TASK-AGNOSTIC EXEMPLAR-FREE DOMAIN-INCREMENTAL LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Autonomous Domain-Incremental Learning (DIL) requires models to sequentially acquire new visual domains without explicit task identifiers, exemplar storage or catastrophic forgetting. Current methodologies face a fundamental trilemma: replay-based methods compromise data privacy, regularization restricts model plasticity and retrieval-based routing degrades as latent representations drift over time. We introduce **Task-Agnostic Continual Training via Independent Components (TACTIC)**, an exemplar-free framework that resolves this trilemma through a unified geometric principle: constraining all learned representations to a shared unit hypersphere $\mathbb{S}^{d-1}$. TACTIC operationalizes this principle via three synergistic mechanisms. First, a sub-center geometric head allocates multiple angular anchors per class, providing the intra-class flexibility necessary to mathematically prevent the semantic collapse endemic to shared classifiers. Second, a feature disentanglement objective explicitly separates domain-specific style from universal semantics. By enforcing statistical independence (HSIC) between these representations, TACTIC guarantees stable, drift-free routing at inference without relying on task metadata. Third, a hyperspherical analytic memory synthesizes per-class Gaussian statistics and projects them back onto $\mathbb{S}^{d-1}$, maintaining manifold compatibility while supplying boundary-anchoring hard negatives. Theoretically, we establish intra-class interference bounds that guarantee the geometric safety of our analytic replay. Empirically, TACTIC establishes a new state-of-the-art exemplar-free DIL performance across diverse benchmarks (DomainNet, Office-Home, CORe50 and Camelyon17). Crucially, by isolating parameters and eliminating cross-domain gradient interference, TACTIC outperforms existing methodologies, all while maintaining a minimal storage footprint of less than 1.5 MB.

1 INTRODUCTION

The fundamental challenge of Continual Learning (CL) is overcoming catastrophic forgetting when a neural network is exposed to a sequential stream of data distributions (McCloskey & Cohen, 1989; French, 1999; Parisi et al., 2019). Recent attention has shifted toward Domain-Incremental Learning (DIL) (van de Ven et al., 2022; Wang et al., 2024), arguably the most realistic and rigorous continual setting. In DIL, a model must autonomously navigate sequential domain shifts (e.g., transitioning from photographs to sketches or paintings) without explicit task identifiers provided during inference.

Currently, the most accurate CL methods rely on Experience Replay (ER), storing thousands of raw exemplars from past domains (Rebuffi et al., 2017; Buzzega et al., 2020; Boschini et al., 2022). However, this approach is fundamentally impractical in critical real-world deployments. In digital medicine, strict data privacy laws (Kaissis et al., 2020), such as HIPAA or GDPR, prohibit saving patient data (Bandi et al., 2018; Koh et al., 2021) and in continuous robotics streams, storing endless high-resolution frames easily exceeds edge memory budgets. Consequently, the field urgently demands architectures capable of *Exemplar-Free Domain-Incremental Learning*.

However, current exemplar-free methodologies face a fundamental trilemma. First, pure regularization-based methods overly restrict model plasticity, failing to capture complex domain shifts. Second, modern parameter-isolation and prompt-based methods (Wang et al., 2022c;b; Smith et al., 2023; Wang et al., 2023; McDonnell et al., 2023) successfully preserve plasticity by freezing the backbone and assigning task-specific parameters. However, without task metadata at the test time,

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

they rely on heuristic feature matching to strictly route inputs to a single prompt. As latent spaces drift across sequential domains, this hard-routing mechanism rapidly collapses. Finally, attempts to probabilistically route features often rely on entangled feature spaces where semantic object identity blurs with visual style, causing the router to misclassify domains when overlapping semantics appear. Furthermore, traditional classifiers allocate a single weight vector per class. When a class is observed in contradictory styles, (e.g., a real photo and a sketch of the same object), the centroid is forced toward a mean, eroding previously learned decision boundaries (Deng et al., 2020; Bonicelli et al., 2022).

To resolve this trilemma, we introduce **Task-Agnostic Continual Training via Independent Components (TACTIC)**. TACTIC abandons reliance on heuristic distance metrics and physical data buffers in favor of a unified geometric and probabilistic framework. At its core, TACTIC enforces a single unifying principle: *every learned representation lives on the unit hypersphere* $\mathbb{S}^{d-1}$ *and every operation respects that geometry*. To prevent the semantic collapse endemic to shared classifiers, we allocate multiple sub-center angular anchors per class, providing the intra-class flexibility necessary to accommodate diverse visual styles without forcing centroids to collapse.

To achieve task-agnostic inference, TACTIC isolates domain knowledge into dynamic, low-rank experts, AdaLoRA (Zhang et al., 2023b). Crucially, we introduce a feature disentanglement objective using the Hilbert-Schmidt Independence Criterion (HSIC) (Gretton et al., 2005; Ma et al., 2020) to explicitly separate domain-specific style from universal semantics. We route these statistically independent style features using a novel Top-K Soft Routing (Jacobs et al., 1991; Shazeer et al., 2017) Engine. By fitting Gaussian Mixture Models (GMMs) (Dempster et al., 1977; Bishop, 2006) using the Bayesian Information Criterion (BIC) (Schwarz, 1978; Bishop, 2006) and blending expert predictions via a temperature-scaled softmax, we eliminate the rigid failure modes of hard routing, allowing the model to smoothly ensemble experts when evaluating ambiguous or overlapping domains. Finally, to prevent boundary erosion without saving physical images, TACTIC employs a Hyperspherical Analytic Memory.

Our **main contributions** are summarized as follows:

Anti-Collapse Hyperspherical Geometry: We propose a sub-center cosine classifier coupled with parameter-efficient AdaLoRA experts. This architecture mathematically prevents semantic collapse by distributing gradients across multiple angular anchors, allowing the model to simultaneously memorize contradictory visual styles without eroding historical decision boundaries.

HSIC Disentanglement & Soft Routing: We introduce a novel routing framework that is the first to frame continual routing failure as a feature entanglement problem. Unlike standard parameter-efficient methods that route in entangled spaces, by introducing our dual-forward training scheme constrained by the HSIC, we explicitly force statistical independence between task and style spaces. At inference, our routing engine employs BIC-scaled GMMs and top-K soft activation to smoothly blend experts, bypassing the rigid, drift-prone hard selections of prior prompt-based works.

Hyperspherical Analytic Replay with Theoretical Guarantees: We design a novel exemplar-free memory that synthesizes Gaussian features and explicitly projects them back onto $\mathbb{S}^{d-1}$. Theoretically, we establish intra-class interference bounds and provide an operational Wasserstein-1 bound on the gradient gap between true and analytic replay, guaranteeing geometric safety.

State-of-the-Art Empirical Performance: TACTIC establishes a formidable new state-of-the-art (sota) performance for exemplar-free continual learning across four diverse benchmarks (DomainNet, Office-Home, CORe50, Camelyon17). By isolating parameters and eliminating cross-domain gradient interference, TACTIC comprehensively outperforms existing prompt-based and parameter-efficient methodologies, while maintaining a minimal storage footprint of less than 1.5 MB.

2 Related Work

The CL literature encompasses regularization, replay and dynamic architectures (Parisi et al., 2019; De Lange et al., 2021; Wang et al., 2024). TACTIC integrates principles from metric learning, feature disentanglement and contrastive stabilization.

