Министерство образования и науки Российской Федерации
Московский физико-технический институт (государственный университет)

Физтех-школа физики и исследований им. Ландау
Кафедра инновационной фармацевтики и биотехнологии
Лаборатория

Выпускная квалификационная работа бакалавра

Обучение локомоции шарнирной системы методами обучения с подкреплением с учетом её структуры

Автор:
Студент 114 группы
Афендульев Кирилл Дмитриевич

Научный руководитель:
кандидат физико-математических наук
Нейчев Радослав Георгиев

Москва 2025

Аннотация

Обучение локомоции шарнирной системы методами обучения с подкреплением с учетом её структуры

Афендульев Кирилл Дмитриевич

Аннотация

В данной работе рассматривается задача обучения политики человекоподобной локомоции в физически достоверной симуляционной среде с использованием методов обучения с подкреплением. Основное внимание уделено трансформерным архитектурам, адаптированным под структуру тела агента. Предлагается метод маскирования внимания на основе графа сочленений, позволяющий учесть априорные знания о физической структуре системы и тем самым повысить устойчивость и интерпретируемость поведения модели.

В качестве алгоритма обучения используется Proximal Policy Optimization (PPO) в актор-критической форме, позволяющий эффективно оптимизировать параметры в условиях непрерывного пространства действий. В работе исследуются три варианта структурного маскирования внимания: жёсткое двоичное, экспоненциально затухающее в зависимости от графового расстояния и data-driven подход с использованием устойчивой когомологии (persistent homology) для выявления значимых паттернов взаимодействия между частями тела.

Эксперименты демонстрируют, что внедрение структурных ограничений существенно ускоряет обучение и повышает качество итоговой политики. Топологический анализ матриц внимания позволяет дополнительно расширить маску на основе эмпирически обнаруженных устойчивых циклов. Полученные результаты подтверждают, что структурно-осведомлённые трансформеры являются перспективным направлением для построения интерпретируемых и вычислительно эффективных систем управления в задачах сложной локомоции.

Abstract

Содержание

1 Введение

Трансформерные архитектуры, изначально созданные для обработки текста и изображений, в последние годы зарекомендовали себя как мощный инструмент для анализа сложных, многомерных данных. Их главное преимущество — в механизме внимания, позволяющем выявлять нетривиальные связи между элементами входных данных и лучше учитывать структуру решаемой задачи.

Расширение применения трансформеров на область обучения с подкреплением (Reinforceme Learning, RL) открывает новые горизонты в построении универсальных стратегий управления. Особенно перспективным оказывается их использование для задач человекоподобной локомоции в реалистичных симуляциях, где агент должен одновременно учитывать положение суставов, контактные силы и собственные скорости, принимая решения в непрерывном пространстве и времени. Однако стандартный механизм *self-attention* в таких сценариях сталкивается с тремя ключевыми проблемами: высокой чувствительностью к настройке гиперпараметров, что может приводить к нестабильному обучению; склонностью фокусироваться на отдалённых, не всегда релевантных связях, что замедляет сходимость и увеличивает риск переобучения; а также высокой вычислительной сложностью, растущей квадратично с числом элементов, что ограничивает масштаб экспериментов.

Справиться с этими сложностями помогает внедрение априорных сведений о физической структуре робота прямо в архитектуру трансформера. Предлагается маскировать матрицу внимания, ориентируясь на граф сочленений: взаимодействия между механически несвязанными звеньями ограничиваются или исключаются. Это уменьшает шум в градиентах, делает оптимизацию устойчивее и снижает вычислительные затраты за счёт разрежённости операций.

Цель работы — подробно изучить эту методику. В симуляционной среде Humanoid Forces с использованием алгоритма **Proximal Policy Optimization (PPO)** обучаются и сравниваются три подхода к маскированию внимания: жёсткое двоичное, экспоненциально убывающее с увеличением графового расстояния и полностью обучаемое (*soft-mask*). Каждая из стратегий оценивается по скорости сходимости и финальному качеству обучения.

Результаты экспериментов дают количественную оценку того, насколько структурное ограничение внимания может выступать в роли полезного индуктивного смещения при обучении высокоразмерных локомоционных политик, повышая эффективность вычислений и улучшая интерпретируемость поведения гуманоидных роботов.

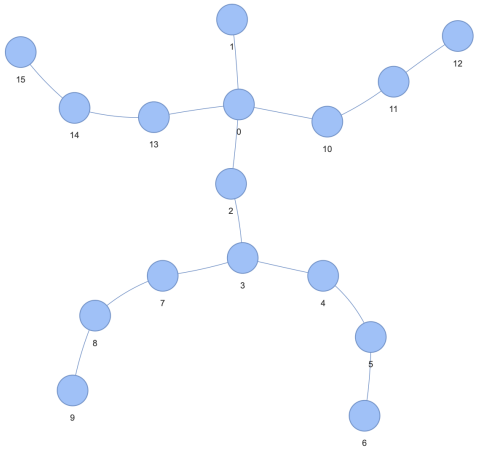
2 Постановка задачи и теоритическое введение

2.1 Среда моделирования и формулировка задачи управления

В данной работе задача человекоподобной локомоции рассматривается в физически достоверной симуляционной среде `HumanoidForces` из набора *IsaacGymEnvs*. Это одна из самых динамически насыщенных сред в составе платформы Isaac Gym, моделирующая движения полноразмерного гуманоидного агента с высокой степенью свободы. Агент обладает десятками суставов и может контактировать с окружающей средой через ступни, ноги и туловище. Физика движения рассчитывается в реальном времени с учётом гравитации, трения, упругости и контактных сил.

