ORTHOGONAL SELECTION OF GRADIENT-FREE TASK POSTERIORS FOR CLASS-INCREMENTAL LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

In class-incremental learning (CIL), a prediction is correct only when both within-task prediction (WP) and task-id prediction (TP) succeed. Methods built on a frozen pre-trained model realize TP through prompt matching, expert routing, or task-identity heads and learn WP within the selected modules. Yet, their modules and classifier heads are trained with a cross-entropy loss restricted to the classes of the current task. This paper shows that this objective is invariant to every logit shift that is constant within each task, and its gradient has no component in the task-level subspace of the logits, which affects not WP but TP. The same subspace is exactly the set of additive corrections that leave WP unchanged for every logit vector. Building on these two results, we propose a method that supplies TP through orthogonal selection: a task posterior, computed in closed form from class-wise first- and second-order moments of frozen features with shared whitening and rank-4 class-specific corrections, is added to the logits without any gradient path. Across 15 methods, five datasets, 14 task splits, and five seeds (1,050 runs under a common training protocol), TP fails in 92.8–97.4% of the errors of every existing method, and the variation of final accuracy across methods is almost entirely a variation of TP. The proposed method raises TP by 16.70 points and final accuracy by 15.74 points while leaving WP unchanged, reaches 78.25% average final accuracy, and its gain increases monotonically with the number of tasks on all five datasets.

1 INTRODUCTION

Class-incremental learning (CIL) trains a model on a sequence of tasks with disjoint classes and requires it to predict, without task identifiers, one class among all classes observed so far (Rebuffi et al., 2017; Wang et al., 2024). A CIL prediction is correct if and only if two events hold: the prediction restricted to the classes of the true task is correct, termed within-task prediction (WP), and the prediction over all classes falls into the true task, termed task-id prediction (TP) (Kim et al., 2022). Good WP and good TP are necessary and sufficient for good CIL, and TP is closely tied to out-of-distribution detection (Kim et al., 2022; 2023). Since WP is an ordinary supervised problem inside each task, the difficulty specific to CIL concentrates in TP, which must separate tasks that are never trained together (Kim et al., 2022; Lin et al., 2024).

Methods built on pre-trained models (PTMs) have become the standard for CIL: a pre-trained vision transformer (ViT) (Dosovitskiy et al., 2021) is frozen, and only lightweight modules such as prompts, adapters, or experts are learned (Wang et al., 2022b;a; Smith et al., 2023; Zhou et al., 2025; 2024a). A large family of these methods consists of a selection rule and a module pool. The selection rule assigns each input to modules by key-query matching, attention over prompt components, task-identity inference, or expert routing, and the selected modules produce the logits (Wang et al., 2022b; Smith et al., 2023; Wang et al., 2023; 2025a). The selection rule is therefore the component through which task identity enters the prediction, whereas within-task discrimination is learned inside the modules and the classifier head. The modules and the classifier head are trained with a cross-entropy loss restricted to the classes of the current task, and the selection parameters with the same loss or with a matching loss on inputs of the current task (Wang et al., 2022b;a; Smith et al., 2023).

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.

However, this objective provides no training signal for the task-level component of the logits, on which TP depends but WP does not. The masked cross-entropy is invariant to every logit shift that is constant within each task, so its gradient has no component in the subspace of task-level shifts (Proposition 1), and with a linear head trained by stochastic gradient descent (SGD), the task sums of the head weights follow a data-independent recursion (Corollary 1). Removing the mask does not resolve the problem without data from earlier tasks, since the gradient then only lowers the logits of old classes and induces a recency bias (Wu et al., 2019; Ahn et al., 2021). Methods that model TP explicitly therefore add a task-identity component that requires an additional training stage for every task (Wang et al., 2023), or estimate task likelihood ratios with samples stored from earlier tasks (Lin et al., 2024), which the rehearsal-free setting excludes.

This paper supplies TP through *orthogonal selection*: task selection confined to the subspace that the masked objective cannot see and computed without any gradient path to or from training. This subspace is exactly the set of additive logit corrections that leave WP unchanged for every logit vector (Proposition 2). Orthogonal selection therefore changes TP while WP stays fixed by construction, and its accuracy is bounded above by the WP of the underlying classifier (Corollary 2). The proposed method, CLASP (CLass-level Anisotropy Shrinkage Prior), instantiates orthogonal selection with a class-wise Gaussian task posterior. It accumulates the first- and second-order moments of frozen features once per class, whitens them with a shrunk pooled covariance, keeps for each class the $r = 4$ directions that deviate most from the pooled structure, marginalizes the class likelihoods within each task, and adds the resulting log task posterior to the logits, without storing exemplars or training any parameter.

The contributions of this paper are as follows.

- A logit-space characterization of TP (Section 3): the masked cross-entropy has no gradient in the subspace of task-level logit shifts, and this subspace is exactly the set of additive corrections that preserve WP. TP can therefore be supplied in a subspace that the training objective never informs, with WP unchanged and accuracy bounded by WP, whereas existing task-identity mechanisms require additional trained components or stored samples.
- A closed-form task posterior from class-wise low-rank Gaussians of frozen features that realizes orthogonal selection without exemplars, gradients, or additional training, at two ViT passes per image for any number of tasks (Section 4). Added without retraining to four prompt-based methods on ImageNet-R, it raises final accuracy by 6.53–11.45 points.
- A verification in 1,050 runs over 15 methods, 14 task splits of five datasets, and five seeds, with the baselines repeated under their original settings (Section 5): TP fails in 92.8–97.4% of the errors of every existing method, and the proposed method raises TP by 16.70 points and final accuracy by 15.74 points with WP unchanged, ranking first in 12 of the 14 settings.

2 Related Work

CIL with pre-trained models. Prompt-based methods learn prompts on a frozen ViT and choose them per input: L2P matches a query against the keys of a shared pool (Wang et al., 2022b), DualPrompt separates general from task-specific prompts (Wang et al., 2022a), CODA-Prompt composes components with input-conditioned attention (Smith et al., 2023), HiDe-Prompt decomposes the objective into within-task prediction, task-identity inference, and task-adaptive prediction (Wang et al., 2023), and CPrompt and VQ-Prompt make the selection train–test consistent or discrete (Gao et al., 2024; Jiao et al., 2024), while PEARL and APT share input-agnostic prompts across tasks (Qin et al., 2025; Chen et al., 2025). Adapter- and expert-based methods ensemble, expand, retrieve, or route parameter-efficient modules (Gao et al., 2023; Zhou et al., 2024b; Sun et al., 2025; Wang et al., 2025a; Liang & Li, 2024), TUNA selecting the adapter with the lowest prediction entropy (Wang et al., 2025b). A further line classifies with prototypes of frozen or first-task-adapted features, possibly after a random projection (Zhou et al., 2025; McDonnell et al., 2023), or aligns the classifier with class-wise Gaussian statistics (Zhang et al., 2023). The proposed method replaces none of these components; it adds a task-level term to the logits of an existing classifier.