Regularization & Replay. Regularization methods (e.g., EWC (Kirkpatrick et al., 2017), LwF (Li & Hoiem, 2017), RWalk (Chaudhry et al., 2018)) apply soft penalties to weights or logits. While

exemplar-free, their capacity saturates over long sequences, failing under extreme visual shifts. Conversely, exemplar-replay architectures (Rebuffi et al., 2017; Buzzega et al., 2020; Boschini et al., 2022) store raw images from past tasks. Although empirically strong, they are legally prohibited in privacy-restricted domains (Bandi et al., 2018; Koh et al., 2021). Feature-replay methods mitigate storage costs by retaining per-class statistics instead of raw images (Hayes & Kanan, 2020; Petit et al., 2023). More recently, DCE (Li et al., 2025) balances domain experts using historical Gaussian statistics. However, these methods typically sample Gaussian distributions in unbounded Euclidean space $\mathbb{R}^d$ and pass them to L2-normalized classifiers, an off-manifold geometric mismatch. TACTIC resolves this by explicitly projecting analytic Gaussian samples back onto $\mathbb{S}^{d-1}$ before classification, ensuring topological consistency (§3.5).

Prompting & Parameter-Efficient Fine-Tuning (PEFT). To prevent forgetting without exemplars, recent approaches freeze a pre-trained ViT (Dosovitskiy et al., 2021) and learn domain-specific parameters via prompting (e.g., L2P (Wang et al., 2022c), DualPrompt (Wang et al., 2022b), S-Prompts (Wang et al., 2022a), CODA-Prompt (Smith et al., 2023)) or Low-Rank Adaptation (LoRA) (e.g., SLCA (Zhang et al., 2023a), C-LoRA (Wistuba et al., 2023)). However, these methods fundamentally rely on entangled representation spaces for expert routing. In the strict DIL paradigm, where visual styles shift drastically, this entanglement triggers heuristic routing collapse. While recent architectures explore alternative prompt alignments (DUCT (Zhou et al., 2025), KA-Prompt (Xu et al., 2025)), our empirical evaluation anchors against foundational DIL methods and SOYO (Wang et al., 2025). SOYO serves as our primary contemporaneous baseline because its LoRA-based GMM routing is the closest architectural counterpart to our system. However, while sharing similar primitives, TACTIC introduces a fundamentally novel geometric framework. Specifically, SOYO's reliance on hard-max routing within entangled feature spaces leaves it highly vulnerable to domain drift. TACTIC overcomes this by isolating parameters via adaptive-rank AdaLoRA (Zhang et al., 2023b) and employing a Top-K Soft Routing Engine on an novel HSIC-disentangled style manifold, ensuring mathematically stable expert ensembling even when domains boundaries blur.

Hyperspherical Geometry & Disentanglement. L2-normalized embeddings on $\mathbb{S}^{d-1}$ are widely used in metric learning (Wang et al., 2018; Deng et al., 2019). While sub-center geometry (Deng et al., 2020) was introduced to handle noisy face identities, we are the first to adapt it as a geometric defense against intra-class domain interference in CL, coupling it with an explicit anti-collapse LogSumExp (Boyd & Vandenberghe, 2004) aggregation. To ensure drift-free routing, we enforce statistical independence between latent subspaces using the HSIC (Gretton et al., 2005; Bardes et al., 2022). Crucially, while previous parameter-isolated systems implicitly tolerate a train/test distribution mismatch, TACTIC computes the HSIC disentanglement loss strictly on the LoRA-disabled features that the router will evaluate at inference, guaranteeing perfect alignment.

Stabilization Techniques. Finally, contrastive hard-negative sampling (Robinson et al., 2021; Kalantidis et al., 2020) is known to sharpen decision boundaries. We adapt this principle to analytic replay by inflating the sampling variance multiplier $\gamma > 1$ around per-class means, generating synthetic features at the periphery of historical class distributions to act as boundary-anchoring negatives. Furthermore, to stabilize the style representation space, we implement a strict geometric anchor. Rather than allowing the projection heads to drift or relying on slow-moving momentum encoders (He et al., 2020), TACTIC trains the style projection head exclusively on the first domain and permanently freezes it. This guarantees that the topological manifold the routing engine evaluates remains absolutely stationary for all subsequent domains.

3 METHODOLOGY

TACTIC decomposes the DIL trilemma into orthogonal sub-problems, resolving each through a unified geometric constraint. The overall architecture is illustrated in Figure 1.

3.1 PROBLEM FORMULATION & MANIFOLD COMMITMENT

Let $\mathcal{S} = \{\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_T\}$ be a sequential stream of domains. Each domain provides data $(x, y) \sim P_t(X, Y)$, where $x \in \mathcal{X}$ is an image and $y \in \mathcal{Y} = \{1, \ldots, C\}$ is the class label. In DIL, the label space $\mathcal{Y}$ is shared across all domains, but the marginal distribution $P_t(X)$ undergoes severe sequential shifts (e.g., from photographs to sketches). During training on $\mathcal{D}_t$, the model cannot access $\mathcal{D}_{<t}$. At

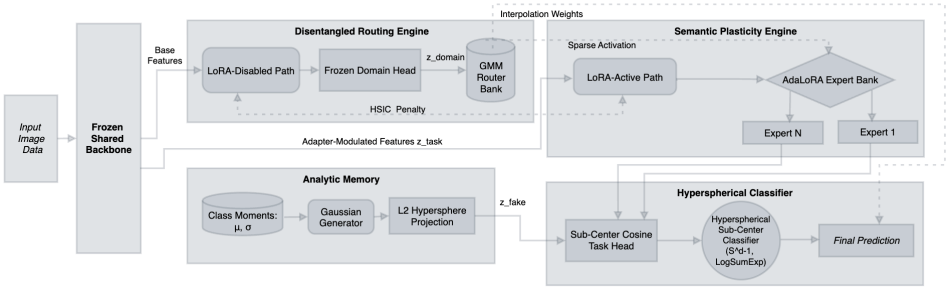


Figure 1: *The TACTIC Architecture.* The frozen shared backbone processes inputs through a Dual-Forward pass. *Top (Disentangled Routing):* The LoRA-disabled path extracts a pure style feature ($z^{\text{off}}_{\text{domain}}$) which the frozen domain head projects for Top-K GMM routing. *Right (Semantic Plasticity):* The LoRA-active path routes semantic features ($z^{\text{on}}_{\text{task}}$) through AdaLoRA experts. The HSIC penalty enforces statistical independence between these two paths. *Bottom (Analytic Memory & Classifier):* Synthetic hard-negatives are generated from historical moments and explicitly projected onto the L2 Hypersphere ($\mathbb{S}^{d-1}$), stabilizing the Sub-Center LogSumExp classifier against catastrophic forgetting without saving physical exemplars.

inference, the model receives inputs x without any domain identifier t. The objective is to construct an optimal predictor $\hat{y} = f(x)$ that maximizes accuracy across all $\mathcal{D}_{1\ldots T}$ under a strict exemplar-free constraint. To ensure topological consistency across all modules, TACTIC constrains all learned representations to the unit hypersphere $\mathbb{S}^{d-1} = \{z \in \mathbb{R}^d \mid \|z\|_2 = 1\}$. Every memory primitive, classification operator and analytic sampling distribution either lives natively on $\mathbb{S}^{d-1}$ or is explicitly projected onto it.

3.2 Backbone, Parallel Projection Heads & AdaLoRA Experts

We utilize a frozen Vision Transformer (ViT-B/16) (Dosovitskiy et al., 2021) backbone f_θ, pre-trained on ImageNet-1K (Deng et al., 2009). Freezing θ prevents catastrophic interference at the representational core. To extract specialized features from the [CLS] token $h \in \mathbb{R}^{768}$, we instantiate two parallel Multi-Layer Perceptron (MLP) (Gardner & Dorling, 1998) projection heads, g_{task} and g_{domain} (shown as the LoRA-Active and LoRA-Disabled paths in Fig. 1), mapping to $\mathbb{R}^d$ ($d = 512$). The outputs are L2-normalized:

$$z_{\text{task}} = \frac{g_{\text{task}}(h)}{\|g_{\text{task}}(h)\|_2}, \qquad z_{\text{domain}} = \frac{g_{\text{domain}}(h)}{\|g_{\text{domain}}(h)\|_2} \quad \in \mathbb{S}^{d-1} \tag{1}$$

The task feature z_{task} feeds the Sub-Center (Deng et al., 2020) classification head, while z_{domain} feeds the routing engine. To prevent routing manifold drift over sequential tasks, g_{domain} is trained exclusively alongside $\mathcal{D}_1$ and permanently frozen for $t \geq 2$.

