representations between PTA and ATA setup.

Figure 3 provide visual evidence of the effectiveness of **ICON** for catastrophic forgetting handling on the ImageNet-100 dataset. In the top rows (without KL divergence), we observe substantial mis-

Table 3: Average classification accuracy (%) of our ablation studies.

	ImageNet-100	CIFAR-100
w/o KL	74.85	75.94
ICON (ours)	88.91	87.07



(a) **W/O KL**: **Task 1** (b) **W/O KL**: **Task 4** (c) **W/O KL**: **Task 7**

(d) **W KL**: **Task 1** (e) **W KL**: **Task 4** (f) **W KL**: **Task 7**

Figure 4: Visualization of latent space distributions across Tasks 1, 4, and 7 on CIFAR-100 benchmark, comparing representations from partial-task aware (PTA, blue) and all-task aware (ATA, red) frameworks. The top row displays results without KL divergence optimization (Eq.14), revealing significant distributional misalignment that indicates catastrophic forgetting. The bottom row demonstrates our **ICON** approach with KL divergence optimization, exhibiting substantially improved alignment between PTA and ATA representations. Each subfigure displays 1,000 uniformly sampled points from the estimated latent representations $\hat{\bar{\mathbf{z}}}^t$ (PTA framework) and $\hat{\tilde{\mathbf{z}}}^t$ (ATA framework) for visualization clarity.

alignments between the PTA (blue) and ATA (red) representations. These misalignments indicates the existence of catastrophic forgetting as the distributions occupy distinctly separate regions of the latent space. In contrast, the bottom rows (with KL divergence optimization) exhibits remarkably improved alignments between the PTA and ATA representations across all three tasks. These alignments confirms the superior classification performance, as quantified in Table 2.

Ablation Studies In this section, we conduct ablation studies to asses the contribution of KL divergence for solving catastrophic forgetting on the ImageNet-100 and CIFAR-100 datasets.

We summarize the results of our ablation study in Table 3. The results demonstrate that our full ICON framework significantly outperforms the variant without KL divergence optimization across both benchmarks. The dramatic performance gap (88.91% versus 74.85% on ImageNet-100, and 87.07% against 75.94% on CIFAR-100) highlights the critical importance of our KL divergence optimization component, which explicitly maximize the shared latent variables $\mathbf{z^t}$ to the end of minimizing the catastrophic forgetting $\mathcal{F}$ in Eq. 3.

7 CONCLUSION

In this paper, we have presented a theoretical framework that characterizes catastrophic forgetting through the lens of identification theory. Upon our identifiability results, we establish a principled approach to mitigating the catastrophic forgetting challenge in continual learning. The empirical results on **ICON** validate our theoretical framework, demonstrating superior performance on both synthetic data and standard continual learning benchmarks.

REFERENCES

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

Elahe Arani, Fahad Sarfraz, and Bahram Zonooz. Learning fast, learning slow: A general continual learning method based on complementary learning system. In *International Conference on Learning Representations*, 2022. URL `https://openreview.net/forum?id=uxxFrDwrE7Y`.

Lorenzo Bonicelli, Matteo Boschini, Angelo Porrello, Concetto Spampinato, and Simone Calderara. On the effectiveness of lipschitz-driven rehearsal in continual learning. *Advances in Neural Information Processing Systems*, 35:31886–31901, 2022.

Lucas Caccia, Rahaf Aljundi, Nader Asadi, Tinne Tuytelaars, Joelle Pineau, and Eugene Belilovsky. New insights on reducing abrupt representation change in online continual learning. In *International Conference on Learning Representations*, 2022.

Sungmin Cha, Hsiang Hsu, Taebaek Hwang, Flavio Calmon, and Taesup Moon. Cpr: Classifier-projection regularization for continual learning. In *International Conference on Learning Representations*, 2021.

Arslan Chaudhry, Puneet K Dokania, Thalaiyasingam Ajanthan, and Philip HS Torr. Riemannian walk for incremental learning: Understanding forgetting and intransigence. In *Proceedings of the European conference on computer vision*, pp. 532–547, 2018.