Task-id prediction. Kim et al. (2022) decompose CIL into WP and TP and relate TP to out-of-distribution detection, and Kim et al. (2023) establish the learnability of CIL on this basis. TPL

predicts task identity from a likelihood ratio estimated with replay data (Lin et al., 2024), HiDe-Prompt trains a task-identity head from preserved statistics of uninstructed representations (Wang et al., 2023), and MOS and TUNA retrieve task-specific adapters at inference (Sun et al., 2025; Wang et al., 2025b): in each case task identity comes from an additional learned component or from stored samples. The proposed method obtains it in closed form without exemplars and with a guarantee that WP is unchanged.

Logit correction and Gaussian discriminants. Bias correction in CIL rescales or shifts the logits of new classes with parameters fitted on exemplars (Wu et al., 2019), aligns the norms of classifier weights (Zhao et al., 2020), or separates the softmax of old and new classes (Ahn et al., 2021), and logit adjustment adds a fixed log-prior in long-tailed recognition (Menon et al., 2021). The post hoc corrections share a single scale or shift across all inputs and separated softmax changes the training loss, whereas the proposed term depends on the input through a task posterior and leaves training unchanged. Class-conditional Gaussians with covariance shrinkage go back to regularized discriminant analysis (Friedman, 1989; Ledoit & Wolf, 2004), and FeCAM classifies with class-wise covariances and Mahalanobis distances in exemplar-free CIL (Goswami et al., 2023). The proposed method uses class-wise Gaussians only to estimate the task posterior, and results on spiked covariance models (Baik et al., 2005; Baik & Silverstein, 2006; Paul, 2007; Benaych-Georges & Nadakuditi, 2011) explain why most sample directions of a class covariance carry estimation noise.

3 Task-Level Structure of CIL Logits

Selection-based CIL. Tasks $t = 1, \dots, T$ arrive sequentially with disjoint class sets $\mathcal{Y}_t$. After task t the model predicts over $\mathcal{Y}_{1:t} = \bigcup_{j \le t} \mathcal{Y}_j$ with $C = |\mathcal{Y}_{1:t}|$ classes, and $\tau(c)$ denotes the task that contains class c. A frozen pre-trained encoder f_θ maps an input x to the query feature $\boldsymbol{z}(x) = f_\theta(x) \in \mathbb{R}^d$. A selection-based method keeps a module pool $\{\phi_k\}_{k=1}^K$ and a selection rule $\kappa(\boldsymbol{z}; \psi) \in \Delta^{K-1}$ that weights the modules, computes the representation $\boldsymbol{h}(x) = f_\theta\big(x; \{\phi_k\}_{k=1}^K, \kappa(\boldsymbol{z}(x); \psi)\big)$, and outputs the logits $\boldsymbol{s}(x) = \boldsymbol{A}\boldsymbol{h}(x) + \boldsymbol{b} \in \mathbb{R}^C$. Key-query matching yields a sparse κ, attention over prompt components a dense one, and a method without selection corresponds to $K = 1$; Section B lists representative instances. The prediction is $\hat{y}(x) = \arg\max_{c \in \mathcal{Y}_{1:t}} s_c(x)$, with ties broken by a fixed order of the classes that every $\arg\max$ below also uses.

Definition 1 (WP and TP)**.** *For a labeled input* (x, y)*,* $\mathrm{WP}(x, y) = \mathbb{I}\big[\arg\max_{c \in \mathcal{Y}_{\tau(y)}} s_c(x) = y\big]$ *and* $\mathrm{TP}(x, y) = \mathbb{I}\big[\hat{y}(x) \in \mathcal{Y}_{\tau(y)}\big]$.

Lemma 1 (Exact factorization)**.** $\mathbb{I}[\hat{y}(x) = y] = \mathrm{TP}(x, y)\,\mathrm{WP}(x, y)$ *for every* (x, y)*. Hence the CIL accuracy satisfies* $\mathrm{Acc} = \Pr(\mathrm{TP} \wedge \mathrm{WP}) \le \min\{\Pr(\mathrm{TP}), \Pr(\mathrm{WP})\}$ *and* $1 - \mathrm{Acc} = \Pr(\neg\mathrm{TP}) + \Pr(\mathrm{TP} \wedge \neg\mathrm{WP})$.

Lemma 1 is the operational form of the decomposition of Kim et al. (2022). It holds pointwise rather than only in expectation and does not require the two events to be independent. The WP and TP values in Section 5 are recorded by an instrument that checks this identity on every evaluation batch.

The task-level subspace. Let $\mathbf{1}_j \in \mathbb{R}^C$ indicate the classes of task j. The logit space decomposes orthogonally as $\mathbb{R}^C = \mathcal{V}_\mathrm{T} \oplus \mathcal{V}_\mathrm{W}$ with

$$\mathcal{V}_\mathrm{T} = \mathrm{span}\{\mathbf{1}_1, \dots, \mathbf{1}_t\}, \qquad \mathcal{V}_\mathrm{W} = \big\{\boldsymbol{v} \in \mathbb{R}^C : \mathbf{1}_j^\top \boldsymbol{v} = 0,\ j = 1, \dots, t\big\}. \tag{1}$$

The projection onto $\mathcal{V}_\mathrm{T}$ replaces every logit by the mean logit of its task, $(\Pi_\mathrm{T}\boldsymbol{s})_c = |\mathcal{Y}_{\tau(c)}|^{-1}\mathbf{1}_{\tau(c)}^\top \boldsymbol{s}$, and $\Pi_\mathrm{W} = \boldsymbol{I} - \Pi_\mathrm{T}$ retains the within-task contrasts. A vector of $\mathcal{V}_\mathrm{T}$ is constant on every task and never changes a within-task $\arg\max$. WP therefore depends on the logits only through $\Pi_\mathrm{W}\boldsymbol{s}$, whereas TP depends on both components, and the task-level component $\Pi_\mathrm{T}\boldsymbol{s}$ is the part of the logits that can move TP without moving WP.

Masked training is blind to the task-level subspace. Rehearsal-free methods train on task t with the cross-entropy restricted to its classes,

$$\ell_t(\boldsymbol{s}; y) = -s_y + \log \sum_{c \in \mathcal{Y}_t} \exp(s_c), \qquad y \in \mathcal{Y}_t, \tag{2}$$

which is equivalent to setting the logits of all other classes to $-\infty$ before a softmax cross-entropy.

