DreamerRL: Um Framework de RL para o Desenvolvimento Autônomo em Robótica Humanoide

Alana de Santana Correia[1], **Paula Dornhofer Paro Costa**[2], **Esther Luna Colombini**[1]

[1]Instituto de Computação – Universidade Estadual de Campinas (Unicamp)
Caixa Postal 6176 – 13083-970 – Campinas – SP – Brazil

[2]Faculdade de Engenharia Elétrica e de Computação
Universidade Estadual de Campinas (Unicamp) – Campinas, SP – Brazil

{alana.correia, esther}@ic.unicamp.br, paulad@unicamp.br

Abstract. *The acquisition of motor and cognitive skills in humanoid robots still heavily relies on task engineering and explicit instructions, limiting them to constrained and predefined scenarios. Our work, the DreamerRL framework, overcomes these limitations by enabling agents to acquire skills in a fully autonomous manner, driven solely by their intrinsic curiosity to explore and understand the environment. Through experiments conducted with the NAO humanoid robot, we demonstrate that DreamerRL advances the state of the art by enabling the spontaneous emergence of complex manipulative behaviors and fundamental cognitive abilities typically observed in children up to three years of age, without the need for extensive human engineering.*

Resumo. *A aquisição de habilidades motoras e cognitivas em robôs humanoides ainda depende fortemente de engenharia de tarefas e instruções explícitas, restringindo-os a cenários limitados e pré-definidos. Nosso trabalho, o framework DreamerRL, supera essas limitações ao permitir que o agente aprenda habilidades de forma totalmente autônoma, guiado unicamente por sua curiosidade em explorar e entender o ambiente. Em experimentos realizados com o robô humanoide NAO, demonstramos que o DreamerRL avança o estado da arte ao possibilitar a emergência espontânea de comportamentos manipulativos complexos e habilidades cognitivas essenciais, tipicamente observadas em crianças de até três anos, sem a necessidade de intensa engenharia humana.*

Este trabalho corresponde ao nível de doutorado, concluído em março de 2025. O orientador e o coorientador estão listados como coautores.

1. Introdução

A aquisição de habilidades em robôs humanoides ainda é um processo manual e supervisionado por humanos. Mesmo com o avanço de técnicas como *transfer learning* [Shah and Kumar 2021], *reinforcement learning* [Sun et al. 2025] e *curriculum learning* [Bengio et al. 2009], a aquisição de habilidades úteis e generalizáveis ainda requer considerável intervenção humana, seja na engenharia de tarefas, na coleta e seleção de dados anotados ou na modelagem de recompensas. Essas abordagens, embora eficazes em cenários limitados e pré-definidos, mostram-se insuficientes quando consideramos empregar robôs humanoides como assistentes em tarefas cotidianas, que demandam adaptação contínua, manipulação complexa e compreensão de contexto.

Enquanto os métodos atuais dependem fortemente de engenharia humana, organismos biológicos, adquirem habilidades motoras e cognitivas de forma completamente autônoma através da exploração ativa do ambiente, aprendendo habilidades antes mesmo de receberem metas explícitas. Entretanto, reproduzir esse tipo de aprendizado aberto em robôs com morfologia semelhante à humana ainda é um desafio em aberto na robótica.

Embora abordagens recentes [Colas et al. 2020] avancem no aprendizado curricular de habilidades em robôs, ainda dependem de tarefas pré-definidas, recompensas específicas ou planejadores externos, longe do aprendizado verdadeiramente autônomo. Nosso trabalho propõe uma abordagem inovadora, ao permitir que um agente humanoide aprenda habilidades complexas de forma completamente autônoma, guiado apenas pela sua curiosidade interna, algo essencial para robôs com alta complexidade corporal.

Para isso, apresentamos o DreamerRL, um *framework* de aprendizado por reforço que permite ao agente adquirir habilidades motoras e cognitivas espontaneamente, explorando e entendendo o ambiente sem supervisão externa. Inspirado no aprendizado infantil, o DreamerRL reduz drasticamente a engenharia manual, aproximando-se de um modelo natural, adaptativo e escalável de aprendizado.

