ChatGPT

State-of-the-art methods for popular continual-learning and RL benchmarks (2024–2025)

Continual Bench

Benchmark – *Continual Bench* (ICML 2025) is a robotic benchmark for continual reinforcement learning (CRL). Six robotic tasks—`peg-unplug`, `faucet-close`, `pick-place`, `door-open`, `window-close` and `button-press`—are placed around a circular rig. A single 7-DOF robot must master each task in sequence while retaining previous skills. The environment removes the physical interference problem of the Continual-World benchmark by allowing the robot to freely rotate and access each task [1].

SOTA method (2025) – Liu et al. (ICML 2025) propose the **FTL Online Agent (OA)**. OA builds an online world model and applies a “fast-tracking layer (FTL)” to adapt the model in real time. The paper compares OA with model-based baselines (Fine-tune, Synaptic Intelligence, Coreset replay and Perfect Memory). OA achieves a final average performance of **72.9 %** with a regret of **27.6 %**, slightly higher average performance and much lower regret than Perfect Memory (72.4 % / 33 % regret) and significantly better than other baselines [2]. OA therefore represents the state of the art on Continual Bench in 2025.

MEAL (Multi-agent Continual RL)

Benchmark – *MEAL* (MEAL, June 2025) is the first continual RL benchmark for **multi-agent** settings. It is built on the “Overcooked-A” environment, where two agents must cooperate across five kitchen layouts to cook dishes sequentially. MEAL measures final performance, forgetting and knowledge transfer while allowing task-shuffling to generate a curriculum.

SOTA methods (2025) – The MEAL paper evaluates Fine-Tuning (FT), Elastic Weight Consolidation (**EWC**), Synaptic Intelligence (SI), Memory Aware Synapses (MAS) and L2 regularization. EWC attains the **highest average performance** and the **lowest forgetting** among tested methods, while FT maintains high plasticity but catastrophically forgets [3]. Thus, **EWC** is the most effective method on MEAL tasks in 2025.

Continual World

Benchmark – Continual World (CW) uses a single robotic manipulator performing multiple tasks in the real world. Sequences like CW5, CW10, GCL10 and CW20 consist of up to twenty contact-rich tasks.

SOTA methods (2024-2025)

Year/paper	Approach & key idea	Evidence of SOTA
t-DGR (arXiv 2024)	Trajectory-based Deep Generative Replay (t-DGR) trains a non-autoregressive generative model to replay entire action trajectories. On CW10 it obtains **81.9 %±3.3** success rate (forgetting 14.4 %); on CW20 it achieves **83.9 %** success, outperforming Fine-tuning, oEWC, PackNet, Generative Replay and CRIL [4].	Highest success rate across CW10, GCL10 and CW20 in 2024. [4]
DISTR (ICLR 2024)	*Diffusion-based Trajectory Replay* uses diffusion models to generate training trajectories and a stable training objective that balances stability and plasticity. On CW5 it achieves an average performance of **84.8** and forward transfer of **16.2**, with forgetting of **4.7**; on CW10 it obtains **81.2** average performance with low forgetting [5].	Outperforms baselines such as Fine-tune, EWC, MAS, PackNet and RePR in both average performance and forgetting [5].

These two methods provide the best results on Continual World sequences in 2024–2025.

MATH-B (MATH-Beyond)

Benchmark – *MATH-Beyond* (ICLR submission 2025) creates an RL setting where language models must solve mathematical reasoning problems far beyond their base model capabilities. The dataset is constructed so that the base model scores **zero** (pass@1024 = 0), forcing methods to generalize and create new reasoning strategies.

SOTA (2025) – The authors evaluate RL fine-tuning (r1-1.5B nemotron v1 and r1-7B skywork or1) and supervised fine-tuning (SFT). RL methods expand the base's capabilities only modestly (7.83 % and 21.21 %), whereas SFT models **Qwen3-4B** and **Qwen3-8B** achieve **58.93 %** and **66.38 %** expansion rates, respectively [6]. Thus the best available approach on MATH-B as of mid-2025 is supervised fine-tuning of larger language models.