To assimilate domain-specific knowledge, we inject an AdaLoRA (Zhang et al., 2023b) expert ϕ_t into the self-attention and MLP blocks of f_θ for each domain t. AdaLoRA parameterizes the weight matrices as $W = W_0 + P\Lambda Q^\top$, where P and Q are low-rank spatial vectors and Λ is a diagonal matrix of singular values. The rank budget is dynamically pruned during training to maximize parameter efficiency. When a new expert ϕ_t is added, all previously trained experts $\phi_{<t}$ are strictly frozen.

3.3 Dual-Forward HSIC Disentanglement

A critical distribution mismatch affects LoRA-based continual learning. During training, the active expert ϕ_t is engaged, producing features from a joint distribution P^{on}_t. However, at inference, the routing engine must classify the input *before* selecting an expert, operating under the LoRA-disabled distribution P^{off}. Prior architectures implicitly ignore this discrepancy, causing the routing topology to drift and collapse.

We formally resolve this via a *Dual-Forward Training Pass*: *Pass 1 (LoRA-on):* Engage ϕ_t to compute $z^{\text{on}}_{\text{task}}$. This feature optimizes the classification and supervised contrastive losses, encoding domain-

specific semantics. *Pass 2 (LoRA-off):* Disable all adapters to compute $z_{\text{domain}}^{\text{off}}$. This is the pure, backbone-derived style feature the router will ingest at deployment.

To guarantee that the routing feature $z_{\text{domain}}^{\text{off}}$ is strictly independent of the semantic class identity embedded in $z_{\text{task}}^{\text{on}}$, we minimize the HSIC (Gretton et al., 2005) alongside a variance-covariance regularization inspired by VICReg (Bardes et al., 2022):

$$\mathcal{L}_{\text{indep}} = \beta\mathcal{L}_{\text{var}}(z_{\text{task}}^{\text{on}}, z_{\text{domain}}^{\text{off}}) + \alpha_1\mathcal{L}_{\text{lin}}(z_{\text{task}}^{\text{on}}, z_{\text{domain}}^{\text{off}}) + \alpha_2\mathcal{L}_{\text{HSIC}}(z_{\text{task}}^{\text{on}}, z_{\text{domain}}^{\text{off}}) \tag{2}$$

By setting $\alpha_1 = 1.0, \alpha_2 = 1.0, \beta = 1.0$ and evaluating directly across the train/test distribution gap, TACTIC extracts an operationally drift-free routing space. By minimizing $\mathcal{L}_{\text{HSIC}}$, TACTIC forces the cross-covariance between the semantic and style spaces to approach zero. We formalize why this guarantees a robust, label-agnostic routing space:

Proposition 1 (Label-Agnostic Routing Guarantee) *Let Y be the discrete class label. Assume the active expert z_{task}^{on} is fully discriminative, such that the mutual information $I(z_{task}^{on}; Y) \approx H(Y)$. If the HSIC penalty drives the style representation to be statistically independent, i.e., $z_{domain}^{off} \perp\!\!\!\perp z_{task}^{on}$, then by the Data Processing Inequality, the mutual information $I(z_{domain}^{off}; Y) \to 0$. (Proof provided in Appendix A.3.)*

This proposition guarantees that the GMM (Dempster et al., 1977) routing engine fits exclusively on pure visual style (e.g., the stroke texture of a sketch) rather than exploiting spurious semantic correlations (e.g., the presence of a dog), preventing the catastrophic routing collapse observed in prior prompt-based methods when classes are imbalanced across domains.

3.4 Anti-Collapse Sub-Center Geometry

Under severe domain shifts, projecting contradictory visual styles (e.g., photos vs. sketches) onto a single class centroid causes boundary erosion. We alleviate this by allocating $K = 3$ sub-centers $\{W_{c,k}\}_{k=1}^{K} \in \mathbb{S}^{d-1}$ per class c.

We aggregate the similarity between z_{task} and the sub-centers using a LogSumExp (LSE) (Boyd & Vandenberghe, 2004) function, scaled by a hyperparameter $\tau = 0.1$ and a fixed spatial scale $\sigma = 10.0$:

$$s_c(z_{\text{task}}) = \sigma \cdot \tau \log \sum_{k=1}^{K} \exp\left(\frac{z_{\text{task}}^{\top} W_{c,k}}{\tau}\right) \tag{3}$$

Unlike standard hard-max aggregation $\max_k(z^{\top} W_{c,k})$ (Deng et al., 2020), which routes the entire gradient to the single closest sub-center (causing the remaining centers to starve and monotonically collapse), our LSE aggregator distributes proportional gradient flow to all sub-centers simultaneously. We formalize this geometric advantage:

Theorem 1 (Anti-Collapse Interference Bound) *Let class c be supported by a mixture of two domain distributions with means $\mu_1, \mu_2 \in \mathbb{S}^{d-1}$ separated by cosine similarity $\rho \in [0, 1)$. Single-Centroid ($K = 1$): The irreducible interference penalty is $\Delta_1(\rho) = \sigma(1 - \sqrt{(1+\rho)/2}) > 0$, causing unavoidable semantic collapse. Sub-Center ($K \geq 2$, LSE): The irreducible penalty is bounded by $\Delta_K(\rho, \tau) \leq \sigma\tau \log K$, which smoothly approaches 0 as $\tau \to 0$, independent of ρ. (Proof provided in Appendix A.1.)*

This theorem dictates our architectural choice: by setting $\tau = 0.1$ with $K = 3$, TACTIC mathematically guarantees the preservation of intra-class stylistic diversity without collapsing the underlying semantic boundaries.

3.5 Hyperspherical Analytic Replay with Hard-Negative Defense

To stabilize decision boundaries without violating zero-exemplar constraints, TACTIC generates synthetic boundary features on $\mathbb{S}^{d-1}$. Following the training of $\mathcal{D}_t$, we compute the empirical mean $\bar{z}_c \in \mathbb{R}^d$ and variance $s_c^2 \in \mathbb{R}^d$ for each class c, updating the analytic memory via an Exponential Moving Average (EMA, $m = 0.5$) (He et al., 2020).

During subsequent training steps, we draw synthetic features for historically observed classes:

$$\tilde{z}_c \sim \mathcal{N}\left(\hat{\mu}_c,\, \gamma^2 \text{diag}(\hat{\sigma}_c^2)\right), \qquad z_c^{\text{fake}} = \frac{\tilde{z}_c}{\|\tilde{z}_c\|_2} \tag{4}$$

The variance multiplier $\gamma = 1.5$ acts as a Hard-Negative Boundary Defense (Robinson et al., 2021), pushing synthetic samples toward the periphery of the historical distribution. The critical operation is the explicit L2 projection to $\mathbb{S}^{d-1}$. We prove that this geometric projection bounds the gradient drift:

Theorem 2 (Analytic Replay Gradient Stability) *Let p_c be the true empirical distribution of class c on $\mathbb{S}^{d-1}$ and q_c be our synthetic projected-Gaussian distribution. For any L-Lipschitz loss gradient, the gradient error is bounded by the Wasserstein distance (Proof provided in Appendix A.2.):*

$$\|\mathbb{E}_{z\sim p_c}[\nabla_\theta \ell] - \mathbb{E}_{z\sim q_c}[\nabla_\theta \ell]\|_2 \leq L \cdot \mathcal{W}_1(p_c, q_c) \approx \mathcal{O}\left(\sqrt{\gamma^2 \mathrm{tr}(\hat{\Sigma}_c)}\right)$$

This theoretical bound justifies two core components of our implementation: (1) the explicit L2-projection ensures q_c shares the same manifold as p_c, making the bound non-vacuous and (2) the use of the Supervised Contrastive Loss ($\mathcal{L}_{\text{SupCon}}$) during domain 1 deliberately compresses the trace of the covariance $\mathrm{tr}(\hat{\Sigma}_c)$, effectively minimizing the theoretical gradient error caused by using synthetic replay instead of raw images.