A base do nosso *framework* está nas teorias de modelos de mundo [Craik 1967], que explicam como os humanos aprendem a agir de forma autônoma no ambiente. Desde a infância, guiados pela curiosidade, buscamos novas experiências que desafiem nosso entendimento atual do mundo. Essa curiosidade nos leva a explorar e testar situações inéditas, com o objetivo de aprender a prever o que vai acontecer. Ao conseguir prever os efeitos das nossas ações, construímos internamente um modelo de mundo, uma espécie de simulador mental, que nos permite antecipar consequências, imaginar cenários futuros, raciocinar e planejar. À medida ajustamos o nosso modelo de mundo para fazer predições corretas, a curiosidade se renova e nos direciona novamente para situações novas, gerando um ciclo contínuo de aprendizado.

Inspirados por esses princípios, desenvolvemos o DreamerRL, um *framework* de aprendizado por reforço em que o agente robótico aprende a prever as consequências de suas ações e busca ativamente experiências inéditas, estimulado pela sua curiosidade interna. Testamos o DreamerRL no robô humanoide NAO [Shamsuddin et al. 2011] em tarefas de manipulação de objetos, que exigem coordenação fina e adaptação constante. Os resultados mostram que o agente desenvolve espontaneamente habilidades complexas sem engenharia de recompensas ou demonstrações humanas, e consegue transferir esse conhecimento para novas tarefas com mínima adaptação supervisionada.

O nosso trabalho tem dois objetivos principais: (i) demonstrar que o DreamerRL permite a aquisição autônoma de habilidades motoras e cognitivas para a manipulação de objetos, sem a necessidade de intensa engenharia humana; e (ii) mostrar que essas habilidades podem ser posteriormente transferidas para uma tarefa específica. As principais contribuições do nosso trabalho são: (i) mostrar que o aprendizado preditivo do ambiente, aliado à curiosidade, é um mecanismo central para o desenvolvimento autônomo de robôs humanoides; (ii) um modelo computacional que simula fases do desenvolvimento infantil, promovendo aprendizado contínuo e progressivo em um robô humanoide complexo; e (iii) o desenvolvimento de um protocolo de adaptação de tarefas com mínima adaptação supervisionada. Além dessas contribuições, este

trabalho também forneceu subsídios conceituais e experimentais para artigos relevantes que desenvolvemos nas áreas da robótica [Cleveston et al. 2025] e da inteligência artificial [de Santana Correia and Colombini 2022] [Santana and Colombini 2021] [Santana et al. 2025].

2. Teorias de Modelos de Mundo

Nosso *framework* é baseado em teorias neurocientíficas que propõem que o cérebro humano constrói um modelo interno do mundo [Craik 1967] [Hawkins et al. 2017]. Esse modelo é baseado em experiências passadas e funciona como um simulador do ambiente externo. Por meio dele, realizamos simulações mentais, prevemos consequências, planejamos ações e nos comunicamos. Para alguns teóricos, esse modelo é a chave para o desenvolvimento da inteligência humana, pois permite o nosso aprendizado autônomo e contínuo [LeCun 2022].

Há evidências de que esse modelo é hierárquico, modular e enraizado na estrutura do neocórtex, onde diferentes neurônios atuam tanto na antecipação de eventos quanto na detecção de discrepâncias entre o esperado e o real. Esses erros de previsão impulsionam ajustes nos pesos neurais, promovendo a atualização contínua do modelo interno, aprimorando-o progressivamente [Hawkins et al. 2017].

Principalmente na infância, a curiosidade e o nosso corpo atuam como forças motrizes na construção do modelo de mundo. Ao impulsionar a busca por experiências que desafiem ou complementem o modelo atual, a curiosidade promove novas entradas sensoriais e favorece o aprendizado contínuo e autônomo. Nesse processo, o corpo exerce um papel central, pois é por meio dos sentidos e das ações motoras que geramos e testamos hipóteses sobre o ambiente [Gottlieb et al. 2013].