CORA (COntinual RL Agents platform)

Benchmark – *CORA* (PMLR 199, 2022) unifies evaluation across four environment families (Atari, Procgen, MiniHack and CHORES) and defines evaluation metrics such as continuous evaluation, isolated forgetting and zero-shot forward transfer.

SOTA – While no major updates were published in 2024/2025, the baseline algorithm **CLEAR** (Continual Learning by Replay) remains a strong point of comparison. CLEAR combines policy distillation with experience replay and outperformed other baselines on Atari and Procgen at the time of publication [7]. Later work builds on CLEAR, but it remains the reference method on CORA tasks.

COOM (Continual Doom)

Benchmark – *COOM* (NeurIPS 2023) presents a 3-D game benchmark for continual RL built on VizDoom. It includes sequences of four (CD4/CO4) and eight (CD8/CO8) tasks, requiring agents to learn new weapons, avoid enemies and recall older skills in visually complex worlds.

SOTA baseline (2023-2025) – The benchmark compares PackNet (progressive pruning), MAS, AGEM, L2 regularization, EWC, VCL, Fine-tuning, ClonEx-SAC and Perfect Memory. **PackNet** achieves the highest average performance (AP ≈ 0.74) with low forgetting and positive forward transfer [8]. ClonEx-SAC has similar performance (AP ≈ 0.73), while perfect memory is unrealistic. Since no newer publications have surpassed these results, PackNet remains the leading method for COOM as of 2025.

AntMaze / MetaWorld

Benchmark – D4RL's *AntMaze* tasks require an agent to navigate a maze to reach a target; trajectories are long and random initial policies can rarely reach the goal. MetaWorld contains 50 robotic manipulation tasks with high-dimensional continuous action spaces.

State-of-the-art methods (2024-2025)

Paper/Year	Method & key idea	Evidence of SOTA
Generative Trajectory Policies (GTP, 2025)	Trains a generative model that outputs entire trajectories (policy is a diffusion model conditioned on start and goal). In behavior-cloning, GTP-BC attains an **average score of 66.3** across AntMaze tasks versus 44.1 for the next-best generative policy [9]. In the actor–critic setting GTP achieves **80.6** average return, surpassing Diffusion-QL (69.6) and QGPO (78.3) and achieving **100** on the antmaze-umaze task 【316969959786800†L617-L649】.	Highest reported scores on AntMaze in 2025, beating diffusion and temporal difference methods 【316969959786800†L617-L649】.
Graph-Assisted Stitching (GAS, ICML 2025)	A hierarchical offline RL method that frames sub-goal selection as a graph search and introduces a temporal-efficiency metric. In a stitching-critical navigation task, GAS achieves **88.3**, whereas the previous state-of-the-art scored **1.0** [10] —a dramatic improvement.	Provides orders-of-magnitude better performance on long-horizon tasks compared with prior HRL methods [10].

Paper/Year	Method & key idea	Evidence of SOTA
Lower Expectile Q-learning (LEQ, 2024/25)	Uses a conservative value estimator based on expectile regression and re-weights updates to lower expectiles, allowing robust exploration. On AntMaze, LEQ achieves success rates around **94 %** on antmaze-umaze and strong performance even on ultra-diverse tasks [11], outperforming previous model-based and model-free baselines.	High success rates across AntMaze tasks and improved generalisation [11].
GO-Skill (Offline Multi-Task RL, ICML 2025)	Goal-Oriented Skill Abstraction extracts reusable skills from offline task-mixed data and learns a hierarchical policy. Experiments on MetaWorld (50 tasks) demonstrate that GO-Skill improves mean episode return across near-optimal and sub-optimal datasets. Though detailed numbers are not released, GO-Skill qualitatively outperforms Decision Transformer and other multi-task baselines.	The ICML 2025 paper reports that GO-Skill significantly improves performance across MetaWorld tasks by learning and re-using skill abstractions.

These methods currently lead performance on AntMaze and MetaWorld tasks.

ProcGen

Benchmark – *ProcGen* is a generalisation benchmark of 16 procedurally generated games requiring reinforcement learners to generalize beyond fixed training seeds.

