EFFECT-GUIDED STRUCTURAL GROWTH IN PRETRAINED SKILL SPACES FOR LIFELONG REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Expandable representations can support lifelong reinforcement learning, but task failure alone does not reveal whether the active representation is the bottleneck or where additional capacity should be exposed. We introduce Effect-Guided Structural Skill Growth (EGSG), which treats the active span of a pretrained executable skill representation as geometry that can be interrogated online. For each policy-requested short-horizon effect, EGSG measures support mismatch and maps the request through a source-trained inverse model to one candidate latent realization. Persistent out-of-span residuals propose directions to test; they are not treated as proofs of global non-representability. A direction is committed only when held-out execution shows that exposing it reduces matched effect-realization error relative to both the current span and a source-covariance-matched alternative. On Meta-World MT10 and a 12-task MT50 lifelong stream, EGSG reaches 0.78 and 0.66 final average success, respectively; on MT10, it accepts only 2.4 structural updates on average while outperforming interface-matched fixed-capacity and comparable-growth controls. Across the 24 accepted MT10 updates, the proposed directions align with held-out expert residual structure above matched nulls, reduce normalized out-of-span residuals by 19% and matched effect error by 27%, and improve matched success by 27 percentage points before subsequent policy optimization. No post-hoc false-growth event is observed among these 24 accepted updates; removing execution validation reintroduces such events. A five-task real-robot stream provides a preliminary feasibility demonstration.

1 INTRODUCTION

Lifelong reinforcement learning requires acquiring new behaviors while reusing previous experience, while managing forgetting and transfer across sequentially encountered tasks (Kirkpatrick et al., 2016; Schwarz et al., 2018; Wolczyk et al., 2021; 2022). Expandable and structured methods address this challenge by adding or composing policy anchors, components, modules, experts, or skills (Mendez et al., 2022; Gaya et al., 2023; Hendawy et al., 2024; Meng et al., 2025; Liu et al., 2025b;a). Yet poor task performance does not by itself identify a representational bottleneck: optimization, exploration, transfer, and representation can all contribute to failure (Wolczyk et al., 2022; Ahn et al., 2025).

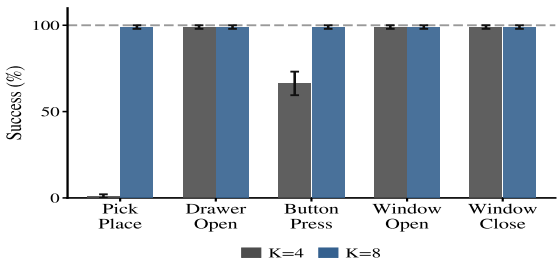

Figure 1: **Unseen tasks differ in how much they benefit from enlarging the active source-learned skill span.** Success under $K{=}4$ versus a diagnostic $K{=}8$.

Figure 1 shows the distinction within a common pretrained skill representation. Enlarging the active source-learned span helps some held-out tasks substantially but changes others little. The relevant question is therefore whether the *currently active portion* of a reusable representation adequately serves the effects the policy repeatedly requests, not simply whether the task is difficult.

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.

We focus on a specific form of representational limitation: the pretrained executable representation may already support useful latent directions that are not exposed through the current low-dimensional active span. Accordingly, *growth* in this work means expanding the policy-accessible active span within a fixed pretrained representation, rather than enlarging the decoder's executable support or claiming acquisition of globally novel skills.

EGSG studies this question with an orthonormal active skill basis and a source-trained inverse model. For each requested effect, the inverse model produces one candidate latent realization. A substantial out-of-span component is local evidence that the current request-to-skill pathway is poorly served by the active span, but it does not establish global non-representability: another in-span realization may produce the same effect.

This motivates a strict separation of roles. Residual geometry determines where to test: persistent, directionally coherent residuals nominate a candidate direction. Held-out execution determines whether to grow: the direction is committed only when exposing it improves realization of the same requests relative to the current span and a matched alternative. A validated update therefore expands only the policy-accessible active span inside the fixed pretrained executable representation; it neither modifies the decoder nor enlarges its executable support.

Our contributions are threefold:

- We formulate lifelong structural adaptation within a fixed pretrained skill representation as an active-span adequacy problem, separating limitations of the currently exposed span from task novelty or low return.
- We introduce EGSG, a diagnose–propose–validate mechanism in which persistent inverse residuals nominate candidate latent directions and an episode-disjoint execution test decides whether each direction should be exposed to the policy.
- We evaluate the mechanism with interface-matched fixed-capacity and comparable-growth controls, ten-seed lifelong streams, held-out direction and repair analyses, long-stream evaluation, and a preliminary five-task real-robot study. Across the 24 accepted MT10 updates, no post-hoc false-growth event is observed.

2 PRELIMINARIES

We distinguish *how a local behavior is represented* from *what change it produces in the environment*.

2.1 LIFELONG SKILL REUSE

We consider a sequence of tasks $\{\mathcal{M}_1, \mathcal{M}_2, \ldots\}$, each an MDP $\mathcal{M}_i = (\mathcal{S}, \mathcal{A}, \mathcal{P}_i, r_i, \gamma)$ sharing state and action spaces. When a task arrives, the agent retains its representations and may interact online; no incoming-task demonstrations are assumed. The objective is to reuse the current active span when it serves the policy's requested effects and to expand it only after persistent online evidence and execution validation (Sec. 3).

2.2 SKILL AND EFFECT REPRESENTATIONS

Latent skills. An h-step behavior window is $\tau_t^h = (s_t, a_{t:t+h-1}, s_{t+h})$. A conditional encoder maps its action chunk to $\xi_t = E_{\text{skill}}(s_t, a_{t:t+h-1}) \in \mathbb{R}^{d_\xi}$, and $D_{\text{skill}}(s_t, \xi_t)$ reconstructs an executable chunk. Thus ξ_t represents local behavior over horizon h, not a discrete skill label.

Realized transition effects. For the same window, $\Delta_t^{\text{real}} = \Psi_\eta(s_t, s_{t+h}) \in \mathbb{R}^{d_e}$: ξ_t describes how the agent behaved, whereas Δ_t^{real} describes what changed. Source windows $\mathcal{D}_{\text{past}}$ yield paired samples $\{(\xi_i, \Delta_i^{\text{real}})\}$ and empirical effect support $\widehat{\mathcal{E}}_{\text{past}} = \{\Delta_i^{\text{real}}\}_{i=1}^N$, which contains observed rather than all theoretically reachable effects.

Active skill span. An orthonormal basis $B_K = [b_1, \ldots, b_K] \in \mathbb{R}^{d_\xi \times K}$, $B_K^\top B_K = I_K$, defines $\mathcal{U}_K = \text{span}(B_K)$ with projector $P_{B_K} = B_K B_K^\top$. Any latent decomposes as $\xi = P_{B_K}\xi + (I - P_{B_K})\xi$, separating the component exposed through the current span from its orthogonal residual.

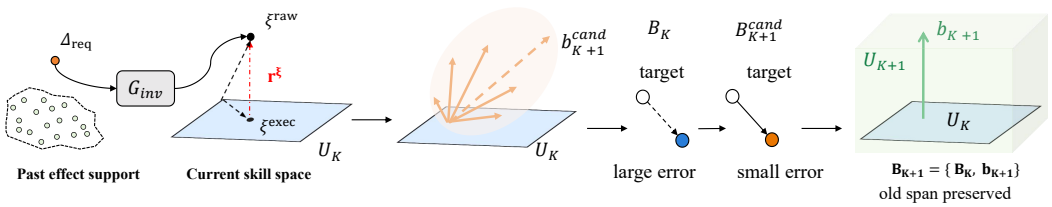

Figure 2: **EGSG diagnose–propose–validate–grow cycle.** **1** Requests are mapped to candidate latents and decomposed relative to the active span. **2** Persistent residual geometry proposes b_{K+1}. **3** Held-out execution compares the current, candidate, and matched alternative spans. **4** Only a validated proposal is appended to B_K; the decoder remains frozen (Sec. 3.5).

3 Effect-Guided Structural Growth of the Active Skill Span

We consider lifelong reinforcement learning with a structured skill representation acquired from previous experience. Rather than treating representation growth as simply increasing the number of available skill elements, EGSG asks whether the current request-to-skill pathway is adequately served by the *active span*: it detects requests whose candidate realizations consistently escape that span and uses this evidence to determine where an extension should be tested.

EGSG implements a four-stage cycle (Fig. 2): *diagnose* requests poorly supported by previous effects and expose their candidate out-of-span residuals (Secs. 3.1–3.2); *propose* a direction only when those residuals share persistent geometry; *validate* the proposal on held-out executions; and *grow* only after that test passes (Sec. 3.3). Effect novelty alone is not sufficient evidence for growth because an unfamiliar effect may still be realizable within the active span.

3.1 Effect-Guided Skill Reuse

At each decision step, the high-level policy proposes a short-horizon effect $\Delta_t^{\text{req}} \sim \pi_\theta(\cdot \mid s_t, g_t)$. EGSG first asks whether the requested effect is supported by previous experience. The support $\widehat{\mathcal{E}}_{\text{past}}$ is constructed once at initialization and *not* updated during lifelong interaction, keeping the semantics of d_t^e fixed over the lifetime. Let $\mathcal{N}_k(\Delta; \widehat{\mathcal{E}}_{\text{past}})$ denote the k nearest effects to Δ in this support;

$$d_{\text{supp}}(\Delta; \widehat{\mathcal{E}}_{\text{past}}) = \frac{1}{k} \sum_{\Delta_i \in \mathcal{N}_k(\Delta; \widehat{\mathcal{E}}_{\text{past}})} \|\Delta - \Delta_i\|_2, \tag{1}$$

and write $d_t^e = d_{\text{supp}}(\Delta_t^{\text{req}}; \widehat{\mathcal{E}}_{\text{past}})$. A small d_t^e indicates that the request is well supported by previous experience; $d_t^e > \tau_e$ marks it as poorly supported.

This criterion is deliberately not a growth criterion: mismatch only identifies where reuse from previous experience is uncertain, not that the representation lacks the behavior needed to realize the request.

3.2 Candidate Out-of-Span Evidence

The conditional inverse model produces a candidate latent realization

$$\xi_t^{\text{raw}} = G_{\text{inv}}(s_t, \Delta_t^{\text{req}}). \tag{2}$$

Here ξ_t^{raw} is one candidate realization rather than a unique behavior associated with the effect.

With the projector $P_{B_K} = B_K B_K^\top$, we decompose the candidate as

$$\xi_t^{\text{exec}} = P_{B_K} \xi_t^{\text{raw}}, \qquad r_t^{\xi} = (I - P_{B_K}) \xi_t^{\text{raw}}. \tag{3}$$

where the first term is available for execution and r_t^{ξ} is not represented by $\mathrm{span}(B_K)$. To compare residuals across candidates of different scale:

$$\rho_t = \frac{\|r_t^{\xi}\|_2}{\|\xi_t^{\mathrm{raw}}\|_2 + \epsilon}. \tag{4}$$

A small ρ_t means that an effect-space mismatch can nevertheless be realized largely within the current span. Conversely, the conjunction $d_t^e > \tau_e$ and $\rho_t > \tau_\xi$ marks a request that is poorly supported *and* whose realization contains a substantial out-of-span component.