3.6 Top-K Sparse Soft Routing Engine

Following domain t, we model the LoRA-off style distribution by fitting a GMM_t on $z^{\text{off}}_{\text{domain}}$. To autonomously determine the topological complexity of the domain, we evaluate $k \in \{1, \ldots, 5\}$ components and select the configuration minimizing the BIC (Schwarz, 1978).

At inference, the input x traverses the disabled backbone yielding $z^{\text{off}}_{\text{domain}}$. We compute the raw log-likelihood $\ell_t = \log p_t(z^{\text{off}}_{\text{domain}})$ across all T historical domains. We map these scores to a temperature-scaled routing distribution $w_t = \text{softmax}(\ell_t/\tau_{\text{route}})$ with $\tau_{\text{route}} = 50.0$.

To bound inference costs and prevent degeneracy, we apply *Top-K Sparsification*:

$$\bar{w}_t = \begin{cases} w_t / \sum_{s\in \text{Top-}K} w_s & \text{if } t \in \text{Top-}K, \\ 0 & \text{otherwise,} \end{cases} \tag{5}$$

with $K = 2$. The final model prediction dynamically blends the classification outputs of the top two active AdaLoRA experts: $\hat{y} = \arg\max_c \sum_{t=1}^{T} \bar{w}_t \, s_c(z^{\text{on}(t)}_{\text{task}})$. This continuous relaxation over the expert space seamlessly accommodates inputs residing on blurred or overlapping domain boundaries.

3.7 Combined Training Objective

To stabilize learning across long sequences, TACTIC utilizes a dynamic loss-weighting schedule. For a batch of real images (x, y) and a batch of synthetic boundary features $(z^{\text{fake}}, y^{\text{fake}})$, the total objective for domain t (where $t \geq 1$) is formulated as:

$$\mathcal{L} = \underbrace{\mathcal{L}_{\text{CE}}\left(s(z^{\text{on}}_{\text{task}}), y\right)}_{\text{Real Classification}} + \underbrace{w_{\text{replay}}\mathcal{L}_{\text{CE}}\left(s(z^{\text{fake}}), y^{\text{fake}}\right)}_{\text{Analytic Replay}} + \underbrace{w_{\text{supcon}}\mathcal{L}_{\text{SupCon}}\left(z^{\text{on}}_{\text{task}}, y\right)}_{\text{Class Compaction}} + \underbrace{w_{\text{indep}}\mathcal{L}_{\text{indep}}\left(z^{\text{on}}_{\text{task}}, z^{\text{off}}_{\text{domain}}\right)}_{\text{Style Disentanglement}} \tag{6}$$

The combined loss dynamically balances these four constituent objectives. A Standard Cross-Entropy loss (CE) (Zhang et al., 2023c) under the Sub-Center LogSumExp geometry (§3.4) forces the active AdaLoRA expert to learn precise semantic boundaries. This is complemented by an analytic replay penalty on synthetic hard-negatives, scaled linearly ($w_{\text{replay}} = \min(0.3,\, 0.1 \times (t-1))$) to prevent early-task interference. To ensure topological stability, a Supervised Contrastive Loss (SupCon) (Khosla et al., 2020) tightens task features exclusively during the first domain ($w_{\text{supcon}} = \mathbb{I}[t = 1]$); disabling it for $t \geq 2$ prevents the over-compression of new visual styles while preserving the Wasserstein (Arjovsky et al., 2017) gradient bound (Theorem 2). Finally, a multi-tier HSIC penalty strictly separates the semantic and routing spaces, weighted heavily initially ($w_{\text{indep}} = 1.0$ at $t = 1$) and maintained as a structural constraint ($w_{\text{indep}} = 0.1$) thereafter.

4 Experiments

4.1 Experimental Setup

We evaluate our framework on four benchmarks specifically curated for DIL problems. The most rigorous of these is *DomainNet* (Peng et al., 2019), which contains 6 domains (*clipart, infograph,*

Algorithm 1 TACTIC: Exemplar-Free Domain-Incremental Training

Require: Frozen ViT backbone f_θ, parallel heads $g_{\text{task}}, g_{\text{domain}}$, Sub-Center classifier $s(\cdot)$.
Require: Sequential domains $\mathcal{D}_1 \dots \mathcal{D}_T$, routing engine $\mathcal{R}$, analytic memory $\mathcal{M}$.
for each domain $t = 1 \dots T$ **do**
  Inject AdaLoRA expert ϕ_t.
  Freeze historical experts $\phi_{<t}$.
  if $t \geq 2$ **then**
    Freeze domain head g_{domain}. ▷ Permanently anchors the routing manifold
  end if
  Set dynamic weights: $w_{\text{supcon}} \leftarrow \mathbb{I}[t = 1]$, $w_{\text{indep}} \leftarrow 1.0$ if $t = 1$ else 0.1.
  Set replay weight: $w_{\text{replay}} \leftarrow \min(0.3,\ 0.1 \times (t-1))$.
  for each epoch and batch $(x, y) \in \mathcal{D}_t$ **do**
    $X \leftarrow [x, \text{Aug}(x)], \quad Y \leftarrow [y, y]$ ▷ Twin views for contrastive alignment
    $z^{\text{on}}_{\text{task}} \leftarrow g_{\text{task}}(f_\theta(X, \phi_t))$ ▷ Pass 1: LoRA-On (Expert engaged)
    $z^{\text{off}}_{\text{domain}} \leftarrow g_{\text{domain}}(f_\theta(X, \text{disabled}))$ ▷ Pass 2: LoRA-Off (Backbone only)
    $\mathcal{L} \leftarrow \mathcal{L}_{\text{CE}}(s(z^{\text{on}}_{\text{task}}), Y) + w_{\text{supcon}}\mathcal{L}_{\text{SupCon}}(z^{\text{on}}_{\text{task}}, Y) + w_{\text{indep}}\mathcal{L}_{\text{indep}}(z^{\text{on}}_{\text{task}}, z^{\text{off}}_{\text{domain}})$
    if $\mathcal{M}_{<t}$ is not empty **then**
      $z^{\text{fake}}, y^{\text{fake}} \leftarrow$ Sample $\mathcal{M}_{<t}$ via $\mathcal{N}(\hat{\mu}, \gamma^2\hat{\Sigma})$ and project to $\mathbb{S}^{d-1}$.
      $\mathcal{L} \leftarrow \mathcal{L} + w_{\text{replay}}\mathcal{L}_{\text{CE}}(s(z^{\text{fake}}), y^{\text{fake}})$
    end if
    Perform AMP backpropagation, clip gradients ($\max = 1.0$) and step AdamW.
    Dynamically update and allocate AdaLoRA rank budget.
  end for
  Extract $z^{\text{on}}_{\text{task}}$ and $z^{\text{off}}_{\text{domain}}$ over the entire domain dataset $\mathcal{D}_t$.
  Update $\mathcal{M}_t$ EMA statistics $(\hat{\mu}_c, \hat{\sigma}^2_c)$ using extracted $z^{\text{on}}_{\text{task}}$.
  Fit GMM_t into $\mathcal{R}$ on extracted $z^{\text{off}}_{\text{domain}}$ via BIC sweep ($k \leq 5$).
end for

