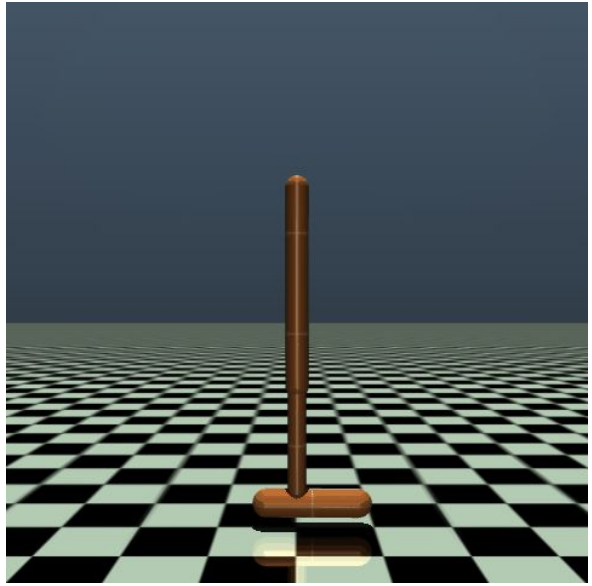
ЦЕНТР МОЛОДЁЖНОЙ
РОБОТОТЕХНИКИ
МГТУ ИМ. Н.Э. БАУМАНА

RL. Применение в робототехнике.

Hopper-v5

Двумерная одноногая фигура, состоящая из туловища, бедра, ноги внизу и одной ступни. Цель по умолчанию — совершать прыжки вперёд путём приложения крутящего момента к трём шарнирам, соединяющим четыре части тела.



$$\text{reward} = \underbrace{w_f \cdot v_x}_{\text{движение вперёд}} + \underbrace{r_{\text{healthy}}}_{\text{бонус за выживание}} - \underbrace{w_c \cdot \|a\|^2}_{\text{штраф за затраты}}$$

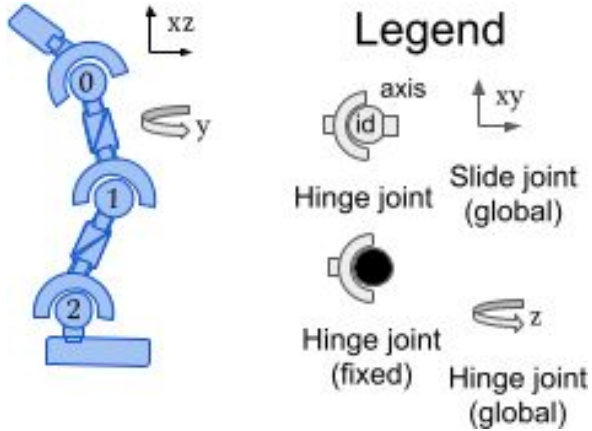
v_x = скорость движения вперед (по оси x)

a = вектор действия (крутящие моменты, приложенные к суставам)