Наблюдение агента включает в себя широкий спектр признаков: относительные углы и угловые скорости суставов, линейные и вращательные скорости туловища, ориентацию тела, а также контактные силы и бинарные индикаторы касания ног с поверхностью. Управляющее воздействие представляет собой вектор непрерывных действий — моментов, прикладываемых к суставам.

(а) Официальная отрисовка агента



(б) Агент в терминах графов

Рис. 1: Обучаемый агент

Среда возвращает скалярное вознаграждение, которое моделирует желаемое поведение. Оно складывается из нескольких компонент: поощрение за скорость поступательного движения вперёд, штраф за избыточное энергопотребление, а также штраф за потерю равновесия или неестественные позы. Таким образом, цель агента — двигаться вперёд устойчиво, плавно и эффективно.

Формально, задача заключается в поиске параметров стохастической политики $\pi_\theta(a_t \mid s_t)$, максимизирующих ожидаемое дисконтированное вознаграждение:

$$J(\theta) = \mathbb{E}_{\tau\sim\pi_\theta}\left[\sum_{t=0}^{\infty}\gamma^t r_t\right], \quad \text{где} \quad 0 < \gamma < 1. \tag{1}$$

2.2 Алгоритм обучения с подкреплением

Для решения задачи используется алгоритм обучения с подкреплением Proximal Policy Optimization (PPO), который зарекомендовал себя как эффективный и надёжный метод для оптимизации параметров в средах с непрерывным пространством дей-

ствий. PPO относится к классу on-policy методов и реализуется в актор-критической форме, где одновременно обучаются две нейросети: одна отвечает за генерацию действий (актор), другая — за оценку ценности состояний (критик).

Ключевая особенность PPO — механизм ограничения обновлений политики, позволяющий избежать деструктивных скачков параметров. Это достигается за счёт введения клиппированной функции потерь:

$$\mathcal{L}_{\text{clip}} = \mathbb{E}_t \left[\min \left(r_t(\theta) A_t,\ \text{clip}(r_t(\theta), 1-\varepsilon, 1+\varepsilon) A_t \right) \right], \tag{2}$$

где $r_t(\theta)$ — отношение новой и старой политики, A_t — оценка преимущества, а ε — параметр, ограничивающий отклонения.

Преимущество вычисляется по методу обобщённой оценки GAE:

$$A_t = \sum_{l=0}^{T-t-1} (\gamma\lambda)^l \delta_{t+l}, \quad \text{где} \quad \delta_t = r_t + \gamma V(s_{t+1}) - V(s_t), \tag{3}$$

что позволяет учитывать будущее вознаграждение с понижающим коэффициентом и сгладить оценки.

Всё обучение осуществляется на GPU с использованием тысяч параллельных копий среды, что позволяет собирать объёмы данных, достаточные для стабильного обучения трансформерных архитектур. Все данные нормализуются, а оптимизация проводится в несколько эпох, что повышает эффективность использования каждого батча.

2.3 Маскированный трансформер и идея структурного внимания

Ключевой исследовательский вклад настоящей работы заключается в разработке и анализе трансформерной архитектуры с маскированным вниманием, адаптированной под задачу локомоции. Классический трансформер использует механизм self-attention, который позволяет каждому элементу входа (в нашем случае — каждой части тела) взаимодействовать со всеми остальными через взвешенную сумму. Это даёт модели гибкость и выразительную мощность, но одновременно приводит к переизбыточности и шуму, особенно в высокоразмерных наблюдениях.

В стандартной формуле внимания:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}} + M \right) V, \tag{4}$$

матрица M служит маской, регулирующей, какие элементы могут взаимодействовать. В нашей работе мы используем эту маску, чтобы внедрить в трансформер знание о структуре тела агента. В частности, внимание ограничивается ближайшими соседями в скелетной структуре: например, стопа взаимодействует с голенью, но не с рукой.

Такая форма структурного ограничения может быть реализована различными способами. Мы исследуем три варианта: жёсткое двоичное маскирование, экспоненциально затухающее внимание в зависимости от расстояния между частями тела, а также обучаемую маску, полученную на основе анализа системы. Даныне подходы позволяют сократить вычислительную сложность внимания и одновременно делают поведение модели более интерпретируемым, физически обоснованным и устойчивым к переобучению.

В результате мы стремимся выяснить, как структура тела агента может быть использована для улучшения качества и стабильности обучения политики локомоции, сохранив при этом выразительность трансформера.

2.4 Топологический анализ матриц внимания

Чтобы дополнить структурное маскирование эмпирическими наблюдениями, мы провели исследование топологической устойчивости матриц внимания, полученных от уже обученного трансформера. Идея заключается в том, чтобы рассматривать каждую матрицу внимания $A_t \in [0,1]^{N\times N}$, сохранённую на шаге t, как полный взвешенный граф, а затем извлекать из него устойчивые циклы первого гомологического измерения.

Построение фильтрации. Узлы графа соответствуют анатомическим звеньям, а веса рёбер интерпретируются как «близость» двух звеньев во внимании. Для применения алгоритма устойчивой когомологии веса переводятся в метрику

$$d_{ij} = 1 - A_{ij},$$

после чего на множестве $\{1,\dots,N\}$ строится фильтрация Вьеториса–Рипса

$$\mathcal{V}(\varepsilon) = \big\{\sigma \subseteq \{1,\dots,N\} \;\big|\; d_{uv} \le \varepsilon \;\; \forall\, u,v \in \sigma\big\},$$

где порог ε непрерывно увеличивается от 0 до 1. При малых значениях ε граф распадается на несвязные компоненты (H_0-особенности); по мере роста порога появляются и умирают циклы (H_1). Для каждой особенности фиксируются моменты «рождения» b и «смерти» d, которые образуют точки (b,d) в диаграмме устойчивости.

