UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN

Reinforcement Learning Gradients as Vitamin for Online Finetuning Decision Transformers

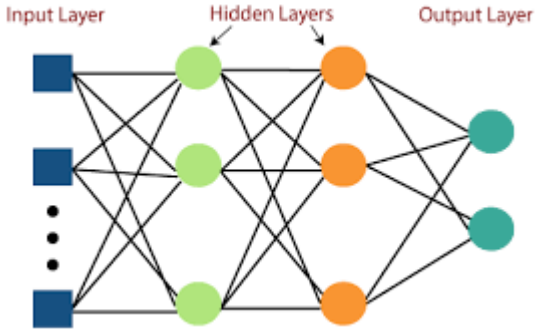


NeurIPS 2024 Spotlight

Kai Yan, Alexander G. Schwing, Yu-Xiong Wang
University of Illinois Urbana-Champaign

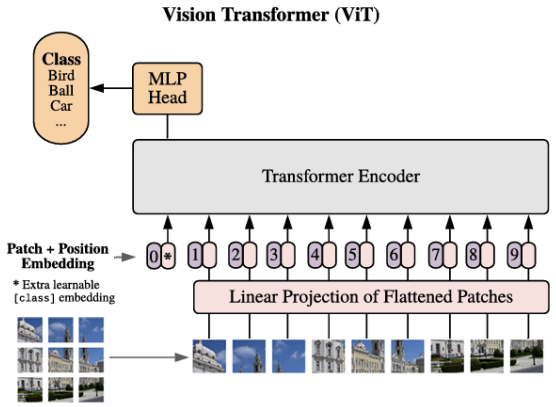
Catching up with the Transformer Age

- Language and Vision communities have much benefitted from scaling…
  - But Reinforcement Learning (RL) is still using small multi-layer perceptrons!



Typical RL Actor / Critic
~100**K** params
Multi-Layer Perceptrons (MLPs)

?



ViT-B-16 [1]
~100**M** params
Transformers

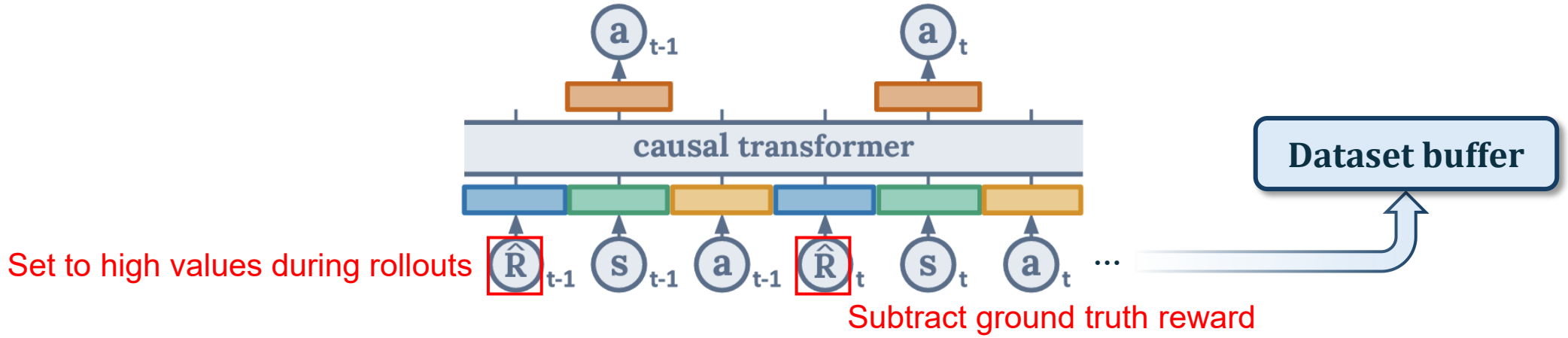
Flagship Large Language Models
~100**B** params
Transformers

[1] A. Dosovitskiy et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In ICLR, 2021.