SOTA methods (2024-2025)

- **Simple 3D-Conv architecture (2024)** – An October 2024 study shows that replacing 2-D convolution with 3-D convolution and stacking 16 frames reduces the optimality gap on ProcGen by **37.9 %** (from 0.58 to 0.36) compared with the VSOP baseline, matching or exceeding previously published state-of-the-art generalisation methods [12].

- **History Compression via Language Models (HELM, ICML 2022)** – HELM uses a frozen pretrained language transformer as a memory module to represent history. It achieved **new state-of-the-art results** on both Minigrid and ProcGen and improved sample efficiency [13]. Though released in 2022, no subsequent work in 2024–2025 has clearly surpassed HELM, so it remains an influential SOTA method.

- **Diffusion-Guided Adaptive Augmentation (DGA^2, ICCV 2025)** – This method uses diffusion models to generate domain-shifted augmentations for training. It is reported to enhance generalisation on

DeepMind Control (GB), ProcGen and Adroit tasks, though specific numbers are not presented. Nevertheless, DGA^2 appears promising for ProcGen in 2025.

Atari / Arcade Learning Environment (ALE)

Benchmark – The ALE includes over 50 classic Atari games. Performance is often measured by human-normalised scores or inter-quartile means (IQM) across games.

SOTA (2024-2025)

- **Beyond The Rainbow (BTR, ICML 2025)** – BTR integrates six improvements into Rainbow DQN (e.g., prioritized replay, distributional RL, multi-step returns). It achieves a **human-normalised inter-quartile mean of 7.6** on 60 Atari games, representing near-state-of-the-art performance [14]. Importantly, BTR runs on a **single high-end desktop PC** and learns from 200 million Atari frames in under **12 hours** [14], making high-performance Atari RL accessible to wider researchers.

- **HackAtari (arXiv 2024)** – This framework introduces controlled novelty into Atari games (object colours, shapes, rewards) to test robustness and interpretability. It shows that existing agents often fail when faced with such variations [15]. Though not a performance-improving algorithm, HackAtari provides an important robustness benchmark for evaluating SOTA agents.

Gym Control (Classic & MuJoCo)

Benchmark – Gym classic control tasks (CartPole, MountainCar, etc.) and D4RL MuJoCo tasks (HalfCheetah, Hopper, Walker2d) are widely used continuous-control benchmarks. Offline variants use static datasets to train agents without online interaction.

SOTA methods (2024-2025)

Paper/Year	Method & highlights	Evidence
Imagination-Limited Q-Learning (ILQ, 2025)	ILQ uses a dynamics model to imagine out-of-distribution action-values and clips the imagined values with the maximum behaviour value. Table 1 shows ILQ achieves the highest normalised scores across MuJoCo tasks; e.g., on `halfcheetah-r` ILQ scores **31.7±0.7** vs. 28.5 for CSVE and 17.5 for CQL; on `hopper-r` ILQ scores **31.6±0.2** vs. 31.8 for CSVE and 8.5 for IQL; across all MuJoCo tasks ILQ attains a total normalised score of **920.4**, surpassing CSVE (873.6) and many other baselines [16]. In Maze2D and Adroit domains, ILQ also yields higher returns than BEAR, CQL, IQL and PlanCP [17].	Highest reported scores on D4RL MuJoCo, Maze2D and Adroit tasks in 2025 [18].

Paper/Year	Method & highlights	Evidence
Selective State-Adaptive Regularization (SSAR, ICML 2025)	Introduces state-dependent regularization coefficients and selects high-quality actions for conservative Q-learning. ICML poster notes that SSAR **significantly outperforms state-of-the-art approaches** on the D4RL benchmark in both offline and offline-to-online settings [19].	Provides robust improvement across MuJoCo tasks over previous algorithms [19].
MOBODY (Model-Based Off-Dynamics Offline RL, arXiv Jun 2025)	MOBODY learns from offline datasets collected under mismatched transition dynamics; it leverages a shared latent representation and Q-weighted behavior cloning. The abstract reports that MOBODY **significantly outperforms state-of-the-art baselines** on standard MuJoCo benchmarks, particularly when source and target dynamics differ [20].	Promising performance on cross-domain MuJoCo tasks where dynamics mismatch; numbers are not yet widely reported.
Selective state-adaptive regularization also outperforms existing offline RL algorithms and facilitates offline-to-online transfer [19].		