3. DreamerRL Framework

O DreamerRL segue os principais princípios das teorias de modelos de mundo [Craik 1967] ao permitir que os agentes adquiram por conta própria, habilidades motoras e cognitivas enquanto tentam prever como o ambiente funciona, sem depender de recompensas externas para aprender cada habilidade. Nosso *framework* possui seis módulos principais: percepção, modelo de mundo, motivação intrínseca, motivação extrínseca, ator e crítico, como ilustrado na Figura 1.

O módulo de **percepção** recebe e processa os dados sensoriais do estado atual do ambiente e os transforma em representações latentes que alimentam o ambiente interno, motivações intrínseca e extrínseca, e o crítico. Esse módulo suporta múltiplas modalidades sensoriais, cada uma com seu próprio codificador, unificando-as em uma única representação latente. Os codificadores podem ser redes neurais artificiais ou criados manualmente para cada modalidade.

O **ambiente interno** é o núcleo cognitivo do agente, ele abriga o modelo de mundo e a motivação intrínseca, responsáveis pelo desenvolvimento autônomo de habilidades. O **modelo de mundo** prevê estados futuros do ambiente a partir da percepção atual do agente. Ele é composto por duas redes neurais, a **StateNet**, que codifica o estado atual do ambiente em um vetor latente $\mathbf{h_t}$, e a **StatePredictor**, que prevê o funcionamento do ambiente no próximo instante de tempo a partir do estado atual do ambiente e da ação

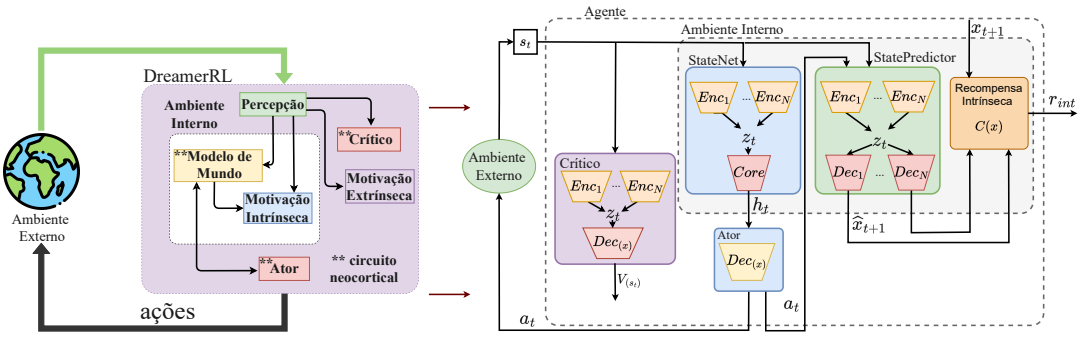


Figura 1. À esquerda, diagrama conceitual do *framework* DreamerRL com os principais módulos interconectados. À direita, implementação prática com redes neurais alocadas funcionalmente em cada módulo.

$\mathbf{a_t}$ escolhida pelo agente. A StatePredictor faz essas previsões usando decodificadores em sua saída, que prevêem o que cada modalidade sensorial do agente irá perceber no próximo instante de tempo.

A **motivação intrínseca** gera recompensas internas baseadas na curiosidade, usando o erro de previsão do modelo de mundo como métrica: quanto mais corretas são as previsões do agente sobre o ambiente, maior o incentivo para explorá-lo de formas diferentes, favorecendo a aquisição de habilidades gerais sem recompensas manuais. Por outro lado, o módulo de **motivação extrínseca** fornece recompensas específicas para tarefas pré-definidas, permitindo avaliar a transferência das habilidades que o agente adquiriu autonomamente por motivação intrínseca no cumprimento dessas tarefas.

