How to Find and Reuse Policies for Continuous Adaptation in Lifelong Reinforcement Learning

Anonymous authors
Paper under double-blind review

Abstract

In lifelong reinforcement learning, retaining previously learned policies is not sufficient for effective transfer to a new task. Useful knowledge may be distributed across several prior policies, and its relevance may change as the learner acquires experience. One hypothesis is that task similarity can be effectively used in a continual learning setting to find and combine previously learned policies. To test it, Adaptive Mask Selection and Composition (AMSC) is designed to estimate similarity from online experience via non-parametric Wasserstein task embeddings from state-action-reward samples. The z-score-normalized sparsemax of the similarity scores are used to derive a variable-size support to periodically choose and weight policies to form a prior when learning a new task. On CT-graph and MiniGrid, AMSC achieves higher mean performance and forward transfer than the evaluated modular composition baselines while exhibiting no forgetting. Results on Continual World suggest that identifying relevant prior knowledge and determining its layer-specific composition may require additional layer-specific tuning. Ablations show that selecting relevant sources and determining how strongly to reuse them are central to these gains. Independently measured pairwise transfer is also positively associated with task-embedding similarity. These results indicate that task similarity can be an effective criterion to select and weight specific knowledge for reuse in lifelong reinforcement learning.

1 Introduction

Lifelong learning requires both retention and reuse of prior knowledge. Although forgetting remains a central challenge, retention alone does not ensure positive transfer to later tasks. As experience accumulates, a model must identify relevant knowledge without allowing irrelevant or conflicting knowledge to cause interference. Prior work has extensively studied retention and increasingly evaluates forward transfer (Rusu et al., 2016; Kirkpatrick et al., 2017; Zenke et al., 2017; Mallya & Lazebnik, 2018; Rolnick et al., 2019; von Oswald et al., 2020), but commonly reuses shared or aggregated knowledge. Selective reuse remains comparatively underexplored.

This distinction is especially consequential in reinforcement learning (RL). Transferred knowledge affects the actions taken and the state-action distribution from which the policy learns. An unsuitable prior can bias exploration and induce negative transfer (Taylor & Stone, 2009; Ahn et al., 2025; Uchendu et al., 2023). Sparse rewards provide little early evidence that a policy is harmful (Gimelfarb et al., 2021). Moreover, useful structure may be distributed unequally across several priors.

Existing lifelong RL methods transfer through replay, modular routing, parameter isolation, and learned composition (Isele & Cosgun, 2018; Kessler et al., 2022; Powers et al., 2022; Mendez et al., 2022; Ben-Iwhiwhu et al., 2023; Zhang et al., 2024; Malagón et al., 2024; Hu et al., 2025; Huang et al., 2025). Modular approaches can preserve completed policies, but reusing the full library may admit interfering priors. Learning how strongly to reuse each policy from the target objective can identify useful policies but may converge only after their opportunity to improve early learning has passed.

Measured task similarity can instead act as a method of estimating the relevance of prior policies without optimization (Carroll & Seppi, 2005; Taylor & Stone, 2009; Li et al., 2019; Gimelfarb et al.,

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.

2021; Wu et al., 2025; Nath et al., 2026). In task-incremental lifelong RL, the availability of prior policies is determined by curriculum history, completed policies are frozen, and the number of useful sources is unknown. A suitable selector must therefore adapt to the current library. As experience is collected on a new task, both the selected knowledge and its contribution to learning may also change. We ask whether online task similarity can determine both *which* policies to reuse and *how strongly* they should contribute under these constraints.

Adaptive Mask Selection and Composition (AMSC) learns task-specific policy masks over a fixed backbone network. It uses online task similarity to select a variable-size support of prior masks and determine their composition weights. Selected and current mask scores are normalized before composition with an ℓ_2 straight-through estimator, and only the current policy-mask scores are optimized. AMSC selects without needing task descriptions, task labels, or transfer probes.

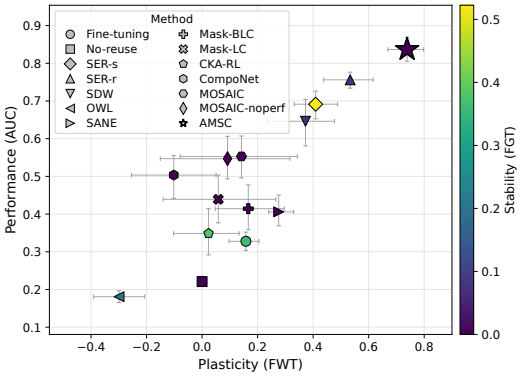

Figure 1: Transfer-retention trade-off on CT28 and MG16. Each marker reports mean FWT and AUC across seven seeds; colour denotes FGT. Error bars are 95% confidence intervals.

AMSC is evaluated on CT-graph (CT28), MiniGrid (MG16), and Continual World (CW10). It obtains AUC/FWT/FGT of $0.94/0.94/0.00$ on CT28 and $0.73/0.54/0.00$ on MG16 (Figure 1). In CT28 and MG16, similarity-derived weights perform comparably to learned composition weights. In CW10, similarity-derived weights underperform with respect to learned mask composition. Learning composition over a sparse selection of prior policies may provide a performance benefit. Ablations show that selecting relevant policies from the agent's library and determining how to reuse each through similarity or learned weights are critical to more effective forward transfer and successful learning in lifelong RL. Results also show that task similarity is positively associated with independently measured pairwise transfer. Lastly, results suggest that masks can be quantized with minimal forgetting and loss in AUC and FWT.

2 Related Work

Selective reuse in lifelong reinforcement learning. Forward transfer and forgetting are distinct objectives in lifelong RL. Preserving earlier tasks does not imply faster learning on later tasks (Wolczyk et al., 2021; 2022). Prior work promotes selective reuse through sparse or compositional subnetworks (Sokar et al., 2022; Ostapenko et al., 2021), related and unrelated task sequences (Ke et al., 2020), selective replay (Isele & Cosgun, 2018; Huang et al., 2025), and adaptive structure allocation (Lee et al., 2021; Dekhovich et al., 2023; Savadikar et al., 2026). These methods select data, capacity, or structures, often to determine what knowledge is worth retaining.

Modular routing and composition. Modular lifelong RL isolates task-specific knowledge and transfers through routing or composition. Prior methods activate modules, select policy heads, infer contexts, or route adapters (Powers et al., 2022; Kessler et al., 2020; Zhang et al., 2024; Julian et al., 2025); others compose masks, modules, policies, or task vectors (Ben-Iwhiwhu et al., 2023; Kessler et al., 2022; Hu et al., 2025; Malagón et al., 2024).

Task similarity and optimal transport. Task similarity predicts transfer and supports policy selection in RL (Carroll & Seppi, 2005; Taylor & Stone, 2009; Li et al., 2019; Gimelfarb et al., 2021), while transferability methods use representations, probe networks, likelihoods, or parameter subspaces (Zamir et al., 2018; Achille et al., 2019; Nguyen et al., 2020; Lin et al., 2022; Wu et al., 2025). Optimal transport compares task distributions through OTDD and Wasserstein embeddings (Courty et al., 2018; Alvarez-Melis & Fusi, 2020; Liu et al., 2025), and supports memory, selection, distillation, context detection, and expert diversity (Wang et al., 2022; 2024; Xu et al., 2023; Dick et al., 2024; Hihn & Braun, 2022).