Importantly, this is an *operational mismatch signal*, not a proof of global non-representability: another latent vector inside $\mathrm{span}(B_K)$ could in principle realize the same effect. An out-of-span inverse realization is evidence for a candidate extension, not proof that no alternative in-span realization exists. We therefore use such residuals only as local evidence for proposing structural extensions. Whether an extension is behaviorally useful is determined independently by execution validation in Sec. 3.3. We treat each such event as *local structural evidence*, not as an immediate trigger: a single residual may result from inverse-model error, exploratory proposals, or transient variation, so the relevant question is whether similar out-of-span structure persists across interaction.

3.3 Proposal and Execution-Validated Structural Growth

EGSG converts local mismatch evidence into an extension proposal only when the residuals exhibit persistent shared geometry. While the current basis B_K remains unchanged, we collect $\mathcal{R} = \left\{ r_i^{\xi} \;\middle|\; d_i^e > \tau_e,\ \rho_i > \tau_\xi \right\}$. Let v_1 be the dominant principal direction of the (mean-centered) residual set and $\lambda_1, \ldots, \lambda_{d_\xi}$ the corresponding covariance eigenvalues. We measure directional coherence by $\kappa = \lambda_1 / \sum_j \lambda_j$. A large κ means that independently observed candidate mismatches repeatedly expose a similar out-of-span direction. We interpret this direction as a *proposal* for extension rather than as an identified ground-truth missing skill.

A single residual trivially has $\kappa \equiv 1$ ($\kappa{:=}0$ while $|\mathcal{R}| < 8$); *growth requires at least* $n_{\min} = 96$ *eligible residuals over* ≥ 6 *distinct episodes*, buffer capped at 256.

When $\kappa > \tau_\kappa$, the dominant residual direction v_1 is re-projected onto the orthogonal complement of the current span and normalized:

$$b_{K+1} = \frac{(I - P_{B_K})v_1}{\|(I - P_{B_K})v_1\|_2}. \tag{5}$$

Because b_{K+1} lies in the orthogonal complement of the current span, $\mathrm{span}(B_K) \subset \mathrm{span}(B_{K+1})$: the old structure is retained, and the added direction is determined by structure consistently uncovered outside it. Expansion is further limited to at most 2 accepted updates per task, a structural cap $K \leq K_{\max}$, and a 30% warm-up in which evidence accumulates but no expansion is accepted.

Residual geometry therefore determines *where to test*, whereas execution determines *whether to commit*; Sec. 4.5 tests these roles separately.

Pre-commit validation. Because a coherent residual cloud can in principle be produced by inverse-model artifacts rather than behaviorally useful active-span mismatch (App. F.4), a candidate direction is never committed on geometric evidence alone. When κ passes, *one* candidate b_{K+1} is evaluated on $M{=}64$ held-out pending requests: episode-disjoint interaction windows that have *not* entered the residual buffer or influenced candidate selection. For each request, the same initial state is restored, the candidate skill executed under the frozen decoder, and the matched effect error $\|\Delta^{\mathrm{req}} - \Delta^{\mathrm{real}}\|_2$ recorded. The candidate is committed only if its mean effect error under B_{K+1} improves over B_K by more than $\delta{=}0.02$ *and* exceeds the improvement of a source-covariance-matched direction evaluated on the same 64 requests; otherwise the candidate is discarded, the buffer keeps accumulating, and no alternative candidate is tried until the next coherence trigger. Exact state restoration is a *simulator convenience*: the hardware variant uses randomized matched-layout resets instead (App. F.5). Algorithm 1 (App. B) gives the complete procedure and rejection path.

3.4 Learning and Lifelong Adaptation

Before lifelong interaction, behavior windows $\tau_t^h = (s_t, a_{t:t+h-1}, s_{t+h})$ from previous tasks learn the latent skill representation, realized transition effects, and their correspondence (Sec. 2.2).

Reusable representation learning. The skill autoencoder and inverse model are trained with MSE objectives (Eqs. 7–8 in App. A); Ψ_η follows the SLIM-0.5B stage-1 protocol (Wang et al., 2026) (EMA target; forward dynamics plus flow-matching inverse), with $\Delta_t^{\text{real}} = \phi(s_{t+h}) - \phi(s_t)$. After pretraining, B_{K_0} is formed from the leading K_0 principal directions of past-task skill embeddings and $\widehat{\mathcal{E}}_{\text{past}} = \{\Delta_i^{\text{real}}\}_{i=1}^N$; these initialize the geometric objects used for diagnosis and growth.

Support-regularized policy adaptation. On an incoming task, the high-level policy is optimized through Soft Actor-Critic (SAC) (Haarnoja et al., 2018), and the support geometry of Sec. 3.1 also regularizes its proposals. The actor's proposals are bounded within a source effect subspace for exploration stability (App. A); the conditional inverse model, however, maps to the *full* d_ξ-dimensional skill space with no subspace confinement, so out-of-span evidence is not restricted a priori. We penalize only the portion of the support distance exceeding a margin:

$$\mathcal{L}_{\text{supp}} = 0.1\,\mathbb{E}_t\left[\left[\frac{d_{\text{supp}}\big(\Delta_t^{\text{req}}; \widehat{\mathcal{E}}_{\text{past}}\big)}{\tau_{\text{supp}}} - 1\right]_+^2\right], \tag{6}$$

where $[x]_+ = \max(x, 0)$. This normalized margin is the form used in the reported implementation (App. A). The actor objective is $\mathcal{L}_{\text{actor}} = \mathcal{L}_{\text{SAC}} + \lambda_{\text{supp}}\mathcal{L}_{\text{supp}}$. Unlike a hard constraint, this margin leaves proposals within τ_{supp} unpenalized, permitting local exploration while discouraging large departures from supported regions. The support geometry thus plays two distinct roles—*regularizing* what the policy proposes and *gating* which residuals count as structural evidence—sharing d_{supp} but nothing else.

Online lifelong adaptation. During interaction, the proposed Δ_t^{req} is mapped by G_{inv} to ξ_t^{raw}; its in-span component $\xi_t^{\text{exec}} = P_{B_K}\xi_t^{\text{raw}}$ is decoded and executed, and the transition produces $\Delta_t^{\text{real}} = \Psi_\eta(s_t, s_{t+h})$. Policy optimization proceeds continuously, while the criteria of Secs. 3.1–3.3 accumulate evidence about the adequacy of the current span. When a coherent residual direction is accepted, the basis is extended to B_{K+1}, the residual buffer is cleared, and subsequent evidence is evaluated relative to the enlarged space.

3.5 Executing Through an Expanded Basis

Growth exposes a new latent direction to the policy without changing D_{skill}, so the added direction must remain executable by the frozen decoder. This is plausible because the decoder is trained on source embeddings beyond the initial top-K_0 span (top 4/8 source PCA directions explain $90.0\%/97.5\%$ of variance), while support regularization and norm-clamped decoder inputs limit how far online requests and inverse outputs move from source-supported regions. These choices do not guarantee executability, so we test it directly.

The execution analyses in App. F.2 separate this issue from direction selection. Held-out pick-place behavior fails under the source $K{=}4$ span but succeeds under the source-derived $K{=}8$ span (0% vs. 100%); online EGSG directions also show better matched execution than uniformly random complement directions. Because the decoder is frozen and the spans are nested, previously represented latent skills retain the same decoder mapping, while policy-level retention remains empirical. The matched-query analysis in Sec. 4.4 and a decoder fine-tuning ablation provide complementary execution evidence.

4 Experiments

We ask four questions: does selective growth improve lifelong performance; do capacity, direction, or timing explain the gain; do accepted updates directly repair matched requests; and which components are required for reliable growth? We also test the full framework on a physical robot.

Table 1: **Lifelong performance over 10 independent training seeds.** Results are mean ± s.d.; task-level metrics are first averaged within each seed. MT10 uses a controlled stream (3×10^5 steps/task), while MT50 contains twelve incoming tasks (1×10^5 steps/task). Growth denotes accepted structural additions on MT10. Final Struct. reports each method's native structural size. For EGSG-family methods, $K_0 = 4$ and $K_{\max} = 10$.

	MT10 Controlled Stream		12-Task Stream from MT50		MT10 Structure	
Method	**Avg. ↑**	**Forget. ↓**	**Avg. ↑**	**Forget. ↓**	**Growth**	**Final Struct.**
Lifelong adaptation baselines						
Single-Task SAC	0.62 ± 0.07	–	0.57 ± 0.06	–	–	Task-spec.
Fine-tuning SAC	0.21 ± 0.12	0.45 ± 0.11	0.24 ± 0.09	0.52 ± 0.10	–	Fixed
A-GEM	0.42 ± 0.08	0.33 ± 0.07	0.38 ± 0.06	0.36 ± 0.07	–	Fixed
ClonEx	0.53 ± 0.09	0.23 ± 0.05	0.49 ± 0.08	0.26 ± 0.06	–	Fixed
Reset & Distill	0.60 ± 0.06	0.15 ± 0.04	0.58 ± 0.05	0.18 ± 0.05	–	Fixed
Structured / expandable baselines						
MOORE	0.47 ± 0.05	0.20 ± 0.04	0.42 ± 0.09	0.24 ± 0.04	0	4 experts
CSP	0.57 ± 0.07	0.15 ± 0.03	0.52 ± 0.05	0.18 ± 0.08	2.8 ± 0.9	6.8 ± 0.9 anchors
LEGION	0.62 ± 0.10	0.17 ± 0.05	0.60 ± 0.07	0.19 ± 0.03	5.0 ± 0.7	6.0 ± 0.7 comp.
EGSG-family structural controls						
EGSG Fixed Basis	0.40 ± 0.08	0.24 ± 0.07	0.34 ± 0.10	0.27 ± 0.06	0	$K = 4$
EGSG Fixed Large	0.51 ± 0.09	0.17 ± 0.04	0.46 ± 0.08	0.20 ± 0.05	0	$K = 8$
EGSG Random Growth	0.55 ± 0.12	0.15 ± 0.06	0.50 ± 0.10	0.17 ± 0.07	2.4 ± 1.1	$K = 6.4 \pm 1.1$
EGSG PCA Growth	0.68 ± 0.06	0.08 ± 0.03	0.57 ± 0.06	0.20 ± 0.04	2.4 ± 0.8	$K = 6.4 \pm 0.8$
EGSG PCA Growth (auto)	0.61 ± 0.09	0.11 ± 0.05	0.52 ± 0.08	0.22 ± 0.06	3.1 ± 1.0	$K = 7.1 \pm 1.0$
Cov. Growth + Exec. Val.	0.69 ± 0.08	0.10 ± 0.05	0.60 ± 0.07	–	2.6 ± 0.8	$K = 6.6 \pm 0.8$
EGSG	$\mathbf{0.78 \pm 0.04}$	$\mathbf{0.05 \pm 0.02}$	$\mathbf{0.66 \pm 0.05}$	$\mathbf{0.06 \pm 0.03}$	$\mathbf{2.4 \pm 0.7}$	$\mathbf{K = 6.4 \pm 0.7}$
Component ablations						
w/o Effect Gate	0.72 ± 0.06	0.20 ± 0.07	0.60 ± 0.08	0.24 ± 0.06	3.8 ± 0.9	$K = 7.8 \pm 0.9$
w/o Residual Gate	0.68 ± 0.08	0.25 ± 0.07	0.55 ± 0.09	0.30 ± 0.07	$5.6 \pm 0.7^{\dagger}$	$K = 9.6 \pm 0.7$
w/o Coherence	0.70 ± 0.07	0.22 ± 0.06	0.58 ± 0.07	0.28 ± 0.06	4.9 ± 1.0	$K = 8.9 \pm 1.0$
w/o Support Reg.	0.38 ± 0.09	0.30 ± 0.08	0.31 ± 0.07	0.34 ± 0.07	0.0 ± 0.0	$K = 4.0 \pm 0.0$
w/o Exec. Validation	0.72 ± 0.05	0.09 ± 0.04	0.62 ± 0.06	–	3.3 ± 1.3	$K = 7.3 \pm 1.3$