No major breakthroughs were reported for simple Gym classic control tasks (CartPole, etc.) because these environments are solved with existing algorithms (DQN, PPO, SAC). For offline RL on MuJoCo tasks, ILQ currently represents the state of the art.

Summary

Across the listed benchmarks, several new methods from 2024–2025 stand out:

- **FTL Online Agent** sets the bar for Continual Bench by learning an online world model and reducing regret [2].
- On the **MEAL** multi-agent benchmark, **EWC** remains the best-performing baseline, balancing plasticity and forgetting [3].
- **t-DGR** and **DISTR** are the top methods for Continual World, delivering high success rates and balanced stability/plasticity [4] [5].
- **Supervised fine-tuning** of large language models (Qwen3-4B/8B) yields the best performance on MATH-B [6].
- For COOM, **PackNet** remains the strongest baseline [8].
- In AntMaze/MetaWorld, **GTP**, **GAS** and **LEQ** provide the highest returns and success rates [9] [10] [11].
- For generalisation on ProcGen, simple 3D convolutions and language-model memory modules continue to deliver state-of-the-art results [12] [13].

- On Atari/ALE, **Beyond The Rainbow** achieves near-SOTA scores while training on a single desktop PC [14].
- In offline continuous control (Gym/MuJoCo), **ILQ** currently shows the best empirical performance, with further improvements from **SSAR** and **MOBODY** [18] [19] [20].

These methods represent the best-performing approaches on their respective benchmarks during 2024–2025, though rapid progress in reinforcement learning means new results may emerge after October 2025.

---

[1] [2] Continual Reinforcement Learning by Planning with Online World Models
https://openreview.net/pdf?id=mQeZEsdODh

[3] [2506.14990] MEAL: A Benchmark for Continual Multi-Agent Reinforcement Learning
https://ar5iv.labs.arxiv.org/html/2506.14990v1

[4] [2401.02576] t-DGR: A Trajectory-Based Deep Generative Replay Method for Continual Learning in Decision Making
https://ar5iv.org/pdf/2401.02576

[5] [2411.10809] Stable Continual Reinforcement Learning via Diffusion-based Trajectory Replay
https://ar5iv.labs.arxiv.org/html/2411.10809

[6] openreview.net
https://openreview.net/pdf?id=RNkErKpCAp

[7] [2110.10067] CORA: Benchmarks, Baselines, and Metrics as a Platform for Continual Reinforcement Learning Agents
https://ar5iv.labs.arxiv.org/html/2110.10067

[8] d61d9f4fe4357296cb658795fd7999f0-Paper-Datasets_and_Benchmarks.pdf
https://proceedings.neurips.cc/paper_files/paper/2023/file/d61d9f4fe4357296cb658795fd7999f0-Paper-Datasets_and_Benchmarks.pdf

[9] Offline Reinforcement Learning with Generative Trajectory Policies
https://arxiv.org/html/2510.11499v1

[10] ICML Poster Graph-Assisted Stitching for Offline Hierarchical Reinforcement Learning
https://icml.cc/virtual/2025/poster/46345

[11] 2407.00699.pdf
https://arxiv.org/pdf/2407.00699.pdf

[12] 2410.10905.pdf
https://arxiv.org/pdf/2410.10905.pdf

[13] [2205.12258] History Compression via Language Models in Reinforcement Learning
https://arxiv.org/abs/2205.12258

[14] ICML Poster Beyond The Rainbow: High Performance Deep Reinforcement Learning on a Desktop PC
https://icml.cc/virtual/2025/poster/45085

[15] 2406.03997.pdf
https://arxiv.org/pdf/2406.03997.pdf

[16] [17] [18] arxiv.org
https://arxiv.org/pdf/2505.12211

[19] ICML Poster Learning to Trust Bellman Updates: Selective State-Adaptive Regularization for Offline RL
https://icml.cc/virtual/2025/poster/44640

[20] MOBODY: Model Based Off-Dynamics Offline Reinforcement Learning
https://arxiv.org/pdf/2506.08460.pdf