Image Source: Internet

(Online) Decision Transformers (ODTs)

- Decision Transformers (DT) [1] were proposed to level the gap
  - Brings sequence modeling & modern transformer architecture into RL
  - Autoregressively completes a states, actions and Returns-To-Go (RTG) sequence

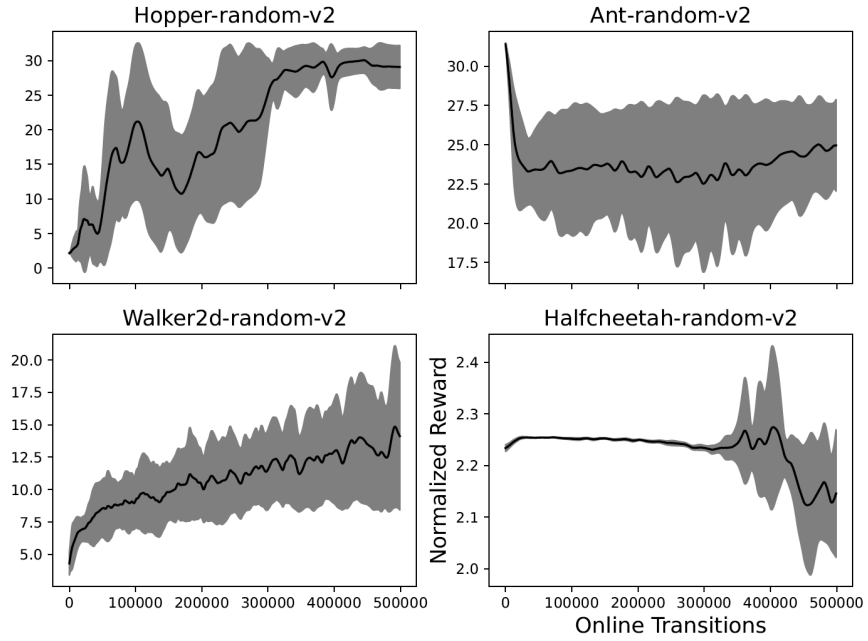


- Online DT (ODT) [2] was proposed for online finetuning
  - Online exploration via entropy term and rollouts **conditioning on high RTG**
  - Surprisingly, few follow-ups on improving online improvement ability

[1] L. Chen et al. Decision Transformer: Reinforcement Learning via Sequence Modeling. In NeurIPS, 2021.
[2] Q. Zheng et al. Online Decision Transformer. In ICML, 2022.

ODTs Online Improvement Deficiency Disease

- ODT fails to improve during online finetuning after pretraining on low RTG data
  - "High RTG" is too out-of-distribution and ODT can't improve its policy locally in action space
  - Check theoretical bounds for this in our paper



Finetune after pretraining on random dataset;
struggle to approach expert-level reward of 100!