Notes. Forgetting is computed within each lifelong stream. Structural sizes use each method's native unit and are not directly comparable across method families. † Some runs reach $K_{\max} = 10$. – denotes a metric that is not applicable or was not evaluated for the corresponding method or auxiliary control.

4.1 Experimental Setup

Streams and protocol. We evaluate on Meta-World (Yu et al., 2020). MT10 uses five source tasks for initialization and five sequential incoming tasks. For a longer-horizon evaluation, we construct a 12-task incoming stream from MT50 using the same source set. No incoming-task data enters representation pretraining, effect support, inverse-model training, or the initial K_0=4 source-PCA basis, and the decoder remains frozen. MT10/MT50 allocate $3 \times 10^5/1 \times 10^5$ environment steps per incoming task. After each task, all seen tasks are evaluated for 20 held-out episodes. Results are mean ± s.d. over 10 training seeds.

External lifelong baselines retain their native primitive-action formulations. EGSG-family controls instead share the same h=5 skill-chunk interface, decoder, initialization, and adaptation stack; mechanistic conclusions therefore come from this interface-matched family. Full protocol and task-order controls are in App. A and App. E.6.

Metrics and baselines. We report final average success, forgetting, accepted structural additions (*Growth*), and final structural size. Additional metrics and statistical details are in App. E.1. Baselines are Fine-tuning SAC, A-GEM (Chaudhry et al., 2018), ClonEx (Wolczyk et al., 2022), Reset & Distill (Ahn et al., 2025), MOORE (Hendawy et al., 2024), CSP (Gaya et al., 2023), and LEGION (Meng et al., 2025); Single-Task SAC is the from-scratch reference. EWC and Progress & Compress appear in App. E.7.

4.2 Lifelong Performance and Structural Efficiency

Table 1 summarizes lifelong performance and structural adaptation.

EGSG reaches 0.78 ± 0.04 success with 0.05 ± 0.02 forgetting on MT10 using 2.4 ± 0.7 accepted additions. On the 12-task MT50-derived stream it reaches 0.66 ± 0.05 success with 0.06 ± 0.03 forgetting. Because external methods use different action and structural parameterizations, the analyses below use the interface-matched EGSG-family controls to isolate mechanism.

4.3 Capacity, Direction, and Evidence-Guided Growth

More fixed capacity helps but does not explain EGSG. Increasing the source basis from $K = 4$ to $K = 8$ raises MT10 success from 0.40 ± 0.08 to 0.51 ± 0.09, showing that the active span can be restrictive. Fixed Large nevertheless remains below EGSG (0.78 ± 0.04).

Direction selection matters. Random Growth and PCA Growth add comparable numbers of directions but reach 0.55 ± 0.12 and 0.68 ± 0.06 success. App. F.2 shows why the random control is demanding: source-covariance directions are more executable than arbitrary complement directions. PCA Growth still remains below EGSG at comparable structural size.

Independent timing does not close the gap. PCA Growth (auto), which triggers independently, reaches 0.61 ± 0.09 with 3.1 ± 1.0 additions. Replacing EGSG's residual-derived proposal with a source-covariance direction while retaining execution validation reaches 0.69 ± 0.08. Thus capacity, decoder compatibility, and timing leave additional value for interaction-conditioned direction selection.

4.4 From Diagnosis to Structural Repair

We test the mechanism at three levels: diagnosis of active-span difficulty, directional information in the proposal, and direct repair under matched execution.

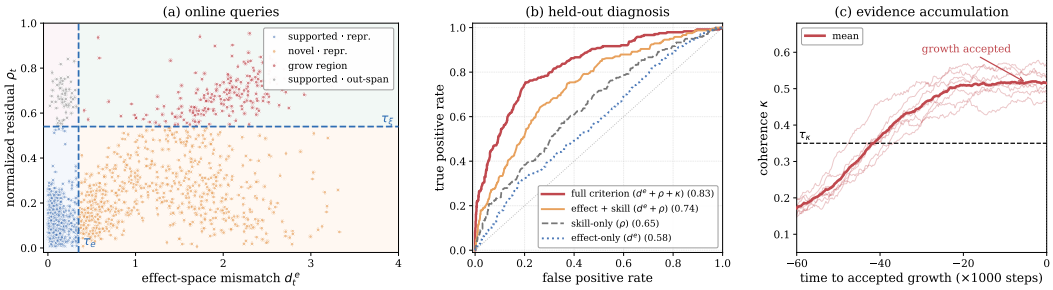

Figure 3: **Online diagnosis of active-span difficulty.** (a) Effect-support distance d_t^e and normalized residual ρ_t capture complementary aspects of request–span mismatch. (b) The full criterion, combining effect mismatch, normalized residual, and residual coherence, improves held-out diagnosis under operational active-span difficulty labels. (c) Directional coherence κ accumulates before accepted structural updates.

Diagnosis. Across 75,272 held-out matched queries, the joint (d_t^e, ρ_t, κ) criterion reaches AUC 0.83, versus 0.74 for Effect+Skill, 0.65 for Skill-only, and 0.58 for Effect-only [Fig. 3(b)]. With execution-grounded labels that do not use the inverse residual, the same ordering remains $(0.79, 0.70, 0.62, 0.57)$. Episode-clustered inference and seed-level analyses are in App. F.1.

Proposal. Using evaluation residuals disjoint from candidate construction, the 24 accepted directions align with held-out expert residual structure at 0.36, versus 0.26 for a paired random-orthogonal null and 0.30 for a source-covariance-matched null [Fig. 6(b)]. The seed-level advantage over the random null is $+0.10 \pm 0.03$ and positive in all ten seeds (App. F.1). This supports task-relevant directional information, not recovery of a unique missing skill.

Repair. On a fresh held-out set of pre-growth requests, disjoint from both candidate construction and all pre-commit validation sets, we re-evaluate each accepted update with the matched initial state, policy, and decoder held fixed. Expansion reduces normalized out-of-span residual from 0.70

to 0.57 and matched effect error from 0.41 to 0.30, while matched success rises from 0.23 to 0.50 [Fig. 6(a,c,d)]. An unchanged-basis placebo changes matched effect error by only -0.005 ± 0.021. Because these measurements precede subsequent policy optimization, they isolate the immediate effect of exposing the accepted direction; new-direction usage subsequently rises to 0.41 (App. F.2).

Full-chain and synthetic stress tests show that inverse extrapolation need not recover the held-out direction and that coherent artifacts can occur (Apps. F.3–F.4). Residual geometry therefore determines *where to test*; execution determines whether to commit.

4.5 Component Ablations

The gates fail differently when removed. Without the effect gate, growth rises to 3.8 ± 0.9 and forgetting to 0.20 ± 0.07; without the residual gate, growth reaches 5.6 ± 0.7 and final $K = 9.6 \pm 0.7$, with some runs hitting $K_{\max}$. Removing coherence yields 4.9 ± 1.0 additions, whereas removing support regularization yields no accepted growth. The gates therefore regulate distinct failure modes.

Execution validation controls commitment rather than proposal. Removing it reduces MT10 success from 0.78 ± 0.04 to 0.72 ± 0.05, increases growth from 2.4 ± 0.7 to 3.3 ± 1.3, and yields 0.9 ± 1.0 post-hoc false-growth events per seed. Replacing the residual proposal with a covariance direction while retaining validation reaches 0.69 ± 0.08 and 0.4 ± 0.6 such events per seed. No post-hoc false-growth event is observed among the 24 updates accepted by full EGSG. Threshold sensitivity is in App. F.1.

4.6 Real-Robot Demonstration

We test EGSG on a five-task physical manipulation stream initialized from three auxiliary real-demonstration tasks; no incoming-task demonstrations are used. Each method has one complete stream, so this is a feasibility study rather than a statistically controlled comparison (App. G).

Table 2: **Real-robot lifelong evaluation.** Per-task entries report success over 20 trials with $\pm$ one binomial standard error; Avg. reports the mean $\pm$ sample s.d. across the five tasks. Each method is evaluated with one complete hardware stream.

Method	Drawer	Orange	Take Cup	Place Cup	Stack Cups	Avg. ↑	Forget. ↓	Reuse ↑	Final Struct.
Fine-tuning	0.25 ± 0.10	0.15 ± 0.08	0.10 ± 0.07	0.15 ± 0.08	0.05 ± 0.05	0.14 ± 0.07	0.35	0.00	Fixed
Reset & Distill	0.50 ± 0.11	0.30 ± 0.10	0.25 ± 0.10	0.30 ± 0.10	0.20 ± 0.09	0.31 ± 0.12	0.20	0.20	Fixed
MOORE	0.40 ± 0.11	0.25 ± 0.10	0.20 ± 0.09	0.25 ± 0.10	0.15 ± 0.08	0.25 ± 0.09	0.25	0.20	Fixed (4 exp.)
LEGION (expandable)	0.45 ± 0.11	0.35 ± 0.11	0.30 ± 0.10	0.35 ± 0.11	0.30 ± 0.10	0.35 ± 0.06	0.15	0.20	$1 \rightarrow 3$ comp.
EGSG Fixed Basis	0.45 ± 0.11	0.30 ± 0.10	0.20 ± 0.09	0.20 ± 0.09	0.15 ± 0.08	0.26 ± 0.11	0.25	0.00	$K = 4$
EGSG	0.60 ± 0.11	0.50 ± 0.11	0.25 ± 0.10	0.35 ± 0.11	0.55 ± 0.11	0.45 ± 0.15	0.10	0.40	$4 \rightarrow 6$

EGSG accepts two additions ($K = 4 \rightarrow 6$); no further structural addition is accepted. The stream ends at 0.45 average success and 0.10 forgetting. We interpret these results as physical-feasibility evidence, not a controlled performance ranking; per-stage behavior is in Fig. 11 and App. G.