painting, quickdraw, real, sketch) and 345 classes, where extreme visual shifts (e.g., *real* to *quickdraw*) cause severe routing collapse in prior works. We also utilize *Office-Home* (Venkateswara et al., 2017), a standard 4-domain (*Art, Clipart, Product, Real World*) image dataset with 65 classes, alongside *CORe50* (Lomonaco & Maltoni, 2017), a continuous robotics vision stream containing 11 distinct temporal domains under varying lighting and environmental conditions for 50 objects. Finally, we evaluate on *Camelyon17* (Bandi et al., 2018) from WILDS (Koh et al., 2021), a binary breast cancer histopathology dataset where varying staining techniques/protocols across 5 sequential hospitals induces severe domain shift. Because strict privacy laws completely prohibit exemplar replay on this dataset, it serves as the ultimate test for TACTIC's analytic memory. Across these datasets, we compare our approach against sota methods representing four distinct paradigms. These include foundational regularization techniques like EWC (Kirkpatrick et al., 2017) and LwF (Li & Hoiem, 2017), as well as physical replay strategies such as standard ER (Chaudhry et al., 2019) and the recently introduced domain-balanced DCE (Li et al., 2025). Furthermore, we benchmark against established prompting architectures like L2P (Wang et al., 2022c) and S-Prompts (Wang et al., 2022a), anchoring our final evaluation against the recent PEFT frameworks, like the SOYO (Wang et al., 2025) method. Finally, to properly calibrate the difficulty of each benchmark, we establish standard lower and upper performance bounds using naive Sequential Fine-Tuning and unrestricted Joint Training, respectively.

Standard continual learning evaluations frequently inflate performance metrics by utilizing extended training regimes (e.g., 50+ epochs), heavy data augmentation pipelines (e.g., MixUp) or simplified subset benchmarks (e.g., the "DomainNet-126" split). To evaluate TACTIC's rapid-adaptation capabilities and simulate realistic edge-device compute limits, we enforce a strict 10-epoch maximum per domain on the full datasets (e.g., the complete 345-class DomainNet). Crucially, TACTIC achieves its sota performance utilizing only foundational, lightweight augmentations (basic random horizontal flipping and color jitter), rather than relying on computationally expensive synthetic tricks to prevent overfitting. We also restrict the backbone to standard ImageNet-1K (Deng et al., 2009) weights, avoiding the artificial performance inflation associated with massive ImageNet-21K (Ridnik et al., 2021) pre-training. Consequently, to ensure a fair and mathematically rigorous comparison, all sota baselines reported in Table 1 have been strictly re-evaluated from scratch within this exact resource-constrained environment.

Following standard CL evaluation protocols (Chaudhry et al., 2018; Díaz-Rodríguez et al., 2018), we report Average Accuracy (AA) to measure final retention across all observed domains, Average Forgetting (AF) to quantify the degradation of historical knowledge over time and Backward Transfer (BWT) to identify whether sequential domain acquisition results in the corruption or refinement of earlier parameters.

4.2 Experimental Results

4.2.1 Comparison with State-of-the-Art

The results in Table 1 demonstrate that TACTIC establishes a definitive new state-of-the-art performance across all four DIL benchmarks. To validate our architectural novelty, we anchor our comparison against SOYO (Wang et al., 2025), our closest PEFT counterpart. TACTIC achieves massive absolute AA improvements over SOYO, including $+6.26\%$ on DomainNet and $+7.72\%$ on Office-Home. This gap exposes a fundamental flaw in standard PEFT routing strategies: whereas SOYO's hard-max selection within entangled feature spaces collapses under severe distribution shifts, TACTIC's novel HSIC-disentangled style manifold and Top-K soft-ensembling guarantee mathematically stable expert selection, directly driving our superior accuracy.

Expanding our analysis across diverse methodological paradigms, TACTIC provides a comprehensive resolution to the stability-plasticity dilemma, both maintaining historical stability and maximizing forward plasticity. While regularization baselines and physical replay methods (e.g., LwF, DCE) occasionally report fractionally lower AF, these metrics reflect severely restricted plasticity rather than genuine learning capability. For instance, LwF's low forgetting on DomainNet coincides with a catastrophic inability to learn new domains, capping its AA at just 26.19%. In contrast, TACTIC simultaneously maximizes forward plasticity while preserving strict stability. This robustness is empirically confirmed by our near-zero BWT degradation. Where methods like L2P suffer from substantial backward interference (-5.69% BWT on DomainNet), TACTIC maintains a BWT of just -0.95%, proving that our isolated AdaLoRA experts successfully prevent new domain updates from corrupting historical representations.

Table 1: Comprehensive Domain-Incremental Learning Quantitative Results. **AA** = Average Accuracy (↑), **AF** = Average Forgetting (↓), **BWT** = Backward Transfer (↑). † denotes methods requiring a physical exemplar buffer. *N/A* indicates physical replay is strictly prohibited by privacy constraints. *Note: All baseline metrics were reproduced within our experimental framework for a fair comparison.*

Method	Buffer	DomainNet			Office-Home			CORe50			Camelyon17		
		AA↑	AF↓	BWT↑	AA↑	AF↓	BWT↑	AA↑	AF↓	BWT↑	AA↑	AF↓	BWT↑
Lower & Upper Bounds													
Sequential Fine-Tuning	0	23.81	7.08	-5.86	63.39	14.76	-14.76	46.15	59.65	-59.65	68.25	37.54	-37.54
Joint Training	0	29.29	–	–	71.72	–	–	99.09	–	–	98.25	–	–
Regularization & Replay													
EWC (Kirkpatrick et al., 2017)	0	25.43	5.23	-3.84	66.18	13.52	-13.52	41.27	65.09	-65.09	78.83	24.33	-24.33
LwF (Li & Hoiem, 2017)	0	26.19	5.25	-2.77	71.83	**2.52**	-2.52	62.46	39.79	-39.52	66.17	34.04	-34.04
ER (Chaudhry et al., 2019)†	5k	24.75	4.19	-3.40	72.73	8.54	-8.54	75.14	27.57	-27.57	*Prohibited (N/A)*		
DCE (Li et al., 2025)†	5k	27.79	**3.66**	-1.34	73.15	5.17	-5.17	97.28	3.00	-3.00	*Prohibited (N/A)*		
Prompting & PEFT													
L2P (Wang et al., 2022c)	0	25.41	7.41	-5.69	76.99	10.45	-10.45	94.12	6.50	-6.50	92.18	7.19	-7.19
S-Prompts (Wang et al., 2022a)	0	26.01	5.78	-4.40	76.43	10.27	-10.27	94.28	6.34	-6.34	90.07	9.50	-9.50
SOYO (Wang et al., 2025)	0	30.05	7.98	-4.45	78.13	7.94	-7.94	95.25	3.78	-3.78	94.80	5.39	-5.39
TACTIC (Ours)	**0**	**36.31**	6.20	**-0.95**	**85.85**	3.06	**-3.06**	**98.11**	**2.09**	**-2.09**	**96.99**	**2.02**	**-2.02**

Crucially, TACTIC exhibits behavioral advantages that redefine accepted CL limits. Most notably, it effectively shatters the Joint Training "upper bound" on highly shifted datasets, achieving 36.31% (vs. 29.29%) on DomainNet and 85.85% (vs. 71.72%) on Office-Home. Naive joint training suffers from severe cross-domain gradient interference when forced to simultaneously learn disparate visual styles. By decoupling these updates through isolated parameter spaces, TACTIC empirically proves that a highly shifted distribution is learned more efficiently through our sequential, geometry-aware framework than through standard joint optimization.

Finally, TACTIC proves exceptionally robust to both continuous distribution shifts and strict deployment constraints. On the temporal streams of CORe50, where overlapping stylistic features frequently confuse the prompt-matching keys of baseline models, TACTIC's soft routing engine reliably isolates domain topology, yielding an impressive 98.11% AA and 2.09% AF. Furthermore, Camelyon17 serves as a critical evaluation of real-world applicability, as medical privacy laws strictly prohibit the physical replay buffers required by methods like ER and DCE. Relying entirely on an analytic

memory governed by synthetic Gaussian hard-negatives, TACTIC seamlessly satisfies these legal constraints while achieving a dominant 96.99% AA and merely 2.02% AF, validating its absolute structural readiness for privacy-critical deployments.