3 PRELIMINARIES

3.1 PROBLEM SETTING

Consider a task-incremental lifelong RL setting with a sequence of K tasks $\{\mathcal{T}_k\}_{k=1}^{K}$. Each task is an MDP $\mathcal{T}_k = (\mathbb{S}, \mathbb{A}, p_k, r_k, \gamma_{\mathrm{rl}})$, where $\mathbb{S}$ and $\mathbb{A}$ are the state and action spaces, p_k is the transition kernel, r_k is the reward function, and γ_{rl} is the discount factor. Tasks arrive sequentially with known boundaries, and only the current task environment is available for interaction. The learner trains on each task for a fixed budget before proceeding to the next task.

A lifelong learning agent aims to maintain performance on previously learned tasks, or improve it through backward transfer, while continuing to learn new tasks through forward transfer. Forward transfer measures whether knowledge acquired from earlier tasks accelerates learning of a new task (Parisi et al., 2019; Khetarpal et al., 2022; Wolczyk et al., 2021).

3.2 MODULAR POLICY MASKS

Score-based masking learns task-specific subnetworks over a shared, fixed backbone. Supermasks in Superposition stores multiple tasks as binary masks over the same weights (Wortsman et al., 2020). Modulating Masks adapts this mechanism to lifelong RL and introduces learned linear combinations of the current and completed task masks (Ben-Iwhiwhu et al., 2023). CKA-RL is a related modular composition method that combines and merges dense task-specific residuals rather than binary masks (Hu et al., 2025).

Let $\{\boldsymbol{W}^{\ell}\}_{\ell=1}^{L}$ be the fixed policy-backbone matrices and $\boldsymbol{S}_k^{\ell}$ the real-valued score matrix for task k at masked layer ℓ. The mask in the forward pass is

$$\boldsymbol{M}_k^{\ell} = g(\boldsymbol{S}_k^{\ell}), \qquad g(x) = \mathbf{1}_{\{x \geq 0\}}, \tag{1}$$

where g is applied elementwise. The effective parameters are $\boldsymbol{W}^{\ell} \odot \boldsymbol{M}_k^{\ell}$. A straight-through estimator propagates gradients through g. Let $\boldsymbol{S}_i^{\ell,*}$ be the frozen score matrix stored after task i, and let $\mathbb{P}(k) \subseteq \{1, \ldots, k-1\}$ contain the priors admitted for task k. A generic convex score composition is

$$\boldsymbol{S}_k^{\ell,\mathrm{comb}} = \sum_{i \in \mathbb{P}(k)} \alpha_{k,i}^{\ell} \boldsymbol{S}_i^{\ell,*} + \alpha_{k,k}^{\ell} \boldsymbol{S}_k^{\ell}, \qquad \sum_{i \in \mathbb{P}(k) \cup \{k\}} \alpha_{k,i}^{\ell} = 1, \tag{2}$$

with non-negative weights. Ben-Iwhiwhu et al. (2023); Hu et al. (2025) use the complete prior library and learn layer-specific weights. For $m > 0$, Ben-Iwhiwhu et al. (2023) initializes the coefficients uniformly in Mask-LC or balances prior and current-mask mass in Mask-BLC. For $m = |\mathbb{P}(k)| > 0$,

$$\begin{aligned} \text{Mask-LC:} \quad & \alpha_{k,k}^{\ell,(0)} = \alpha_{k,i}^{\ell,(0)} = \frac{1}{m+1}, \\ \text{Mask-BLC:} \quad & \alpha_{k,k}^{\ell,(0)} = \frac{1}{2}, \qquad \alpha_{k,i}^{\ell,(0)} = \frac{1}{2m}, \qquad i \in \mathbb{P}(k). \end{aligned} \tag{3}$$

When $m = 0$, the current mask has an initial weight of 1. Prior scores remain frozen, and the final composition is consolidated, $\boldsymbol{S}_k^{\ell,*} \leftarrow \boldsymbol{S}_k^{\ell,\mathrm{comb}}$, for later tasks.

4 ADAPTIVE MASK SELECTION AND COMPOSITION

AMSC consists of three operations: normalized sparsemax retrieval, scale-normalized mask scores, and composition from the retrieval weights.

4.1 MASK LIBRARY AND ONLINE TASK REPRESENTATIONS

At task k, AMSC initializes a new current score matrix $\boldsymbol{S}_k^{\ell}$ and the policy library contains the frozen score matrices $\{\boldsymbol{S}_i^{\ell,*}\}_{i<k,\ell=1}^{L}$ following the formulation defined in Section 3.2. Let $\mathbb{D}_k$ be the current-task experience buffer and $\boldsymbol{X}_k^{(t)}$ a sampled batch matrix. AMSC uses an embedding

map $\psi(\boldsymbol{X}_k^{(t)}) = \boldsymbol{e}_k^{(t)} \in \mathbb{R}^{d_e}$ that places online experience from different tasks in a common vector space. The main experiments use a non-parametric linear Wasserstein embedding (WTE) of state–action–reward samples (Liu et al., 2025) which has been demonstrated in RL tasks (Nath et al., 2026).

Discrete actions are one-hot encoded, states are standardized within each sampled batch, and rewards are retained as observed. Given a fixed reference matrix $\boldsymbol{X}_0$ with R samples and an optimal transport matrix $\boldsymbol{\Gamma}^\star$, the WTE batch embedding is

$$\boldsymbol{e}_k^{(t)} = \mathrm{vec}\left(\frac{(R\boldsymbol{\Gamma}^\star)^\top \boldsymbol{X}_k^{(t)} - \boldsymbol{X}_0}{\sqrt{R}}\right). \tag{4}$$

The current representation is smoothed online as

$$\bar{\boldsymbol{e}}_k^{(t)} = \mathrm{normalize}\Big(\rho \bar{\boldsymbol{e}}_k^{(t-1)} + (1-\rho)\boldsymbol{e}_k^{(t)}\Big), \tag{5}$$

where the first update initializes the normalized representation and $\rho = 0.5$ thereafter. Completed task representations remain fixed. Appendix B.1 formalizes the optimal transport plan and the empirical distributions represented by $\boldsymbol{X}_k^{(t)}$ and $\boldsymbol{X}_0$.

4.2 Sparse Retrieval

Cosine similarity has been used to evaluate task relationships through gradient alignment, and task embeddings (Lopez-Paz & Ranzato, 2017; Achille et al., 2019; Yu et al., 2020a; Nath et al., 2026). Following this, AMSC compares the current representation of task k with every completed task using cosine similarity,

$$s_{k,i} = \cos(\bar{\boldsymbol{e}}_k, \bar{\boldsymbol{e}}_i), \qquad i < k. \tag{6}$$

Let $\boldsymbol{s}_k = [s_{k,1}, \ldots, s_{k,k-1}]^\top$. AMSC retrieves

$$\boldsymbol{w}_k = \mathrm{sparsemax}(\mathrm{zscore}(\boldsymbol{s}_k)/\tau), \qquad \mathbb{P}(k) = \{i < k : w_{k,i} > 0\}. \tag{7}$$

The z-score has zero mean and unit variance within the available library, and $\tau = 1$ in all main experiments. Sparsemax projects these relative scores onto the probability simplex and assigns exact zeros, so $\sum_{i\in\mathbb{P}(k)} w_{k,i} = 1$ (Martins & Astudillo, 2016).

4.3 Scale-Normalized Fixed Composition

Positive rescaling of a task's scores changes its influence when scores from several tasks are combined before thresholding. Large magnitude scores become more difficult to alter. Consequently, raw score magnitude can override the linear combination weights. We therefore normalize each selected prior and current score to unit ℓ_2 norm in the forward pass, using a straight-through estimator to preserve the current score's gradient,

$$\widehat{\boldsymbol{S}}_i^\ell = \boldsymbol{S}_i^\ell + \mathrm{sg}\left(\frac{\boldsymbol{S}_i^\ell}{\max(\|\boldsymbol{S}_i^\ell\|_F, \varepsilon_m)} - \boldsymbol{S}_i^\ell\right), \tag{8}$$

where sg denotes stop-gradient.