$w_f \rightarrow$ вес награды за движение вперёд

$-w_c = 1e^{-3} \rightarrow$ вес штрафа за затраты

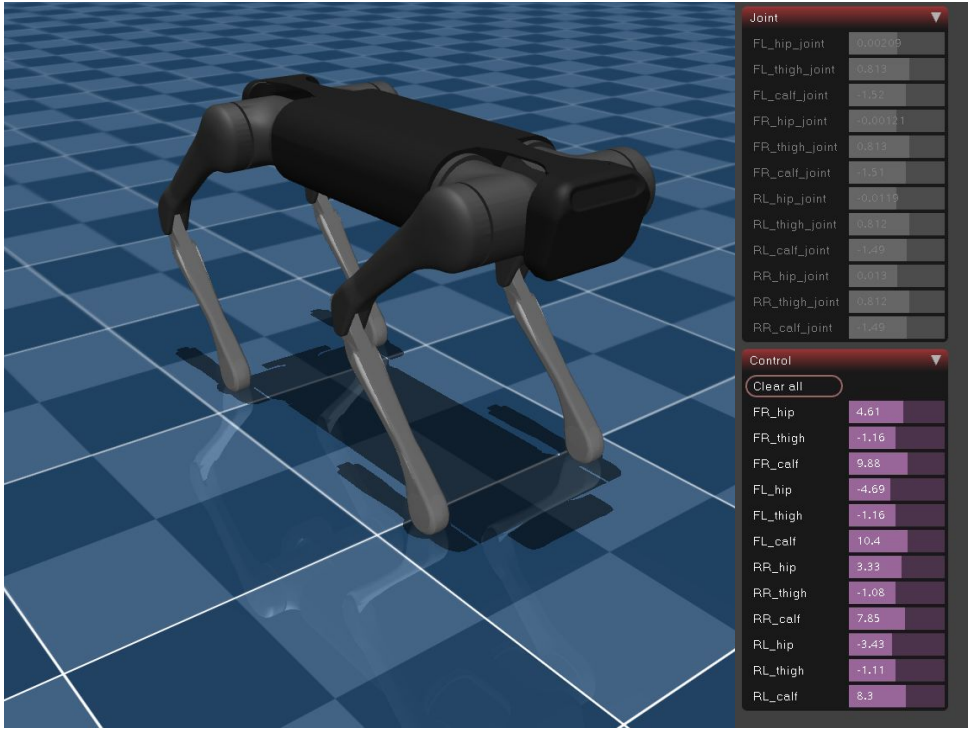
$-r_{\text{healthy}} = 1.0$ если робот "жив", иначе 0



MuJoCo

MuJoCo — это физический движок общего назначения.

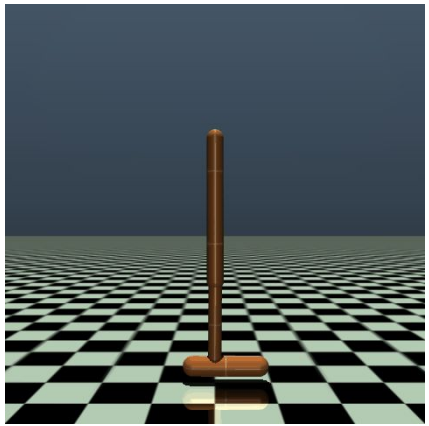
- mjModel содержит описание модели и, как ожидается, останется постоянным.
- mjData содержит все динамические переменные и промежуточные результаты
  - **qpos** представляет положение каждого сустава в симуляции
  - **qvel** обозначает скорости этих суставов



```
        x_pos = self.unwrapped.data.qpos[0]       # горизонтальное
положение
        x_vel = self.unwrapped.data.qvel[0]       # горизонтальная скорость
        z_vel = self.unwrapped.data.qvel[1]       # вертикальная скорость
```