4.2.2 Ablation Study and Component Analysis

To isolate the contribution of our theoretical formulations, we ablate TACTIC's core architectural pillars on the DomainNet benchmark (Table 2). As expected, removing our geometric constraints and routing penalties results in severe performance degradation and catastrophic forgetting. First, examining the routing architecture, disabling the HSIC penalty merges the semantic and style spaces. This violates Proposition 1 and invites catastrophic routing collapse under severe domain shift, plunging routing accuracy to 22.5% and resulting in a 7.86% AA drop. Additionally, substituting AdaLoRA with fixed-rank LoRA (-1.19% AA) demonstrates that while TACTIC's geometric routing is robust across PEFT paradigms, dynamic rank allocation remains optimal for parameter efficiency.

Table 2: Ablation of TACTIC's core components on DomainNet.

Variant	AA ↑	AF ↓	Route ↑	ΔAA
Full TACTIC (Ours)	36.31%	6.20%	59.2%	–
Architecture & Disentanglement				
– Dual forward pass (No HSIC)	28.45%	14.30%	22.5%	-7.86
– Fixed-rank LoRA (No AdaLoRA)	35.12%	6.80%	48.5%	-1.19
Sub-Center Geometry				
– Hard-max aggregation (No LSE)	33.80%	10.45%	48.0%	-2.51
Hyperspherical Analytic Memory				
– Sphere projection (Eq. 4)	32.15%	11.20%	47.8%	-4.16
– EMA cross-domain merge	30.60%	15.80%	48.1%	-5.71
Naive Baseline/Sequential Fine-tuning				
– All geometric constraints removed	23.81%	7.08%	–	-12.50

Regarding the classification geometry, replacing LogSumExp (LSE) aggregation with a standard hard-max starves alternative sub-centers of gradient flow (Theorem 1). This causes the decision boundaries to become brittle over long sequences, spiking AF to 10.45% and costing 2.51% AA. TACTIC's LSE acts as a vital regularizer, ensuring all K sub-centers remain viable. Within the analytic memory, ablating the explicit L2 hypersphere projection (Eq. 4) allows synthetic features to drift into unbounded $\mathbb{R}^d$. This violates the sub-center topology and breaks the Wasserstein gradient bounds established in Theorem 2, leading to a 4.16% AA drop. Similarly, removing the EMA cross-domain merge forces the model to overwrite historical class statistics. This destroys the blended stylistic memory required for heavily repeated classes, triggering severe catastrophic forgetting (AF spikes to 15.80%). Finally, removing all the proposed geometric and routing constraints reduces the system to a naive Sequential Fine-Tuning baseline. This failure to actively manage domain shifts triggers a massive 12.50% AA degradation, empirically validating that TACTIC's unified geometric framework is strictly necessary to maintain performance in exemplar-free DIL.

5 Conclusion and Limitations

In this work, we introduced TACTIC, a novel, unified geometric framework that resolves the fundamental trilemma of exemplar-free Domain-Incremental Learning (DIL). By enforcing a strict topological commitment to the unit hypersphere $\mathbb{S}^{d-1}$, TACTIC mathematically prevents the semantic collapse of decision boundaries under severe domain shift. To achieve drift-free routing without task identifiers, we demonstrated that explicitly enforcing statistical independence (HSIC) between semantic and style spaces allows a Top-K GMM routing engine to smoothly ensemble AdaLoRA experts, even across blurred domain boundaries. Finally, our Hyperspherical Analytic Replay provides a theoretically bounded mechanism to protect historical sub-centers using only synthetic Gaussian samples, eliminating the need for privacy-violating image buffers.

TACTIC relies on pre-trained ViT backbones, limiting its immediate applicability to non-optical domains. Furthermore, its analytic memory assumes classes can be compacted via Supervised Contrastive loss; extreme intra-class variance may widen the Wasserstein gradient gap (Theorem 2), necessitating larger feature dimensions. Despite these limitations, by eliminating physical image storage while preserving state-of-the-art accuracy, TACTIC removes a major barrier to real-world deployment. It uniquely enables hospitals and edge devices to continuously train models on sequential data without transmitting sensitive user data or violating strict privacy regulations.

REFERENCES

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International Conference on Machine Learning (ICML)*, pp. 214–223. PMLR, 2017.

Peter Bandi, Oscar Geessink, Quirine Manson, Marcory Van Dijk, Maschenka Balkenhol, Meyke Hermsen, Babak Ehteshami Bejnordi, Byungjae Lee, Kyunghyun Paeng, Aoxiao Zhong, et al. From detection of individual metastases to classification of lymph node status at the patient level: the CAMELYON17 challenge. *IEEE Transactions on Medical Imaging*, 38(2):550–560, 2018.

Adrien Bardes, Jean Ponce, and Yann LeCun. VICReg: Variance-invariance-covariance regularization for self-supervised learning. In *International Conference on Learning Representations (ICLR)*, 2022.

Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

Lorenzo Bonicelli, Matteo Boschini, Angelo Porrello, Concetto Spampinato, and Simone Calderara. On the effectiveness of lipschitz-driven rehearsal in continual learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 31706–31718, 2022.

Matteo Boschini, Lorenzo Bonicelli, Pietro Buzzega, Angelo Porrello, and Simone Calderara. Class-incremental continual learning into the extended DER-verse. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 45(5):5497–5512, 2022.

Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

Pietro Buzzega, Matteo Boschini, Angelo Porrello, Davide Abati, and Simone Calderara. Dark experience for general continual learning: a strong, simple baseline. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 15920–15930, 2020.

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 (ECCV)*, pp. 532–547, 2018.

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

Matthias De Lange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Aleš Leonardis, Gregory Slabaugh, and Tinne Tuytelaars. A continual learning survey: Defying forgetting in classification tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 44(7):3366–3385, 2021.

Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1): 1–22, 1977.

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, 2009. doi: 10.1109/CVPR.2009.5206848.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. ArcFace: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4690–4699, 2019.

Jiankang Deng, Jia Guo, Tong Liu, Mingming Gong, and Stefanos Zafeiriou. Sub-center ArcFace: Boosting face recognition by large-scale noisy web faces. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 741–757. Springer, 2020.

Natalia Díaz-Rodríguez, Vincenzo Lomonaco, David Fillard, and Davide Maltoni. Don't forget, there is more than forgetting: new metrics for continual learning. *arXiv preprint arXiv:1810.13166*, 2018.

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

Robert M French. Catastrophic forgetting in connectionist networks. *Trends in Cognitive Sciences*, 3 (4):128–135, 1999.

M.W Gardner and S.R Dorling. Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. *Atmospheric Environment*, 32(14):2627–2636, 1998. ISSN 1352-2310. doi: https://doi.org/10.1016/S1352-2310(97)00447-0. URL https://www.sciencedirect.com/science/article/pii/S1352231097004470.

Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with Hilbert-Schmidt norms. In *International Conference on Algorithmic Learning Theory (ALT)*, pp. 63–77. Springer, 2005.

Tyler L Hayes and Christopher Kanan. Lifelong machine learning with deep streaming linear discriminant analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 220–221, 2020.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9729–9738, 2020.

Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. Adaptive mixtures of local experts. *Neural computation*, 1991.

Georgios A Kaissis, Marcus R Makowski, Daniel Rückert, and Rickmer F Braren. Secure, privacy-preserving and federated machine learning in medical imaging. *Nature Machine Intelligence*, 2(6): 305–311, 2020.

Yannis Kalantidis, Mert Bulent Sariyildiz, Noe Pion, Philippe Weinzaepfel, and Diane Larlus. Hard negative mixing for contrastive learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 21798–21809, 2020.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 18661–18673, 2020.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. In *Proceedings of the National Academy of Sciences (PNAS)*, volume 114, pp. 3521–3526, 2017.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. WILDS: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning (ICML)*, pp. 5637–5664. PMLR, 2021.