Извлечение устойчивых циклов. В каждый момент времени вычислялась диаграмма для H_1 и выбирался цикл с максимальной длительностью $\ell = d - b$. Его коцикл содержит конечный набор рёбер, формирующих представителя цикла. Вместо исчерпывающих сводных статистик мы аккумулировали: (1) частоту участия каждой вершины в самых «живучих» циклах; (2) количество появлений каждого ребра внутри этих циклов; (3) сами пары (b,d), позволяющие оценить характерные масштабы внимания.

Интеграция результата в граф маски. Отобранные рёбра с наибольшей накопленной частотой добавлялись в исходный кинематический граф агента, расширяя область разрешённого внимания. Тем самым получается мягкая, но физически осмысленная корректировка маски, основанная не на априорных, а на наблюдаемых паттернах взаимодействия звеньев во время инференса. Предполагается, что такое *data-driven* расширение маски сможет облегчить передачу информации между наиболее коррелированными частями тела, сохраняя при этом вычислительную экономию и интерпретируемость структурно-осведомлённого внимания.

3 Обзор существующих решений

Body Transformer (BoT): краткий обзор и критическая оценка

Одним из наиболее заметных недавних подходов, близких к поставленной задаче, является архитектура *Body Transformer* (BoT) — трансформер с маскированием, зависящим от морфологии робота. Авторы рассматривают тело агента как граф с узлами–датчиками и узлами–приводами, а затем подавляют элементы self-attention, связывающие топологически далёкие пары узлов. Такой «телесно-индуцированный» индиктивный bias позволяет, по их отчётам, опередить vanilla-трансформер и многослойный перцептрон как в имитационном, так и в онлайн-RL-режиме, причём при меньших FLOPs за счёт экстремально разреженных матриц внимания.

С точки зрения требований, сформулированных в разд. **??** (on-policy обучение PPO, среда `HumanoidForces`, сохранение интерпретируемости и пост-hoc анализ внимания), классическая версия BoT удовлетворяет лишь части ограничений:

- **Алгоритм RL.** В оригинальной работе BoT тестировался как в имитации, так и в off-policy RL; строгой привязки к PPO там нет, поэтому перенос требует адаптации.
- **Среды и набор наблюдений.** Авторы фокусируются главным образом на задачах с умеренным числом степеней свободы (Unitree A1, Adroit Hand). Среда `HumanoidForces`, где критично учитывать большие контактные силы и штрафы за энергопотребление, в их эксперименты не входит.
- **Форма маски.** BoT опирается на фиксированную двоичную маску «ближайших соседей», тогда как в нашей работе исследуются более гибкие схемы (экспоненциальное затухание, обучаемый bias, а также расширение маски на основе топологической устойчивости внимания).
- **Пост-hoc анализ.** BoT не предусматривает извлечение устойчивых циклов или какую-либо топологическую диагностику внутренних матриц внимания; следовательно, выводы о том, *как* сеть использует своё внимание, остаются поверхностными.

Таким образом, Body Transformer демонстрирует, что статическое маскирование само по себе способно улучшить качество и ускорить обучение, однако остаётся открытым вопрос, насколько эти преимущества сохранятся в более жёстких условиях задачи `HumanoidForces` и каким образом можно дополнительно повысить интерпретируемость модели. Настоящая работа развивает направление BoT в двух аспектах: во-первых, адаптирует маскирование и сам алгоритм обучения к специфике высокоэнергетической человекоподобной локомоции под PPO; во-вторых, вводит *data-driven* расширение маски через анализ устойчивых H_1-циклов в динамике внимания, что позволяет выявлять и закреплять важные межзвеньевые связи, обнаруженные самой моделью во время инференса.

Graph Embodiment Transformer (GET-Zero): краткий обзор и критическая оценка

GET-Zero — недавний подход, направленный на решение задачи **zero-shot управления** роботами с изменяющейся морфологией. Архитектура GET-Zero строится на трёх ключевых идеях: (i) модифицированный **Graph Embodiment Transformer (GET)**,

в котором механизм внимания снабжён *обучаемым биасом*, зависящим от графовых расстояний между звеньями тела и их иерархических (родитель–потомок) отношений; (ii) **distillation-aware обучение**, при котором знания множества PPO-экспертов конденсируются в одну универсальную политику посредством поведенческого клонирования; (iii) **self-modeling loss**, включающий вспомогательное задание на предсказание прямой кинематики, что усиливает способность модели к обобщению.

Базовые эксперименты выполнены в симуляции задачи манипуляции (LEAP Hand), где агент должен вращать кубик внутри ладони четырёхпальцевой руки. Авторы сгенерировали 236 различных морфологий (графов руки), обучили PPO-экспертов для 44 из них, а затем обучили GET на основе их траекторий. Результаты показали, что модель демонстрирует уверенное обобщение на ранее не виданные морфологии (например, варианты с удалёнными пальцами или удлинёнными фалангами), превосходя трансформерные и графовые базлайны на 10–20% без необходимости в дообучении. Также показана успешная переносимость политики на реальное аппаратное обеспечение.

Сравнивая GET-Zero с задачей, решаемой в нашей работе, можно отметить следующее:

- **Алгоритм RL.** GET-Zero реализован в off-policy парадигме: политика обучается на демонстрациях, полученных от ранее обученных PPO-экспертов. В отличие от этого, наша работа использует *он-policy PPO*, где трансформер обучается непосредственно во взаимодействии с симуляцией. Это делает наш подход более универсальным и пригодным для условий, где сбор демонстраций затруднён или невозможен.