Por fim, o **ator** seleciona ações a partir dos estados latentes do modelo de mundo, buscando maximizar recompensas intrínsecas ou extrínsecas, e retroalimenta o modelo de mundo com suas escolhas para que ele possa simular as consequências das ações do agente antes da execução no ambiente. Simultaneamente, o **crítico** estima o valor de cada estado, orientando a política do ator. Em nossas implementações o ator e o crítico são redes neurais artificiais.

O nosso framework pode operar em dois modos distintos: modo-1 (intrínseco) e modo-2 (extrínseco). O **modo-1** é utilizado exclusivamente para aquisição de habilidades de forma autônoma, enquanto o **modo-2** é utilizado para avaliar a transferência de habilidades aprendidas de forma autônoma para outras tarefas específicas.

No **modo-1**, apenas a motivação intrínseca está ativa. O agente aprende autonomamente, guiado pela curiosidade sobre as predições do ambiente. A cada passo de tempo t, o estado $\mathbf{s_t}$ do ambiente é processado pelo módulo de percepção, que o codifica para a StateNet gerar o vetor latente $\mathbf{h_t}$. Com base em $\mathbf{h_t}$, o ator escolhe a ação $\mathbf{a_t}$, que, junto ao estado atual, é usada pela StatePredictor para prever as próximas observações sensoriais $\hat{\mathbf{x}}_{\mathbf{t+1}}$ que o agente espera ter do ambiente. Após essa predição, a ação é executada no ambiente, e a nova observação $\mathbf{x_{t+1}}$ é coletada para calcular a recompensa intrínseca de curiosidade do agente:

$$r_{\text{int}} = \sum_{i=1}^{N} w_i \mathcal{L}_i\left(\hat{\mathbf{x}}_{\mathbf{i,t+1}}, \mathbf{x}_{i,t+1}\right),$$

onde N é o número de modalidades sensoriais, w_i são pesos normalizados, $\mathcal{L}_i$ mede o erro entre a predição $\hat{\mathbf{x}}_{\mathbf{i,t+1}}$ e a observação real $\mathbf{x}_{i,t+1}$ da modalidade i. Recompensas baixas indicam que o agente domina aquele estado do ambiente, motivando-o a se desenvolver para explorar novos estados através de um ciclo contínuo de aprendizado e evolução motora e cognitiva.

No **modo-2**, o agente é guiado exclusivamente por recompensas extrínsecas específicas da tarefa. A StatePredictor é desativada e o foco é realizar o *fine-tuning* do agente previamente treinado no modo-1, adaptando as habilidades autônomas à tarefa específica. Esse modo permite avaliar a reutilização e o aprimoramento das habilidades adquiridas autonomamente para acelerar o aprendizado em tarefas pré-definidas.

O treinamento do nosso *framework* é principalmente por reforço: a percepção, o ator, o crítico e a StateNet são treinados com o algoritmo PPO [Schulman et al. 2017]. Paralelamente, a StatePredictor é treinada de forma supervisionada, usando os dados reais do ambiente coletados durante o aprendizado por reforço como rótulos para a descida de gradiente. Esse treinamento supervisionado ocorre junto com a otimização dos outros módulos.

4. Materiais e Métodos

Em todos os experimentos, utilizamos o robô humanoide NAO [Shamsuddin et al. 2011] simulado no CoppeliaSim [Rohmer et al. 2013], escolhido por sua morfologia semelhante à humana, rica base sensorial e capacidade de executar movimentos complexos. Nos cenários experimentais, o NAO está sentado diante de uma mesa com alguns objetos, com as juntas dos membros inferiores e tronco travadas e 28 juntas de braços, mãos e cabeça disponíveis para controle (Figura 2). O controle foi realizado com a biblioteca PyRep e código em Python (PyTorch 1.3.1, Numpy, Scipy, Matplotlib), executado em um computador com CPU Intel Core i7-13700KF, GPU Nvidia RTX 4090 (24 GB, Cuda 12.2), 32 GB de RAM, HD de 2 TB e Ubuntu 22.04.