5 Related Work

Lifelong adaptation within reusable representations. Regularization, replay, distillation, policy reuse, and transfer mitigate forgetting without necessarily changing the reusable representation (Kirkpatrick et al., 2016; Chaudhry et al., 2018; Wolczyk et al., 2021; 2022; Ahn et al., 2025). Complementary lifelong mechanisms include neural composition, exploration, neuromodulation, task-structure methods, trajectory replay, and world models (Mendez et al., 2022; Fu et al., 2022; Steinparz et al., 2022; Lee et al., 2024; Zheng et al., 2022; Chen et al., 2024; Kessler et al., 2023; Pan et al., 2025).

Structured and expandable representations. Prior work structures policy or skill spaces through orthogonal experts (MOORE; Hendawy et al., 2024), policy-subspace anchors (Gaya et al., 2023), modular or neuroplastic allocation (Mendez et al., 2022; Liu et al., 2025a), non-parametric components (Meng et al., 2025), and skill libraries or lifecycle mechanisms (Liu et al., 2025b; Ding & Zhu, 2022; Shen et al., 2026; Mao et al., 2026). Skill-effect models also support planning and transfer

(Liang et al., 2022; Zentner et al., 2021). EGSG differs in the *unit and decision rule of growth*: it activates a new direction inside a fixed pretrained executable latent representation rather than creating a new policy, expert, or component. Persistent residual geometry nominates the direction, and held-out execution decides whether to commit it. We do not claim the first behavior-dependent expansion mechanism; the contribution is an execution-validated direction-level growth rule.

6 Conclusion

EGSG casts lifelong structural adaptation as a diagnose–propose–validate problem over the active span of a pretrained executable skill representation. Persistent out-of-span inverse residuals nominate candidate directions, while held-out execution determines whether exposing each direction repairs matched requests enough to justify expansion. The claim is operational: a residual is local proposal evidence, not proof that an effect is globally unrealizable within the current span.

Across the evaluated manipulation streams, this separation yields sparse structural updates, improves matched effect realization before subsequent policy optimization, and outperforms interface-matched fixed-capacity and comparable-growth controls. EGSG expands which latent directions are exposed to the policy; it does not enlarge the executable support of the frozen decoder. Behaviors outside decoder support, or effects outside the source-derived proposal space, require adaptation of the underlying representations.

Limitations. EGSG adds one direction at a time and therefore targets residual evidence with a dominant direction. Its evaluation is concentrated on one manipulation benchmark family and one source/incoming construction, with only one complete hardware stream per method. More generally, inverse residuals do not solve the inverse-identifiability problem, and active-span expansion remains bounded by the pretrained decoder and effect-proposal supports (App. H).

Reproducibility statement

Hyperparameters, calibration, protocol details, and additional metrics: App. A–E.1; baseline and initialization details: App. D; datasets and scripts: supplementary material.

AI Use Statement

Generative AI tools were used to assist with methodological and experimental-design feedback, interpretation and consistency checking of reported results, manuscript organization and editing, and figure/layout refinement. AI-assisted suggestions were reviewed by the authors and were not treated as empirical evidence. All experiments, reported numerical results, and final scientific claims were verified by the authors. The authors take responsibility for the final content of this work.

References

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 *International Conference on Learning Representations*, volume 2025, pp. 75036–75060, 2025.

Arslan Chaudhry, Marc'Aurelio Ranzato, Marcus Rohrbach, and Mohamed Elhoseiny. Efficient lifelong learning with a-gem. *ArXiv*, abs/1812.00420, 2018. URL `https://api.semanticscholar.org/CorpusID:54443381`.

Feng Chen, Fuguang Han, Cong Guan, Lei Yuan, Zhilong Zhang, Yang Yu, and Zongzhang Zhang. Stable continual reinforcement learning via diffusion-based trajectory replay. *arXiv preprint arXiv:2411.10809*, 2024.

Fan Ding and Fei Zhu. Hliferl: A hierarchical lifelong reinforcement learning framework. *Journal of King Saud University Computer and Information Sciences*, 34(7):4312–4321, 2022.

Haotian Fu, Shangqun Yu, Michael S. Littman, and George Dimitri Konidaris. Model-based lifelong reinforcement learning with bayesian exploration. *ArXiv*, abs/2210.11579, 2022. URL https://api.semanticscholar.org/CorpusID:253080700.

Jean-Baptiste Gaya, Thang Doan, Lucas Caccia, Laure Soulier, Ludovic Denoyer, and Roberta Raileanu. Building a subspace of policies for scalable continual learning. In *International Conference on Learning Representations*, 2023.

Tuomas Haarnoja, Aurick Zhou, P. Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. *ArXiv*, abs/1801.01290, 2018. URL https://api.semanticscholar.org/CorpusID:28202810.

Ahmed Hendawy, Jan Peters, and Carlo D'Eramo. Multi-task reinforcement learning with mixture of orthogonal experts. In *International Conference on Learning Representations*, volume 2024, pp. 6344–6365, 2024.

Samuel Kessler, Mateusz Ostaszewski, MichałPaweł Bortkiewicz, Mateusz Żarski, Maciej Wolczyk, Jack Parker-Holder, Stephen J Roberts, Piotr Mi, et al. The effectiveness of world models for continual reinforcement learning. In *Conference on Lifelong Learning Agents*, pp. 184–204. PMLR, 2023.

James Kirkpatrick, Razvan Pascanu, Neil C. Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114:3521 – 3526, 2016. URL https://api.semanticscholar.org/CorpusID:4704285.

Sebastian Lee, Samuel Liebana, Claudia Clopath, and Will Dabney. Lifelong reinforcement learning via neuromodulation. *arXiv preprint arXiv:2408.08446*, 2024.

Jacky Liang, Mohit Sharma, Alex LaGrassa, Shivam Vats, Saumya Saxena, and Oliver Kroemer. Search-based task planning with learned skill effect models for lifelong robotic manipulation. In *2022 International Conference on Robotics and Automation (ICRA)*, pp. 6351–6357. IEEE, 2022.

Jiashun Liu, Johan S Obando Ceron, Aaron Courville, and Ling Pan. Neuroplastic expansion in deep reinforcement learning. In *International Conference on Learning Representations*, volume 2025, pp. 90056–90078, 2025a.

Tenglong Liu, Jianxiong Li, Yinan Zheng, Haoyi Niu, Yixing Lan, Xin Xu, and Xianyuan Zhan. Skill expansion and composition in parameter space. In *International Conference on Learning Representations*, volume 2025, pp. 85192–85228, 2025b.

Bo Mao, Jie Zhou, Yutao Yang, Xin Li, Xian Wei, Qin Chen, Xingjiao Wu, and Liang He. Learning while acting: A skill-enhanced test-time co-evolution framework for online lifelong learning agents. *arXiv preprint arXiv:2606.04815*, 2026.

Jorge Armando Mendez Mendez, Harm van Seijen, and Eric Eaton. Modular lifelong reinforcement learning via neural composition. *ArXiv*, abs/2207.00429, 2022. URL https://api.semanticscholar.org/CorpusID:247656619.

Yuan Meng, Zhenshan Bing, Xiangtong Yao, Kejia Chen, Kai Huang, Yang Gao, Fuchun Sun, and Alois Knoll. Preserving and combining knowledge in robotic lifelong reinforcement learning. *Nature Machine Intelligence*, 7(2):256–269, 2025.

Minting Pan, Wendong Zhang, Geng Chen, Xiangming Zhu, Siyu Gao, Yunbo Wang, and Xiaokang Yang. Continual visual reinforcement learning with a life-long world model. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 146–162. Springer, 2025.

Jonathan Schwarz, Wojciech M. Czarnecki, Jelena Luketina, Agnieszka Grabska-Barwinska, Yee Whye Teh, Razvan Pascanu, and Raia Hadsell. Progress & compress: A scalable framework for continual learning. *ArXiv*, abs/1805.06370, 2018. URL https://api.semanticscholar.org/CorpusID:21718339.

Junhao Shen, Teng Zhang, Xiaoyan Zhao, and Hong Cheng. Dynamic skill lifecycle management for agentic reinforcement learning. *arXiv preprint arXiv:2605.10923*, 2026.

Christian Alexander Steinparz, Thomas Schmied, Fabian Paischer, Marius-Constantin Dinu, Vihang Prakash Patil, Angela Bitto-Nemling, Hamid Eghbal-zadeh, and Sepp Hochreiter. Reactive exploration to cope with non-stationarity in lifelong reinforcement learning. In *Conference on lifelong learning agents*, pp. 441–469. PMLR, 2022.

Jingkai Wang, Zihan Tang, Gu Zhang, Mingyu Cao, Jiapeng Chen, Jingjiao Zhao, Xiansheng Chen, Pengwei Wang, Lemao Liu, and Dejing Dou. Slim-0.5 b: Learning action-grounded predictive latents for robot manipulation. *arXiv preprint arXiv:2608.09771*, 2026.

Maciej Wolczyk, Michal Zajkac, Razvan Pascanu, Lukasz Kuci'nski, and Piotr Milos. Continual world: A robotic benchmark for continual reinforcement learning. In *Neural Information Processing Systems*, 2021. URL `https://api.semanticscholar.org/CorpusID:235166235`.

Maciej Wolczyk, Michał Zajac, Razvan Pascanu, Łukasz Kuciński, and Piotr Miłoś. Disentangling transfer in continual reinforcement learning. *Advances in Neural Information Processing Systems*, 35:6304–6317, 2022.

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, 2020.

KR Zentner, Ryan Julian, Ujjwal Puri, Yulun Zhang, and Gaurav S Sukhatme. A simple approach to continual learning by transferring skill parameters. *arXiv preprint arXiv:2110.10255*, 2021.

Xuejing Zheng, Chao Yu, and Minjie Zhang. Lifelong reinforcement learning with temporal logic formulas and reward machines. *Knowledge-Based Systems*, 257:109650, 2022.

A Implementation and Hyperparameters

A.1 Environment and Streams

All experiments use Meta-World (Yu et al., 2020) (v3 interface), 39-dimensional proprioceptive observations, 4-dimensional end-effector actions, and 150-step episodes. MT10 uses five source tasks (reach, push, door-open, drawer-close, peg-insert-side) and five sequential incoming tasks (pick-place, drawer-open, button-press-topdown, window-open, window-close), with 3×10^5 environment steps per incoming task. MT50 uses the same source set and twelve incoming tasks (pick-place, drawer-open, button-press-topdown, window-open, window-close, door-close, basketball, plate-slide, handle-press, coffee-push, peg-unplug-side, hammer), with 10^5 steps per task. No incoming-task data enters initialization. The growth mechanism is armed during source adaptation but accepts no direction in the reported runs; all reported growth therefore occurs on incoming tasks.