Learning composition weights from the RL objective requires the agent to discover useful sources through task interaction. AMSC instead can use the same online similarity signal for both retrieval and weighting, making source contributions available at the first retrieval event without optimization. Let $m = |\mathbb{P}(k)|$. Instead of learning the composition, AMSC applies weights computed from task similarity which,

$$\boldsymbol{C}_k^\ell = \begin{cases} \widehat{\boldsymbol{S}}_k^\ell, & m = 0, \\ \dfrac{1}{m+1}\widehat{\boldsymbol{S}}_k^\ell + \dfrac{m}{m+1}\displaystyle\sum_{i\in\mathbb{P}(k)} w_{k,i}\widehat{\boldsymbol{S}}_i^\ell, & m > 0, \end{cases} \qquad \boldsymbol{M}_k^\ell = g(\boldsymbol{C}_k^\ell). \tag{9}$$

For $m > 0$, the current mask receives mass $1/(m+1)$, while the selected priors share $m/(m+1)$ according to their retrieval weights. These weights are shared across layers.

Each task begins without retrieved priors. Once sufficient experience is available, AMSC updates the current embedding and recomputes equation 7–equation 9 between training iterations to reflect new data. At the task boundary, the final composed scores $\boldsymbol{C}_k^\ell$ are consolidated as $\boldsymbol{S}_k^{\ell,*} \leftarrow \boldsymbol{C}_k^\ell$, frozen, and added to the reusable mask library. Appendix B.2 Algorithm 1 gives the complete training pseudo-code.

5 RESULTS

5.1 EXPERIMENTAL SETUP

Benchmarks. We evaluate AMSC on CT28, a sparse-reward decision-tree curriculum with hierarchical tasks (Soltoggio et al., 2023); MG16, sparse-reward MiniGrid navigation tasks with shuffled observations (Chevalier-Boisvert et al., 2018); and CW10, Continual World continuous robotic control tasks (Wolczyk et al., 2021). In CT28 and MG16, tasks within each family share a common observation space and increase in complexity. Families differ in their representations while sharing input dimensions and actions. These curricula test reuse when prior knowledge varies in relevance in relation to the next task. Details and pairwise-transfer matrices are reported in Appendix C.

Baselines. We compare against replay methods (SER (Isele & Cosgun, 2018) and SDW (Huang et al., 2025)), modular methods (Mask-LC/BLC (Ben-Iwhiwhu et al., 2023), CKA-RL (Hu et al., 2025), CompoNet (Malagón et al., 2024), OWL (Kessler et al., 2022), SANE (Powers et al., 2022), MOSAIC (Nath et al., 2026), and ProgNet (Rusu et al., 2016)), fine-tuning, and learning from scratch. We also evaluate retrieval guided by independently measured pairwise transfer. Tables 1 and 2 report comparisons to baselines on CT28, MG16 and CW10. Appendix E provides details of the hyperparameters for AMSC and baselines.

Protocol and Metrics. All empirical results are evaluated with seven shared seeds. CT28 and MG16 use PPO with $51,200$ and $512,000$ steps per task, respectively. CW10 uses SAC with 1M steps per task. We report mean task area under the evaluation curve (AUC), forward transfer (FWT), and forgetting (FGT) from normalized performance between $[0, 1]$, following (Wolczyk et al., 2021). We report these metrics together because zero forgetting can reflect strong retention but not necessarily strong policies. Appendix E specifies training and evaluation hyperparameters, Appendix F.3 reports hyperparameter sensitivity, and Appendix G reports the details of how the metrics are computed.

5.2 COMPARISON TO BASELINES

Selective reuse improves learning. AMSC obtains strong AUC and FWT with zero measured forgetting on CT28 and MG16 (Table 1), occupying a favorable region of the transfer-retention tradeoff (Figure 1) with respect to baselines. AMSC achieves AUC/FWT of $0.94/0.94$ and $0.73/0.54$ on CT28 and MG16, respectively. From-scratch learning also avoids forgetting but obtains only 0.22 and 0.23 AUC. Surprise-prioritized SER slightly exceeds AMSC on CT28 ($0.96/0.95$) but incurs 0.54 FGT, while reservoir SER obtains $0.63/0.20$ on MG16 with 0.07 FGT.

Comparison with modular reuse. AMSC achieves the highest mean AUC and FWT among modular methods on CT28 and MG16 (Table 1). These methods differ in how they reuse prior knowledge: Mask-LC/BLC learn weights over all prior masks, CompoNet learns to combine policy-unit outputs, and MOSAIC selects masks through absolute similarity and performance gates before optimizing composition. AMSC instead periodically updates both its selected sources and their weights using library-relative similarity. On CT28, AMSC achieves AUC/FWT of $0.94/0.94$, compared with $0.81/0.79$ for MOSAIC and $0.69/0.64$ for CompoNet. On MG16, both baselines have negative FWT, whereas AMSC achieves 0.54. AMSC's gains over Mask-LC, CompoNet, and MOSAIC are significant on both metrics and curricula ($p \leq 0.023$). Its MG16 advantage over Mask-BLC ($0.73/0.54$ versus $0.55/0.23$, $p = 0.016$ and $p = 0.016$, respectively). These results associate explicit, relative selection and composition with stronger reuse than dense composition or fixed absolute selection.

Table 1: Evaluation on CT28 and MG16. Each cell reports the mean AUC/FWT/FGT with an 95% confidence interval (Colas et al., 2018). AVG is computed per seed by averaging CT28 and MG16, then bootstrapping the seven paired averages. Bold and underlining denote the best and second-best non-oracle result for each metric. Significance tests are reported in Appendix H.