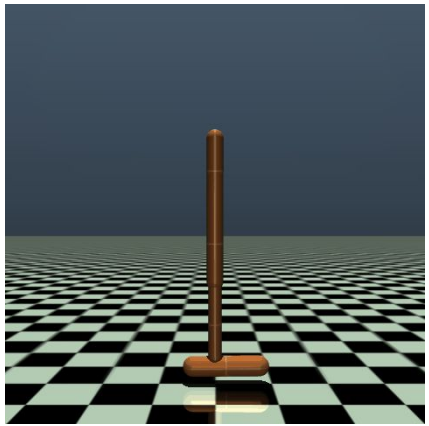
Вперёд

- Награда за x_vel



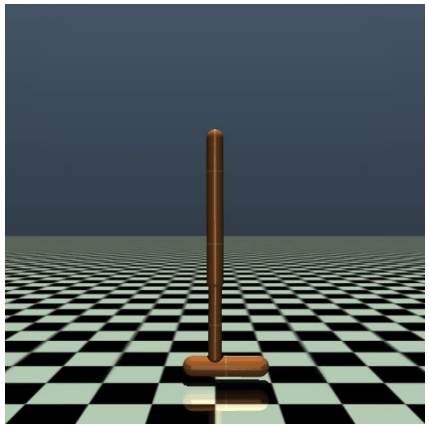
Прыжки на месте

- Штраф за x_vel
- Награда за z_vel
- Штраф за изменение x_pos



Назад

- Награда за -x_vel



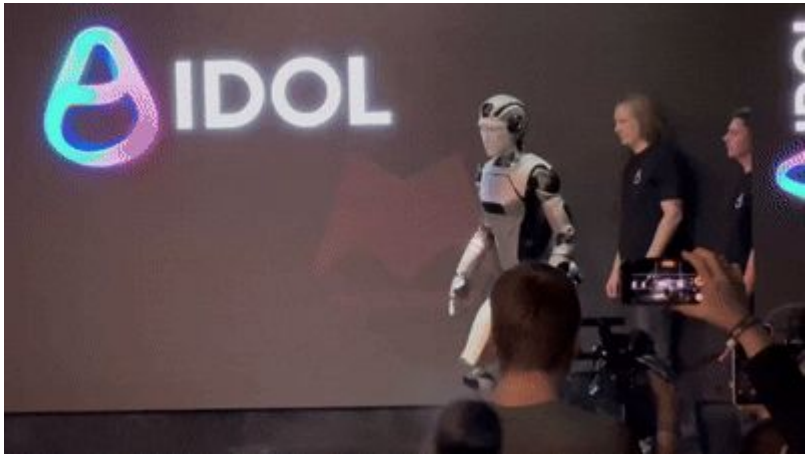
Награда робособаки

Для обучения Aliengo ходьбе вперед используется награда со множеством компонентов, здесь приведена только часть из них

Награда	Формула	Описание
Forward	$\min(v_x, 0.35)$	Поощряет движение вперед, ограничивая чрезмерную скорость.
Lateral Movement and Rotation	$-\|\mathbf{v}_y^t\|^2 - \|\omega_z^t\|^2$	Наказывает боковое (из стороны в сторону) движение и вращение вокруг вертикальной оси для поощрения прямолинейного движения.
Work	$-\|\mathbf{q}^t - \mathbf{q}^{t-1}\|^2$	Наказывает большие изменения положения суставов между шагами для поощрения плавного движения.
Ground Impact	$-\|\mathbf{f}^t - \mathbf{f}^{t-1}\|^2$	Снижает вознаграждение за резкие изменения сил реакции земли для обеспечения стабильного контакта стопы.
Smoothness	$-\|\boldsymbol{\tau}^t - \boldsymbol{\tau}^{t-1}\|^2$	Наказывает быстрые изменения крутящего момента для более плавного срабатывания с течением времени.
Action Magnitude	$-\|\mathbf{a}^t\|^2$	Препятствует большим управляющим действиям (например, крутящим моментам) для повышения энергоэффективности.
Joint Speed	$-\|\dot{\mathbf{q}}^t\|^2$	Штрафует высокие скорости сочленений для снижения рывковых или нестабильных движений.
Orientation	$-\|\theta_{roll}, \theta_{pitch}\|^2$	Поощряет робота сохранять вертикальную ориентацию, штрафуя отклонения крена и тангажа.
Z Acceleration	$-\|\dot{v}_z^t\|^2$	Штрафует вертикальное ускорение для предотвращения подпрыгивания и повышения устойчивости.
Foot Slip	$-\|\text{diag}(\mathbf{g}^t) \cdot \mathbf{v}_f^t\|^2$	Штрафует проскальзывание стопы во время контакта с землей для обеспечения стабильных моделей походки.

Трудности применения РЛ в реальности

- Сложность разработки системы вознаграждений
- Проблема sim2real



Sim2real

- Потери на трение: момент мотора на выходном валу различен в симуляции и реальности
- Цифровая модель робота отличается от реальной
- Люфты и эластичность: конструкция робота со временем изнашивается
- Задержки в управлении
- Прошивка моторов
- …

