Chunking the Critic: A Transformer-based Soft Actor-Critic with N-Step Returns

Dong Tian, Onur Celik and Gerhard Neumann

Karlsruher Institut für Technologie

Autonomous Learning Robots Lab

ICLR

Why standard SAC struggles on long-horizon, sparse-reward control

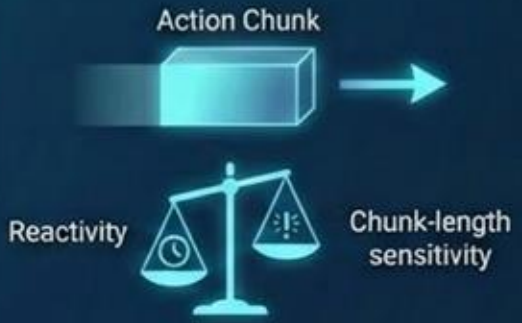
Step-wise critics miss temporal structure

scoring (s_t, a_t) "in isolation" is weak when success depends on **a short action sequence** (multi-phase tasks).

N-step returns & Importance Sampling (IS)

help credit assignment, but **off-policy** N-step typically needs **importance sampling** (IS) → high variance / instability → in practice limits usable horizon.

Actor-side action chunking



helps long horizons but trades off **reactivity + chunk-length sensitivity**, so it's not consistently robust in online off-policy settings.

Chunk the critic, not the actor

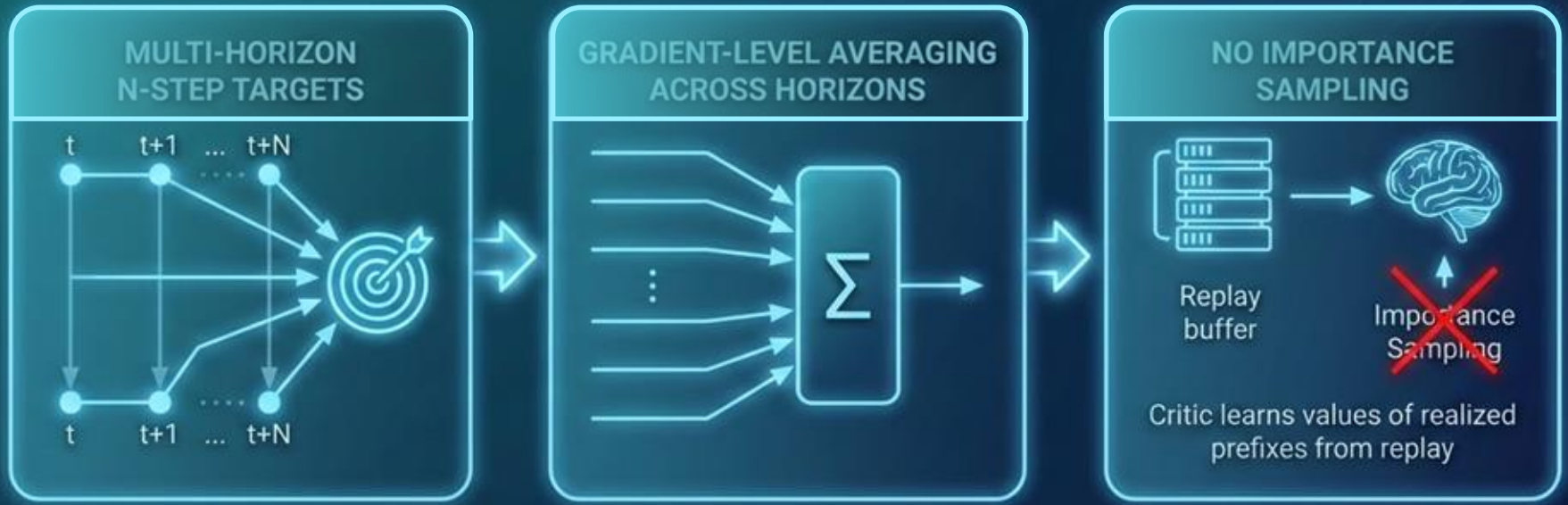
Methods

- Replace SAC's MLP critic with a ***causal Transformer critic***
- Train on short realized state-action windows with ***multi-horizon N-step targets***
- Get better long-horizon credit assignment while the ***policy stays one-step***

Takeaway

Temporal abstraction lives in the critic, not in an open-loop chunked policy

How training works



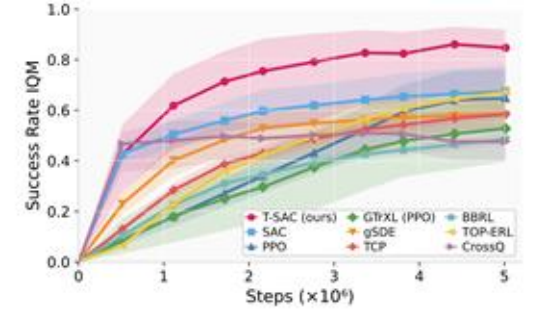
Stability trick

critic-parameter freezing schedule

Takeaway

snapshot critic targets, freeze them for a short interval, and train without a target network

Experimental Results



(b) Aggregated success on Meta–World ML1 (50 tasks) comparing T–SAC to step-based and episodic baselines. Curves show IQM success rate with 95% bootstrap confidence intervals; unless noted, results average 8 seeds and count success only at the final timestep.

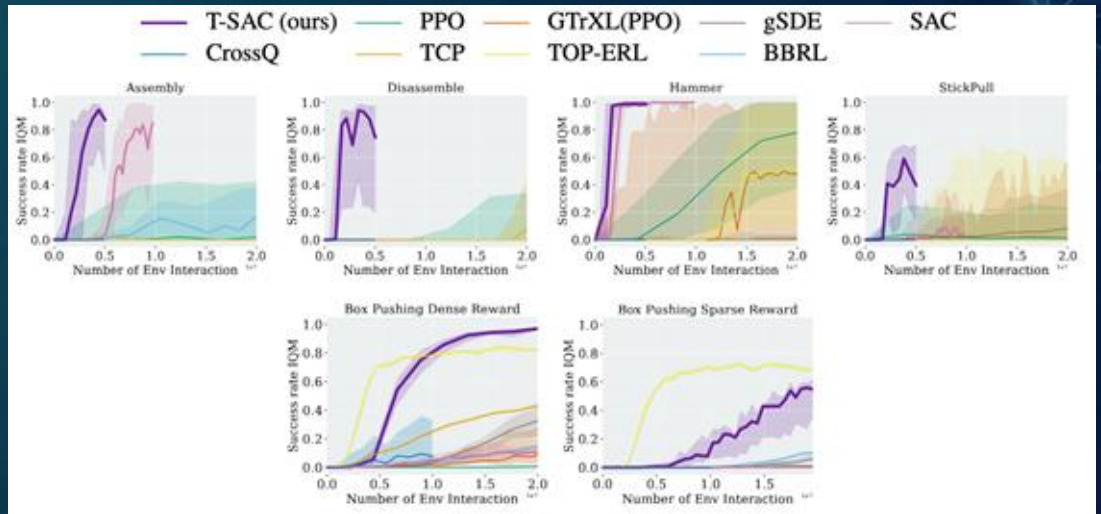


Figure 2: Success-rate IQM vs. environment interactions on challenging Meta-World ML1 tasks and FANCY-GYM Box-Pushing. Panels show Assembly, Disassemble, Hammer, and Stick-Pull, plus Box-Pushing under dense and sparse rewards. Success is counted only at the final timestep.

Takeaway

critic-side sequence modeling helps long-horizon RL without changing the policy into an open-loop chunking method