Arslan Chaudhry, Naeemullah Khan, Puneet Dokania, and Philip Torr. Continual learning in low-rank orthogonal subspaces. *Advances in Neural Information Processing Systems*, 33:9900–9911, 2020.

Guangyi Chen, Yifan Shen, Zhenhao Chen, Xiangchen Song, Yuewen Sun, Weiran Yao, Xiao Liu, and Kun Zhang. Caring: Learning temporal causal representation under non-invertible generation process. *arXiv preprint arXiv:2401.14535*, 2024.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Mohammad Mahdi Derakhshani, Enrique Sanchez, Adrian Bulat, Victor G Turrisi da Costa, Cees GM Snoek, Georgios Tzimiropoulos, and Brais Martinez. Bayesian prompt learning for image-language model generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15237–15246, 2023.

Mehrdad Farajtabar, Navid Azizan, Alex Mott, and Ang Li. Orthogonal gradient descent for continual learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 3762–3773. PMLR, 2020.

Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision*, 132(2):581–595, 2024.

Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pp. 297–304. JMLR Workshop and Conference Proceedings, 2010.

Christian Henning, Maria Cervera, Francesco D'Angelo, Johannes Von Oswald, Regina Traber, Benjamin Ehret, Seijin Kobayashi, Benjamin F Grewe, and João Sacramento. Posterior meta-replay for continual learning. *Advances in Neural Information Processing Systems*, 34:14135–14149, 2021.

Saihui Hou, Xinyu Pan, Chen Change Loy, Zilei Wang, and Dahua Lin. Learning a unified classifier incrementally via rebalancing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 831–839, 2019.

Ching-Yi Hung, Cheng-Hao Tu, Cheng-En Wu, Chien-Hung Chen, Yi-Ming Chan, and Chu-Song Chen. Compacting, picking and growing for unforgetting continual learning. *Advances in Neural Information Processing Systems*, 32, 2019.

Saurav Jha, Dong Gong, and Lina Yao. CLAP4CLIP: Continual learning with probabilistic finetuning for vision-language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL `https://openreview.net/forum?id=rF1YRtZfoJ`.

Ta-Chu Kao, Kristopher Jensen, Gido van de Ven, Alberto Bernacchia, and Guillaume Hennequin. Natural continual learning: success is a journey, not (just) a destination. *Advances in neural information processing systems*, 34:28067–28079, 2021.

Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 19113–19122, 2023.

Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvarinen. Variational autoencoders and nonlinear ica: A unifying framework. In *International conference on artificial intelligence and statistics*, pp. 2207–2217. PMLR, 2020.

Lingjing Kong, Shaoan Xie, Weiran Yao, Yujia Zheng, Guangyi Chen, Petar Stojanov, Victor Akinwande, and Kun Zhang. Partial disentanglement for domain adaptation. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 11455–11472. PMLR, 17–23 Jul 2022.

Lingjing Kong, Guangyi Chen, Petar Stojanov, Haoxuan Li, Eric P. Xing, and Kun Zhang. Towards understanding extrapolation: a causal lens. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL `https://openreview.net/forum?id=2squ766Iq4`.

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

Xilai Li, Yingbo Zhou, Tianfu Wu, Richard Socher, and Caiming Xiong. Learn to grow: A continual structure learning framework for overcoming catastrophic forgetting. In *International Conference on Machine Learning*, pp. 3925–3934. PMLR, 2019.

Yuke Li, Guangyi Chen, Ben Abramowitz, Stefano Anzellotti, and Donglai Wei. Learning causal domain-invariant temporal dynamics for few-shot action recognition. In *Forty-first International Conference on Machine Learning*, 2024. URL `https://openreview.net/forum?id=LvuuYqU0BW`.

Zijian Li, Ruichu Cai, Guangyi Chen, Boyang Sun, Zhifeng Hao, and Kun Zhang. Subspace identification for multi-source domain adaptation. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL `https://openreview.net/forum?id=BACQLWQW8u`.

Yan-Shuo Liang and Wu-Jun Li. Loss decoupling for task-agnostic continual learning. *Advances in Neural Information Processing Systems*, 36, 2023.

