Return-Aligned Decision Transformer

Tsunehiko Tanaka[1] Kenshi Abe[2] Kaito Ariu[2] Tetsuro Morimura[2] Edgar Simo-Serra[1]

[1]Waseda University [2]CyberAgent

arXiv

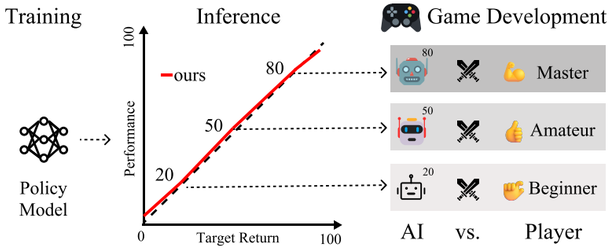
Code

TL;DR **RADT matches target returns, not just maximizes returns.**

Problem: **Return Alignment**

Example: **Game developers need AI opponents with controllable performance**.

A return-aligned model changes opponent performance by changing target return.



Objective: **Minimize alignment error between target return and cumulative reward.**

- Target return specifies the desired performance.
- Cumulative reward should match it.

$$\left| R^{\text{target}} - \sum_{t=1}^{T} r_t \right|$$

Why does Decision Transformer miss Target Returns?

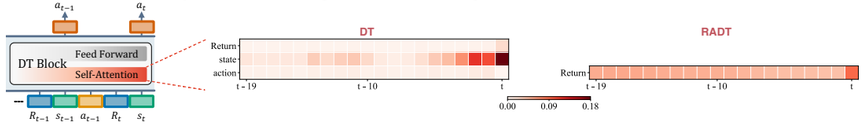
Return-to-go tokens receive little attention.

DT uses a Transformer to predict actions conditioned on return-to-go.

Return-to-go is the remaining return to reach the target return: $\hat{R}_t = R^{\text{target}} - \sum_{t'=1}^{t-1} r_{t'}$

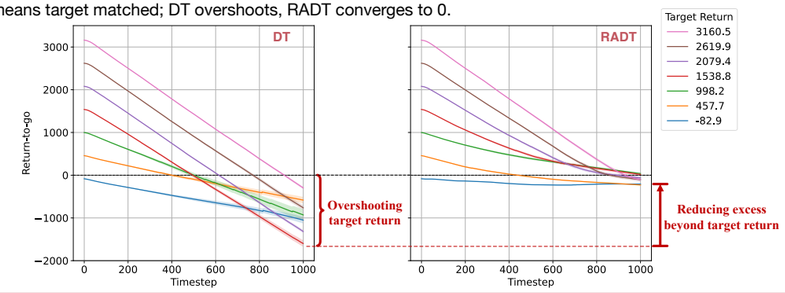
where R^{target} is the desired sum of rewards over an episode, given at the start.

DT allocates hardly any attention weight to return-to-go tokens.



Return-to-go over an episode:

0 means target matched; DT overshoots, RADT converges to 0.

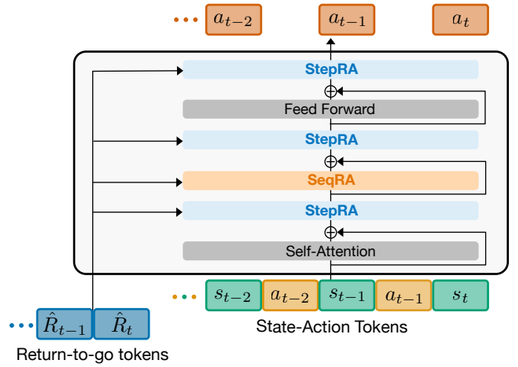


Return-Aligned Decision Transformer

Key Idea: **Return-focused model architecture.**

Separate return-to-go tokens from state-action tokens

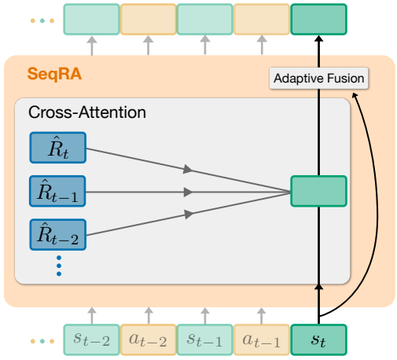
- DT: one interleaved sequence - return-to-go tokens lose the attention competition.
- RADT: return-to-go tokens form their own sequence; two aligners inject it into state-action tokens.



Sequence Return Aligner (SeqRA)

- Long-term dependencies

Puts all attention on the return tokens, capturing the whole return history



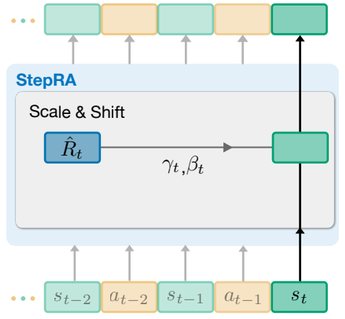
Cross-Attention(Q=state-action, KV=return):
Select relevant returns across timesteps

Adaptive Fusion:
Merge the selected return context into each state/action token

Stepwise Return Aligner (StepRA)

- Stepwise response

Conditions each state/ action token on only its same-timestep return for an immediate response

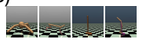


Scale & Shift:
Modulates state/action tokens with scale $1+\gamma_t$ and shift β_t, inferred from the same-timestep return-to-go

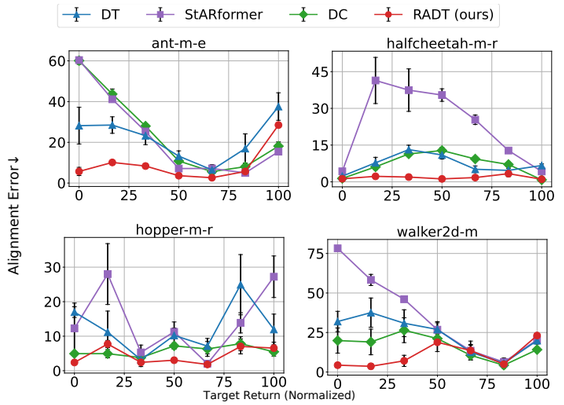
Results

Alignment error ↓44.6% (MuJoCo) / **↓65.5%** (Atari) vs DT

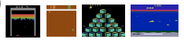
MuJoCo - ↓44.6% vs DT (12-task avg: 19.5→10.8)



- Lowest error on all 12 tasks.
- RADT tracks the full target range; baselines bias to common returns.



Atari - ↓65.5% vs DT (4-task avg: 74.6→25.7)



- Best on all 4 games
- Works with image inputs & discrete actions, even with delayed rewards.