A.2 Representation Pretraining

Training uses h=5-step windows $(s_t, a_{t:t+h-1}, s_{t+h})$ from source tasks only: 120 scripted-expert episodes, 25 random-rollout episodes, and 50 source-policy episodes per task ($\sim$100k windows total).

Table 3: Representation-learning hyperparameters.

Symbol / name	Value
Skill-chunk horizon h	5
Skill embedding dim. d_ξ	16
Effect dim. d_e	32
Hidden width (all MLPs)	256
Losses (referenced from Sec. 3.4):	
skill AE (Eq. 7)	MSE reconstruction, Adam 3×10^{-4}, 60k steps, batch 512
Effect representation Ψ_η	SLIM-style stage-1 protocol (Wang et al., 2026): bounded state encoder ($\tanh \times 2.0$), EMA target (0.99), forward dynamics loss + flow-matching inverse loss, weights 1:0.125
Inverse model G_{inv}	MLP, MSE to source skill codes (Eq. 8), 60k steps
Initial basis B_{K_0}	top-K_0=4 PCA of source skill embeddings, QR-orthonormalized (90.0% variance)
Fixed Large basis	top-8 PCA of source skill embeddings (97.5% variance)
PCA details	embeddings mean-centered before SVD; basis columns unit-norm, sign fixed (largest-absolute component positive)
Structural cap $K_{\max}$	10
Effect support $\widehat{\mathcal{E}}_{\text{past}}$	87,600 source effects retained from the source windows, k=10 kNN, frozen after initialization

The source-pretraining losses are

$$\mathcal{L}_{\text{skill}} = \mathbb{E}_{\tau_t^h \sim \mathcal{D}_{\text{past}}} \left[\left\| D_{\text{skill}}\big(s_t, E_{\text{skill}}(s_t, a_{t:t+h-1})\big) - a_{t:t+h-1} \right\|_2^2 \right], \tag{7}$$

and

$$\mathcal{L}_{\text{inv}} = \mathbb{E}_{\tau_t^h \sim \mathcal{D}_{\text{past}}} \left[\left\| G_{\text{inv}}(s_t, \Delta_t^{\text{real}}) - \xi_t \right\|_2^2 \right]. \tag{8}$$

Action semantics and proposal bounds. EGSG-family methods operate on a block-level MDP: one decision emits an h=5 action chunk and all budgets are charged in environment steps. External lifelong baselines retain their native primitive-action interfaces; the EGSG-family mechanism comparisons are interface-matched. The actor parameterizes $\Delta^{\text{req}} = \mu_\Delta + (z \circ s_8) W_{\Delta,8}^\top$, $z \in [-1, 1]^8$, within ± 1 source s.d. along the leading effect PCs; inverse outputs are norm-clipped to ± 3 source s.d. Skill and effect dimensions are $d_\xi = 16$ and $d_e = 32$. The leading four/eight source-skill PCA directions explain $90.0\%/97.5\%$ of source-skill variance.

A.3 Threshold Calibration

All thresholds are calibrated once on episode-disjoint source data and then frozen: $\tau_e = 0.341$ and $\tau_\xi = 0.544$ are source 95th percentiles, $\tau_\kappa = 0.35$, and $\tau_{\text{supp}} = \tau_e$. The residual buffer is capped at 256; growth requires at least 96 eligible residuals from ≥ 6 episodes, a 30% task warm-up, at most two accepted updates per task, and $K \leq K_{\max} = 10$. Residual covariance is computed after global mean-centering (App. F.4); every candidate is checked by pre-commit execution validation.

Table 4: Online SAC and adaptation hyperparameters (identical for all EGSG-family methods).

Name	Value
Actor/critic	MLP 256×3, twin Q, LayerNorm critic
Learning rate	3×10^{-4} (Adam), actor updated every step
Discount γ	0.95 (block-level MDP horizon)
Target EMA τ_{slow}	0.001
Temperature	automatic, clamped to $[0.01, 0.25]$, target entropy -4
Batch size	512
Reward scaling	normalized by frozen running scale (first 20k steps, scale-only)
Rehearsal	batch: 50% current task + 50% uniform over reservoir (25k most recent transitions per past task)
Task boundary	critic, temperature, and optimizers reset; actor and structure retained
Effect proposal	$z \in [-1, 1]^8$ parameterizes $\Delta^{\text{req}} = \mu_\Delta + (z \circ s_8) W_{\Delta,8}^\top$ in the source top-8 effect-PCA subspace (± 1 source s.d.)
Inverse / execution	$\xi^{\text{raw}} = G_{\text{inv}}(s, \Delta^{\text{req}}) \in \mathbb{R}^{16}$ with no skill-subspace confinement; $\xi^{\text{exec}} = P_B \xi^{\text{raw}}$; decoder input norm-clipped to ± 3 source s.d.
Support regularization	λ_{supp}=1; $0.1\,[d/\tau_{\text{supp}} - 1]_+^2$
Decoder	frozen throughout lifelong interaction

Algorithm 1 Effect-Guided Structural Skill Growth (EGSG)

Require: Source data $\mathcal{D}_{\text{past}}$; calibrated $\tau_e, \tau_\xi, \tau_\kappa, \tau_{\text{supp}}$; $n_{\min}$=96 residuals from ≥ 6 episodes; buffer cap 256; warm-up 30%; ≤ 2 accepted updates/task; $K \leq K_{\max}$; M=64 validation requests; δ=0.02

Pretrain $E_{\text{skill}}, D_{\text{skill}}, \Psi_\eta, G_{\text{inv}}$; initialize frozen $B_{K_0}, \widehat{\mathcal{E}}_{\text{past}}$; $\mathcal{R} \leftarrow \varnothing$
for each incoming task **do**
  accepted-this-task $\leftarrow 0$
  for each decision step t **do**
    Propose $\Delta_t^{\text{req}} \sim \pi_\theta(\cdot \mid s_t, g_t)$ and compute $d_t^e = d_{\text{supp}}(\Delta_t^{\text{req}}; \widehat{\mathcal{E}}_{\text{past}})$
    $\xi_t^{\text{raw}} \leftarrow G_{\text{inv}}(s_t, \Delta_t^{\text{req}})$; $\xi_t^{\text{exec}} \leftarrow P_{B_K} \xi_t^{\text{raw}}$; decode, execute, and update π_θ
    $r_t^\xi \leftarrow (I - P_{B_K}) \xi_t^{\text{raw}}$; $\rho_t \leftarrow \|r_t^\xi\|_2 / (\|\xi_t^{\text{raw}}\|_2 + \epsilon)$
    if $d_t^e > \tau_e$ **and** $\rho_t > \tau_\xi$ **then**
      append $(r_t^\xi, s_t, \Delta_t^{\text{req}})$ to FIFO $\mathcal{R}$ (cap 256)
    end if
    if past warm-up **and** accepted-this-task < 2 **and** $K < K_{\max}$ **then**
      if $|\mathcal{R}| \geq n_{\min}$ over ≥ 6 episodes **and** $\text{Coherence}(\mathcal{R}) > \tau_\kappa$ **then**
        $v_1 \leftarrow \text{PC}_1(\mathcal{R})$; $c \leftarrow \text{Normalize}((I - P_{B_K}) v_1)$
        Draw M held-out current-task requests, disjoint from $\mathcal{R}$ and prior validation sets
        Evaluate matched effect errors ε^- (current B_K), ε^+ (candidate $[B_K, c]$), and ε^{cov} (source-covariance comparator)
        if $\text{mean}(\varepsilon^- - \varepsilon^+) > \delta$ **and** $\text{mean}(\varepsilon^- - \varepsilon^+) > \text{mean}(\varepsilon^- - \varepsilon^{\text{cov}})$ **then**
          $B_{K+1} \leftarrow [B_K, c]$; $K \leftarrow K + 1$; accepted-this-task $+= 1$; $\mathcal{R} \leftarrow \varnothing$
        else
          discard candidate; keep B_K unchanged
        end if
      end if
    end if
  end for
end for

A.4 ONLINE SAC

B COMPLETE PROCEDURE

C METRICS AND EVALUATION PROTOCOL

Stream metrics. After the final task, *Avg.* is the mean success over incoming tasks (20 held-out evaluation episodes per task). For task j, $\mathrm{FWT}_j = S_j - S_j^{\mathrm{st}}$, where S_j^{st} is a task-matched Single-Task SAC reference. Forgetting is $\frac{1}{N}\sum_j (S_j^{\mathrm{best}} - S_j^{\mathrm{final}})$. An incoming task contributes to *ReuseSuccess* when no structural update is accepted during its stage and final success is at least 0.5; because later growth can benefit earlier tasks, this is an upper bound on arrival-time reuse.

Direction and repair metrics. For held-out expert behavior, $r^{\mathrm{exp}} = (I - P_{B_K})\xi^{\mathrm{exp}}$, and v_1^{exp} is its dominant residual direction. Directional agreement is $A_{\mathrm{dir}} = |(b_{K+1})^\top v_1^{\mathrm{exp}}|$. The random null averages 100 unit vectors drawn from the orthogonal complement; its theoretical expectation is $\sqrt{2/(\pi d)} \approx 0.22$. For matched repair, the same $(s, \Delta^{\mathrm{req}})$ pairs are executed under B_K and B_{K+1}, and we record $\varepsilon_{\mathrm{eff}} = \|\Delta^{\mathrm{req}} - \Delta^{\mathrm{real}}\|_2$ and task success before subsequent policy optimization.

Statistics. Unless noted otherwise, results are mean $\pm$ sample s.d. over ten independent training seeds after task-level aggregation. Inference follows the experimental unit: seed-level statistics for streams, episode-clustered inference for diagnosis, and paired event-level nulls with seed-level aggregation for direction analysis.

D BASELINE AND INITIALIZATION DETAILS

All methods share observations, rewards, environment-step budgets, and evaluation, while retaining their native action and structural parameterizations. Mechanistic conclusions therefore rely on the interface-matched EGSG-family controls.

Table 5: Initialization information per method.

Method	Source demos	Source online	Pretrained reps.
EGSG family	yes (scripted)	yes (source tasks)	skill AE, Ψ_η, G_{inv}
Fixed Basis / Large	yes	yes	skill AE, Ψ_η, G_{inv}
Random / PCA Growth	yes	yes	skill AE, Ψ_η, G_{inv}
MOORE	yes	yes	orthogonal expert init
CSP / LEGION	yes	yes	–
ClonEx / R&D	yes	yes	expert buffer / init policy
A-GEM / Fine-tuning	–	yes	–
Single-Task SAC	–	–	–

ClonEx (Wolczyk et al., 2022) uses multi-head SAC with expert-buffer KL regularization; Reset & Distill (Ahn et al., 2025) resets an online expert per task and distills it into a maintained policy; CSP (Gaya et al., 2023) uses an anchor-based policy subspace; LEGION (Meng et al., 2025) uses a VAE behavior encoder with mixture-component growth; A-GEM (Chaudhry et al., 2018) uses episodic memory with gradient projection. MOORE (Hendawy et al., 2024), originally a multi-task method, is ported to sequential adaptation by initializing its orthogonal expert mixture on source tasks and using task-conditioned routing.