Sen Lin, Li Yang, Deliang Fan, and Junshan Zhang. Trgp: Trust region gradient projection for continual learning. In *International Conference on Learning Representations*, 2022.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Arun Mallya and Svetlana Lazebnik. Packnet: Adding multiple tasks to a single network by iterative pruning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 7765–7773, 2018.

Hiroshi Morioka and Aapo Hyvarinen. Causal representation learning made identifiable by grouping of observational variables. In *Forty-first International Conference on Machine Learning*, 2024.

Pingbo Pan, Siddharth Swaroop, Alexander Immer, Runa Eschenhagen, Richard Turner, and Mohammad Emtiyaz E Khan. Continual deep learning by functional regularisation of memorable past. *Advances in Neural Information Processing Systems*, 33:4453–4464, 2020.

Jingyang Qiao, Xin Tan, Chengwei Chen, Yanyun Qu, Yong Peng, Yuan Xie, et al. Prompt gradient projection for continual learning. In *The Twelfth International Conference on Learning Representations*, 2024.

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

Tim GJ Rudner, Freddie Bickford Smith, Qixuan Feng, Yee Whye Teh, and Yarin Gal. Continual learning via sequential function-space variational inference. In *International Conference on Machine Learning*, pp. 18871–18887. PMLR, 2022.

Gobinda Saha, Isha Garg, and Kaushik Roy. Gradient projection memory for continual learning. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=3AOj0RCNC2`.

Fahad Sarfraz, Elahe Arani, and Bahram Zonooz. Error sensitivity modulation based experience replay: Mitigating abrupt representation drift in continual learning. In *The Eleventh International Conference on Learning Representations*, 2023.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021. doi: 10.1109/JPROC.2021.3058954.

Joan Serra, Didac Suris, Marius Miron, and Alexandros Karatzoglou. Overcoming catastrophic forgetting with hard attention to the task. In *International conference on machine learning*, pp. 4548–4557. PMLR, 2018.

Xiangchen Song, Weiran Yao, Yewen Fan, Xinshuai Dong, Guangyi Chen, Juan Carlos Niebles, Eric Xing, and Kun Zhang. Temporally disentangled representation learning under unknown nonstationarity. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Peter Sorrenson, Carsten Rother, and Ullrich Köthe. Disentanglement by nonlinear ica with general incompressible-flow networks (gin). In *International Conference on Learning Representations*, 2020. URL `https://openreview.net/forum?id=rygeHgSFDH`.

Vishal Thengane, Salman Khan, Munawar Hayat, and Fahad Khan. Clip model is an efficient continual learner. *arXiv preprint arXiv:2210.03114*, 2022.

Michalis K. Titsias, Jonathan Schwarz, Alexander G. de G. Matthews, Razvan Pascanu, and Yee Whye Teh. Functional regularisation for continual learning with gaussian processes. In *International Conference on Learning Representations*, 2020. URL `https://openreview.net/forum?id=HkxCzeHFDB`.

Runqi Wang, Xiaoyue Duan, Guoliang Kang, Jianzhuang Liu, Shaohui Lin, Songcen Xu, Jinhu Lü, and Baochang Zhang. 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, 2023a.

Shipeng Wang, Xiaorong Li, Jian Sun, and Zongben Xu. Training networks in null space of feature covariance for continual learning. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pp. 184–193, 2021.

Zhenyi Wang, Enneng Yang, Li Shen, and Heng Huang. A comprehensive survey of forgetting in deep learning beyond continual learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

Zifeng Wang, Zheng Zhan, Yifan Gong, Yucai Shao, Stratis Ioannidis, Yanzhi Wang, and Jennifer Dy. Dualhsic: Hsic-bottleneck and alignment for continual learning. In *International Conference on Machine Learning*, 2023b.

Mingqing Xiao, Qingyan Meng, Zongpeng Zhang, Di He, and Zhouchen Lin. Hebbian learning based orthogonal projection for continual learning of spiking neural networks. In *The Twelfth International Conference on Learning Representations*, 2024.

Dingling Yao, Danru Xu, Sebastien Lachapelle, Sara Magliacane, Perouz Taslakian, Georg Martius, Julius von Kügelgen, and Francesco Locatello. Multi-view causal representation learning with partial observability. In *The Twelfth International Conference on Learning Representations*, 2024.