	CT28			MG16			AVG		
Method	**AUC ↑**	**FWT ↑**	**FGT ↓**	**AUC ↑**	**FWT ↑**	**FGT ↓**	**AUC ↑**	**FWT ↑**	**FGT ↓**
Fine-tuning	$.36^{+.02}_{-.02}$	$.22^{+.03}_{-.03}$	$.38^{+.02}_{-.02}$	$.30^{+.05}_{-.05}$	$.10^{+.08}_{-.09}$	$.34^{+.08}_{-.07}$	$.33^{+.02}_{-.02}$	$.16^{+.05}_{-.06}$	$.36^{+.03}_{-.03}$
From-scratch	$.22^{+.01}_{-.01}$	$.00^{+.00}_{-.00}$	$\underline{.00}^{+.00}_{-.00}$	$.23^{+.03}_{-.02}$	$.00^{+.00}_{-.00}$	$\mathbf{.00}^{+.00}_{-.00}$	$.22^{+.01}_{-.01}$	$.00^{+.00}_{-.00}$	$\mathbf{.00}^{+.00}_{-.00}$
SER-s	$\mathbf{.96}^{+.02}_{-.03}$	$\mathbf{.95}^{+.03}_{-.03}$	$.54^{+.17}_{-.18}$	$.42^{+.06}_{-.07}$	$-.13^{+.17}_{-.17}$	$.51^{+.10}_{-.09}$	$.69^{+.03}_{-.04}$	$.41^{+.08}_{-.08}$	$.52^{+.09}_{-.11}$
SER-r	$.88^{+.05}_{-.05}$	$.87^{+.06}_{-.06}$	$.09^{+.08}_{-.07}$	$\underline{.63}^{+.04}_{-.03}$	$.20^{+.14}_{-.20}$	$\underline{.07}^{+.11}_{-.09}$	$\underline{.76}^{+.02}_{-.02}$	$\underline{.53}^{+.08}_{-.10}$	$.08^{+.09}_{-.06}$
SDW	$.76^{+.11}_{-.10}$	$.73^{+.13}_{-.13}$	$\mathbf{-.01}^{+.01}_{-.01}$	$.53^{+.05}_{-.06}$	$.02^{+.12}_{-.18}$	$.15^{+.13}_{-.12}$	$.65^{+.06}_{-.07}$	$.37^{+.10}_{-.14}$	$\underline{.07}^{+.06}_{-.06}$
OWL	$.15^{+.01}_{-.01}$	$-.31^{+.17}_{-.20}$	$.18^{+.03}_{-.04}$	$.21^{+.03}_{-.02}$	$-.29^{+.18}_{-.19}$	$.22^{+.03}_{-.02}$	$.18^{+.02}_{-.02}$	$-.30^{+.09}_{-.09}$	$.20^{+.02}_{-.02}$
SANE	$.46^{+.02}_{-.03}$	$.37^{+.03}_{-.03}$	$\underline{.00}^{+.00}_{-.00}$	$.36^{+.09}_{-.08}$	$.18^{+.11}_{-.09}$	$\mathbf{.00}^{+.00}_{-.00}$	$.41^{+.04}_{-.04}$	$.28^{+.05}_{-.04}$	$\mathbf{.00}^{+.00}_{-.00}$
Mask-BLC	$.27^{+.02}_{-.02}$	$.10^{+.06}_{-.06}$	$\underline{.00}^{+.00}_{-.00}$	$.55^{+.10}_{-.10}$	$\underline{.23}^{+.21}_{-.19}$	$\mathbf{.00}^{+.00}_{-.00}$	$.41^{+.06}_{-.06}$	$.17^{+.13}_{-.12}$	$\mathbf{.00}^{+.00}_{-.00}$
Mask-LC	$.48^{+.03}_{-.03}$	$.35^{+.06}_{-.07}$	$\underline{.00}^{+.00}_{-.00}$	$.40^{+.14}_{-.13}$	$-.24^{+.41}_{-.37}$	$\mathbf{.00}^{+.00}_{-.00}$	$.44^{+.06}_{-.06}$	$.06^{+.21}_{-.20}$	$\mathbf{.00}^{+.00}_{-.00}$
CKA-RL	$.36^{+.05}_{-.04}$	$.25^{+.05}_{-.05}$	$.36^{+.05}_{-.05}$	$.34^{+.11}_{-.09}$	$-.21^{+.20}_{-.24}$	$.40^{+.14}_{-.12}$	$.35^{+.07}_{-.06}$	$.02^{+.11}_{-.13}$	$.38^{+.08}_{-.06}$
CompoNet	$.69^{+.08}_{-.10}$	$.64^{+.09}_{-.12}$	$\underline{.00}^{+.00}_{-.00}$	$.31^{+.07}_{-.08}$	$-.85^{+.24}_{-.25}$	$\mathbf{.00}^{+.00}_{-.00}$	$.50^{+.05}_{-.06}$	$-.10^{+.15}_{-.15}$	$\mathbf{.00}^{+.00}_{-.00}$
MOSAIC	$.81^{+.03}_{-.03}$	$.79^{+.03}_{-.03}$	$\underline{.00}^{+.00}_{-.00}$	$.29^{+.11}_{-.11}$	$-.50^{+.40}_{-.45}$	$\mathbf{.00}^{+.00}_{-.00}$	$.55^{+.05}_{-.06}$	$.14^{+.20}_{-.22}$	$\mathbf{.00}^{+.00}_{-.00}$
MOSAIC $\neg\mathbb{I}_{\text{perf}}$	$.81^{+.03}_{-.03}$	$.78^{+.03}_{-.03}$	$\underline{.00}^{+.00}_{-.00}$	$.29^{+.13}_{-.11}$	$-.60^{+.47}_{-.50}$	$\mathbf{.00}^{+.00}_{-.00}$	$.55^{+.06}_{-.05}$	$.09^{+.22}_{-.24}$	$\mathbf{.00}^{+.00}_{-.00}$
Transfer oracle	$.82^{+.00}_{-.01}$	$.72^{+.07}_{-.13}$	$.00^{+.00}_{-.00}$	$.66^{+.08}_{-.09}$	$.54^{+.10}_{-.10}$	$.00^{+.00}_{-.00}$	$.74^{+.04}_{-.04}$	$.63^{+.08}_{-.09}$	$\mathbf{.00}^{+.00}_{-.00}$
AMSC	$\underline{.94}^{+.02}_{-.03}$	$\underline{.94}^{+.02}_{-.03}$	$\underline{.00}^{+.00}_{-.00}$	$\mathbf{.73}^{+.06}_{-.06}$	$\mathbf{.54}^{+.12}_{-.14}$	$\mathbf{.00}^{+.00}_{-.00}$	$\mathbf{.84}^{+.03}_{-.03}$	$\mathbf{.74}^{+.06}_{-.07}$	$\mathbf{.00}^{+.00}_{-.00}$

Early learning across reuse mechanisms. Over the first 20% of CT28 tasks 1–27, AMSC obtains $\mathrm{AUC}_{20} = 0.86$, compared with 0.35 for MOSAIC ($p = 0.016$). The first positive training reward appears after 2.34k versus 5.83k steps ($p = 0.016$). SER-surprise reaches 0.91 early AUC and SER-reservoir 0.86; neither differs significantly from AMSC on this metric. Thus AMSC's early advantage is clear relative to MOSAIC, while replay can match or exceed its early learning. Appendix Tables 8 and 9 report aggregate and per-task results.

Table 2: CW10 performance across seven seeds. Entries report means with 95% confidence intervals.

Method	**AUC**	**FWT**	**FGT**
Fine-tuning	$.34^{+.17}_{-.18}$	$-7.12^{+7.20}_{-13.38}$	$.39^{+.19}_{-.21}$
From-scratch	$.41^{+.04}_{-.04}$	$.00^{+.00}_{-.00}$	$.00^{+.00}_{-.00}$
PNN	$.34^{+.02}_{-.03}$	$-.31^{+.14}_{-.14}$	$.00^{+.00}_{-.00}$
Mask-LC	$.47^{+.06}_{-.06}$	$-.22^{+.34}_{-.40}$	$.00^{+.00}_{-.00}$
Mask-BLC	$.53^{+.04}_{-.04}$	$.12^{+.19}_{-.20}$	$.00^{+.00}_{-.00}$
CKA-RL	$.39^{+.10}_{-.14}$	$-2.12^{+1.76}_{-2.24}$	$.00^{+.00}_{-.00}$
MOSAIC	$.26^{+.04}_{-.03}$	$-4.76^{+3.74}_{-6.41}$	$.00^{+.00}_{-.00}$
AMSC	$.48^{+.05}_{-.05}$	$-1.21^{+.79}_{-.98}$	$.00^{+.00}_{-.00}$

5.3 Limits of Similarity Weighting

Does fixed similarity weighting extend to continuous control? On CW10, AMSC obtains 0.48 AUC with zero measured forgetting, but its FWT is -1.21; Mask-BLC obtains 0.53 AUC and 0.12 FWT (Table 2). Their composition weights overlap on some tasks, yet AMSC often allocates more mass to prior masks later in the curriculum. At task 9, mean actor prior mass is 0.78 for AMSC versus 0.46 for Mask-BLC (Appendix F.2). AMSC applies the same similarity-derived proportions across masked layers, whereas Mask-BLC learns layer-specific weights. This difference suggests that identifying relevant prior policies and determining how to compose their knowledge are distinct challenges, even when learned composition addresses both through a single set of weights.