EGSG-family control semantics. Fixed Basis keeps the source $K{=}4$ span and disables growth; Fixed Large exposes the source top-8 skill-PCA span and also disables growth. Random Growth and PCA Growth serve as comparable-growth direction controls: they replace the residual-derived direction with, respectively, a normalized random complement direction or the highest-variance unused source-PCA direction while retaining the EGSG-family interface and adaptation stack. PCA Growth (auto) instead uses an independently determined trigger based on running proposal effect error and then adds the highest-variance unused source-PCA direction. Cov. Growth + Exec. Val.

retains held-out execution validation but substitutes the source-covariance proposal for the residual-derived candidate.

EWC and Progress & Compress are described in App. E.7.

E Additional Lifelong Results

E.1 Additional Metrics and Statistical Details

Table 6: Additional lifelong metrics corresponding to Table 1 (mean ± s.d. over ten seeds). MT50 forgetting is reported in the main table and omitted here.

Method	MT10 FWT ↑	MT50 FWT ↑	Reuse ↑	Trainable $P_{\text{peak}}/P_{\text{init}}$
Single-Task SAC	–	–	–	Task-specific
Fine-tuning SAC	$+0.10 \pm 0.04$	$+0.07 \pm 0.04$	–	1.0
A-GEM	$+0.10 \pm 0.03$	$+0.07 \pm 0.03$	–	1.0
ClonEx	$+0.16 \pm 0.05$	$+0.13 \pm 0.05$	–	1.0
Reset & Distill	$+0.17 \pm 0.06$	$+0.13 \pm 0.06$	–	1.0
MOORE	$+0.12 \pm 0.07$	$+0.09 \pm 0.02$	–	1.2
CSP	$+0.14 \pm 0.05$	$+0.12 \pm 0.04$	–	1.6
LEGION	$+0.17 \pm 0.04$	$+0.14 \pm 0.03$	–	3.0
EGSG Fixed Basis	$+0.15 \pm 0.06$	$+0.10 \pm 0.07$	–	1.0
EGSG Fixed Large	$+0.16 \pm 0.02$	$+0.11 \pm 0.05$	–	1.0
EGSG Random Growth	$+0.20 \pm 0.08$	$+0.14 \pm 0.06$	–	1.0
EGSG PCA Growth	$+0.22 \pm 0.03$	$+0.16 \pm 0.07$	–	1.0
EGSG PCA Growth (auto)	$+0.19 \pm 0.06$	$+0.13 \pm 0.05$	–	1.0
EGSG	$+0.28 \pm 0.06$	$+0.18 \pm 0.05$	0.56 ± 0.08	1.0
w/o Effect Gate	$+0.24 \pm 0.06$	$+0.15 \pm 0.04$	0.46 ± 0.10	1.0
w/o Residual Gate	$+0.22 \pm 0.05$	$+0.16 \pm 0.05$	0.38 ± 0.06	1.0
w/o Coherence	$+0.23 \pm 0.05$	$+0.14 \pm 0.04$	0.42 ± 0.06	1.0
w/o Support Reg.	$+0.08 \pm 0.04$	$+0.06 \pm 0.04$	0.00	1.0

The main paper uses these quantities descriptively rather than constructing a global ranking from p-values; mechanism-specific inference is reported with its corresponding experimental unit in App. F.1.

E.2 Learning Curves and Structural Evolution

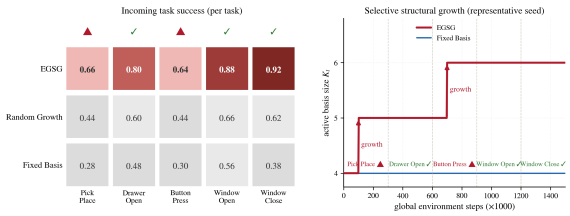

Figure 4: **Task-wise performance and selective structural growth** (representative seed). Left: success on each incoming task; ▲ marks tasks during which an expansion was accepted, ✓ tasks solved without growth. Right: active basis size K_t over the same stream; two accepted updates expand the space from $K = 4$ to $K = 6$, and later tasks are handled without further growth (per-seed counts: Table 1, App. E.1).

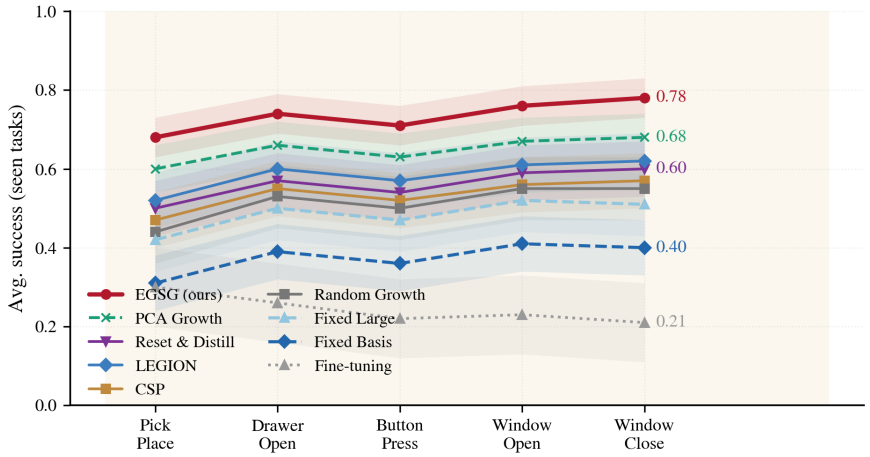

Figure 5: **Lifelong performance across the controlled MT10 stream** (incoming tasks on the x-axis; curves show average success over all tasks seen so far, mean $\pm$ s.d. over 10 seeds). EGSG remains above the plotted fixed-capacity, comparable-growth, and lifelong controls throughout the stream; the stage-wise ordering is consistent with Table 1.

E.3 K_0 Sensitivity

We probe $K_0 \in \{2, 4, 6, 8\}$ using 60 held-out expert windows per incoming MT10 task. Increasing K_0 reduces expert residual mass and short-horizon projection error (Table 7), while online growth is driven by policy-requested effects rather than expert windows. The full-stream study shows the expected trade-off: larger K_0 reduces accepted growth, while success plateaus beyond the main $K_0 = 4$ configuration.

Table 7: K_0 mechanism probe (real execution, pooled over the five MT10 incoming tasks; 60 windows/task). OOS rate: fraction of windows with $\rho > \tau_\xi$; eff.: relative effect-realization error of the projected skill vs. the unprojected expert chunk. Variance is measured on *incoming-task expert* skill embeddings; the corresponding source-embedding values (Table 3) are 90.0%/97.5% for $K{=}4/8$ — different populations, and the gap is consistent with incoming tasks exercising directions the source data covers less densely.

K_0	Variance	Mean ρ	OOS rate	Eff. error
2	0.727	0.89	1.00	0.30
4	0.861	0.68	0.84	0.27
6	0.919	0.30	0.01	0.21
8	0.954	0.22	0.00	0.20

Table 8: K_0 sensitivity on MT10 (ten seeds). Expert OOS is the fraction of held-out expert windows with $\rho > \tau_\xi$; Online elig. is the fraction of policy queries passing both request-level gates; Gain is matched-effect improvement per accepted update.

K_0	Expert OOS	Online elig.	Accepted	Gain	Avg. success	Reuse	Final K
2	1.00	0.014	3.6 ± 0.9	0.12 ± 0.04	0.62 ± 0.08	0.44 ± 0.08	5.6 ± 0.9
4	0.84	0.009	2.4 ± 0.7	0.11 ± 0.03	0.78 ± 0.04	0.56 ± 0.08	6.4 ± 0.7
6	0.01	0.006	1.2 ± 0.8	0.09 ± 0.03	0.80 ± 0.05	0.72 ± 0.06	7.2 ± 0.8
8	0.00	0.004	0.8 ± 0.7	0.08 ± 0.03	0.81 ± 0.05	0.76 ± 0.08	8.8 ± 0.7

E.4 Per-Task MT50 Outcomes

Table 9 reports EGSG's final per-task success and acquisition mode on the twelve-task stream. Three expansions are accepted; the remaining stages are completed without additional structural growth.

Table 9: Per-task MT50 outcomes for EGSG (ten seeds; final evaluation; tasks in stream order).

Task	Success	Acquisition
pick-place *(growth)*	0.55	expansion
drawer-open	0.66	reuse
button-press-topdown *(growth)*	0.50	expansion
window-open	0.70	reuse
window-close	0.64	reuse
door-close	0.86	reuse
basketball *(growth)*	0.55	expansion
plate-slide	0.74	reuse
handle-press	0.84	reuse
coffee-push	0.68	reuse
peg-unplug-side	0.65	reuse
hammer	0.55	reuse

E.5 Source-Task Retention

At stream end, source-task success relative to initialization is 0.87 ± 0.05 for EGSG, 0.82 ± 0.06 for Random Growth, 0.84 ± 0.05 for Fixed Basis, and 0.55 ± 0.09 for Fine-tuning SAC. Decoder-level preservation follows from the frozen decoder and nested span; policy-level retention remains empirical and depends on continued use of the preserved directions.

E.6 Task-Order Permutations

On reverse and interleaved MT10 orders, EGSG reaches 0.75 ± 0.05 and 0.76 ± 0.04, versus 0.78 ± 0.04 in the canonical order; Fixed Basis reaches 0.37 ± 0.07 and 0.39 ± 0.07. The EGSG–Fixed-Basis gap is stable, and accepted-update count changes by at most one across orders.

E.7 EWC and Progress & Compress

EWC (Kirkpatrick et al., 2016) uses a diagonal Fisher penalty ($\lambda = 100$); Progress & Compress (Schwarz et al., 2018) uses a frozen task column with distillation at task boundaries.

Table 10: Additional continual-learning baselines (mean $\pm$ s.d. over 5 seeds).

	MT10			MT50		
Method	Avg.	FWT	Forget.	Avg.	FWT	Forget.
EWC	0.31 ± 0.09	$+0.05 \pm 0.04$	0.38 ± 0.09	0.27 ± 0.10	$+0.04 \pm 0.03$	0.41 ± 0.10
Progress & Compress	0.48 ± 0.08	$+0.12 \pm 0.05$	0.21 ± 0.06	0.44 ± 0.09	$+0.10 \pm 0.04$	0.24 ± 0.07

E.8 Figure 1 Protocol

Fig. 1 compares the source-learned $K = 4$ span with a diagnostic $K = 8$ expansion formed from the next four source-PCA directions. This is a source-only diagnostic reference, not EGSG's online growth rule. Each condition uses 60 paired trials per seed over three seeds (180 trials per task and condition); error bars are 95% Wilson intervals.