Yujia Zheng and Kun Zhang. Generalizing nonlinear ica beyond structural sparsity. *Advances in Neural Information Processing Systems*, 36, 2024.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022.

A Technical Appendices and Supplementary Material

A.1 Proof of Theorem 4.1

Proof: We proceed our proof by contradiction, focusing on the latent space $\tilde{\mathbf{z}}^{\mathbf{t}}$ in the ATA setting. Suppose that $\mathbf{z}^{\mathbf{t}} \in \mathcal{Z}^{\mathbf{t}}$ simultaneously resides on two distinct manifolds $\tilde{\mathcal{Z}}_1^{\mathbf{t}}$ and $\tilde{\mathcal{Z}}_2^{\mathbf{t}}$. Under this assumption, there exist points $\tilde{\mathbf{z}}_1^{\mathbf{t}} \in \tilde{\mathcal{Z}}_1^{\mathbf{t}}$ and $\tilde{\mathbf{z}}_2^{\mathbf{t}} \in \tilde{\mathcal{Z}}_2^{\mathbf{t}}$ such that we can establish the following:

$$g(\tilde{\mathbf{z}}_1^{\mathbf{t}}) - g(\mathbf{z}^{\mathbf{t}}) = \left(\int_0^1 J_g(\lambda \mathbf{z}^{\mathbf{t}} + (1-\lambda)\tilde{\mathbf{z}}_1^{\mathbf{t}}) d\lambda \right) h_1 \tag{15}$$

$$g(\tilde{\mathbf{z}}_2^{\mathbf{t}}) - g(\mathbf{z}^{\mathbf{t}}) = \left(\int_0^1 J_g(\lambda \mathbf{z}^{\mathbf{t}} + (1-\lambda)\tilde{\mathbf{z}}_2^{\mathbf{t}}) d\lambda \right) h_2 \tag{16}$$

where $h_1 = \tilde{\mathbf{z}}_1^{\mathbf{t}} - \mathbf{z}^{\mathbf{t}}$, $h_2 = \tilde{\mathbf{z}}_2^{\mathbf{t}} - \mathbf{z}^{\mathbf{t}}$. The L.H.S of Eq.15 and 16 uses the fact that the shared observation can be mapped through either g or $g^{\mathbf{t}}$ from $\mathbf{z}^{\mathbf{t}}$.

We take the substraction of Eq.15 and Eq.16:

$$\begin{aligned} g(\tilde{\mathbf{z}}_1^{\mathbf{t}}) - g(\tilde{\mathbf{z}}_2^{\mathbf{t}}) = &\left(\int_0^1 J_g(\lambda \mathbf{z}^{\mathbf{t}} + (1-\lambda)\tilde{\mathbf{z}}_1^{\mathbf{t}}) d\lambda \right) h_1 \\ &- \left(\int_0^1 J_g(\lambda \mathbf{z}^{\mathbf{t}} + (1-\lambda)\tilde{\mathbf{z}}_2^{\mathbf{t}}) d\lambda \right) h_2 \end{aligned} \tag{17}$$

Let us denote $\Lambda_1 = \left(\int_0^1 J_g(\lambda \mathbf{z}^{\mathbf{t}} + (1-\lambda)\tilde{\mathbf{z}}_1^{\mathbf{t}}) d\lambda \right) h_1$, and $\Lambda_2 = \left(\int_0^1 J_g(\lambda \mathbf{z}^{\mathbf{t}} + (1-\lambda)\tilde{\mathbf{z}}_2^{\mathbf{t}}) d\lambda \right) h_2$. Eq.17 implies that:

$$\begin{aligned} &||\Lambda_2 - \Lambda_1|| \geq \mathcal{D}(\tilde{\mathcal{Z}}_1^{\mathbf{t}}, \tilde{\mathcal{Z}}_2^{\mathbf{t}}) \\ \implies &||\Lambda_2|| + ||\Lambda_1|| \geq \mathcal{D}(\tilde{\mathcal{Z}}_1^{\mathbf{t}}, \tilde{\mathcal{Z}}_2^{\mathbf{t}}) \\ \implies &J_{\tilde{\mathbf{u}}}(||h_2|| + ||h_1||) \geq \mathcal{D}(\tilde{\mathcal{Z}}_1^{\mathbf{t}}, \tilde{\mathcal{Z}}_2^{\mathbf{t}}) \\ \implies &\max(||h_2||, ||h_1||) \geq \frac{\mathcal{D}(\tilde{\mathcal{Z}}_1^{\mathbf{t}}, \tilde{\mathcal{Z}}_2^{\mathbf{t}})}{2J_{\tilde{\mathbf{u}}}} \end{aligned} \tag{18}$$

where $J_{\tilde{\mathbf{u}}}$ denotes the spectral norm of the Jacobian matrices $\left(J_g(\tilde{\mathbf{z}}_1^{\mathbf{t}}), J_g(\tilde{\mathbf{z}}_2^{\mathbf{t}})\right)$. This directly contradicts our assumption that $\mathcal{D}(\tilde{\mathbf{z}}^{\mathbf{t}}, \mathbf{z}^{\mathbf{t}}) \leq \frac{\mathcal{D}(\tilde{\mathcal{Z}}_1^{\mathbf{t}}, \tilde{\mathcal{Z}}2^{\mathbf{t}})}{2J\tilde{\mathbf{u}}}$. Therefore, $\mathbf{z}^{\mathbf{t}}$ can only be explained by a single manifold within $\tilde{\mathcal{Z}}^{\mathbf{t}}$.

Given the injectiveness of g in Eq.2, the correct estimate $\hat{\tilde{\mathbf{z}}}^{\mathbf{t}}$ is feasible for $\tilde{\mathbf{z}}^{\mathbf{t}}$ as Eq.7 suggests. Utilizing an analogous constraint to Assumption 5, we can bound the distance between the estimated latent variables as follows: $\mathcal{D}(\hat{\tilde{\mathbf{z}}}^{\mathbf{t}}, \hat{\mathbf{z}}^{\mathbf{t}}) \leq \frac{\mathcal{D}(\hat{\tilde{\mathcal{Z}}}_1^{\mathbf{t}}, \hat{\tilde{\mathcal{Z}}}_2^{\mathbf{t}}))}{2J_{\hat{\mathbf{u}}}}$. $J_{\hat{\mathbf{u}}}$ denotes the spectrum norm of $\left(J_{\hat{g}}(\hat{\tilde{\mathbf{z}}}_1^{\mathbf{t}}), J_g(\hat{\tilde{\mathbf{z}}}_2^{\mathbf{t}})\right)$.

Extending our contradiction statement, let us denote the difference $\hat{\tilde{\mathbf{z}}}_1^{\mathbf{t}} - \hat{\mathbf{z}}^{\mathbf{t}}$ as $\hat{h}$. Any incorrect estimate $\hat{\tilde{\mathbf{z}}}^{\mathbf{t}}$ would lead to:

$$||\hat{h}|| \geq \frac{\mathcal{D}(\hat{\tilde{\mathcal{Z}}}_1^{\mathbf{t}}, \hat{\tilde{\mathcal{Z}}}_2^{\mathbf{t}}))}{2J_{\hat{\mathbf{u}}}} \tag{19}$$

This directly contradicts our established constraint assumption. Therefore, such incorrect estimates are excluded from the feasible solution space.

B Implementation Details

We detail the network architectures for both synthetic and real-world experiments in Section 6.1 and 6.2, respectively.

For both synthetic and real-world experiments, we optimize our network using AdamW (Loshchilov & Hutter, 2019) with an initial learning rate of 0.002 and weight decay of 10^{-2}, employing a cosine annealing schedule for learning rate decay. For the contrastive learning objective, we set the temperature parameter $\tau = 0.07$ in the NCE loss $\mathcal{L}^{\mathbf{c}}$. We implement our framework in PyTorch and conduct all experiments on a single NVIDIA GeForce RTX 3090 GPU with 24GB memory.

C THE USE OF LARGE LANGUAGE MODELS (LLMS)

We use LLMs to detect and correct grammatical errors throughout the manuscript.