- **Целевая задача.** GET-Zero сосредоточен на задачах манипуляции с умеренным числом степеней свободы и преимущественно локальными контактами. В нашей работе решается задача *человекоподобной локомоции* в среде `HumanoidForces`, требующая учёта сложной динамики, взаимодействия с опорной поверхностью и глобального баланса тела. Эти различия предъявляют иные требования к архитектуре, устойчивости и интерпретируемости моделей.

- **Механизм структурного внимания.** В GET-Zero внимание регулируется через *обучаемый графовый биас* — мягкий, непрерывный сдвиг logits в зависимости от расстояний по телесному графу. Такой подход удобен при работе с множеством вариаций морфологии. В нашей работе внимание маскируется явно, причём рассматриваются сразу несколько вариантов маски: жёсткое двоичное, экспоненциально-затухающее и адаптивное расширение на основе топологического анализа. Это позволяет напрямую встраивать априорные знания о сочленениях и гибко настраивать степень локальности взаимодействий.

Таким образом, несмотря на разницу в задачах, наша работа логически развивает идеи, заложенные в GET-Zero и BoT: использование телесной структуры агента в трансформере, внедрение графовой информации в self-attention, и ориентация на обобщаемость. Однако в отличие от GET-Zero, мы делаем акцент на *обучение локомоции*, на *он-policy* взаимодействие с физической средой и на *интерпретируемость* через топологические методы. Это делает предложенную архитектуру более пригодной для высокоэнергетических, динамически насыщенных задач управления.

4 Исследование и построение решения задачи

4.1 Декомпозиция задачи

Поставленная цель — получение устойчивой, энергоэффективной и интерпретируемой политики локомоции в среде с физически достоверной симуляцией — является слишком комплексной, чтобы решать её напрямую. Для последовательного приближения к решению задача была декомпозирована на ряд независимых и воспроизводимых этапов. Каждый из них формулируется таким образом, чтобы его корректность и влияние можно было оценить экспериментально либо количественно сравнить с базовыми сценариями. Итоговая структура декомпозиции включает следующие шаги:

1. Реализация и валидация *базовой трансформерной* политики без маскирования, служащей опорной точкой качества. Дополнительно проводится обучение модели на основе *MLP*, что позволяет прояснить вклад архитектуры внимания в общем качестве поведения.

2. Внедрение *жёсткого двоичного* маскирования, соответствующего структуре скелетного графа. Обнуляются элементы внимания между механически несвязанными звеньями. Анализируется влияние такой структуры на устойчивость обучения, скорость сходимости и итоговую награду.

3. Разработка *экспоненциально-затухающих* масок на основе расстояний в графе сочленений. Для множества значений коэффициента затухания α проводится серия обучений и выбирается оптимальный диапазон с точки зрения метрик качества.

4. Проведение *топологического анализа внимания* после завершения обучения. С использованием методов устойчивой гомологии извлекаются повторяющиеся циклические структуры в графе внимания. Результаты анализа интерпретируются и используются для построения новой маски на основе статистически значимых рёбер.

5. Обучение новой политики с использованием маски, расширенной по результатам TDA. Сравнение её эффективности и вычислительных свойств с предыдущими вариантами.

6. Финальное сравнение всех протестированных методов по ключевым метрикам: скорость сходимости, средняя эпизодическая награда, интерпретируемость структуры внимания и вычислительная эффективность. На основе результатов формулируются рекомендации по использованию структурного маскирования в задачах локомоции.

Каждый из указанных этапов документирован в следующих разделах, снабжён иллюстрациями, количественными оценками и выводами, необходимыми для сопоставления исследуемых подходов.

4.2 Базовая трансформерная политика

На первом этапе эксперимента была реализована и обучена базовая трансформерная архитектура, в которой механизм внимания функционирует в стандартном виде без какого-либо структурного ограничения. Такая модель используется в качестве опорной точки — она позволяет оценить вклад различных методов маскирования внимания, рассматриваемых в последующих экспериментах.

Трансформер обучался в течение 12 часов на видеокарте NVIDIA RTX 4090 при фиксированных значениях генераторов случайных чисел (сиды). Использовалась симуляционная среда `HumanoidForces` из пакета `IsaacGymEnvs`. В процессе обучения средняя эпизодическая награда стабилизировалась на уровне примерно 9300. Это значение принято в данной работе в качестве базового ориентира, с которым сравниваются все последующие модификации архитектуры.

На рис. 2 представлена кривая обучения модели, демонстрирующая динамику накопления награды. Видно, что после фазы начальной нестабильности кривая постепенно выравнивается, что свидетельствует о сходимости модели к устойчивой стратегии поведения. Тем не менее, итоговое качество политики не является удовлетворительным с точки зрения ни точности движений, ни их интерпретируемости.

Для дополнительного сравнения была также обучена модель, основанная на многоуровневом перцептроне (MLP), использующем те же входные признаки, но не обладающем механизмом внимания. Это позволяет оценить относительную пользу от использования трансформерной архитектуры как таковой. Как показали предварительные эксперименты, MLP-политика демонстрировала существенно худшие результаты, как по стабильности, так и по максимальной достигаемой награде.

Таким образом, базовая трансформерная политика без маскирования служит фундаментом для анализа эффективности предложенных модификаций, направленных на улучшение обучения за счёт структурных ограничений внимания.