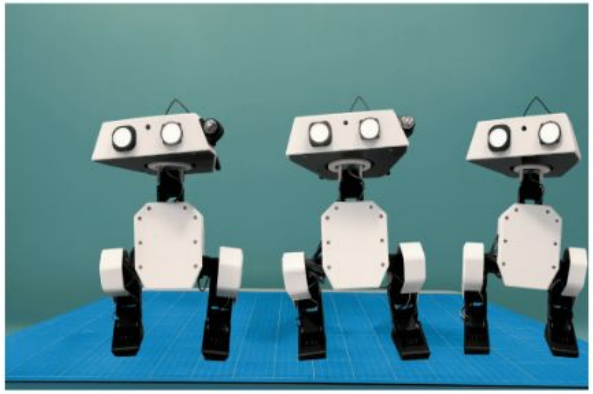
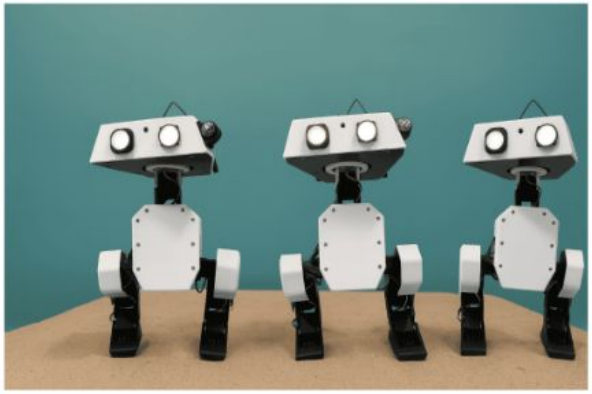
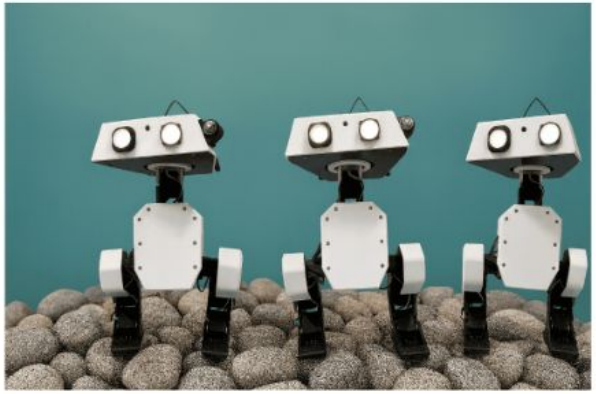
Способы решения проблемы sim2real

- Рандомизация домена (DR)
- Обучение в реальном мире, а не симуляторе
  - Риски: Поломка робота из-за резких и неправильных действий, долгое обучение
- Комбинация обучения в реальности и симуляторе

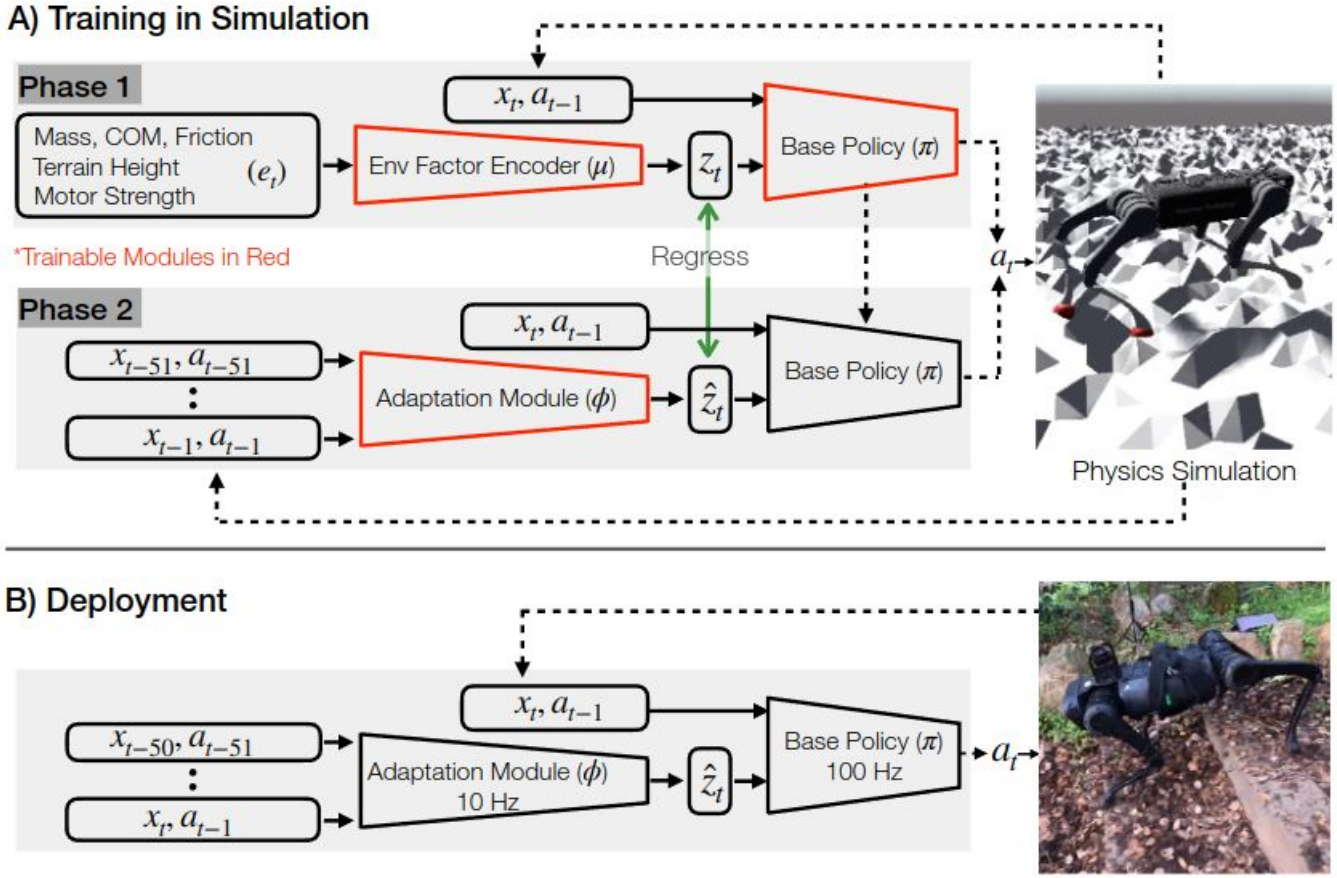
Рандомизация домена (DR)