A metodologia foi organizada em quatro etapas principais: (i) primeiro, configuramos o robô, os ambientes e criamos as cenas; (ii) em seguida, realizamos três experimentos (Experimentos 1–3) com o *framework* no modo-1, no qual apenas a curiosidade guiou o aprendizado, permitindo a aquisição de habilidades de forma autônoma. Nesses experimentos, os principais objetivos foram validar o *framework* e analisar como o aumento da complexidade preditiva influencia a imersão e o surgimento de comportamentos mais sofisticados; (iii) posteriormente, realizamos o experimento 4, onde utilizamos o modo-2, com o objetivo de adaptar o agente a uma tarefa pré-definida (*capturingBall*) através do *fine-tuning* do aprendizado intrínseco; (iv) e por fim, comparamos o desempenho do agente adaptado com *fine-tuning* com o de outro treinado do zero para a mesma tarefa, a fim de medir os ganhos de eficiência e generalização resultantes do pré-treinamento intrínseco.

Em todos os experimentos, utilizamos os seguintes hiperparâmetros: doze *rollouts*, comprimento de trajetória de trinta e dois passos, taxa de aprendizado de 1×10^{-4} em todos os módulos treináveis, vinte épocas de treinamento por atualização da política, fator de desconto igual a $0,99$, parâmetro de corte do PPO configurado para $0,2$ e desvio padrão da política igual a $0,5$.

O desempenho do agente foi avaliado por métricas quantitativas e qualitativas. No

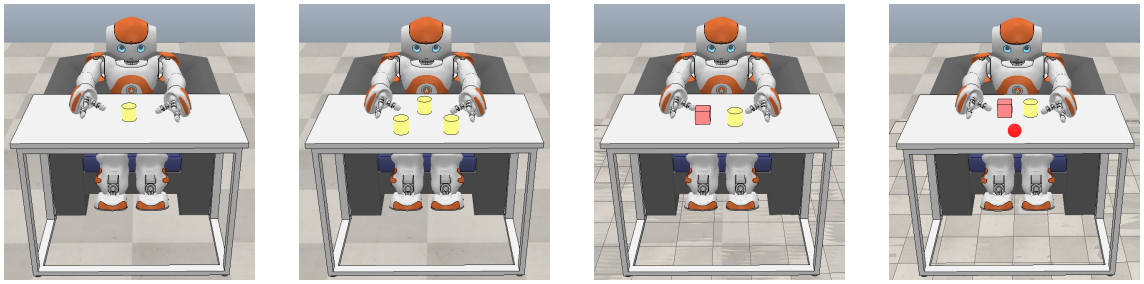


Figura 2. Amostras de cenas do simulador com o robô humanoide NAO sentado em frente a uma mesa com diferentes objetos, como cubos, esferas e cilindros.

modo-1, utilizamos a intensidade média de interação dos dedos do agente com os objetos e o erro residual médio de predição em cada modalidade sensorial (MSE, SAD, SSIM e L1) [LeCun et al. 2015, Wang et al. 2017], além da análise visual das predições e dos comportamentos emergentes. No modo-2, a comparação entre o agente com *fine-tuning* e aquele treinado do zero baseou-se na recompensa média por episódio obtida na tarefa e no desvio padrão das recompensas nos ambientes de teste.

A transferência do aprendizado intrínseco para a tarefa extrínseca foi realizada removendo a rede neural *StatePredictor*, mantendo fixos os pesos das demais arquiteturas previamente treinadas e ajustando apenas as camadas finais da política por meio de *fine-tuning*. O agente, inicialmente motivado apenas por curiosidade, foi então exposto à tarefa *CapturingBall*, recebendo um sinal de recompensa extrínseca para guiar a adaptação e avaliar a reutilização do conhecimento adquirido de forma autônoma.

5. Experimentos e Resultados

Nesta seção, descrevemos os experimentos realizados e os resultados obtidos. Todos os experimentos seguem a metodologia previamente detalhada na Seção 4.