4.3 Жёсткая двоичная маска

На втором этапе был реализован подход, в котором механизм внимания трансформера ограничивается заранее заданной бинарной маской. Маска строится на основе графа сочленений агента, отражающего механическую структуру тела: каждое звено представляет собой вершину графа, а наличие физического соединения между двумя звеньями задаёт ребро. Соответственно, если элементы i и j не соединены в этом графе, соответствующий элемент матрицы внимания A_{ij} принудительно зануляется:

$$A_{ij} = 0, \quad \text{если } (i,j) \notin E,$$

где E — множество рёбер в графе сочленений.

Мотивация введения такого ограничения связана с устранением избыточных связей, не соответствующих физической структуре системы. Полная матрица self-attention содержит $\mathcal{O}(n^2)$ связей между n узлами, однако значительная часть этих связей является физически бессмысленной. Такие связи не только не способствуют формированию корректной политики, но и вводят шум в процесс оптимизации, затрудняя обучение модели.

Использование бинарной маски позволяет добиться двух эффектов. Во-первых, это снижает вычислительную нагрузку, поскольку разреженность матрицы внимания может быть эффективно учтена в тензорных операциях. Во-вторых, устраняется влияние нерелевантных компонент, что способствует концентрации внимания модели на локально значимых элементах. Это особенно важно в задачах локомоции, где пространственная близость частей тела отражает и функциональную связанность.

Практически, обучение с жёстким маскированием показало улучшение как по скорости сходимости, так и по финальному качеству поведения. Уже к шагу $6 \cdot 10^3$ средняя эпизодическая награда превысила значение, достигнутое базовым трансформером, а к финалу обучения она стабилизировалась на уровне 9400. Также наблюдалась повышенная устойчивость градиентных норм, что свидетельствует о более предсказуемом процессе оптимизации.

Таким образом, простое структурное ограничение внимания даёт значительный прирост в качестве политики, подтверждая гипотезу о важности априорных связей в моделировании поведения агента.

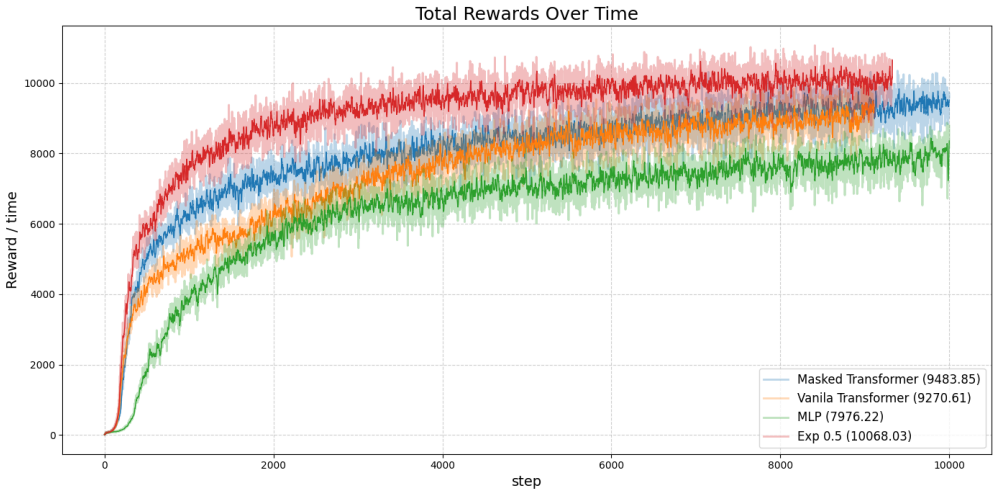


Рис. 2: Динамика средней награды для базового трансформера без маскирования.

4.4 Экспоненциально-затухающая маска

Жёсткое бинарное маскирование, описанное в предыдущем разделе, обладает очевидным ограничением: оно полностью блокирует передачу информации между механически несвязанными звеньями. Хотя такое локальное внимание повышает интерпретируемость и снижает переобучение, оно также препятствует агрегации глобального контекста, который может быть необходим для выработки координированного поведения, особенно в высокоуровневых паттернах локомоции.

Для более гибкого контроля степени локальности внимания была реализована экспоненциально затухающая маска, задающая вес взаимодействия между узлами на основе графового расстояния между ними. Вес внимания между вершинами i и j задавался по формуле:

$$w_{ij} = \exp(-\alpha\, d_{ij}),$$

где d_{ij} — длина кратчайшего пути между узлами i и j в графе сочленений, а параметр $\alpha > 0$ управляет скоростью затухания.

Такой подход позволяет непрерывно модулировать вклад удалённых компонентов: соседние звенья получают вес, близкий к единице, в то время как влияние удалённых элементов экспоненциально уменьшается. Это не только снимает ограничение на передачу глобальной информации, но и позволяет сохранять локальный приоритет внимания, необходимый для точной координации движений.

Для подбора оптимального значения параметра α был проведён сеточный поиск по значениям $\alpha \in \{0.3, 0.5, 0.7\}$. Эксперименты показали, что наилучший баланс между скоростью сходимости и итоговым качеством демонстрировала модель с $\alpha^\star = 0.5$. В этом случае средняя эпизодическая награда достигала значения 9950, а выход на стадию стабильного поведения происходил быстрее, чем при других значениях параметра.

На рис. 3 приведена динамика обучения для всех протестированных значений α. Можно видеть, что слабое затухание ($\alpha = 0.3$) приводит к более шумной и нестабильной траектории, в то время как чрезмерное усиление локальности ($\alpha = 0.7$) замедляет обучение и снижает итоговую награду.

Таким образом, плавная экспоненциальная маска даёт возможность учесть как локальные, так и удалённые зависимости в рамках единой архитектуры, улучшая адап-

тивность модели и обеспечивая компромисс между структурными ограничениями и выразительностью внимания. Однако такое маскирование не зануляет веса, следовательно не происходит оптимизация вычислений. Также подобное маскирование все равно почти не "допускает"влияние друг на друга удаленных вершин графа.