5.4 Task Representation Sensitivity

One question is whether the performance of AMSC is dependent on the specific representation approach, or whether alternatives to WTE can be used with similar performance. Table 3 compares WTE (Liu et al., 2025), sliced-Wasserstein embeddings (Naderializadeh et al., 2021), an empirical-Fisher adaptation of Task2Vec (Achille et al., 2019), recurrent encoders motivated by PEARL and VariBAD (Rakelly et al., 2019; Zintgraf et al., 2021), and FID-inspired Fréchet moments (Heusel et al., 2017). All comparisons receive complete (s, a, r, s', d) as Task2Vec-style and learned context encoder variants require the reward and next state as part of the objective. The main AMSC experiments use only (s, a, r) (also reported in Table 3 for reference).

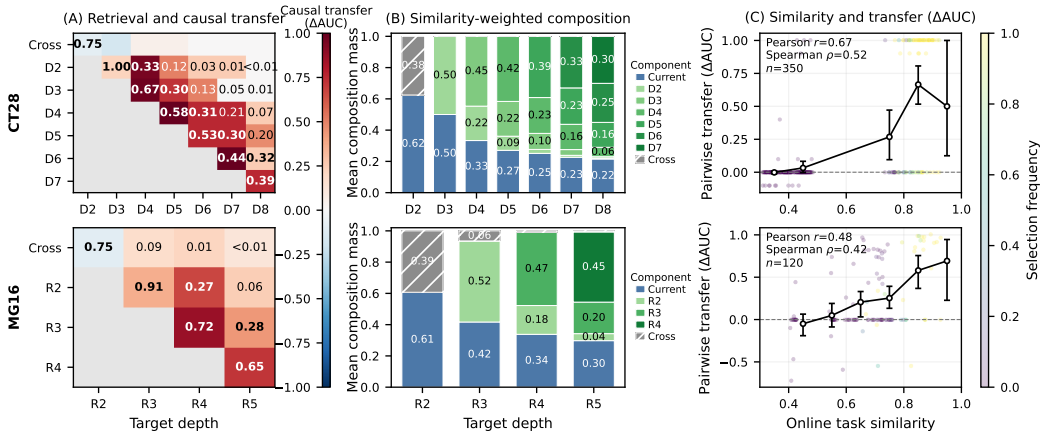

Figure 2: AMSC on CT28 (top) and MG16 (bottom). Same-family source policies share the target task's observation space. (A) Mean pairwise ΔAUC by source category. Annotations give the mean total sparsemax retrieval mass assigned to that category. (B) Mean composition mass after applying Equation 9 to selected sources. (C) Online similarity versus pairwise transfer; color denotes selection frequency and the line shows bin means with target-cluster 95% confidence intervals. CT28 includes depths D4–D8, whereas MG16 includes all causal non-self pairs. Retrieval uses seven seeds; pairwise transfer uses one source-policy seed.

Representation choice is task-dependent but AMSC remains agnostic. Among the variants, WTE–transition has the highest mean AUC/FWT on CT28 (0.93/0.92), while Fréchet moments lead on MG16 (0.70/0.47). SWE and random recurrent features also yield positive transfer on both curricula. AMSC therefore accommodates several representations, although performance depends on the specific distributions being measured and specific configurations used and no single approach is universally best.

Table 3: Task-representation sensitivity of AMSC. Reports AUC/FWT with 95% confidence intervals across seven seeds. T denotes complete transitions; R and L denote random and online-learned. Bold and underlining indicate the best and second-best results.

Method	CT28 AUC↑/FWT↑	MG16 AUC↑/FWT↑
WTE (T)	$\underline{.93}^{+.04}_{-.04}/\underline{.92}^{+.04}_{-.04}$	$.61^{+.07}_{-.08}/.38^{+.12}_{-.14}$
SWE (T)	$.92^{+.04}_{-.03}/.91^{+.05}_{-.05}$	$.67^{+.04}_{-.04}/.39^{+.15}_{-.16}$
Fisher (T2V)	$.87^{+.06}_{-.07}/.85^{+.08}_{-.10}$	$.45^{+.12}_{-.10}/.21^{+.17}_{-.18}$
Context (R)	$.92^{+.03}_{-.03}/.91^{+.03}_{-.04}$	$.64^{+.06}_{-.07}/.41^{+.10}_{-.12}$
Context (L)	$.90^{+.03}_{-.03}/.87^{+.06}_{-.06}$	$.44^{+.09}_{-.10}/.14^{+.21}_{-.28}$
Fréchet	$.90^{+.03}_{-.03}/.87^{+.06}_{-.07}$	$\underline{.70}^{+.07}_{-.07}/\underline{.47}^{+.13}_{-.16}$
WTE (SAR)	$\mathbf{.94^{+.02}_{-.03}/.94^{+.02}_{-.03}}$	$\mathbf{.73^{+.06}_{-.06}/.54^{+.12}_{-.14}}$

5.5 Retrieval, Composition, and Causal Transfer

We test whether similarity identifies useful prior knowledge using independently measured pairwise transfer. For each source–target pair, we measure ΔAUC (target task AUC with a source minus AUC without it).

Does task similarity rank useful priors? Figure 2C presents the correlation between pairwise transfer and similarity. Online similarity correlates with pairwise transfer on CT28 depths 4–8 (Pearson $r = 0.67$, 95% CI $[0.57, 0.77]$ and MG16 ($r = 0.48$, $[0.31, 0.62]$) with Spearman $\rho = 0.52$ and 0.42. These correlations suggest that similarity is informative of which prior masks are likely to transfer.

Does retrieval exploit this ranking? We define a pair as regularly selected if it is retrieved in at least half of its available selection events. On CT28, regularly selected pairs have mean $\Delta\text{AUC} = 0.66$, versus 0.02 for other pairs. These pairs cover 81% of positive-transfer pairs and include no negative-transfer pairs. On MG16, corresponding means are 0.49 and 0.15, but retrieval covers only 19 of 47 positive-transfer pairs as a result of the sparse selection.

How is reuse allocated within priors? On CT28, support grows from one prior at depth 3 to approximately four at depth 8, reducing current-mask mass from 0.50 to 0.22. At depth 8, predecessors

at depths 7 and 6 receive 0.30 and 0.25 of the mass. MG16 shows a similar pattern: at depth-5 (R5), the current mask and depth-4 (R4) and depth-3 (R3) masks receive 0.30, 0.45, and 0.20 and below 0.01 to selected cross-family masks. Reuse therefore concentrates on related hierarchical sub-task masks obtained earlier in learning, while retaining a current-task component (Figure 2B).

5.6 Ablation Studies

Controls. Table 4 tests whether the performance gains of AMSC are attributed to the combination of components. *Random* permutes selected priors while preserving how many are selected and their specific similarity-derived weights. *Learned-α* optimizes the retrieved weights, whereas *No Sim.-α* tests a uniform composition over selected prior masks. *No Periodicity* retains the first retrieval for the entire learning duration. *Top-1* retains the highest-similarity prior after sparsemax selection. *Transf. Oracle* uses independent pair-wise ΔAUC measurements to guide retrieval of multiple masks but is not an upper bound on joint composition.

Table 4: Ablation of AMSC's selection and composition mechanism. All runs use WTE embeddings of state–action–reward tuples. Values are means with 95% confidence intervals. All configurations obtain zero FGT.