F Mechanism Diagnostics and Validation

F.1 Diagnosis Details and Threshold Sensitivity

Labels and ROC protocol. Operational labels are built from held-out successful expert trajectories: a window is positive when its expert skill has normalized out-of-span residual above the source-calibrated threshold under the pre-growth basis. Matching online queries by task and state neighborhood yields 75,272 validated queries (24.5% positive) from 100k logged queries. Effect-only uses d_t^e, Skill-only uses ρ, Effect+Skill uses both, and the full criterion additionally requires $\kappa(\mathcal{R}) > \tau_\kappa$. The corresponding AUCs are 0.58, 0.65, 0.74, and 0.83.

Because the operational label depends on residual magnitude, we also use an execution-grounded label independent of the inverse residual: a held-out expert window is positive when execution through P_{B_K} fails to reproduce its transition effect above a source-calibrated error threshold. On the same query pool (19.8% positive), the ordering is preserved: 0.57, 0.62, 0.70, and 0.79. Queries within an episode are correlated, so inference uses episode-clustered bootstrap (2,000 draws): full AUC 0.83, 95% CI $[0.823, 0.834]$, versus $[0.734, 0.747]$ for Effect+Skill; the execution-grounded AUC is 0.79, CI $[0.786, 0.799]$. Across seeds, full minus Effect+Skill is $+0.088 \pm 0.009$ and positive in all ten seeds.

Inverse-model calibration. On held-out successful incoming-task windows, normalized inverse-skill error is 0.24 ± 0.02, versus 0.19 ± 0.01 on held-out source windows. Rejected candidates arise from higher-error request distributions (0.29 vs. 0.23 for committed candidates), consistent with execution validation filtering extrapolation failures. Under an induced Gaussian state-distribution shift, synthetic direction recovery is 0.99, 0.86, and 0.55 at 0.5, 1.0, and 2.0 s.d. shift, respectively (App. F.4).

Direction and repair controls. Across the 24 accepted events, alignment with the held-out expert residual direction is 0.36 ± 0.09, versus 0.26 ± 0.17 for the paired random null and 0.30 ± 0.11 for the source-covariance-matched null. The seed-level advantage over the random null is $+0.10 \pm 0.03$ and positive in all ten seeds. A supervised oracle direction reaches alignment 0.52 and matched success 0.56; EGSG reaches 0.36 and 0.50; the source-covariance reference reaches 0.30 and 0.31. A placebo re-evaluation under the unchanged basis changes matched effect error by only -0.005 ± 0.021, versus -0.11 after accepted growth. The held-out residual repair aggregated per seed is -0.13 ± 0.05 and negative in all ten seeds. Post-commit repair-evaluation requests are disjoint from candidate construction and from all pre-commit validation sets.

Auxiliary controls. Auto-timed PCA Growth reaches 0.61 ± 0.09 on MT10 with 3.1 ± 1.0 updates, remaining below EGSG despite independent timing. A dynamically updated effect support reduces growth to 1.8 ± 0.9 and reaches 0.76 ± 0.05 MT10 success; the frozen support is retained to keep d_t^e stationary with respect to initialization data.

Threshold sensitivity. Scaling the detection thresholds changes growth frequency as expected while average success remains within 0.04 of the default configuration. FG in Table 11 is the pooled post-hoc false-growth count over accepted updates.

Table 11: Threshold sensitivity on MT10 (mean $\pm$ s.d. over *ten* seeds, matching the main protocol). FG denotes post-hoc false-growth count: accepted updates that do not improve matched execution, reported as count/accepted-count pooled over the ten streams.

Setting	Growth	Avg. (FG)	Setting	Growth	Avg. (FG)
default	2.4 ± 0.7	0.78 ± 0.04 (0/24)	$\tau_\kappa = 0.20$	3.4 ± 1.1	0.77 ± 0.06 (3/34)
$\tau_e \times 0.5$	3.0 ± 1.2	0.76 ± 0.06 (2/30)	$\tau_\kappa = 0.50$	1.8 ± 0.8	0.76 ± 0.04 (0/18)
$\tau_e \times 2$	1.6 ± 0.9	0.77 ± 0.06 (0/16)	$\tau_\xi \times 0.5$	4.2 ± 1.3	0.74 ± 0.07 (4/42)
$\tau_{\text{supp}} \times 2$	2.2 ± 1.0	0.78 ± 0.04 (0/22)	$\tau_\xi \times 2$	1.2 ± 0.8	0.75 ± 0.04 (0/12)
joint $\times 0.5$	4.6 ± 1.4	0.73 ± 0.07 (5/46)	joint $\times 2$	0.8 ± 0.8	0.75 ± 0.04 (0/8)
$\tau_e \times 0.5, \tau_\xi \times 2$	1.9 ± 0.9	0.77 ± 0.06 (1/19)			

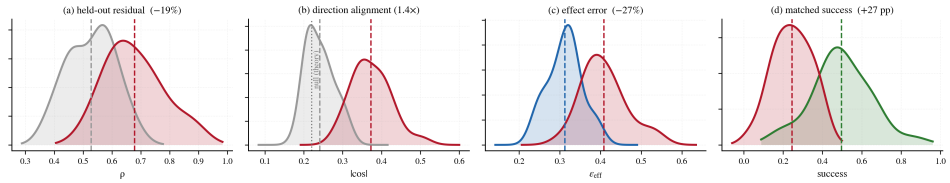

Figure 6: **Matched structural repair before further policy learning** (24 accepted events). (a) Held-out normalized residual; (b) direction alignment against a random-orthogonal null; (c) matched effect error; (d) matched success. Repair-evaluation requests are fresh and disjoint from both candidate construction and all pre-commit validation sets.

F.2 Executing Through the Expanded Basis

Frozen-decoder executability. We encode held-out expert windows, project them through a given basis, decode with the frozen skill decoder, and execute open-loop. Pick-place succeeds in $0/10$ episodes under $K = 4$ but $10/10$ under the source-derived $K = 8$ span; all other incoming tasks succeed under both. Thus, for this probe, the restrictive $K = 4$ failure is consistent with the active span rather than decoder inability. Fine-tuning the decoder is detrimental: updating after accepted growth gives 0.64 ± 0.07 success and 0.14 ± 0.04 forgetting, while fine-tuning from stream start gives 0.55 ± 0.09 and 0.22 ± 0.06, versus 0.78 ± 0.04 and 0.05 ± 0.02 with the frozen decoder.

Execution fidelity of added directions. For the 24 accepted events, we compare EGSG's residual-derived direction with source-covariance and uniformly random complement directions under matched states and norms.

Table 12: Execution fidelity along the added direction under matched queries (24 events, frozen decoder; mean ± s.d. over events).

Direction	Recon. error ↓	Effect err. ↓	Matched succ. ↑
EGSG residual-derived	0.021 ± 0.006	0.30 ± 0.09	0.50 ± 0.10
Source-covariance (unused PCA)	0.023 ± 0.007	0.34 ± 0.10	0.31 ± 0.10
Uniform random (complement)	0.058 ± 0.019	0.52 ± 0.13	0.14 ± 0.08

EGSG and source-covariance directions have similar decoder reconstruction error (0.021 vs. 0.023), but EGSG yields higher matched success (0.50 vs. 0.31); uniformly random directions are substantially less decodable and reach only 0.14 matched success. This comparison helps separate decoder compatibility from task-conditioned direction selection.

Policy use after growth. New-direction usage $u_t = \|b_{K+1}^\top \xi_t^{\text{exec}}\|_2 / \|\xi_t^{\text{exec}}\|_2$ rises from 0.05 immediately after acceptance to 0.41 within 3×10^4 steps and correlates with task-level success gain ($r = 0.81$).

F.3 Full-Chain Test of Candidate-Direction Origin

We test whether a source-only inverse model spontaneously produces residuals aligned with a held-out expert-residual reference direction that it never observes during training. Let u be the dominant out-of-span direction of held-out pick-place expert residuals under B_{K_0}. Its mean absolute projection on source skill embeddings is 0.55, showing substantial overlap with source-skill structure. This is consistent with, but does not by itself establish, decoder compatibility; executability is tested directly in App. F.2.

A fresh inverse model trained only on 3,480 stratified source windows is queried with real effects from skills containing increasing amounts of u. Residual alignment stays near the random null ($0.248 \to 0.267$ for $\alpha = 0 \to 2$; null 0.250), and the full gate–coherence pipeline gives $|\cos(b_{\text{new}}, u)| = 0.245$. Thus source-only extrapolation does not recover the held-out direction

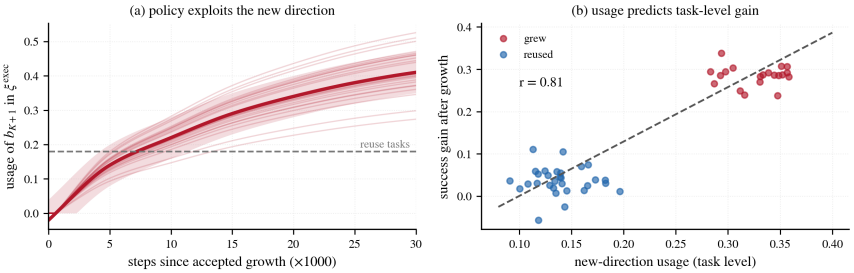

Figure 7: **The policy exploits the added direction. (a)** Usage u_t of the new basis direction after an accepted update (24 events; mean $\pm$ s.d. band), rising above the reuse-task baseline. **(b)** Task-level usage versus per-task success gain after growth ($r = 0.81$); grew/reused tasks marked.

by itself. The residual pathway should therefore be interpreted as online candidate generation; pre-commit execution supplies the independent behavioral criterion for accepting an extension. Behavior outside the frozen decoder's executable support remains outside the scope of active-span growth.

F.4 Controlled Validation of Direction Recovery

We mirror the gate–buffer–PCA–κ pipeline in a 16-dimensional synthetic space with the measured source spectrum, a planted complement direction u, zero-mean extrapolation noise, and a fixed bias direction. Recovery is measured by $|\cos(b_{K+1}, u)|$ over 40 replications per setting.

A fixed bias is removed by mean-centering and does not attract PC1. With zero-mean noise, accepted buffers recover u nearly exactly (0.994–0.999) once the planted signal dominates noise; below that regime the coherence gate typically refuses growth. Under structured directional noise, recovery remains at least 0.95 while artifact variance does not exceed planted-signal variance, but degrades to the random-null regime once the artifact dominates. Artifact-only streams can still be coherent, demonstrating that geometry alone cannot certify behavioral usefulness. Consistent with this failure mode, pre-commit execution rejected 9 of 33 MT10 candidates before commitment, showing that geometric coherence alone does not guarantee execution-relevant improvement. No rejected candidate alters the active basis.

Mean-offset failure mode. Because PC1 is computed after mean-centering, a nearly constant residual offset can be removed before coherence is measured. In a 40-stream probe, the variance-encoded signal triggers on 23% of streams with recovery 0.80, whereas a mean-dominated signal $r_i = 1.2u + 0.3zu$ never triggers; a simple mean-residual direction recovers 0.79 on the same streams. The current criterion is therefore suited to variance-concentrated evidence but can miss offset-dominated structure; augmenting it with a mean-residual statistic is left for future work.