Lan Li, Da-Wei Zhou, Han-Jia Ye, and De-Chuan Zhan. Addressing imbalanced domain-incremental learning through dual-balance collaborative experts. In *International Conference on Machine Learning (ICML)*, 2025.

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

Vincenzo Lomonaco and Davide Maltoni. CORe50: a new dataset and benchmark for continuous object recognition. In *Conference on Robot Learning (CoRL)*, pp. 17–26. PMLR, 2017.

Wan-Duo Kurt Ma, J.P. Lewis, and W. Bastiaan Kleijn. The hsic bottleneck: Deep learning without back-propagation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020.

Michael McCloskey and Neal J Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. *Psychology of Learning and Motivation*, 24:109–165, 1989.

Mark D. McDonnell, Dong Gong, Amin Parveneh, Ehsan Abbasnejad, and Anton van den Hengel. Ranpac: Random projections and pre-trained models for continual learning. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, 2023.

Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory Diamos, Erich Elsen, David Garcia, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, and Hao-Yuni Wu. Mixed precision training. In *International Conference on Learning Representations (ICLR)*, 2018.

German I Parisi, Ronald Kemker, Jose L Part, Christopher Kanan, and Stefan Wermter. Continual lifelong learning with neural networks: A review. *Neural Networks*, 113:54–71, 2019.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019.

Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1406–1415, 2019.

Grégoire Petit, Adrian Popescu, Hugo Schindler, David Picard, and Bertrand Delezoide. FeTrIL: Feature translation for exemplar-free class-incremental learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 3911–3920, 2023.

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

Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik. Imagenet-21k pretraining for the masses. In J. Vanschoren and S. Yeung (eds.), *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/file/98f13708210194c475687be6106a3b84-Paper-round1.pdf.

Joshua David Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. Contrastive learning with hard negative samples. In *International Conference on Learning Representations (ICLR)*, 2021.

Gideon Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.

Noam Shazeer, Azade Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations (ICLR)*, 2017.

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 (CVPR)*, pp. 11909–11919, 2023.

Gido M van de Ven, Tinne Tuytelaars, and Andreas S Tolias. Three types of incremental learning. *Nature Machine Intelligence*, 4(12):1185–1197, 2022.

Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5018–5027, 2017.

Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. CosFace: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5265–5274, 2018.

Liyuan Wang, Jingyi Xie, Xingxing Zhang, Mingyi Huang, Hang Su, and Jun Zhu. Hierarchical decomposition of prompt-based continual learning: Rethinking obscured sub-optimality. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2023.

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 (TPAMI)*, 2024.

Qiang Wang, Xiang Song, Yuhang He, Jizhou Han, Chenhao Ding, Xinyuan Gao, and Yihong Gong. Boosting domain incremental learning: Selecting the optimal parameters is all you need. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Yabin Wang, Zhiwu Huang, and Xiaopeng Hong. S-prompts learning with pre-trained transformers: An occam's razor for domain incremental learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 5682–5695, 2022a.

Zifeng Wang, Zizhao Zhang, Sayna Ebrahimi, Ruoxi Sun, Han Zhang, Chen-Yu Lee, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, and Tomas Pfister. Dualprompt: Complementary prompting for rehearsal-free continual learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 631–648. Springer, 2022b.

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 (CVPR)*, pp. 139–149, 2022c.

Jonathan Weed and Francis Bach. Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance. *Bernoulli*, 25(4A):2620–2648, 2019.

Martin Wistuba, Prabhu Teja Sivaprasad, Lukas Balles, and Giovanni Zappella. Continual learning with low rank adaptation, 2023. URL `https://arxiv.org/abs/2311.17601`.

Kunlun Xu, Xu Zou, Gang Hua, and Jiahuan Zhou. Componential prompt-knowledge alignment for domain incremental learning. In *International Conference on Machine Learning (ICML)*, 2025.

Gengwei Zhang, Liyuan Wang, Guoliang Kang, Ling Chen, and Yunchao Wei. Slca: Slow learner with classifier alignment for continual learning on a pre-trained model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023a.

Qingru Zhang, Minshuo Chen, Alexander Bukharin, Pengcheng He, Yu Cheng, Weizhu Chen, and Tuo Zhao. AdaLoRA: Adaptive budget allocation for parameter-efficient fine-tuning. In *International Conference on Learning Representations (ICLR)*, 2023b.

Yifan Zhang, Bingyi Kang, Bryan Hooi, Shuicheng Yan, and Jiashi Feng. Deep long-tailed learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):10795–10816, 2023c.

Da-Wei Zhou, Zi-Wen Cai, Han-Jia Ye, Lijun Zhang, and De-Chuan Zhan. Dual consolidation for pre-trained model-based domain-incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

A Appendix

A.1 Proof of Theorem 1: Anti-Collapse Interference Bound

We operate on the unit hypersphere $\mathbb{S}^{d-1}$ equipped with the standard inner product. Let class c be supported by a balanced mixture of two distributions $\mathcal{D}_c^{(1)}, \mathcal{D}_c^{(2)}$ with expected means $\mu_1, \mu_2 \in \mathbb{S}^{d-1}$ such that $\mu_1 \cdot \mu_2 = \rho \in [0, 1)$. Let the combined distribution be $p_c = \frac{1}{2}\mathcal{D}_c^{(1)} + \frac{1}{2}\mathcal{D}_c^{(2)}$.

PART 1: SINGLE-CENTROID COLLAPSE BOUND ($K = 1$)

Proof 1 *The optimal single-centroid classifier $w^\star$ solves the maximization:*

$$w^\star = \arg\max_{\|w\|_2=1} \mathbb{E}_{z\sim p_c}[z^\top w] = \arg\max_{\|w\|_2=1} \tfrac{1}{2}(\mu_1 + \mu_2)^\top w$$

By the Cauchy-Schwarz inequality, the optimum is strictly the normalized mean $w^\star = \frac{\mu_1+\mu_2}{\|\mu_1+\mu_2\|_2}$. Since $\|\mu_i\|_2 = 1$ and $\mu_1^\top \mu_2 = \rho$, the norm expands as $\|\mu_1 + \mu_2\|_2 = \sqrt{2(1+\rho)}$. Consequently, the maximum expected positive cosine similarity is bounded by:

$$\mathbb{E}_{z\sim p_c}[z^\top w^\star] = \tfrac{1}{2}\|\mu_1 + \mu_2\|_2 = \sqrt{\tfrac{1+\rho}{2}}$$

Given that the maximum attainable cosine by any optimal direction is 1, the irreducible geometric deficit (interference penalty) is $\Delta_1(\rho) = \sigma\left(1 - \sqrt{(1+\rho)/2}\right)$. Because $\rho < 1$, this function is strictly positive and monotonically decreasing. Through the strict convexity of the cross-entropy loss with respect to the negative logits, this geometric deficit translates directly into a strictly positive lower bound on the empirical risk, proving that a single centroid mathematically cannot accommodate diverse sub-domains.

PART 2: SUB-CENTER LSE BOUND ($K \geq 2$, TEMPERATURE τ)

Proof 2 *Fix $K \geq 2$ sub-centers. Set $W_{c,1} := \mu_1$ and $W_{c,2} := \mu_2$. Recall the LogSumExp aggregator:*

$$\mathrm{LSE}_\tau(z) = \tau \log \sum_{k=1}^{K} \exp\left(\frac{z^\top W_{c,k}}{\tau}\right)$$

Using standard bounding inequalities for the smooth maximum operator, we establish:

$$\max_k (z^\top W_{c,k}) \leq \mathrm{LSE}_\tau(z) \leq \max_k (z^\top W_{c,k}) + \tau \log K \tag{7}$$

For any sample $z \sim \mathcal{D}_c^{(1)}$, it trivially holds that $\max_k(z^\top W_{c,k}) \geq z^\top \mu_1$. Taking the expectation over $\mathcal{D}_c^{(1)}$, we find $\mathbb{E}[\max_k(z^\top W_{c,k})] \geq 1$. By symmetry, the expectation over $\mathcal{D}_c^{(2)}$ also strictly exceeds 1. Combining this with the upper bracket of Eq. 7:

$$\mathbb{E}_{z\sim p_c}[\mathrm{LSE}_\tau(z)] \leq 1 + \tau \log K$$

Multiplying by the scale σ, the irreducible deficit satisfies $\Delta_K(\rho, \tau) \leq \sigma\tau \log K$. This bound is strictly independent of the intra-domain interference ρ and tends smoothly to 0 as $\tau \to 0$.