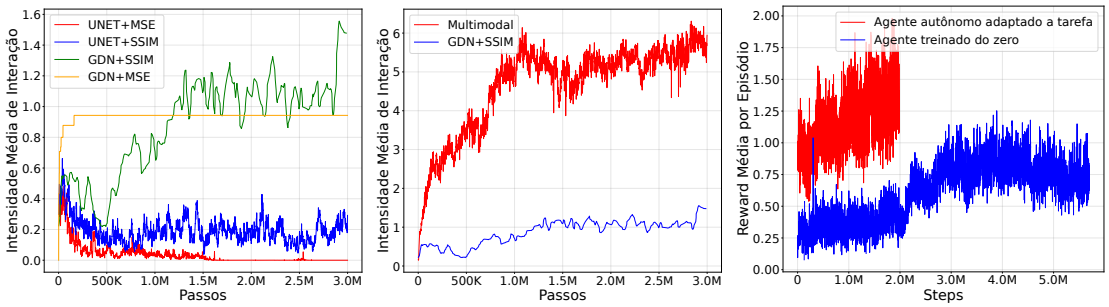
5.1. Experimento 1: Descoberta de Habilidades de Manipulação

Neste experimento, avaliamos nossa **primeira hipótese**, que propõe que um agente robótico complexo pode desenvolver habilidades de manipulação de objetos apenas por curiosidade em entender o ambiente, sem recompensas extrínsecas. Para isso, testamos quatro variações do agente, combinando duas arquiteturas (UNET e GDN) com duas funções de recompensa intrínseca (SSIM e MSE): UNET+SSIM, UNET+MSE, GDN+SSIM e GDN+MSE. Nesse experimento, treinamos todos os agentes por 3 milhões de passos e, durante todo o treinamento, o único objetivo dos agentes era prever o próximo *frame* $\hat{\mathbf{f}}$ do ambiente, com a previsão das demais modalidades sensoriais desativadas para simplificar o experimento.

O espaço de observação $\mathbf{x_t}$ é multimodal, composto pelo vetor de propriocepção $\mathbf{p}$ (posições das juntas em radianos) e duas imagens RGB em terceira pessoa, uma da vista superior $\mathbf{i^t}$ e uma da vista frontal $\mathbf{i^f}$, que, empilhadas verticalmente, formam o *frame* $\hat{\mathbf{f}}$ a ser previsto. Assim, $\mathbf{x_t} = [\mathbf{p}, \mathbf{i^t}, \mathbf{i^f}]$. O estado $\mathbf{s_t}$ empilha as três últimas observações $\mathbf{s_t} = [\mathbf{x_t}, \mathbf{x_{t-1}}, \mathbf{x_{t-2}}]$ para evitar *aliasing* perceptivo e incluir contexto temporal. A ação $\mathbf{a_t}$ contém os 26 valores angulares das juntas dos braços e mãos, com a cabeça fixa para simplificar o controle. A recompensa intrínseca r_{int} é calculada por

Tabela 1. Resumo dos principais aspectos do ambiente aprendidos e dos comportamentos autônomos adquiridos por cada agente no Experimento 1.

Agente	Objetos			Dinâmica do Corpo		Forma do Corpo				Comportamentos					
	Cores	Forma	Dinâmica	Braços	Mãos	Cabeça	Tronco	Braços	Mãos	Toque	Segurar	Levantar	Arrastar	Atirar	Put
UNET+MSE	✓	✓	✓	○	○	✓	✓	✓		✓					
UNET+SSIM	✓	✓	✓	✓	○	✓	✓	○		✓	✓		✓		
GDN+SSIM	✓	✓	○	✓	○	✓	✓	✓	○	✓	✓	✓	✓	✓	
GDN+MSE	✓	✓	○	✓	○	✓	✓	○	○	✓	✓	✓	✓	✓	



(a) Experimento 1. **(b) Experimento 2.** **(c) Experimento 4.**

Figura 3. Curva de interação dos agentes sobre os objetos durante o treinamento.