Proposition 1. *For every task t and label $y \in \mathcal{Y}_t$: (i) $\ell_t(\boldsymbol{s} + \boldsymbol{v}; y) = \ell_t(\boldsymbol{s}; y)$ for all $\boldsymbol{v} \in \mathcal{V}_{\mathrm{T}}$; (ii) $\nabla_{\boldsymbol{s}} \ell_t(\boldsymbol{s}; y) \in \mathcal{V}_{\mathrm{W}}$; (iii) for a linear head $\boldsymbol{s} = \boldsymbol{A}\boldsymbol{h} + \boldsymbol{b}$ with rows $\boldsymbol{a}_c^\top$, $\sum_{c \in \mathcal{Y}_j} \nabla_{\boldsymbol{a}_c} \ell_t = \boldsymbol{0}$ and $\sum_{c \in \mathcal{Y}_j} \partial \ell_t / \partial b_c = 0$ for every task j.*

Corollary 1. *Assume every row of a linear head, including rows added when the head expands, is initialized independently of the data, and train the head on any sequence of masked losses (2) with SGD without coordinate-wise normalization, momentum μ, weight decay ω, and a learning-rate schedule η_k that does not depend on the data. Then the task sums $\bar{\boldsymbol{a}}_j = \sum_{c \in \mathcal{Y}_j} \boldsymbol{a}_c$ and $\bar{b}_j = \sum_{c \in \mathcal{Y}_j} b_c$ follow a linear recursion determined by their initialization and by (η_k, μ, ω) alone, and the task-level component of the logits,*

$$(\Pi_{\mathrm{T}} \boldsymbol{s}(x))_c = |\mathcal{Y}_{\tau(c)}|^{-1} \big(\bar{\boldsymbol{a}}_{\tau(c)}^\top \boldsymbol{h}(x) + \bar{b}_{\tau(c)} \big), \tag{3}$$

projects the representation onto data-independent directions. With a zero-initialized head, $\Pi_{\mathrm{T}} \boldsymbol{s}(x) = \boldsymbol{0}$ for every input.

Proofs of all statements are given in Section A. Proposition 1(i)–(ii) formalizes the shift invariance of a softmax restricted to one task, which also underlies separated softmax training (Ahn et al., 2021), and makes precise the statement that the training objective contains no term for TP: the loss (2) constrains only $\Pi_{\mathrm{W}} \boldsymbol{s}$, while TP is decided jointly by $\Pi_{\mathrm{W}} \boldsymbol{s}$ and a task-level component that training leaves undetermined. The component in Equation (3) still varies with x through $\boldsymbol{h}(x)$, which the modules shape, but only along directions $\bar{\boldsymbol{a}}_j$ that the objective never informs. Three remarks delimit the result. Without the mask and without old data, $\partial \ell / \partial s_c = p_c > 0$ for every old class c, so the gradient pushes the logits of old classes down at every step; this is the recency bias that Wu et al. (2019) and Zhao et al. (2020) correct post hoc. If weight decay acts on the whole head, the rows of earlier tasks, which receive no gradient, shrink geometrically in the number of later steps, which lowers their logits independently of the data. Adaptive optimizers such as Adam normalize each coordinate, so their updates no longer inherit the zero sums of Proposition 1(iii) and the recursion of Corollary 1 fails; the task-level component then changes, but through the optimizer rather than through the objective.

4 Orthogonal Selection

4.1 Definition and Guarantees

Definition 2 (Orthogonal selection)**.** *A task-selection term is a map $\boldsymbol{g} : \mathcal{X} \to \mathcal{V}_{\mathrm{T}}$, i.e., $g_c(x) = \gamma_{\tau(c)}(x)$ for task scores $\boldsymbol{\gamma}(x) \in \mathbb{R}^t$. It is orthogonal to the classifier when it is computed without any gradient path to or from the training objective. The selected logits are $\boldsymbol{s}'(x) = \boldsymbol{s}(x) + \beta\, \boldsymbol{g}(x)$ with $\beta \geq 0$.*

Proposition 2. *(i) For every $\boldsymbol{g}$ with values in $\mathcal{V}_{\mathrm{T}}$ and every $\beta \in \mathbb{R}$, WP computed from $\boldsymbol{s} + \beta \boldsymbol{g}$ equals WP computed from $\boldsymbol{s}$ for all (x, y). (ii) Conversely, if a correction $\boldsymbol{v} \in \mathbb{R}^C$ satisfies $\mathrm{WP}(\boldsymbol{s} + \boldsymbol{v}) = \mathrm{WP}(\boldsymbol{s})$ for all logits $\boldsymbol{s} \in \mathbb{R}^C$ and all labels, then $\boldsymbol{v} \in \mathcal{V}_{\mathrm{T}}$.*

Proposition 2 identifies $\mathcal{V}_{\mathrm{T}}$ as the set of additive corrections that are safe for WP without any knowledge of the classifier, and Proposition 1 identifies the same subspace as the one on which the masked objective is silent. Orthogonal selection therefore places the TP estimate exactly where training leaves the logits undetermined and cannot interfere with what training has learned.

Corollary 2 (Accuracy ceiling)**.** *For any orthogonal selection, $\mathrm{Acc}(\boldsymbol{s}') \leq \Pr(\mathrm{WP})$, where WP is that of the base logits $\boldsymbol{s}$, and $\mathrm{Acc}(\boldsymbol{s}') - \mathrm{Acc}(\boldsymbol{s}) = \Pr(\mathrm{WP} \wedge \mathrm{TP}') - \Pr(\mathrm{WP} \wedge \mathrm{TP}) \leq \Pr(\mathrm{WP} \wedge \neg \mathrm{TP})$, where TP' is computed from $\boldsymbol{s}'$.*

Lemma 2 (Task competition)**.** *Let $m_j(x) = \max_{c \in \mathcal{Y}_j} s_c(x)$. The prediction from $\boldsymbol{s}'$ lies in task $\arg\max_j \big[m_j(x) + \beta \gamma_j(x) \big]$, with ties resolved by the order $\prec$ of the block maximizers. If $\gamma_j = \log \pi_j$ for a task posterior $\boldsymbol{\pi}(x)$ with a unique maximizer $j^\star$, the selected task is $j^\star$ whenever $\beta > \big(\max_j m_j - \min_j m_j \big) / \big(\log \pi_{j^\star} - \max_{j \neq j^\star} \log \pi_j \big)$.*

By Lemma 2, β weighs the log task posterior against the block maxima of the logits, which by Corollary 1 carry task-level offsets that training does not control. In the idealized case $s_c = \log p(c \mid x, \tau(c)) + a_{\tau(c)}(x)$, the identity $\log p(c \mid x) = \log p(c \mid x, \tau(c)) + \log p(\tau(c) \mid x)$ shows that exact

CIL posteriors require the offsets $a_j(x)$ to equal $\log p(j \mid x)$ up to a constant shared by all tasks. With constant offsets, $\beta = 1$ and the true posterior would recover the Bayes rule; the selection strength $\beta = 0.3$ used below instead weighs an estimated log-posterior against offsets that training does not control, and makes no claim of calibration.