Чем разнообразнее сценарии в ходе обучения, тем устойчивее политика в условиях разнообразной динамики окружающей среды.

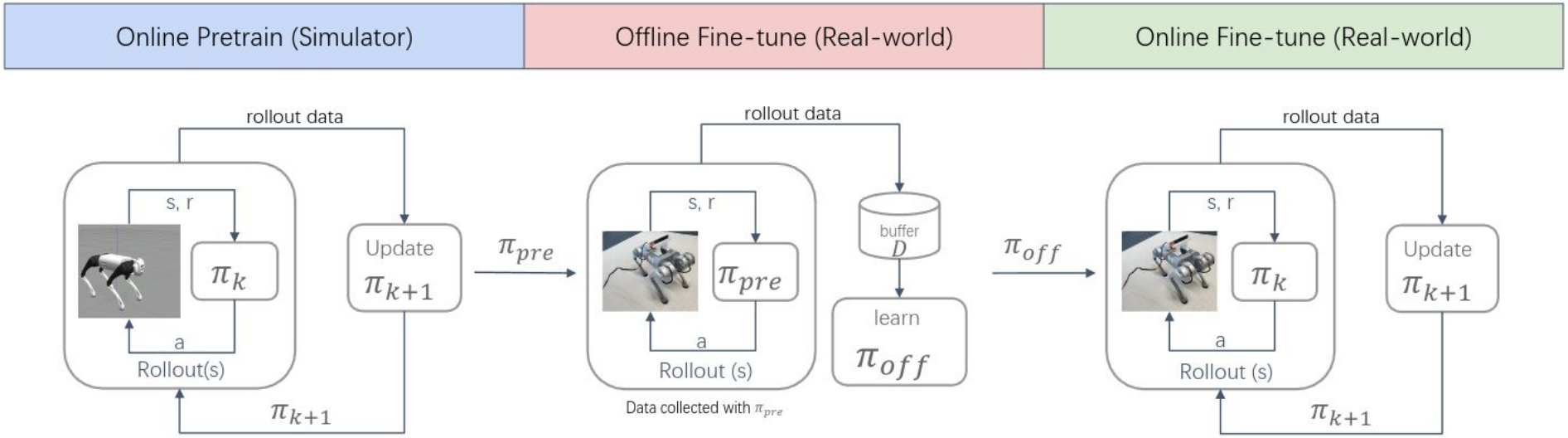
Например, можно рандомизировать коэффициент трения поверхности в задаче передвижения, или центр масс объекта для задачи манипуляции.