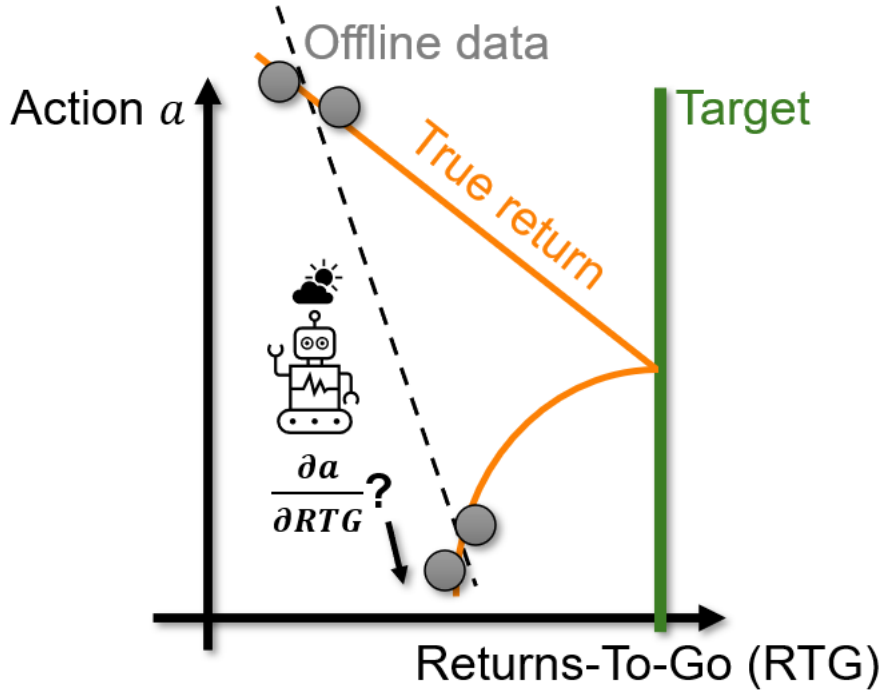
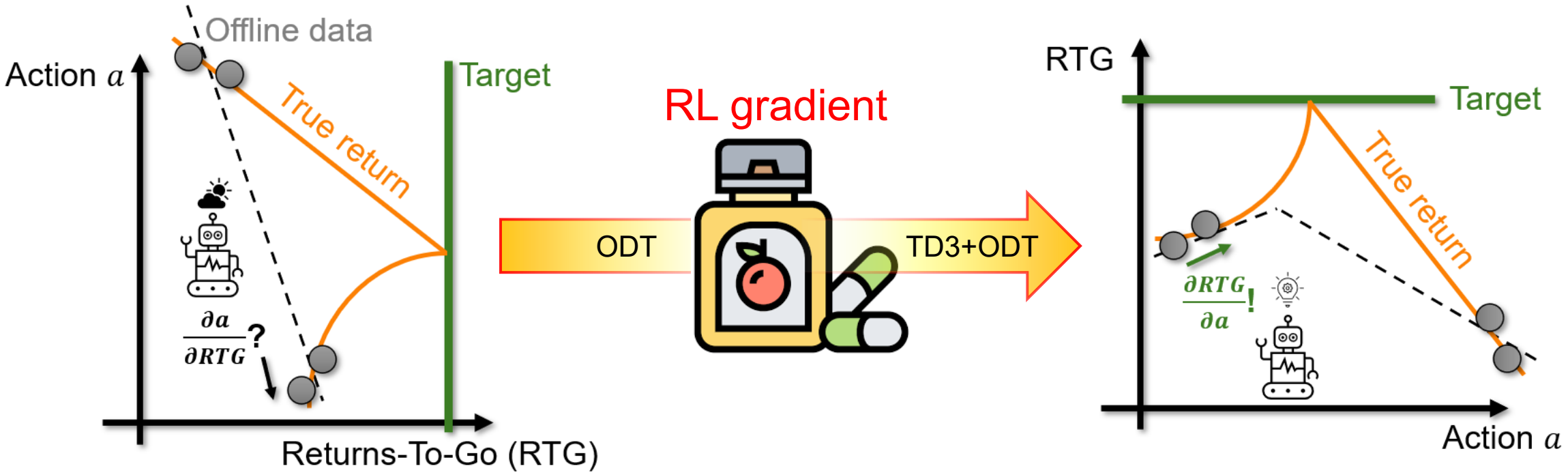