Relation to selection rules. Orthogonal selection differs from the selection rules of Section 3 in where it acts and how it is obtained. A selection rule $\kappa(\boldsymbol{z}; \psi)$ acts on the representation: it decides which modules process x, and its parameters ψ are fitted with objectives defined on the current task, such as key-query matching losses or the masked cross-entropy itself (Wang et al., 2022b; Smith et al., 2023). Its effect on the logits therefore mixes $\Pi_{\mathrm{T}}\boldsymbol{s}$ and $\Pi_{\mathrm{W}}\boldsymbol{s}$, and a wrong selection changes the within-task contrasts through the selected modules. Orthogonal selection acts on the logits only through $\Pi_{\mathrm{T}}\boldsymbol{s}$ and has no parameter that any loss can update, so it complements an existing selection rule rather than replacing it.

4.2 Class-Wise Low-Rank Gaussian Task Posterior

The proposed method (CLASP) estimates $p(j \mid x)$ from the frozen query feature $\boldsymbol{z} = f_\theta(x)$, which selection-based methods compute anyway, and sets $\gamma_j(x) = \log p(j \mid x)$.

Moment accumulation. For each class c, during the task in which it appears, the method accumulates the count n_c, the mean $\boldsymbol{\mu}_c$, and the scatter matrix $\boldsymbol{S}_c = \sum_{i:y_i=c}(\boldsymbol{z}_i - \boldsymbol{\mu}_c)(\boldsymbol{z}_i - \boldsymbol{\mu}_c)^\top$ from running first and second moments, and discards the samples. No sample of an earlier task is revisited, so the method is exemplar-free by construction; storing class-wise moments is common in exemplar-free CIL (Zhang et al., 2023; Goswami et al., 2023).

Shared whitening. The pooled within-class covariance over all classes observed so far is $\bar{\boldsymbol{\Sigma}} = \sum_c \boldsymbol{S}_c / \sum_c (n_c - 1)$. Since $\mathrm{rank}(\boldsymbol{S}_c) \le n_c - 1$ and every class in the benchmarks has fewer training samples than $d = 768$ (a median of 20 in ImageNet-A), $\bar{\boldsymbol{\Sigma}}$ is singular early in a sequence and ill-conditioned afterwards. The method therefore shrinks it toward a scaled identity with a fixed intensity, in the spirit of Ledoit & Wolf (2004), $\boldsymbol{P} = (1-\rho)\bar{\boldsymbol{\Sigma}} + \rho\, d^{-1}\mathrm{tr}(\bar{\boldsymbol{\Sigma}})\, \boldsymbol{I}$ with $\rho = 0.05$, and uses the whitening $\boldsymbol{W} = \boldsymbol{P}^{-1/2}$, a single $d \times d$ matrix shared by all classes.

Class-specific low-rank correction. In whitened coordinates, the class covariance $\boldsymbol{\Sigma}_c = \boldsymbol{S}_c/(n_c - 1)$ is shrunk toward the pooled one as in regularized discriminant analysis (Friedman, 1989), $\boldsymbol{Z}_c = \boldsymbol{W}\big(\alpha\boldsymbol{\Sigma}_c + (1-\alpha)\bar{\boldsymbol{\Sigma}}\big)\boldsymbol{W}$ with $\alpha = 0.1$. With the eigendecomposition $\boldsymbol{Z}_c = \sum_{k=1}^{d} \lambda_k \boldsymbol{u}_k \boldsymbol{u}_k^\top$, the method keeps the $r = 4$ eigenpairs with the largest $|\lambda_k - 1|$ in $\boldsymbol{U}_c \in \mathbb{R}^{d\times r}$ and $\boldsymbol{\Lambda}_c = \mathrm{diag}(\lambda_k)$, and assigns unit variance to the remaining $d - r$ directions.[1] Directions in which the class follows the pooled covariance map to eigenvalues of $\boldsymbol{W}\bar{\boldsymbol{\Sigma}}\boldsymbol{W}$, which equal $\sigma_i/((1-\rho)\sigma_i + \rho\, d^{-1}\mathrm{tr}\bar{\boldsymbol{\Sigma}})$ for the pooled eigenvalues σ_i and lie near 1 wherever the pooled variance dominates the ridge term. A large $|\lambda_k - 1|$ therefore marks a class-specific deviation from the pooled structure or a direction in which the ridge dominates. The resulting precision matrix is $\boldsymbol{\Omega}_c = \boldsymbol{W}\big(\boldsymbol{I} + \boldsymbol{U}_c(\boldsymbol{\Lambda}_c^{-1} - \boldsymbol{I})\boldsymbol{U}_c^\top\big)\boldsymbol{W}$.

Lemma 3. *Let $\lambda_k > 0$ for the retained eigenpairs, $\boldsymbol{e}_c = \boldsymbol{W}(\boldsymbol{z} - \boldsymbol{\mu}_c)$, and $\boldsymbol{p} = \boldsymbol{U}_c^\top \boldsymbol{e}_c$. Then*

$$-2\log\mathcal{N}\big(\boldsymbol{z}; \boldsymbol{\mu}_c, \boldsymbol{\Omega}_c^{-1}\big) = \|\boldsymbol{e}_c\|^2 + \sum_{k=1}^{r}\Big(\frac{p_k^2}{\lambda_k} - p_k^2 + \log\lambda_k\Big) + d\log 2\pi - 2\log\det\boldsymbol{W}, \qquad (4)$$

where the last two terms do not depend on c.