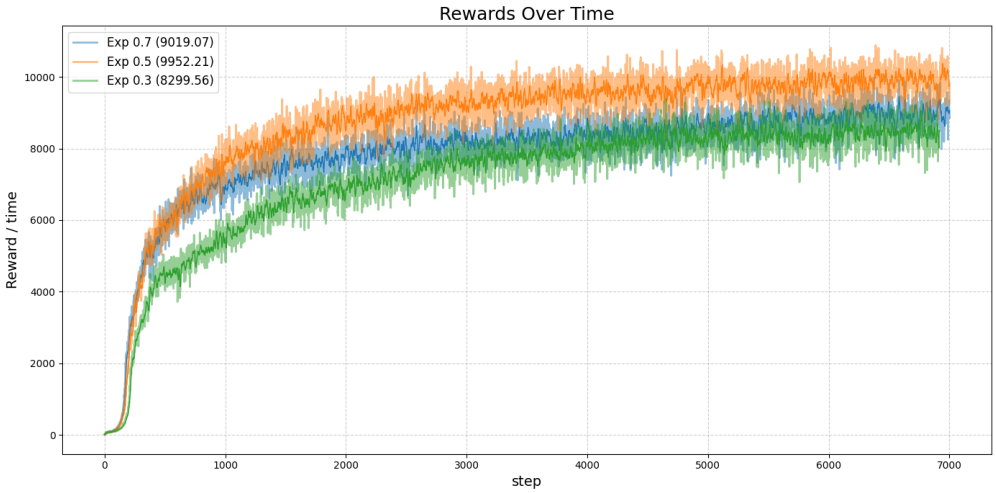


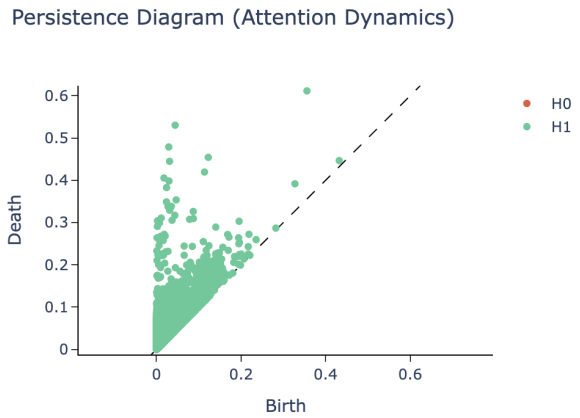
Рис. 3: Влияние коэффициента затухания α на динамику обучения.

4.5 Топологический анализ внимания

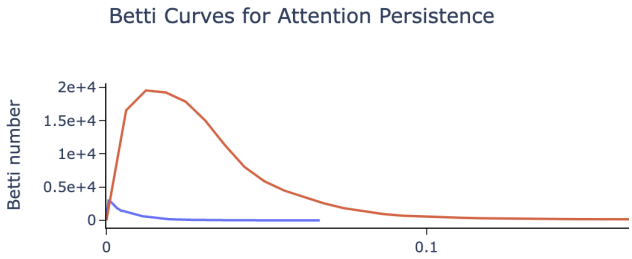
Для углубленного анализа поведения механизма внимания после обучения трансформерных политик была применена методология топологического анализа данных (TDA). Целью являлось выявление устойчивых глобальных структур в матрицах внимания, которые могут указывать на формирование повторяющихся и значимых паттернов взаимодействий между различными частями тела агента.

Были сохранены тензоры внимания для 1000 последовательных шагов инференса. Для каждого временного среза строилась симметризованная матрица расстояний $D_{ij} = 1 - A_{ij}$, где A_{ij} — соответствующий элемент матрицы внимания. Далее на этих данных применялась фильтрация Вьеториса–Рипса, в рамках которой строились комплексные графы и вычислялись диаграммы устойчивости для гомологий нулевого и первого порядков.

Особое внимание уделялось компонентам первого порядка (H_1), соответствующим нетривиальным циклам в графе внимания. Для каждого временного шага извлекался наиболее устойчивый (долго живущий) цикл, и на его основе определялись вершины и рёбра, наиболее часто входящие в состав этих циклов. Примеры характерных диаграмм устойчивости и характеристик Бетти приведены на рис. 4.



(а) Диаграмма устойчивости



(б) Характеристика Бетти

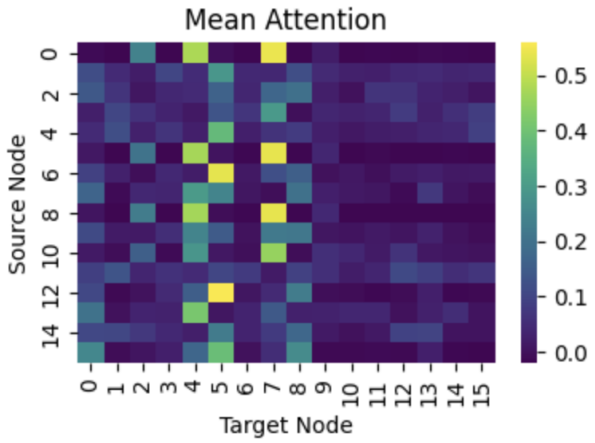
Рис. 4: Топологические характеристики внимания на этапе инференса

Результаты анализа показали, что граф внимания не распадается на изолированные компоненты связанности: все звенья остаются взаимосвязанными, что указывает на наличие глобальной координации в выработанной политике. В то же время устойчивые циклы в H_1 свидетельствуют о формировании сложных замкнутых паттернов взаимодействий, поддерживающих поведение агента.