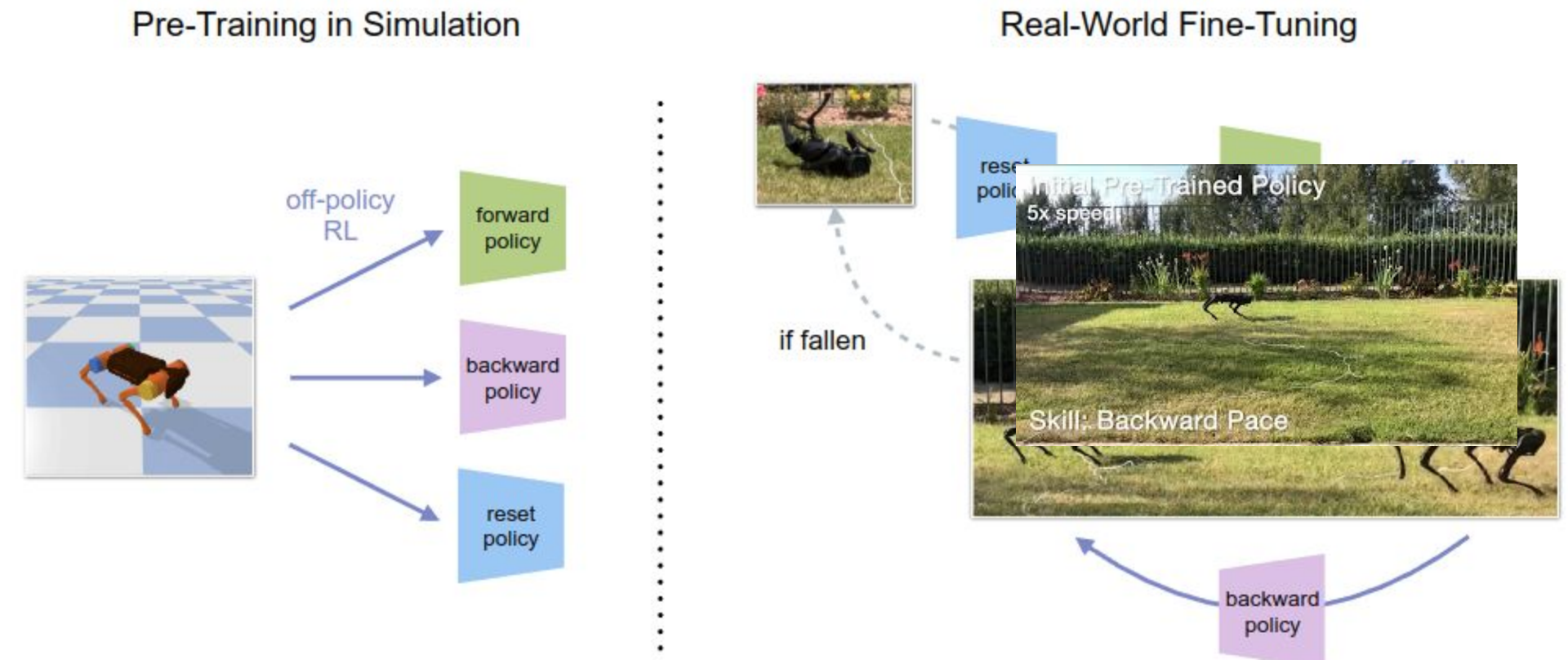
RMA


A) Training in Simulation
Phase 1
Mass, COM, Friction
Terrain Height
Motor Strength
(e_t)
Env Factor Encoder (μ)
z_t
x_t, a_{t-1}
Base Policy (π)
*Trainable Modules in Red
Regress
a_t
Phase 2
x_{t-51}, a_{t-51}
x_{t-1}, a_{t-1}
Adaptation Module (ϕ)
$\hat{z}_t$
x_t, a_{t-1}
Base Policy (π)
Physics Simulation
B) Deployment
x_{t-50}, a_{t-51}
x_t, a_{t-1}
Adaptation Module (ϕ)
10 Hz
$\hat{z}_t$
x_t, a_{t-1}
Base Policy (π)
100 Hz
a_t

Uni-O4: Unifying Online and Offline Deep Reinforcement Learning with Multi-Step On-Policy Optimization