Why few directions. For $d > n$, the spiked covariance model limits what a sample covariance can estimate. Under isotropic unit-variance noise, a population variance ℓ produces a sample eigenvalue separated from the noise bulk, and a sample eigenvector with non-vanishing overlap with the

[1] All d eigenpairs are ranked by $|\lambda_k - 1|$, and the retained eigenvalues are clamped below at 10^{-8}, so $\lambda_k > 0$ in Lemma 3. The null space of $\boldsymbol{Z}_c$ is that of $\bar{\boldsymbol{\Sigma}}$, on which $\lambda_k = 0$ and $|\lambda_k - 1| = 1$, so such a direction is retained only when fewer than r directions deviate from the pooled structure by more than one.

population direction, only if $\ell > 1 + \sqrt{d/n}$; below this threshold the sample eigenvector is asymptotically uninformative, and noise alone produces sample eigenvalues up to $(1 + \sqrt{d/n})^2$ (Baik & Silverstein, 2006; Benaych-Georges & Nadakuditi, 2011). For $d = 768$ the threshold is about 3.8 at $n_c = 100$ and 7.2 at $n_c = 20$, so most eigen-directions of a class covariance are dominated by estimation noise, and a full class covariance carries this noise into the likelihood. Retaining a few directions limits the noise, although a fixed r does not guarantee that every retained direction exceeds the threshold.

Task posterior and selection. Each task is modeled as a mixture of its classes under a uniform class prior,

$$p(j \mid x) = \frac{\sum_{c \in \mathcal{Y}_j} \mathcal{N}(\boldsymbol{z}; \boldsymbol{\mu}_c, \boldsymbol{\Omega}_c^{-1})}{\sum_{c' \in \mathcal{Y}_{1:t}} \mathcal{N}(\boldsymbol{z}; \boldsymbol{\mu}_{c'}, \boldsymbol{\Omega}_{c'}^{-1})}, \qquad s'_c(x) = s_c(x) + \beta \log p\big(\tau(c) \mid x\big), \tag{5}$$

evaluated with log-sum-exp and $\beta = 0.3$. The added term lies in $\mathcal{V}_{\mathrm{T}}$ by construction, so Proposition 2 holds for every β. Modeling each class rather than each task is important: a task is a union of distinct classes and is multimodal in feature space, so a single Gaussian per task places its mass between the class modes (Section 5.5).

Cost. All statistics are computed without gradients and need no ViT pass beyond the frozen query pass, so inference uses two passes per image for any number of tasks, with a per-image overhead of $O(d^2 + Cd(r+1))$. Since $\boldsymbol{W}$ is recomputed as classes arrive, the accumulator carried across tasks keeps the class moments, $C(d^2 + d + 1)$ values (118M for $C = 200$), as do methods that store class covariances (Zhang et al., 2023; Goswami et al., 2023); inference itself keeps only $\boldsymbol{W}$ and, per class, $\boldsymbol{W}\boldsymbol{\mu}_c$, $\boldsymbol{U}_c$, and $\boldsymbol{\Lambda}_c$, i.e., $d^2 + C(d + rd + r)$ values (1.36M for $C = 200$), about $1/87$ of a full Gaussian per class. Algorithm 1 in Section D summarizes the procedure.

5 Experiments

5.1 Setup

Benchmarks. Five datasets are used: CIFAR-100 (Krizhevsky, 2009), ImageNet-R (Hendrycks et al., 2021a), ImageNet-A (Hendrycks et al., 2021b), CUB-200 (Wah et al., 2011), and VTAB (Zhai et al., 2019), following the benchmark of Zhou et al. (2025). Each dataset is split into tasks by halving the number of classes per task down to five, which gives 14 settings: CIFAR-100 with 5, 10, and 20 tasks; ImageNet-R, ImageNet-A, and CUB-200 with 10, 20, and 40 tasks; and VTAB with 5 and 10 tasks. Every setting is run with five seeds.

Protocol. All methods use ViT-B/16 pre-trained on ImageNet-21k (Dosovitskiy et al., 2021), frozen unless a method trains it by design. The original training budgets of the compared methods differ widely (e.g., 50 epochs for LAE, 20 for EASE, and 5 for the proposed method), so the main results use a common protocol: 5 epochs, batch size 64, SGD with momentum 0.9, a constant learning rate of 0.01, and weight decay 5×10^{-4}. The selected learning rate coincides with the one used by APER-SSF and PEARL in their original papers. The hyperparameters of the proposed method ($\rho = 0.05$, $\alpha = 0.1$, $r = 4$, $\beta = 0.3$) are shared by all settings. Detailed protocol described in (Section C). Because no single protocol is neutral for every method, the 910 runs of the 13 baselines were repeated under the original settings of each (Section E).

Methods and metrics. Fifteen methods are compared: L2P (Wang et al., 2022b), DualPrompt (Wang et al., 2022a), CODA-Prompt (Smith et al., 2023), HiDe-Prompt (Wang et al., 2023), PEARL (Qin et al., 2025), APT (Chen et al., 2025), LAE (Gao et al., 2023), EASE (Zhou et al., 2024b), SimpleCIL and APER with adapter (Houlsby et al., 2019), VPT (Jia et al., 2022), and SSF (Lian et al., 2022) modules (Zhou et al., 2025), MEMO (Zhou et al., 2023), the control, and the proposed method. The control is identical to the proposed method except that the task posterior is removed ($\beta = 0$); both use the prompt-based classifier described in Section F. MEMO is the only method that stores exemplars (20 per class) and trains the backbone. The metrics are the accuracy after the last task (Final), the average incremental accuracy (AIA) (Rebuffi et al., 2017), forgetting (Fgt) (Chaudhry et al., 2018), and the final WP and TP of Definition 1, all averaged over the 14 settings and five seeds unless stated otherwise; means and differences are computed before rounding.

Table 1: Results averaged over 14 settings and five seeds (%) under the common protocol. The control is the proposed method with $\beta = 0$. †Stores exemplars and trains the backbone. Standard deviations over seeds are reported for every setting in Section E.

Family	Method	Final ↑	AIA ↑	Fgt ↓	WP ↑	TP ↑
Prompt	L2P (Wang et al., 2022b)	53.23	64.06	11.28	90.67	54.44
	DualPrompt (Wang et al., 2022a)	62.59	72.29	11.12	93.02	63.62
	CODA-Prompt (Smith et al., 2023)	61.58	72.27	11.30	93.06	62.56
	HiDe-Prompt (Wang et al., 2023)	60.14	69.90	11.32	90.90	61.19
	PEARL (Qin et al., 2025)	63.85	74.06	11.50	93.54	64.81
	APT (Chen et al., 2025)	67.32	76.11	9.08	**94.45**	68.18
Adapter	LAE (Gao et al., 2023)	61.67	71.82	10.78	92.91	62.65
	EASE (Zhou et al., 2024b)	72.90	79.77	9.35	92.72	74.66
Prototype	SimpleCIL (Zhou et al., 2025)	72.92	79.75	7.48	92.20	74.86
	APER-Adapter (Zhou et al., 2025)	73.55	80.31	7.36	92.54	75.40
	APER-VPT (Zhou et al., 2025)	73.46	80.30	6.89	92.27	75.31
	APER-SSF (Zhou et al., 2025)	75.15	81.78	6.99	93.61	76.85
Exemplar	MEMO† (Zhou et al., 2023)	43.45	52.17	25.19	70.87	47.16
Proposed	Control ($\beta = 0$)	62.51	72.69	11.16	93.15	63.50
	CLASP (proposed)	**78.25**	**81.80**	**5.83**	93.15	**80.20**