Image Source: Internet

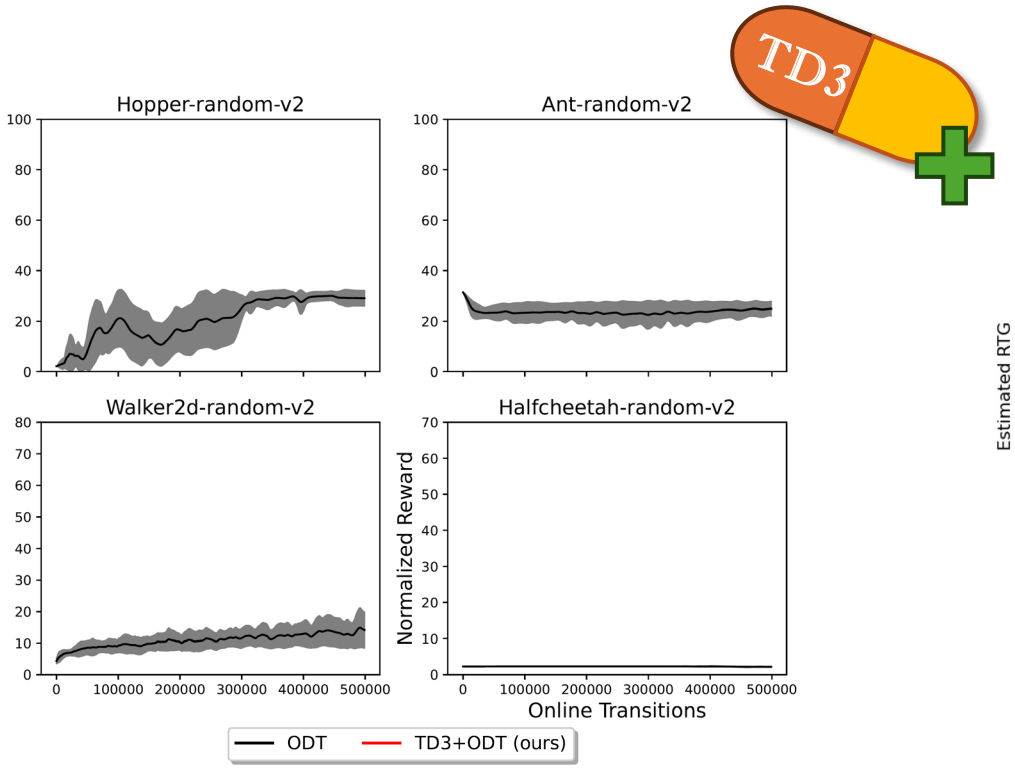
Can a Small Dose of RL Gradients Cure?

- We need to enable ODT to improve RTG in local action space
  - This is exactly what RL does!
  - We found TD3 to be the best choice

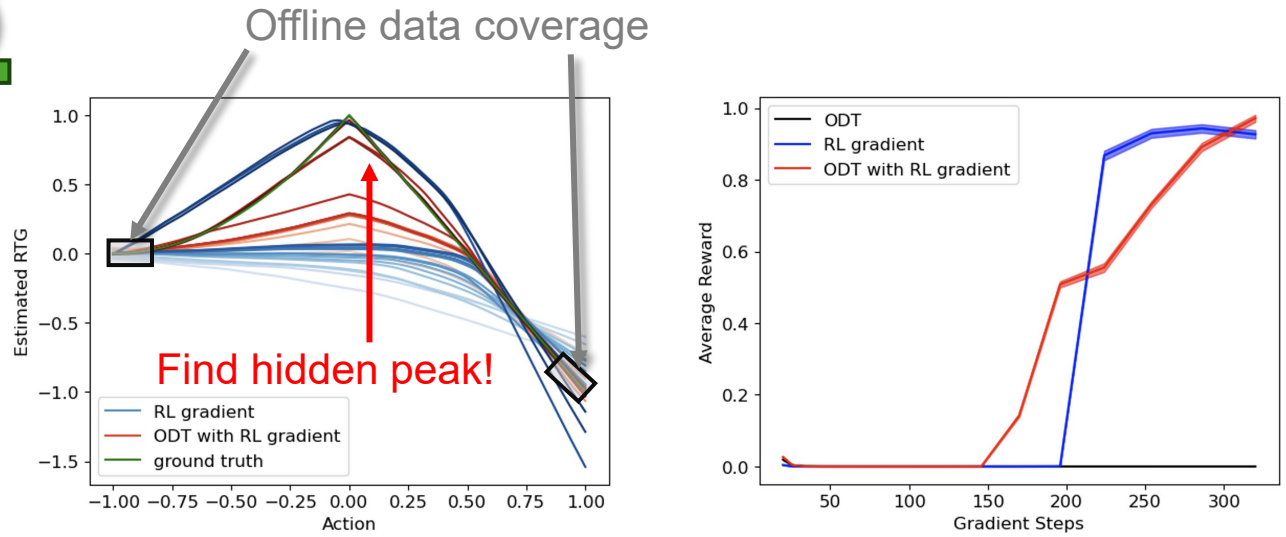


How Well Does Our Prescription Work?

- TD3+ODT outperforms many baselines on a variety of tasks
  - Especially after pretraining on low-RTG data!