Анализ распределения внимания по различным частям тела агента показал выраженную неоднородность в активации различных сегментов скелета. Наибольшую значимость, с точки зрения участия в формировании паттернов внимания, продемонстрировали узлы, соответствующие нижним конечностям — в частности, вершины 4, 5, 7 и 8, отвечающие за бёдра и колени. Именно на эти области проецируется наибольшее количество внимания со стороны остальных сегментов, что подчёркивает их ключевую роль в стабилизации и управлении движением.

В противоположность этому, сегменты, связанные с руками (вершины 10–15), получили существенно меньшее внимание со стороны остальных компонентов. В ряде случаев можно наблюдать почти полное отсутствие входящих направлений внимания, что указывает на их маргинальную роль в задаче локомоции. При этом руки сами активно отслеживают состояние нижних конечностей, что можно интерпретировать как проявление компенсаторной стратегии стабилизации — возможно, в целях балансировки или удержания центра масс.

Таким образом, структура внимания демонстрирует логично интерпретируемую асимметрию: основная часть тела «ориентируется» на ноги как на опорный и управляющий элемент, в то время как конечности верхнего пояса выполняют скорее реактивную роль. Более детальное распределение направлений внимания между сегментами визуализировано на рис:



Среднее внимание между узлами

Статистический анализ частоты появления вершин и рёбер в устойчивых циклах позволил выделить наиболее значимые структурные элементы. Как видно из табл. 1, наиболее активными оказались звенья, соответствующие области таза, бедра и коленей, что согласуется с физиологическим значением этих узлов для устойчивости и баланса при локомоции.

Таблица 1: Анализ значимости узлов и связей в графе внимания

Топ вершин по в.с.	**Суммарный вес**
Вершина 7 (бедро)	1.786
Вершина 5 (колено)	1.730
Вершина 4 (бедро)	1.724
Вершина 8 (колено)	1.212
Вершина 0 (тело)	1.133

Пара вершин	**Описание**	τ **(порог слияния)**
(0, 8)	тело–колено	0.491
(5, 8)	колено–колено	0.491
(0, 5)	тело–колено	0.492
(4, 5)	колено–бедро	0.604
(0, 7)	бедро–тело	0.609
(5, 7)	колено–бедро	0.615
(7, 8)	бедро–колено	0.615

На основании этих наблюдений была построена новая маска внимания, в которую добавлялись связи, часто встречающиеся в устойчивых H_1-циклах. В результате была сформирована более разреженная, но топологически обоснованная структура внимания, которая служит компромиссом между полной связностью и жёсткой структурной

маской. Обучение политики с использованием такой маски дало результаты, сопоставимые с экспоненциально-затухающей схемой, как показано на рис. 5. При этом новая маска обладала меньшей плотностью, что позволило ускорить обучение и сократить объём вычислений благодаря применению fast-path оптимизаций в библиотеке PyTorch.

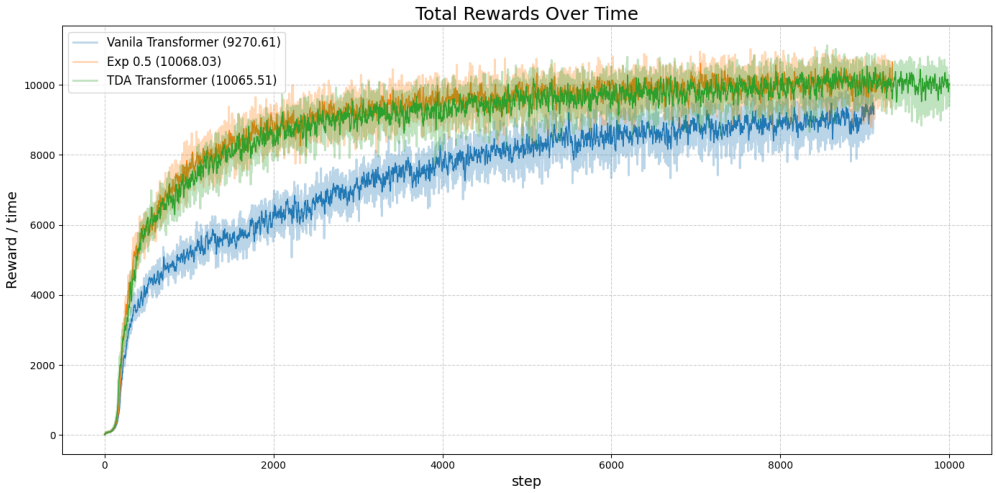


Рис. 5: Динамика средней награды при использовании маски на основе TDA-анализа

4.6 Разработка программных компонентов

В рамках выполнения работы были реализованы и протестированы четыре варианта трансформерных политик, различающихся стратегией формирования маски внимания:

- **TransformerPolicy** — базовая реализация трансформера без какого-либо структурного маскирования. Служит опорной точкой для сравнения с модифицированными архитектурами.
- **MaskedTransformer** — вариант, в котором используется фиксированная *жёсткая двоичная маска*, построенная на основе графа сочленений агента. Маска обнуляет элементы внимания между парами узлов, не связанных в графе.
- **ExpDecayTransformer** — архитектура с *экспоненциально-затухающим* маскированием. Маска формируется по формуле $w_{ij} = \exp(-\alpha d_{ij})$, где d_{ij} — длина кратчайшего пути в графе сочленений, а α — гиперпараметр, определяющий степень локальности взаимодействий. Поддерживается настройка маски по предобработанным графам или на лету.
- **CustomMaskTransformer** — трансформер, в котором используется маска, расширенная по результатам топологического анализа. В неё входят как связи, заданные исходной топологией тела, так и рёбра, выявленные как устойчивые циклы H_1 в динамике внимания.