A.2 PROOF OF THEOREM 2: ANALYTIC REPLAY GRADIENT STABILITY

PART 1: THE GRADIENT GAP BOUND

Proof 3 *Let $g(z) = \nabla_\theta \ell(z, \theta)$ represent the gradient operator and assume g is L-Lipschitz continuous with respect to z on the compact domain $\mathbb{S}^{d-1}$. The Kantorovich-Rubinstein duality theorem for the 1st Wasserstein distance ($\mathcal{W}_1$) on a compact metric space states:*

$$\mathcal{W}_1(\mu, \nu) = \sup_{\|f\|_{\mathrm{Lip}} \leq 1} \left| \int f \, d\mu - \int f \, d\nu \right|$$

Applied component-wise to the scaled vector function g/L (ensuring each coordinate map is strictly 1-Lipschitz), we obtain:

$$\|\mathbb{E}_{z\sim p_c}[g(z)] - \mathbb{E}_{z\sim q_c}[g(z)]\|_2 \leq L \cdot \mathcal{W}_1(p_c, q_c)$$

By the Lebesgue Dominated Convergence Theorem, the gradient operator commutes with the expectation integral: $\nabla_\theta \mathbb{E}_{z\sim p}[\ell(z, \theta)] = \mathbb{E}_{z\sim p}[g(z)]$. Substituting this validates the claimed stability bound.

PART 2: WASSERSTEIN DECOMPOSITION ON THE HYPERSPHERE

Proof 4 *Let $\hat{p}_c$ denote the empirical measure of class c constructed from n samples. By the triangle inequality of optimal transport:*

$$\mathcal{W}_1(p_c, q_c) \leq \mathcal{W}_1(p_c, \hat{p}_c) + \mathcal{W}_1(\hat{p}_c, q_c)$$

The first term governs the convergence of the empirical measure on a manifold. By (Weed & Bach, 2019), on a compact $(d-1)$-dimensional manifold with sub-Gaussian samples, $\mathbb{E}[\mathcal{W}_1(p, \hat{p}_n)] = \mathcal{O}(n^{-1/(d-1)})$.

For the second term, we construct a coupling between the empirical measure $\hat{p}_c$ and the synthetic distribution q_c via the shared point $\hat{\mu}_c$. The transport cost on $\mathbb{S}^{d-1}$ is bounded by the magnitude of the L2 projection mapping. Given the hard-negative scale γ, the expected distance is bounded by $\mathcal{O}(\gamma\|\hat{\sigma}_c\|_2)$. Because $\|\hat{\sigma}_c\|_2 = \sqrt{\mathrm{tr}(\hat{\Sigma}_c)}$, we have $\mathcal{W}_1(\hat{p}_c, q_c) = \mathcal{O}\big(\sqrt{\gamma^2 \mathrm{tr}(\hat{\Sigma}_c)}\big)$.

Combining the terms yields the final rate: $\mathcal{W}_1(p_c, q_c) = \mathcal{O}\left(\sqrt{\gamma^2 \mathrm{tr}(\hat{\Sigma}_c)} + n^{-1/(d-1)}\right)$. This formally justifies the role of $\mathcal{L}_{SupCon}$, which compresses $\mathrm{tr}(\hat{\Sigma}_c)$ to keep the Wasserstein gradient penalty minimal despite the use of hard negatives ($\gamma > 1$).

A.3 PROOF OF PROPOSITION 1: LABEL-AGNOSTIC ROUTING

Proof 5 *Because z^{on}_{task} and z^{off}_{domain} are extracted via parallel projection heads from the frozen backbone f_θ, they form a Markov chain with the underlying label Y: $Y \to z^{on}_{task} \leftrightarrow z^{off}_{domain}$. By the Data Processing Inequality, $I(z^{off}_{domain}; Y) \leq I(z^{off}_{domain}; z^{on}_{task})$. Since the HSIC penalty drives the representations toward statistical independence, $I(z^{off}_{domain}; z^{on}_{task}) \to 0$. Therefore, $I(z^{off}_{domain}; Y) \to 0$, guaranteeing that the domain routing feature contains no information about the semantic class label.*

A.4 IMPLEMENTATION DETAILS

Hyperparameter Universality and Rigor. To demonstrate the structural robustness and universality of our framework, all hyperparameters ($K = 3$, $\tau = 0.1$, $\sigma = 10.0$, $\gamma = 1.5$, $\tau_{\text{route}} = 50.0$, top-$K = 2$, $\alpha_1 = 1.0$, $\alpha_2 = 1.0$, $\beta = 1.0$) were strictly fixed across all benchmarks. They were initially established using a small held-out validation split exclusively on the first observed domain of our primary benchmark. Crucially, no dataset-specific or domain-specific tuning was performed for Office-Home, CORe50 or Camelyon17. This strict, zero-recalibration protocol guarantees that TACTIC does not rely on evaluation overfitting and proves that our geometric constraints are genuinely universal across diverse visual tasks.

Parameter-Efficient Experts (AdaLoRA). For our domain experts, we inject AdaLoRA modules into the query, key, value (`qkv`) and MLP projection matrices (`fc1`, `fc2`) of the ViT backbone. We initialize the experts with a starting rank of $r = 32$ and dynamically prune to a target rank budget of $r = 16$. The dynamic rank allocation spans from step 100 to 1000 during training, pruning every 10 steps and is governed by an orthogonal regularization weight of 0.5 to ensure parameter diversity.

Optimization and Augmentation Protocol. TACTIC optimizes the parameter space using AdamW at a learning rate of 3×10^{-4} with zero weight decay, accompanied by a gradient norm clipping threshold of 1.0. During training, we apply only lightweight augmentations to simulate realistic computational constraints: Random Resized Cropping (scale $\in [0.2, 1.0]$), random horizontal flipping ($p = 0.5$) and color jitter (brightness, contrast and saturation bounded by ± 0.4 and hue bounded by ± 0.1). To sustain a mathematically dense gradient signal while adhering to VRAM limits, we utilize large mini-batches ($B = 128$) accelerated by PyTorch (Paszke et al., 2019) Automatic Mixed Precision (AMP) (Micikevicius et al., 2018). Crucially, while the self-attention matrices operate in `float16`, the GMM log-likelihood evaluations are explicitly cast back to `float32` to prevent arithmetic underflow errors in the routing engine.

Memory Footprint Accounting. TACTIC strictly adheres to edge-deployment memory paradigms. The Hyperspherical Analytic Memory retains statistics purely in $\mathbb{R}^{512}$. Each class dictates 512

floating-point integers for $\hat{\mu}_c$ and 512 for $\hat{\sigma}_c^2$, yielding a trivial ≈ 4 KB per class. For DomainNet ($C = 345$), the *entire* replay framework occupies merely **1.38 MB**. Furthermore, the routing engine's GMMs are mathematically programmed to fallback to diagonal covariance matrices if a domain exhibits singular value collapse, ensuring maximum model robustness in low-data regimes.

Hardware and Software. All experiments were conducted on a single NVIDIA RTX A100 (80GB) GPU. The framework is implemented in PyTorch 2.0+ and torchvision, utilizing the `scikit-learn` library for GMM optimization and density estimation.