Configuration	CT28 AUC↑/FWT↑	MG16 AUC↑/FWT↑
No Sim.-α	$.90^{+.03}_{-.03}/.88^{+.04}_{-.04}$	$.61^{+.10}_{-.10}/.41^{+.15}_{-.15}$
No L2-STE	$.89^{+.03}_{-.03}/.86^{+.03}_{-.03}$	$.59^{+.09}_{-.10}/.21^{+.26}_{-.29}$
No Periodicity	$.87^{+.04}_{-.05}/.85^{+.05}_{-.06}$	$.63^{+.11}_{-.13}/.42^{+.14}_{-.15}$
Top-1	$.88^{+.04}_{-.04}/.86^{+.05}_{-.04}$	$.63^{+.09}_{-.08}/.39^{+.16}_{-.16}$
Learned-α	$\underline{.92}^{+.03}_{-.03}/\underline{.91}^{+.04}_{-.04}$	$.60^{+.11}_{-.11}/.33^{+.15}_{-.14}$
Random	$.35^{+.02}_{-.02}/.13^{+.07}_{-.06}$	$.36^{+.05}_{-.06}/.08^{+.13}_{-.13}$
Transf. Oracle	$.82^{+.00}_{-.01}/.72^{+.07}_{-.13}$	$\underline{.66}^{+.08}_{-.09}/\underline{.54}^{+.10}_{-.10}$
AMSC	$\mathbf{.94}^{+.02}_{-.03}/\mathbf{.94}^{+.02}_{-.03}$	$\mathbf{.73}^{+.06}_{-.06}/\mathbf{.54}^{+.12}_{-.14}$

Does source relevance matter beyond sparsity? Randomizing which prior masks are selected significantly reduces AUC/FWT to $0.35/0.13$ on CT28 and $0.36/0.08$ on MG16 ($p = 0.016$). Because support size and weights are preserved, these reductions show that sparse composition alone is insufficient and selecting relevant priors is central to the gains.

Do retrieval updates and multiple sources improve transfer? Freezing the first retrieval lowers AUC/FWT to $0.87/0.85$ on CT28 and $0.63/0.42$ on MG16. These decreases suggest that revising the support and weights as experience accumulates contributes beyond the initial selection. In AMSC, selection remains stable throughout learning. Consecutive-selection Jaccard similarity is 0.94–1.00 on CT28 and 0.97–1.00 on MG16. AMSC successfully identifies relevant prior masks early on in learning then gradually refines its selection. Restricting retrieval to one prior mask results in $0.88/0.86$ and $0.63/0.39$, respectively. Since total prior mass is unchanged, this comparison supports composing multiple sources rather than assigning the entire reuse allocation to the highest-ranked prior.

Transfer-informed retrieval. Transfer oracle also obtains positive FWT, obtaining AUC/FWT of $0.82/0.72$ on CT28 and $0.66/0.54$ on MG16. It nevertheless falls below AMSC on CT28 and matches only its FWT on MG16. Selection is guided by pair-wise transfer, therefore it does not provide an upper bound on joint composition.

Do composition weights need to be learned? Learning weights within the retrieved support yields AUC/FWT of $0.92/0.91$ on CT28 and $0.60/0.33$ on MG16, compared with $0.94/0.94$ and $0.73/0.54$ for AMSC. Optimization thus provides no improvement in mean performance in these experiments. Replacing similarity-derived source proportions with uniform weights also lowers performance, to $0.90/0.88$ and $0.61/0.41$. These comparisons suggest that similarity provides useful information about how strongly to reuse selected masks, beyond determining which masks enter the composition; however, task similarity in AMSC cannot determine how knowledge should be uniquely composed at every layer of the network as indicated by the earlier CW10 results.

Does mask-score normalization matter? Composition weights determine source contributions only relative to the magnitudes of the scores being combined. Removing L2-STE lowers AUC/FWT to $0.89/0.86$ on CT28 and $0.59/0.21$ on MG16. These results support normalizing scores before composition so that differences in score magnitude do not override the intended similarity-derived allocation.

5.7 Addressing Per-Task Mask Memory Costs

Continuous-value masks are memory intensive. We test whether masks can be quantized with minimal performance loss. Ternary quantization maps scores to $\{-1, 0, 1\}$, preserving their sign while suppressing near-zero scores. On CT28, ternary quantization achieves AUC/FWT/FGT of $0.94/0.91/0.04$ compared to $0.941/0.935/0.000$ for Float32 masks. Ternary scores also reduce per-task storage by $16\times$ relative to Float32, while retaining better AUC/FWT (Appendix A).

6 Discussion and Future Work

The results suggest that task similarity can guide effective policy reuse without directly optimizing retrieval or composition weights when relevant prior knowledge is available. On CT28 and MG16, AMSC achieves competitive AUC and FWT relative to replay methods while avoiding forgetting through per-task masks. The random-selection ablation indicates that sparsity alone is insufficient. Periodic retrieval further improves reuse as more experience is collected and task representations are updated.

The differing performance of transition embeddings and Fréchet moments also suggests that representation choice depends on the distributions. Future work could evaluate combinations of priors beyond pairwise transfer to determine how reliably embedding similarity predicts joint transfer. The CW10 results highlight a distinction between identifying relevant knowledge and determining how to reuse it. Similarity-derived weights shared across layers may be too restrictive for effective composition. Future work could test whether task embeddings can guide layer-specific reuse or whether similarity-based selection could be combined with learned composition, including an adaptive balance between prior and current knowledge. Another direction is determining when an agent should abstain from reuse when no relevant prior knowledge exists.

This work examines how sparsity and selectivity can enable effective forward transfer as the policy library grows, while managing that growth remains a complementary challenge. Since available knowledge influences what can be effectively learned, consolidation and compression should be evaluated for their effects on future transfer as well as retained performance. Finally, extending AMSC beyond known task boundaries requires detecting distribution changes and deciding when to retrieve, adapt, or create a new mask.

7 Conclusion

This work introduced Adaptive Mask Selection and Composition (AMSC) to address how lifelong agents can selectively reuse prior knowledge when learning new tasks. AMSC uses task similarity to periodically select and compose a sparse set of policy masks, without directly optimizing selection or composition. On CT28 and MG16, it outperforms modular baselines in mean AUC and forward transfer and remains competitive with replay methods while retaining prior policies without forgetting. These findings show that task similarity can provide an effective alternative to learning compositions, with ablations supporting the importance of relevant selection and updating reuse during learning. The advantage is not universal, however, as the CW10 results suggest that identifying relevant prior knowledge and adapting its composition across layers may require different mechanisms. The effectiveness of similarity-guided reuse depends on how prior knowledge is composed per-layer. Together, the results position task similarity as a practical approach for selective reuse.

AI USE STATEMENT

In this work, we used generative AI tools to provide feedback on research methodology. We have not used generative AI tools for proposing or refining hypotheses, implementing methods, support qualitative and thematic data analyses, or interpreting results. Generating synthetic datasets, helping develop theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting the writing of proofs, assisting with translation, and cleaning and reformatting of datasets are not applicable to this work.

Additionally, we used generative AI tools for editing software code, editing the research paper to improve readability and for proofreading, identifying and summarizing relevant literature. AI assisted code was limited to edits for individual ablations, and scripts used to calculate metrics and produce figures. We have reviewed all AI-assisted work. Code implementations were verified and tested for correctness by 2 authors, and relevant literature was verified by papers. Errors flagged by LLM were corrected manually. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We provide anonymized source-code in the supplementary materials, including experiment configurations, launch scripts, and scripts for computing metrics and generating the results used in this paper. Section 4 specifies how AMSC works, and Appendix B provides implementation details and pseudocode. Benchmark configurations and task sequences are described in Appendix C, with hyperparameters for AMSC and baselines in Appendix E. Section 5.1 describes the training and evaluation protocol, while Appendices G and H define the metrics and statistical procedures used to summarize results across seeds.