Кроме основной архитектуры политики, была реализована поддержка следующих ключевых компонентов:

- *Модуль загрузки и предобработки скелетных графов* агента на основе структуры среды `HumanoidForces`, с возможностью вычисления кратчайших расстояний и построения масок различных типов.
- *Сбор и сохранение тензоров внимания* в процессе инференса модели, с логированием всех слоёв и голов self-attention для последующего анализа.
- *TDA-модуль анализа* внимания на основе библиотеки `Giotto-TDA` и `ripser`. Реализована фильтрация Вьеториса–Рипса, вычисление диаграмм устойчивости, извлечение главных циклов и подсчёт частотных характеристик рёбер.
- *Интеграция с RL-фреймворком* `rl_games`, в том числе модификация конфигурации и циклов обучения PPO под кастомные модели политики.

Вся реализация выполнена на языке `Python` с использованием библиотек `PyTorch`, `rl-games`, `networkx`, `Giotto-TDA`, `Matplotlib` и `Isaac Gym`. Выбор платформы обусловлен широким распространением этих инструментов в задачах обучения с подкреплением, развитой экосистемой RL-фреймворков и поддержкой вычислений на GPU, критически важной для ускоренного обучения и инференса.

5 Заключение

В данной работе была исследована задача обучения человекоподобной локомоции в физически достоверной симуляционной среде `HumanoidForces` при помощи методов обучения с подкреплением. Основное внимание было уделено архитектурам трансформерного типа, адаптированным под морфологию тела агента посредством маскирования внимания.

Были реализованы и проанализированы три варианта структурного ограничения self-attention:

- **Жёсткая двоичная маска**, отражающая граф сочленений тела, позволила существенно повысить устойчивость обучения и качество политики, устраняя нерелевантные связи и снижая вычислительную нагрузку;
- **Экспоненциально-затухающая маска** обеспечила плавное ослабление взаимодействий между удалёнными звеньями, сохранив при этом локальную интерпретируемость. Оптимальный коэффициент затухания был подобран экспериментально;
- **TDA-расширенная маска**, сформированная на основе анализа устойчивых H_1-циклов в матрицах внимания, выявила эмпирически значимые паттерны взаимодействий между частями тела, позволив дополнительно усилить политические связи в модели.

На основе этих трёх подходов была построена и протестирована серия трансформерных политик, показавших устойчивый прирост качества по сравнению с базовой немаскированной моделью. Наилучшие результаты достигнуты при использовании TDA-расширенной маски, сочетающей априорные и data-driven компоненты.

Особое внимание в работе было уделено *топологическому анализу внимания* — новому направлению, которое позволяет рассматривать динамику self-attention как сложную, но интерпретируемую структуру. Применение устойчивой когомологии (persistent homology) позволило извлечь повторяющиеся циклы в графе внимания, охарактеризовать вклад отдельных звеньев и рёбер, и использовать эти данные для модификации архитектуры.

Тем не менее, это направление остаётся недостаточно исследованным. Методика TDA в контексте анализа внимания требует дальнейшей адаптации: повышения устойчивости к шуму, более тонкой фильтрации значимых циклов и интеграции с процессом обучения на более глубоком уровне. Также актуальными являются вопросы интерпретации найденных паттернов с точки зрения физики управления и морфологии агента. В перспективе возможна разработка гибридных методов, сочетающих TDA с обучаемыми механизмами внимания и анализом траекторий.

Список литературы

[1] Body Transformer: Leveraging Robot Embodiment for Policy Learning / Carmelo Sferrazza, Dun-Ming Huang, Fangchen Liu et al. // *arXiv preprint arXiv:2408.06316*. — 2024.

[2] *Patel, Austin.* GET-Zero: Graph Embodiment Transformer for Zero-shot Embodiment Generalization / Austin Patel, Shuran Song // *arXiv preprint arXiv:2407.15002*. — 2024.

[3] Masked Sensory-Temporal Attention for Sensor Generalization in Quadruped Locomotion / Dikai Liu, Tianwei Zhang, Jianxiong Yin, Simon See // *arXiv preprint arXiv:2409.03332*. — 2024.

[4] Real-World Humanoid Locomotion with Reinforcement Learning / Ilija Radosavovic, Tete Xiao, Bike Zhang et al. // *arXiv preprint arXiv:2303.03381*. — 2023.

[5] *Gu, Xinyang.* Humanoid-Gym: Reinforcement Learning for Humanoid Robot with Zero-Shot Sim2Real Transfer / Xinyang Gu, Yen-Jen Wang, Jianyu Chen // *arXiv preprint arXiv:2404.05695*. — 2024.

[6] Isaac Gym: High Performance GPU-Based Physics Simulation For Robot Learning / Viktor Makoviychuk, Lukasz Wawrzyniak, Yunrong Guo et al. // *arXiv preprint arXiv:2108.10470*. — 2021.

[7] *Kushnareva, Laida.* Betti numbers of attention graphs is all you really need / Laida Kushnareva, Dmitri Piontkovski, Irina Piontkovskaya // *arXiv preprint arXiv:2207.01903*. — 2022.

[8] *Snopov, Petar.* Vulnerability Detection via Topological Analysis of Attention Maps / Petar Snopov, Anton N. Golubinsky // *arXiv preprint arXiv:2410.03470*. — 2024.

[9] Proximal Policy Optimization Algorithms / John Schulman, Filip Wolski, Prafulla Dhariwal et al. // *arXiv preprint arXiv:1707.06347*. — 2017.