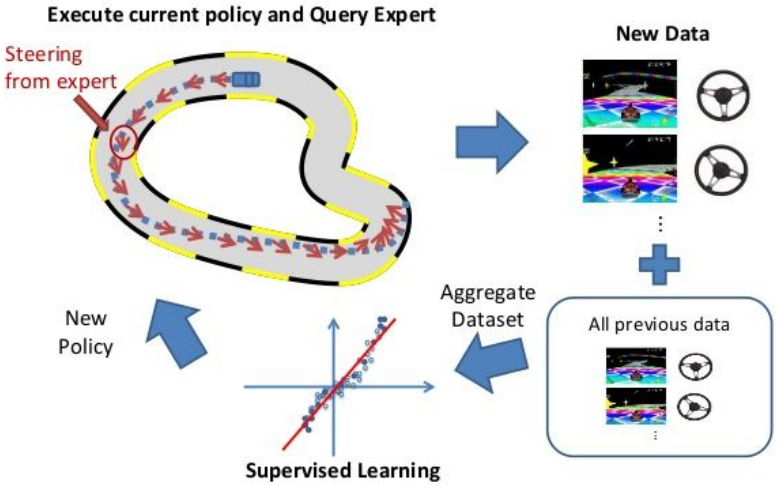
Legged Robots that Keep on Learning: Fine-Tuning Locomotion Policies in the Real World



Imitation Learning (IL)

Проблемы использования RL:

- Сложность создания награды
- Высокая стоимость эпизодов, Sim2Real gap
- Для решения сложных задач требуется огромное количество эпизодов



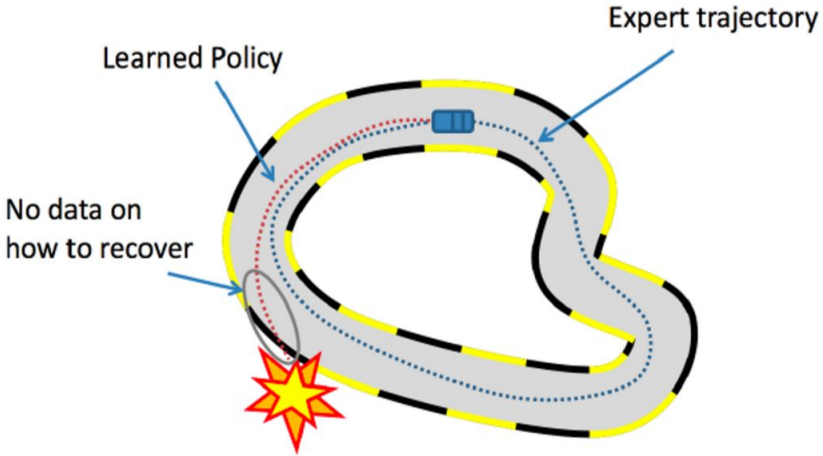
Подходы имитационного обучения позволяют учиться по демонстрациям экспертов, без необходимости задавать награду: $\pi_\theta(a|s) \approx \pi_E(a|s)$

Однако, IL ограничено качеством эксперта и агент не может превзойти его

Imitation Learning (IL)

Преимущества IL:

+ Не требует разработки награды
+ Не требует времени на исследования, начальная политика уже не случайна
+ Обучение быстрее и дешевле (Supervised Learning)



Недостатки:

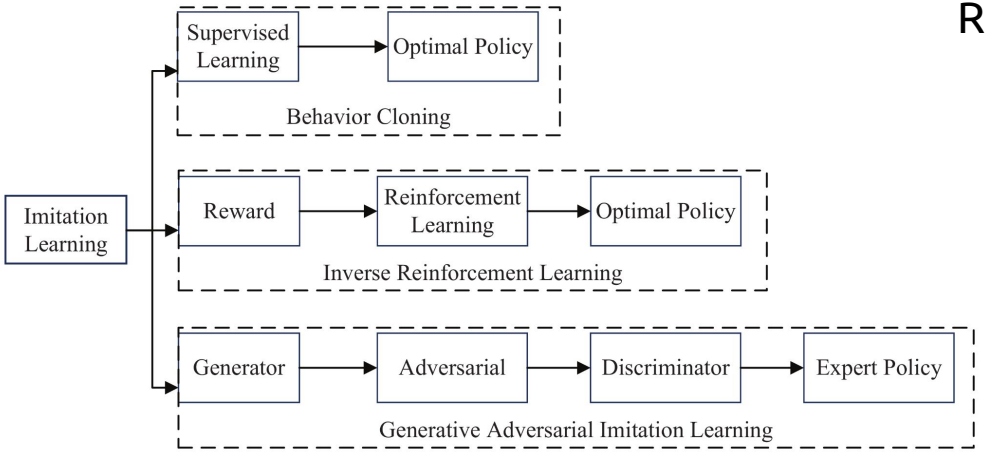
- Не лучшее решение задачи, максимум на уровне эксперта
- Проблема неизученных состояний, длинных задач
- Нет исследования и “понимания” цели задачи