REFERENCES

Alessandro Achille, Michael Lam, Rahul Tewari, Avinash Ravichandran, Subhransu Maji, Charless Fowlkes, Stefano Soatto, and Pietro Perona. Task2vec: Task embedding for meta-learning. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6429–6438, 2019. doi: 10.1109/ICCV.2019.00653.

Hongjoon Ahn, Jinu Hyeon, Youngmin Oh, Bosun Hwang, and Taesup Moon. Prevalence of negative transfer in continual reinforcement learning: Analyses and a simple baseline. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=KAIqwkB3dT.

David Alvarez-Melis and Nicolo Fusi. Geometric dataset distances via optimal transport. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 21428–21439. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/f52a7b2610fb4d3f74b4106fb80b233d-Paper.pdf.

Eseoghene Ben-Iwhiwhu, Saptarshi Nath, Praveen Kumar Pilly, Soheil Kolouri, and Andrea Soltoggio. Lifelong Reinforcement Learning with Modulating Masks. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=V7tahqGrOq.

J.L. Carroll and K. Seppi. Task similarity measures for transfer in reinforcement learning task libraries. In *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.*, volume 2, pp. 803–808 vol. 2, 2005. doi: 10.1109/IJCNN.2005.1555955.

Maxime Chevalier-Boisvert, Lucas Willems, and Suman Pal. Minimalistic Gridworld Environment for OpenAI Gym. https://github.com/maximecb/gym-minigrid, 2018.

Cédric Colas, Olivier Sigaud, and Pierre-Yves Oudeyer. How Many Random Seeds? Statistical Power Analysis in Deep Reinforcement Learning Experiments. *arXiv preprint arXiv:1806.08295*, 2018.

Nicolas Courty, Rémi Flamary, and Mélanie Ducoffe. Learning wasserstein embeddings. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=SJyEH91A-.

Aleksandr Dekhovich, David M. J. Tax, Marcel H. F Sluiter, and Miguel A. Bessa. Continual prune-and-select: class-incremental learning with specialized subnetworks. *Applied Intelligence*, 53 (14):17849–17864, 2023. doi: 10.1007/s10489-022-04441-z. URL https://doi.org/10.1007/s10489-022-04441-z.

Jeffery Dick, Saptarshi Nath, Christos Peridis, Eseoghene Benjamin, Soheil Kolouri, and Andrea Soltoggio. Statistical Context Detection for Deep Lifelong Reinforcement Learning, 2024.

Michael Gimelfarb, Scott Sanner, and Chi-Guhn Lee. Contextual policy transfer in reinforcement learning domains via deep mixtures-of-experts. In Cassio de Campos and Marloes H. Maathuis (eds.), *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*, volume 161 of *Proceedings of Machine Learning Research*, pp. 1787–1797. PMLR, 27–30 Jul 2021. URL https://proceedings.mlr.press/v161/gimelfarb21a.html.

Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. Deep Reinforcement Learning That Matters. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), Apr. 2018. doi: 10.1609/aaai.v32i1.11694. URL https://ojs.aaai.org/index.php/AAAI/article/view/11694.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/8a1d694707eb0fefe65871369074926d-Paper.pdf.

Heinke Hihn and Daniel Alexander Braun. Mixture-of-variational-experts for continual learning. In *ICLR Workshop on Agent Learning in Open-Endedness*, 2022. URL https://openreview.net/forum?id=HhzzNQPZLb9.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Jinwu Hu, ZiHao Lian, Zhiquan Wen, ChenghaoLi, Guohao Chen, Xutao Wen, Bin Xiao, and Mingkui Tan. Continual knowledge adaptation for reinforcement learning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=QRlVickNdN.

Zhiyi Huang, Xiaohan Shan, and Jianmin Li. Lifelong reinforcement learning with similarity-driven weighting by large models, 2025. URL https://arxiv.org/abs/2503.12923.

David Isele and Akansel Cosgun. Selective Experience Replay for Lifelong Learning. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18. AAAI Press, 2018. ISBN 978-1-57735-800-8.

Jack Julian, Yun Sing Koh, and Albert Bifet. Building adaptive knowledge bases for evolving continual learning models. *npj Artificial Intelligence*, 1(1):26, 2025. doi: 10.1038/s44387-025-00028-4. URL https://doi.org/10.1038/s44387-025-00028-4.

Zixuan Ke, Bing Liu, and Xingchang Huang. Continual learning of a mixed sequence of similar and dissimilar tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

Samuel Kessler, Jack Parker-Holder, Philip Ball, Stefan Zohren, and Stephen J Roberts. Unclear: A straightforward method for continual reinforcement learning. In *Proceedings of the 37th International Conference on Machine Learning*, 2020.

Samuel Kessler, Jack Parker-Holder, Philip Ball, Stefan Zohren, and Stephen J. Roberts. Same state, different task: Continual reinforcement learning without interference, 2022.

Khimya Khetarpal, Matthew Riemer, Irina Rish, and Doina Precup. Towards continual reinforcement learning: A review and perspectives. *Journal of Artificial Intelligence Research*, 75:1401–1476, 12 2022. doi: 10.1613/jair.1.13673.

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. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017.

Nils Koster, Oliver Grothe, and Achim Rettinger. Signing the Supermask: Keep, Hide, Invert. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=e0jtGTfPihs.

Seungwon Lee, Sima Behpour, and Eric Eaton. Sharing less is more: Lifelong learning in deep networks with selective layer transfer. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 6065–6075, 2021.

Siyuan Li, Fangda Gu, Guangxiang Zhu, and Chongjie Zhang. Context-aware policy reuse. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '19, pp. 989–997, Richland, SC, 2019. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 9781450363099.

Sen Lin, Li Yang, Deliang Fan, and Junshan Zhang. TRGP: Trust region gradient projection for continual learning. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=iEvAf8i6JjO.

Xinran Liu, Yikun Bai, Yuzhe Lu, Andrea Soltoggio, and Soheil Kolouri. Wasserstein Task Embedding for Measuring Task Similarities. *Neural Networks*, 181:106796, 2025. ISSN 0893-6080. doi: https://doi.org/10.1016/j.neunet.2024.106796. URL https://www.sciencedirect.com/science/article/pii/S0893608024007202.

David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. *Advances in neural information processing systems*, 30, 2017.

Mikel Malagón, Josu Ceberio, and Jose A. Lozano. Self-composing policies for scalable continual reinforcement learning. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Arun Mallya and Svetlana Lazebnik. Packnet: Adding Multiple Tasks to a Single Network by Iterative Pruning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 7765–7773, 2018.

Andre Martins and Ramon Astudillo. From softmax to sparsemax: A sparse model of attention and multi-label classification. In Maria Florina Balcan and Kilian Q. Weinberger (eds.), *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 1614–1623, New York, New York, USA, 20–22 Jun 2016. PMLR. URL https://proceedings.mlr.press/v48/martins16.html.

Jorge A Mendez, Harm van Seijen, and Eric Eaton. Modular Lifelong Reinforcement Learning via Neural Composition. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=5XmLzdslFNN.

Navid Naderializadeh, Joseph F Comer, Reed Andrews, Heiko Hoffmann, and Soheil Kolouri. Pooling by sliced-wasserstein embedding. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 3389–3400. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/1bc2029a8851ad344a8d503930dfd7f7-Paper.pdf.