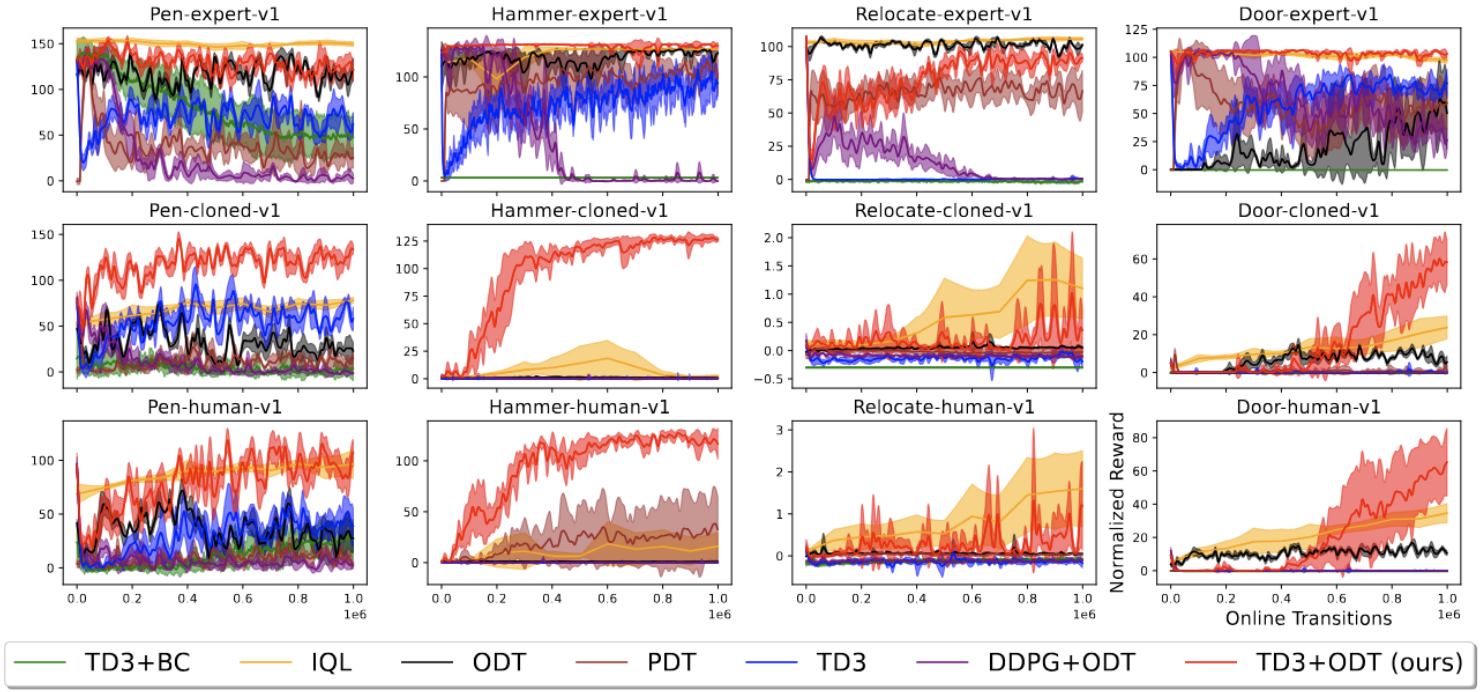
Ours vs. ODT on finetune with pretraining on MuJoCo random dataset



Reward curves on solving the simple single-state MDP

How Well Does Our Prescription Work?

- TD3+ODT outperforms many baselines on a variety of tasks
  - Tasks: Adroit, MuJoCo, antmaze & maze2d (30+ different dataset-environment pairs)
  - See our paper for 10+ detailed ablation studies



Reward curves on adroit (Higher is better); our method highlighted in red

Thank you!

Feel free to contact kaiyan3@illinois.edu for any questions!

Code repository

https://github.com/KaiYan289/RL_as_Vitamin_for_Online_Decision_Transformers

Website

https://t.ly/LR7pE

PDF of our paper

arXiv: 2410.24108