5.2 Verification: Where the Errors Occur

The 1,050 runs locate the errors of every existing method in TP (Table 1). By Lemma 1, the marginal accuracies determine the joint events: of the $100 - \mathrm{Final}$ points of error, $\mathrm{TP} - \mathrm{Final}$ are WP failures with a correct task, $b = (100 - \mathrm{WP}) - (\mathrm{TP} - \mathrm{Final})$ are joint failures, and $(100 - \mathrm{TP}) - b$ are TP failures with a correct WP. Apart from MEMO, TP fails while WP holds in 70.9–83.0% of the errors, both fail in 14.3–22.2%, and WP alone fails in only 2.6–7.2%; TP therefore fails in 92.8–97.4% of the errors of every baseline, MEMO included, and of the control. The WP loss with a correct task, $\mathrm{TP} - \mathrm{Final} = \Pr(\mathrm{TP} \wedge \neg \mathrm{WP})$, varies between 0.86 and 1.94 points across the baselines other than MEMO, whereas TP varies over 22.4 points (54.4–76.9). The ranking of the 13 baselines by final accuracy therefore coincides with their ranking by TP (Spearman $\rho = 1.00$) and not with their ranking by WP ($\rho = 0.43$, or 0.27 without MEMO). WP itself varies little: over the 12 baselines other than MEMO, which trains the backbone and is the only method whose WP collapses (70.87), the sample standard deviation is 1.07 points for WP and 7.36 points for TP. These statistics locate the errors but do not isolate the mechanism of Section 3: prototype heads, to which Proposition 1 does not apply, show the same pattern, and TP also becomes harder as the number of tasks grows, independently of the objective. The controlled test, in which only the task-level term changes, is Section 5.4.

The same structure appears within a sequence (Figure 2): on ImageNet-R with 20 tasks, the WP curves of the five methods shown stay within a narrow band while TP falls with every new task. Across sequence lengths, the average TP drops from 79.6% to 57.4% as the number of tasks grows from 5 to 40, while the average WP stays between 89.6% and 93.3%, and TP falls with every doubling within each dataset as well (Tables 17 and 18).

5.3 Main Results

Table 1 compares the methods under the common protocol. Among the methods reported in Table 1, the proposed method attains the highest final accuracy (78.25%), the highest TP (80.20%), and the lowest forgetting (5.83), and it ranks first in 12 of the 14 settings, with a mean rank of 1.2 (Section E). The strongest baselines, the three APER variants, SimpleCIL, and EASE, classify with prototypes computed on frozen, first-task-adapted, or task-specific features and reach TP above 74%. A prototype head is computed from class means rather than trained with (2), so Proposition 1 does not apply to it, and its task-level component carries information about all classes observed so far. The prompt-based methods reach WP above 90% but remain between 54.4% and 68.2% in TP, so their differences in final accuracy follow their differences in TP. The proposed method adds to

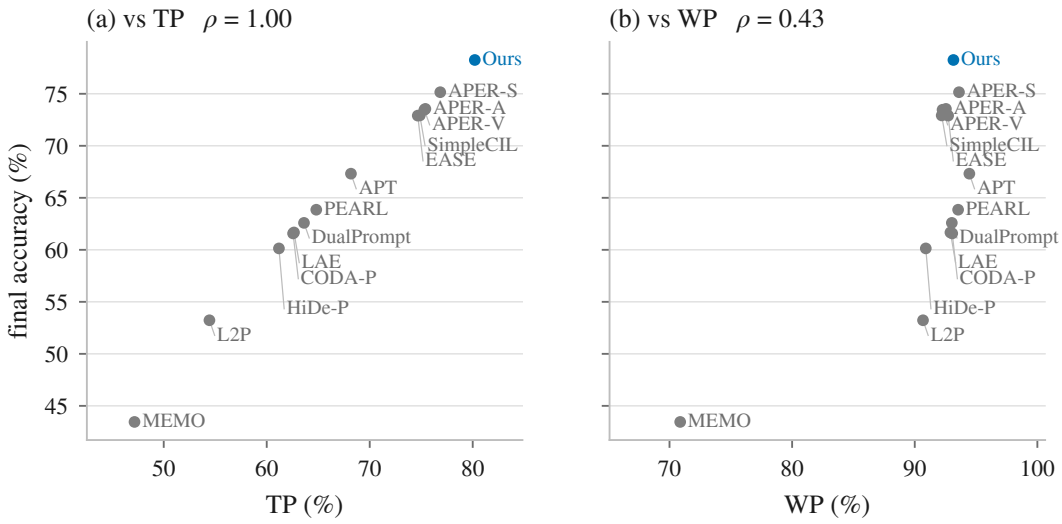


Figure 1: Final accuracy against TP (a) and WP (b) for the 14 methods of Table 1 other than the control. The panels share the vertical axis, so the names in (a) identify the points of (b); Spearman ρ is computed over the 13 baselines (Section 5.2).

Table 2: Final accuracy by dataset (%), each column averaged over the task splits of that dataset and five seeds; the last column repeats the Final column of Table 1. The best value of each column is in **bold**, excluding the control. Per-setting values are in Section E.

Method	CIFAR-100	ImageNet-R	ImageNet-A	CUB-200	VTAB	All 14
L2P	75.76	59.69	31.76	47.72	50.20	53.23
DualPrompt	79.77	64.54	49.82	58.47	59.24	62.59
CODA-Prompt	77.86	64.40	49.62	58.25	55.83	61.58
HiDe-Prompt	78.01	64.88	43.13	54.25	60.55	60.14
PEARL	81.32	66.61	50.35	60.56	58.73	63.85
APT	84.44	70.88	56.44	61.25	61.72	67.32
LAE	80.59	64.81	48.62	55.74	57.04	61.67
EASE	83.71	64.82	56.51	80.44	82.08	72.90
SimpleCIL	80.47	64.49	58.54	81.98	82.23	72.92
APER-Adapter	83.17	64.72	58.54	81.98	82.22	73.55
APER-VPT	82.85	67.56	54.51	82.25	83.48	73.46
APER-SSF	83.84	70.09	57.98	82.51	84.44	75.15
MEMO†	51.83	53.71	14.67	67.91	22.00	43.45
Control ($\beta = 0$)	79.02	65.36	48.73	60.56	57.09	62.51
CLASP (proposed)	**85.83**	**72.13**	**60.36**	**87.06**	**89.69**	**78.25**

the prompt-based control a task-level term built from second-order class statistics and exceeds the prototype-based methods in TP (80.20% against 76.85% for APER-SSF). It is also first on every dataset (Table 2), by margins from 1.25 points on ImageNet-R to 5.25 points on VTAB.