Saptarshi Nath, Christos Peridis, Eseoghene Benjamin, Xinran Liu, Soheil Kolouri, Peter Kinnell, Zexin Li, Cong Liu, Shirin Dora, and Andrea Soltoggio. Policy search, retrieval, and composition via task similarity in collaborative agentic systems. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(29):24504–24512, Mar. 2026. doi: 10.1609/aaai.v40i29.39633. URL https://ojs.aaai.org/index.php/AAAI/article/view/39633.

Cuong Nguyen, Tal Hassner, Matthias Seeger, and Cedric Archambeau. LEEP: A new measure to evaluate transferability of learned representations. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 7294–7305. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/nguyen20b.html.

Oleksiy Ostapenko, Pau Rodríguez, Massimo Caccia, and Laurent Charlin. Continual Learning via Local Module Composition. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA, 2021. Curran Associates Inc. ISBN 9781713845393.

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. ISSN 0893-6080. doi: https://doi.org/10.1016/j.neunet.2019.01.012. URL https://www.sciencedirect.com/science/article/pii/S0893608019300231.

Sam Powers, Eliot Xing, and Abhinav Gupta. Self-activating neural ensembles for continual reinforcement learning. In Sarath Chandar, Razvan Pascanu, and Doina Precup (eds.), *Proceedings of The 1st Conference on Lifelong Learning Agents*, volume 199 of *Proceedings of Machine Learning Research*, pp. 683–704. PMLR, 22–24 Aug 2022. URL https://proceedings.mlr.press/v199/powers22a.html.

Kate Rakelly, Aurick Zhou, Chelsea Finn, Sergey Levine, and Deirdre Quillen. Efficient off-policy meta-reinforcement learning via probabilistic context variables. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5331–5340. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/rakelly19a.html.

David Rolnick, Arun Ahuja, Jonathan Schwarz, Timothy Lillicrap, and Gregory Wayne. Experience replay for continual learning. *Advances in Neural Information Processing Systems*, 32, 2019.

Andrei A Rusu, Neil C Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, and Raia Hadsell. Progressive Neural Networks. *arXiv preprint arXiv:1606.04671*, 2016.

Chinmay Savadikar, Michelle Dai, and Tianfu Wu. Cheem: Continual learning by reuse, new, adapt and skip - a hierarchical exploration-exploitation approach. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25066–25076, June 2026.

Ghada Sokar, Decebal Constantin Mocanu, and Mykola Pechenizkiy. Avoiding forgetting and allowing forward transfer in continual learning via sparse networks. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2022, Grenoble, France, September 19–23, 2022, Proceedings, Part III*, pp. 85–101, Berlin, Heidelberg, 2022. Springer-Verlag. ISBN 978-3-031-26408-5. doi: 10.1007/978-3-031-26409-2_6. URL https://doi.org/10.1007/978-3-031-26409-2_6.

Andrea Soltoggio, Eseoghene Ben-Iwhiwhu, Christos Peridis, Pawel Ladosz, Jeffery Dick, Praveen K. Pilly, and Soheil Kolouri. The configurable tree graph (ct-graph): measurable problems in partially observable and distal reward environments for lifelong reinforcement learning, 2023. URL https://arxiv.org/abs/2302.10887.

Matthew E. Taylor and Peter Stone. Transfer learning for reinforcement learning domains: A survey. *Journal of Machine Learning Research*, 10(56):1633–1685, 2009. URL http://jmlr.org/papers/v10/taylor09a.html.

Ikechukwu Uchendu, Ted Xiao, Yao Lu, Banghua Zhu, Mengyuan Yan, Joséphine Simon, Matthew Bennice, Chuyuan Fu, Cong Ma, Jiantao Jiao, Sergey Levine, and Karol Hausman. Jump-Start Reinforcement Learning. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 34556–34583. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/uchendu23a.html.

Johannes von Oswald, Christian Henning, Benjamin F. Grewe, and João Sacramento. Continual learning with hypernetworks. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SJgwNerKvB.

Yifan Wang, Yafei Liu, Chufan Shi, Haoling Li, Chen Chen, Haonan Lu, and Yujiu Yang. Inscl: A data-efficient continual learning paradigm for fine-tuning large language models with instructions, 2024. URL https://arxiv.org/abs/2403.11435.

Zhenyi Wang, Li Shen, Le Fang, Qiuling Suo, Tiehang Duan, and Mingchen Gao. Improving task-free continual learning by distributionally robust memory evolution. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 22985–22998. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/wang22v.html.

Anderson M. Winkler, Gerard R. Ridgway, Matthew A. Webster, Stephen M. Smith, and Thomas E. Nichols. Permutation inference for the general linear model. *NeuroImage*, 92:381–397, 2014. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2014.01.060. URL https://www.sciencedirect.com/science/article/pii/S1053811914000913.

Maciej Wolczyk, Michał Zajac, Razvan Pascanu, Łukasz Kuciński, and Piotr Miłoś. Continual world: A robotic benchmark for continual reinforcement learning. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=5qsptDcsdEj.

Maciej Wolczyk, MichałZajac, Razvan Pascanu, Ł ukasz Kuciński, and Piotr Mił oś. Disentangling transfer in continual reinforcement learning. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 6304–6317. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/2938ad0434a6506b125d8adaff084a4a-Paper-Conference.pdf.

Mitchell Wortsman, Vivek Ramanujan, Rosanne Liu, Aniruddha Kembhavi, Mohammad Rastegari, Jason Yosinski, and Ali Farhadi. Supermasks in Superposition. *Advances in Neural Information Processing Systems*, 33:15173–15184, 2020.

Yanru Wu, Jianning Wang, Xiangyu Chen, Enming Zhang, Yang Tan, Hanbing Liu, and Yang Li. Exploiting task relationships in continual learning via transferability-aware task embeddings. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=V8FnYzDX35.

Zihao Xu, Xuan Tang, Yufei Shi, Jianfeng Zhang, Jian Yang, Mingsong Chen, and Xian Wei. Continual learning via manifold expansion replay. In *2023 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pp. 733–738, 2023. doi: 10.1109/SMC53992.2023.10393967.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 5824–5836. Curran Associates, Inc., 2020a. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/3fe78a8acf5fda99de95303940a2420c-Paper.pdf.

Tianhe Yu, Deirdre Quillen, Zhanpeng He, Ryan Julian, Karol Hausman, Chelsea Finn, and Sergey Levine. Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning. In *Conference on robot learning*, pp. 1094–1100. PMLR, 2020b.

Amir Zamir, Alexander Sax, William Shen, Leonidas Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling task transfer learning, 2018. URL https://arxiv.org/abs/1804.08328.

Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence. In *International Conference on Machine Learning*, pp. 3987–3995. PMLR, 2017.

Tiantian Zhang, Zichuan Lin, Yuxing Wang, Deheng Ye, Qiang Fu, Wei Yang, Xueqian Wang, Bin Liang, Bo Yuan, and Xiu Li. Dynamics-adaptive continual reinforcement learning via progressive contextualization. *IEEE Transactions on Neural Networks and Learning Systems*, 35(10):14588–14602, 2024. doi: 10.1109/TNNLS.2023.3280085.

Luisa Zintgraf, Sebastian Schulze, Cong Lu, Leo Feng, Maximilian Igl, Kyriacos Shiarlis, Yarin Gal, Katja Hofmann, and Shimon Whiteson. VariBAD: Variational Bayes-Adaptive Deep RL via Meta-Learning. *J. Mach. Learn. Res.*, 22(1), January 2021. ISSN 1532-4435.