Imitation Learning - обучение политике $\pi_\theta(\alpha|s)$
по примерам поведения эксперта $\pi_E(\alpha|s)$

Основные подходы в Imitation Learning

Основные подходы

- Behavioral Cloning (BC):
- Inverse Reinforcement Learning (IRL):



Прямое обучение политике эксперта (ошибка накапливается по мере отклонения от траектории эксперта)

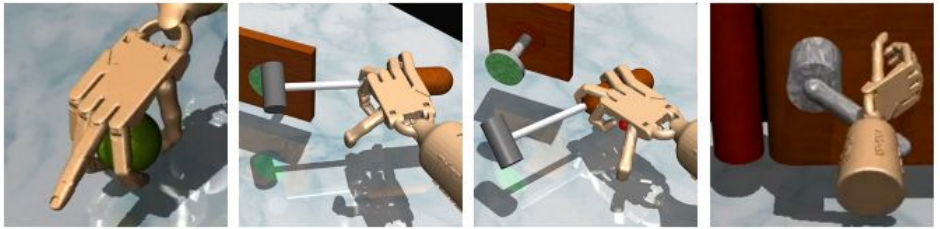
Восстанавливает функцию награды, после ее восстановления применяется RL

BC: DAgger (DAtaset Aggregation), Generative BC, Diffusion Policy

IRL: MaxEnt IRL, AIRL, (GAIL)

Комбинация RL и IL

DAPG:



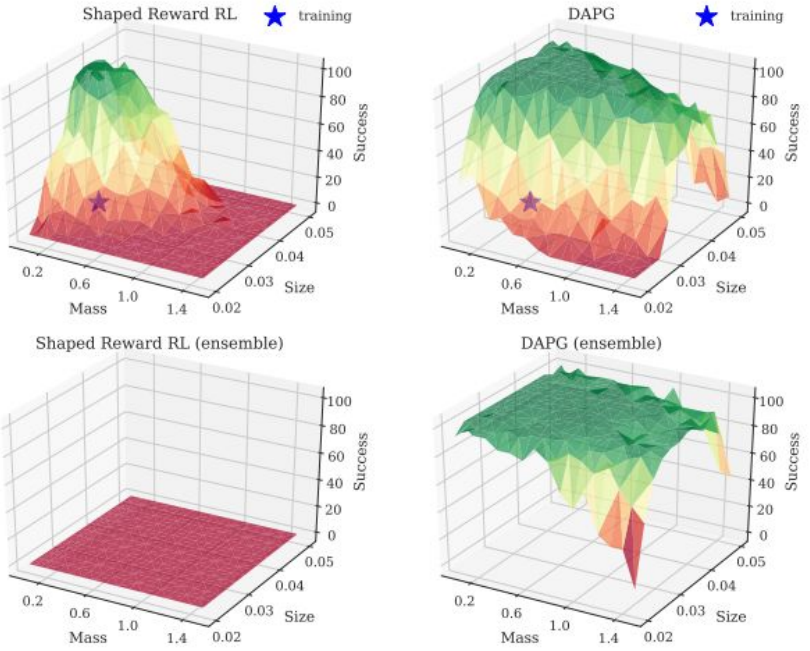
Современные методы обучения используют IL (для быстрого старта) и RL (для оптимального дообучения)

Стратегии:
1. Pretraining + Fine-tuning - Инициализация политики через IL, RL дообучение

2. Interactive Imitation - с вмешательством “эксперта” или человека во время обучения

3. Reward Learning from Demonstrations

IRL, GAIL, AIRL — оценка награды через дискриминатор



Начальная инициализация моделей

Главный урок за последние годы в RL - ни в коем случае не пытаться в учить с нуля RL на сложных задачах. Нужна хорошая базовая модель.