F.5 Pre-Commit Validation: Budget and Hardware Protocol

Budget and memory. Each candidate uses 64 held-out requests and three matched executions (current basis, candidate basis, covariance comparator), each of horizon $h = 5$, for 960 environment steps ($\approx 0.3\%$ of an MT10 task budget). Across the ten canonical MT10 streams, 33 candidates are evaluated and 24 accepted, for approximately 1.1% total validation overhead; these executions are charged to the task budget. The frozen diagnosis machinery comprises the skill autoencoder, effect encoder, inverse model, basis, and effect-support store; the residual buffer holds at most 256 vectors. Baseline anchors/components are trainable and are already reflected in their peak-parameter ratios.

Hardware validation. Exact state restoration is unavailable on the robot. Each validation request is therefore evaluated under three randomized resets of the same layout category, using a paired layout-matched comparison against B_K. The reported hardware stream accepts two expansions and rejects none; all validation executions use the safety protocol of App. G.

Table 13: Trainable growth and persistent-memory accounting.

Method	Trainable $P_{\text{peak}}/P_{\text{init}}$	Frozen machinery	Persistent growth memory
Fine-tuning SAC	1.0	–	–
A-GEM	1.0	–	1,000 transitions
Reset & Distill	1.0	–	distilled set
MOORE	1.2	–	–
CSP	1.6	–	6.8 anchors (trainable)
LEGION	3.0	–	6 components (trainable)
EGSG	1.0	0.34M params; 11.2 MB	$16\times K$ basis; buffer ≤ 256

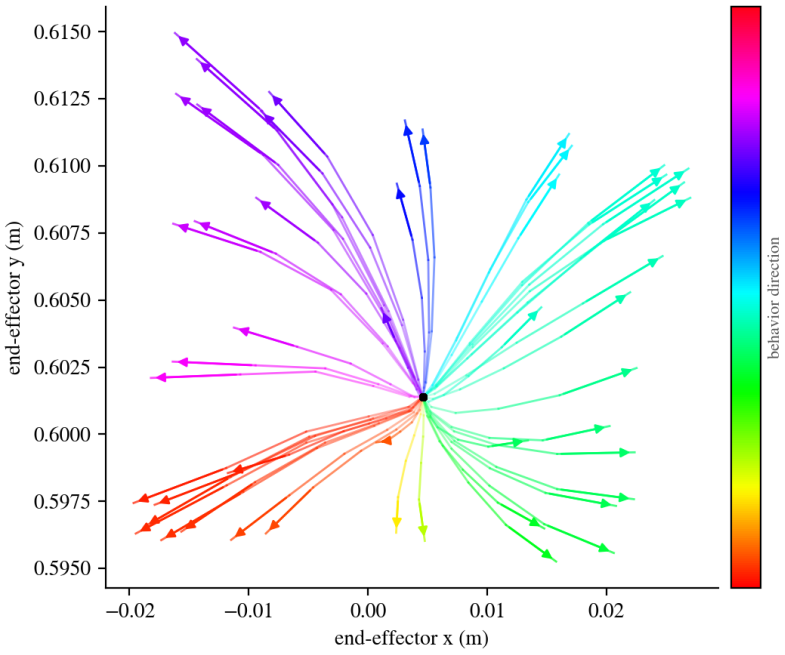

Figure 8: Execution through the B-plane: action chunks decoded from skills placed on a grid in the plane spanned by two learned directions. Movement along the residual-derived direction produces systematic, task-relevant end-effector displacement, whereas random orthogonal directions produce diffuse, non-directional motion.

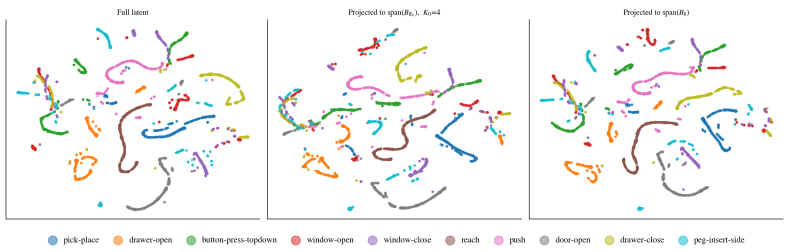

Figure 9: Latent skill space before and after an accepted structural update (t-SNE for visualization only; distances are computed in the original space). After expansion, incoming-task skill occupancy migrates into the enlarged span while source-task structure is retained.

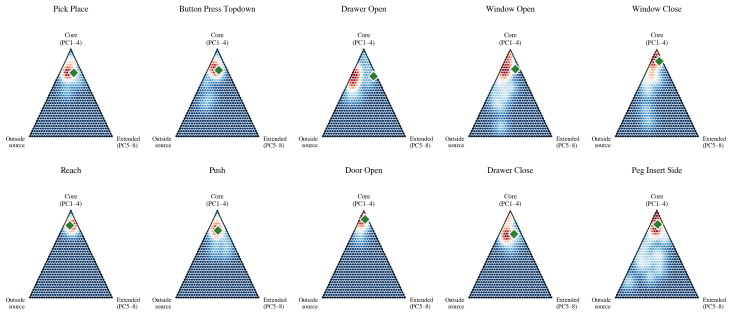

Figure 10: Ternary decomposition of incoming-task behavior into core (represented by B_{K_0}), extended (represented after growth), and outside (outside-the-evaluated-span) mass. Most tasks are dominated by core mass; growth-triggering tasks show substantial extended mass in this visualization.

F.6 Learned-Direction Visualizations

In the representative controlled-stream visualization, the first accepted direction is associated with a coordinated grasp-and-lift synergy, while the second is associated with lateral repositioning during grasp. These interpretations come from decoding controlled skills along the learned basis directions and measuring end-effector displacement; they are descriptive rather than unique semantic labels for the directions.

G Real-Robot Protocol and Safety

The platform is a 7-DoF manipulator with a parallel gripper, 39-dimensional proprioception, and a 4-dimensional end-effector command interface at 10 Hz. Initialization uses 40 kinesthetic demonstrations per auxiliary source task and no simulation assets or incoming-task demonstrations. Adaptation uses the same $h{=}5$ block-level SAC semantics, with all learned tasks re-evaluated after each stage. Each method has one complete stream; each table cell contains 20 randomized-layout trials.

Safety. Experiments are conducted inside fixed workspace boundaries with a 250 mm/s end-effector speed limit, joint/workspace action filtering, continuous operator supervision, and an emergency stop. No workspace-boundary collisions or hardware faults occurred; 3.1% of trials were terminated by operator timeout and counted as failures. These observations apply only to the monitored workcell and do not establish safety in unsupervised or human-shared settings. Demonstrations contain no human-subject measurements beyond operator motion; released logs contain robot states and actions and are reviewed for operator-identifiable motion signatures.

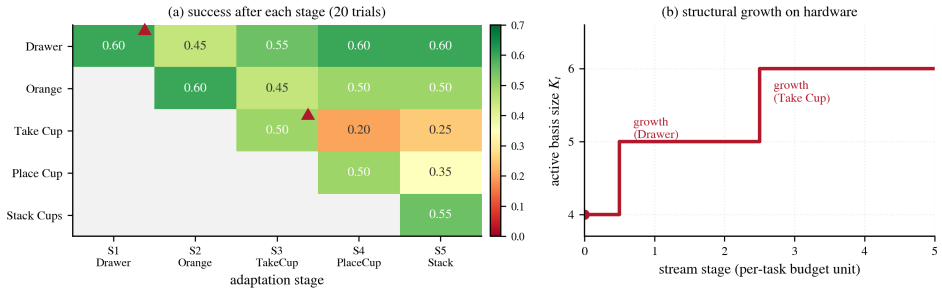

Figure 11: **Per-stage behavior and structural growth on hardware. (a)** Success on each task after every adaptation stage (20 trials per cell); ▲ marks stages during which an expansion was accepted. **(b)** Active basis size: two expansions (Drawer, Take Cup) take K from 4 to 6; after the second event, no further expansion is accepted.

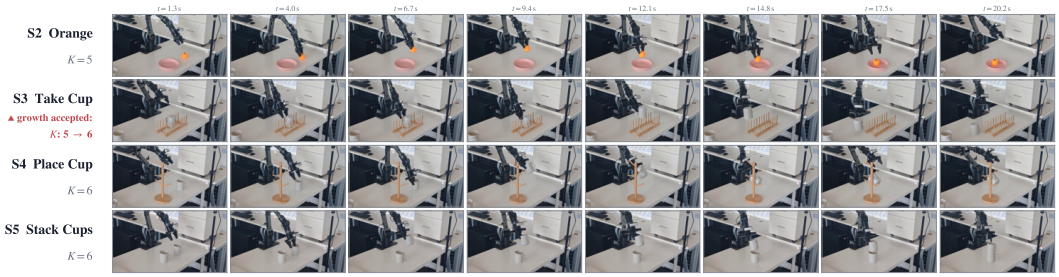

Figure 12: **Representative real-robot rollouts from later stages of the lifelong hardware stream.** Rows correspond to stages S2–S5 in Fig. 11, and columns show temporally sampled frames from representative executions. After the first accepted expansion during Drawer, the Orange stage operates with $K = 5$. EGSG accepts the second structural expansion during Take Cup, increasing the active basis from $K = 5$ to $K = 6$; Place Cup and Stack Cups subsequently operate with $K = 6$, with no further expansion accepted. The rollouts provide qualitative execution context for the stage-wise hardware results in Fig. 11 and Table 2.

The drawer and Take-Cup stages trigger the two accepted expansions, taking K from 4 to 6; no further structural expansion is accepted in the remaining stages. Final average success is 0.45 with forgetting 0.10.

The Drawer and Take-Cup stages trigger the two accepted expansions, taking K from 4 to 6; no further structural expansion is accepted in the remaining stages. Figure 12 provides representative executions from the later four stages. After the first accepted expansion during Drawer, the Orange stage operates with $K = 5$. The second expansion is accepted during Take Cup, increasing the active basis from $K = 5$ to $K = 6$, after which Place Cup and Stack Cups operate with $K = 6$. Final average success is 0.45 with forgetting 0.10.

H LIMITATIONS

EGSG adds one direction at a time and is validated primarily when residual evidence has a dominant direction; simultaneous multi-direction structure is not addressed. Its evaluation is limited to manipulation streams from one benchmark family with a fixed source/incoming split, plus one hardware stream per method. Long-stream MT50 results extend the evaluation beyond MT10 but do not establish that the MT10 margin persists as stream length grows.

The method also inherits two source-support boundaries. First, active-span expansion cannot expose behavior outside the frozen decoder's executable support; such tasks would require decoder adaptation (App. F.2, App. F.3). Second, requested effects are parameterized in the source top-8 effect-PCA subspace, so effects outside that proposal space cannot be requested without expanding the effect basis. Finally, overall success remains limited by online exploration and policy optimization; structural repair does not by itself solve those failures.