5.4 Only TP Moves

Adding the task posterior to the control changes TP alone (Table 3). WP is unchanged up to rounding: within a run, WP before and after adding the task posterior is identical by Proposition 2, and between the separate control and CLASP runs the largest difference in any setting is 0.003 points, which is attributable to nondeterminism; TP rises by 16.70 points and final accuracy by 15.74 points, with an improvement in all 14 settings. With every baseline trained under its original settings instead, the proposed method, whose own setting is the common protocol, remains first (78.25% against 75.68% for APER-Adapter, Table 19), so its lead is not an artifact of the protocol. Relative to the ceiling of Corollary 2, the control leaves 30.64 points between its WP and its final accuracy,

Table 3: Effect of the task posterior (%). Δ is the difference of the means over 14 settings and five seeds.

	WP	TP	Final
Control	93.15	63.50	62.51
CLASP	93.15	80.20	78.25
Δ	0.00	+16.70	+15.74

Table 4: Final-accuracy gain of CLASP over the control by dataset and number of tasks (points).

Dataset	5	10	20	40
CIFAR-100	+3.91	+6.03	+10.49	–
ImageNet-R	–	+3.81	+6.71	+9.79
ImageNet-A	–	+7.40	+10.88	+16.60
CUB-200	–	+17.81	+26.51	+35.16
VTAB	+20.87	+44.32	–	–

and orthogonal selection closes 51.4% of this gap. The gain increases monotonically with the number of tasks on every dataset (Table 4), consistent with the concentration of errors in TP for longer sequences in Table 17.

5.5 Analysis of Experiments

Class mixture versus task Gaussian. Replacing the class mixture with one Gaussian per task reduces the TP gain from 7.49 to 4.61 points and the final-accuracy gain from 6.77 to 4.13 (Table 21). One mean and covariance per task place mass between the class modes, whereas the class mixture keeps the modes apart and separates the tasks more sharply.

Rank. The final-accuracy gain varies by 0.36 points over $r \in \{1, \dots, 16\}$, and on the 20-task split, the only one run with the full class covariance, it is 1.36 points higher at $r = 4$ than at $r = d$ (Tables 4 and 21). This is consistent with the spiked-covariance argument of Section 4.2: beyond the leading directions, the sample eigenvectors of a class covariance estimated from about a hundred images in $d = 768$ dimensions carry estimation noise, so retaining them lowers the accuracy. Between $r = 4$ and $r = 16$ the choice is a storage trade-off: $r = 16$ adds 0.25 points and multiplies the statistics kept at inference by about 2.4.

Selection strength. By Proposition 2, no finite β can change WP, and Table 21 confirms $\Delta\mathrm{WP} = +0.00$ for every one of them. The final-accuracy gain rises from 3.14 points at $\beta = 0.03$ to 6.77 at $\beta = 0.3$ and changes by 0.21 points up to $\beta = 1$. The only setting that moves WP is $\beta \to \infty$, which selects the maximum a posteriori task first (Lemma 2) and masks every other task. This correction is not of the form $\beta \boldsymbol{g}$ with $\boldsymbol{g} \in \mathcal{V}_{\mathrm{T}}$, so Proposition 2 no longer applies: WP falls by 15.49 points, and the gain drops to 5.96.

Cost and other classifiers. Inference stays at two ViT passes per image for any number of tasks, whereas adapter-expansion methods such as EASE evaluate one adapter per task and need up to $T + 1$ passes (41 at $T = 40$). The task posterior also applies to other classifiers without retraining them: on ImageNet-R with 20 tasks, it raises the final accuracy of L2P, DualPrompt, CODA-Prompt, and HiDe-Prompt by 11.45, 7.77, 6.53, and 6.59 points with WP unchanged (Table 20).

6 Conclusion

This paper showed that the masked cross-entropy of rehearsal-free CIL is blind to the task-level subspace of the logits, that this subspace is exactly the set of WP-preserving additive corrections, and that supplying TP there through a closed-form, class-wise Gaussian task posterior raises final accuracy by 15.74 points with WP unchanged. Corollary 2 also sets the direction of future work: the 14.90 points between the final accuracy of the proposed method and the WP of its classifier can be narrowed by sharper task posteriors with WP fixed, and the ceiling itself rises with the WP of the underlying classifier; Lemma 2 further suggests replacing the fixed β with an input-dependent selection strength.

AI USE STATEMENT

Large language models were used solely for grammar correction, final proofreading (checking typos), and verifying notation consistency across the manuscript. The authors verified all text, derivations, code, and results and take full responsibility for the content of the paper.

ETHICS STATEMENT

The authors have read and adhere to the ICLR Code of Ethics. This work involves no human subjects and collects no new data; all experiments use publicly available research benchmarks, CIFAR-100, ImageNet-R, ImageNet-A, CUB-200, and VTAB, under their respective terms of use. The proposed method retains no training sample: across tasks it keeps only class-level statistics of frozen features (counts, means, and second-moment sums), and it requires no additional training stage. As in every compared method, the frozen ViT-B/16 backbone is pre-trained on ImageNet-21k, so biases present in its features carry over to the predictions. The work concerns a general-purpose classification method and does not target an application with a foreseeable direct harm.

REPRODUCIBILITY STATEMENT

Proofs of all statements are given in Section A, the procedure in Algorithm 1, and the training protocol, the protocol search, the hyperparameter audit, and every per-method deviation in Section C. All runs use five fixed seeds that also fix the class order. The code for both protocols (1,960 runs) and the scripts that generate every table and figure are provided as anonymous supplementary material at `https://anonymous.4open.science/r/CLASP_ICLR`.

REFERENCES

Hongjoon Ahn, Jihwan Kwak, Subin Lim, Hyeonsu Bang, Hyojun Kim, and Taesup Moon. Ss-il: Separated softmax for incremental learning. In *2021 IEEE/CVF international conference on computer vision (ICCV)*, pp. 824–833. IEEE, 2021.

Jinho Baik and Jack W Silverstein. Eigenvalues of large sample covariance matrices of spiked population models. *Journal of multivariate analysis*, 97(6):1382–1408, 2006.

Jinho Baik, Gérard Ben Arous, and Sandrine Péché. Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *The Annals of Probability*, 33(5):1643–1697, 2005.

Florent Benaych-Georges and Raj Rao Nadakuditi. The eigenvalues and eigenvectors of finite, low rank perturbations of large random matrices. *Advances in Mathematics*, 227(1):494–521, 2011.

Arslan Chaudhry, Puneet K Dokania, Thalaiyasingam Ajanthan, and Philip HS Torr. Riemannian walk for incremental learning: Understanding forgetting and intransigence. In *European conference on computer vision*, pp. 556–572. Springer, 2018.

Haoran Chen, Ping Wang, Zihan Zhou, Xu Zhang, Zuxuan Wu, and Yu-Gang Jiang. Achieving more with less: Additive prompt tuning for rehearsal-free class-incremental learning. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 340–349. IEEE, 2025.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=YicbFdNTTy`.

Jerome H Friedman. Regularized discriminant analysis. *Journal of the American statistical association*, 84(405):165–175, 1989.

Qiankun Gao, Chen Zhao, Yifan Sun, Teng Xi, Gang Zhang, Bernard Ghanem, and Jian Zhang. A unified continual learning framework with general parameter-efficient tuning. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11449–11459. IEEE, 2023.

Zhanxin Gao, Jun Cen, and Xiaobin Chang. Consistent prompting for rehearsal-free continual learning. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28463–28473. IEEE, 2024.

Dipam Goswami, Yuyang Liu, Bartłomiej Twardowski, and Joost Van De Weijer. Fecam: Exploiting the heterogeneity of class distributions in exemplar-free continual learning. *Advances in Neural Information Processing Systems*, 36:6582–6595, 2023.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8320–8329. IEEE, 2021a.

Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15262–15271, 2021b.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pp. 2790–2799. PMLR, 2019.

Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European conference on computer vision*, pp. 709–727. Springer, 2022.

Li Jiao, Qiuxia Lai, Yu Li, and Qiang Xu. Vector quantization prompting for continual learning. *Advances in Neural Information Processing Systems*, 37:34056–34076, 2024.

Gyuhak Kim, Changnan Xiao, Tatsuya Konishi, Zixuan Ke, and Bing Liu. A theoretical study on solving continual learning. *Advances in neural information processing systems*, 35:5065–5079, 2022.

Gyuhak Kim, Changnan Xiao, Tatsuya Konishi, and Bing Liu. Learnability and algorithm for continual learning. In *International conference on machine learning*, pp. 16877–16896. PMLR, 2023.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Olivier Ledoit and Michael Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of multivariate analysis*, 88(2):365–411, 2004.

Dongze Lian, Daquan Zhou, Jiashi Feng, and Xinchao Wang. Scaling & shifting your features: A new baseline for efficient model tuning. *Advances in Neural Information Processing Systems*, 35: 109–123, 2022.

Yan-Shuo Liang and Wu-Jun Li. Inflora: Interference-free low-rank adaptation for continual learning. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23638–23647. IEEE, 2024.

Haowei Lin, Yijia Shao, Weinan Qian, Ningxin Pan, Yiduo Guo, and Bing Liu. Class incremental learning via likelihood ratio based task prediction. In *International Conference on Learning Representations*, volume 2024, pp. 50989–51028, 2024.

Mark D McDonnell, Dong Gong, Amin Parvaneh, Ehsan Abbasnejad, and Anton Van den Hengel. Ranpac: Random projections and pre-trained models for continual learning. *Advances in Neural Information Processing Systems*, 36:12022–12053, 2023.

Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=37nvvqkCo5`.

Debashis Paul. Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statistica Sinica*, pp. 1617–1642, 2007.

Yongchun Qin, Pengfei Fang, and Hui Xue. Pearl: Input-agnostic prompt enhancement with negative feedback regulation for class-incremental learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 20051–20059, 2025.

Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 5533–5542. IEEE, 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 *2023 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 11909–11919. IEEE, 2023.

Hai-Long Sun, Da-Wei Zhou, Hanbin Zhao, Le Gan, De-Chuan Zhan, and Han-Jia Ye. Mos: Model surgery for pre-trained model-based class-incremental learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 39, pp. 20699–20707, 2025.

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

Huiyi Wang, Haodong Lu, Lina Yao, and Dong Gong. Self-expansion of pre-trained models with mixture of adapters for continual learning. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10087–10098. IEEE, 2025a.

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

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

Yan Wang, Da-Wei Zhou, and Han-Jia Ye. Integrating task-specific and universal adapters for pre-trained model-based class-incremental learning. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 806–816. IEEE, 2025b.

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

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 *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 139–149. IEEE, 2022b.

Yue Wu, Yinpeng Chen, Lijuan Wang, Yuancheng Ye, Zicheng Liu, Yandong Guo, and Yun Fu. Large scale incremental learning. In *2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 374–382. IEEE, 2019.

Xiaohua Zhai, Joan Puigcerver, Alexander Kolesnikov, Pierre Ruyssen, Carlos Riquelme, Mario Lucic, Josip Djolonga, Andre Susano Pinto, Maxim Neumann, Alexey Dosovitskiy, et al. A large-scale study of representation learning with the visual task adaptation benchmark. *arXiv preprint arXiv:1910.04867*, 2019.

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 *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 19091–19101. IEEE, 2023.

Bowen Zhao, Xi Xiao, Guojun Gan, Bin Zhang, and Shu-Tao Xia. Maintaining discrimination and fairness in class incremental learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13205–13214. IEEE, 2020.

Da-Wei Zhou, Qi-Wei Wang, Han-Jia Ye, and De-Chuan Zhan. A model or 603 exemplars: Towards memory-efficient class-incremental learning. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=S07feAlQHgM.

Da-Wei Zhou, Hai-Long Sun, Jingyi Ning, Han-Jia Ye, and De-Chuan Zhan. Continual learning with pre-trained models: a survey. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, IJCAI '24, 2024a. ISBN 978-1-956792-04-1. doi: 10.24963/ijcai.2024/924. URL https://doi.org/10.24963/ijcai.2024/924.

Da-Wei Zhou, Hai-Long Sun, Han-Jia Ye, and De-Chuan Zhan. Expandable subspace ensemble for pre-trained model-based class-incremental learning. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23554–23564. IEEE, 2024b.

Da-Wei Zhou, Zi-Wen Cai, Han-Jia Ye, De-Chuan Zhan, and Ziwei Liu. Revisiting class-incremental learning with pre-trained models: Generalizability and adaptivity are all you need. *International Journal of Computer Vision*, 133(3):1012–1032, 2025.