$r_{\text{int}} = 1 - \frac{1}{1+\text{MSE}(\hat{\mathbf{f}}_{\mathbf{t+1}}, \mathbf{f}_{\mathbf{t+1}})}$ ou $r_{\text{int}} = 1 - \frac{1}{1+\text{SSIM}(\hat{\mathbf{f}}_{\mathbf{t+1}}, \mathbf{f}_{\mathbf{t+1}})}$, onde o MSE mede o erro quadrático médio e SSIM a similaridade estrutural entre o *frame* previsto $\hat{\mathbf{f}}_{\mathbf{t+1}}$ e o *frame* real $\mathbf{f}_{\mathbf{t+1}}$.

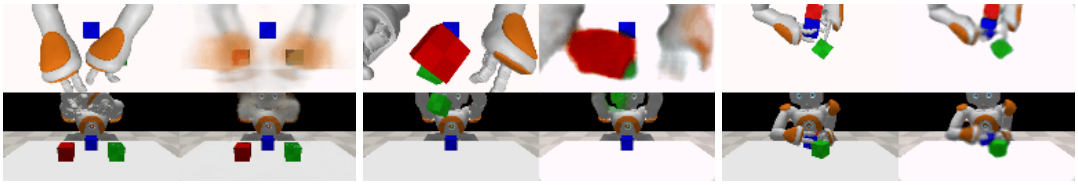
Os resultados obtidos evidenciam uma evolução autônoma clara, especialmente no agente GDN+SSIM, conforme ilustrado no gráfico verde da Figura 3 (a). Inicialmente, as interações com os objetos eram esporádicas e restritas a toques simples. Com o avanço no aprendizado da previsão do ambiente, o agente passou a focar cada vez mais nos objetos, aumentando tanto a frequência quanto a complexidade das interações. Progressivamente, desenvolveu habilidades mais sofisticadas, como segurar, levantar, arrastar e até arremessar objetos com as duas mãos (Tabela 1 e Figura 4 (b)).

Esse agente destacou-se também pela qualidade superior nas previsões visuais e corporais (Figura 4 (b)), apresentando maior estabilidade. Em contraste, os demais agentes, com menor precisão preditiva em regiões complexas como braços e mãos, ficaram presos na curiosidade sobre esses elementos, realizando movimentos repetitivos de levantar os braços para cima para tentar entender sua dinâmica, o que limitou sua evolução comportamental (Figura 4 (a)).

Esses resultados confirmam a nossa hipótese de que a curiosidade intrínseca em aprender sobre o mundo leva o agente a aprender habilidades autonomamente, e que seu progresso depende da capacidade preditiva de prever o ambiente. Assim, agentes mais avançados que o agente GDN+SSIM podem evoluir para comportamentos ainda mais complexos.

5.2. Experimento 2: Desenvolvimento da Destreza e Precisão para manipular

Neste experimento, aumentamos a complexidade do agente, para verificar se isso favorece a aquisição autônoma de habilidades mais avançadas que as observadas no experimento 1. A nossa **hipótese** é que incluir as modalidades propriocepção e tato na função de recom-



(a) Agente UNET+MSE. (b) Agente GDN+SSIM. (c) Agente Multimodal.

Figura 4. Previsões dos agentes nos treinamentos intrínsecos. Em (a) e (b) temos agentes do experimento 1, e (c) do experimento 2.

pensa e nas predições do agente intensifica sua curiosidade e aprimora o desenvolvimento motor. Para isso, adicionamos um vetor binário de colisões (16 posições) à observação do agente, e além de prever o próximo *frame visual*, agora o agente também deverá prever o próximo vetor de colisões e de propriocepção, aumentando a complexidade sensorial das previsões, pois agora a próxima observação completa $\mathbf{x_{t+1}}$ deverá ser prevista pelo agente.

Alteramos a função de recompensa curiosa r_t para uma soma ponderada de medidas específicas para cada modalidade. Continuamos usando a SSIM para a visão, MSE para propriocepção e SAD para colisões. Na função de recompensa, a colisão recebeu peso maior (0,5), por ser a modalidade mais difícil, devido à sua natureza binária e às frequentes oclusões das mãos que dificultam sua previsão via visão.

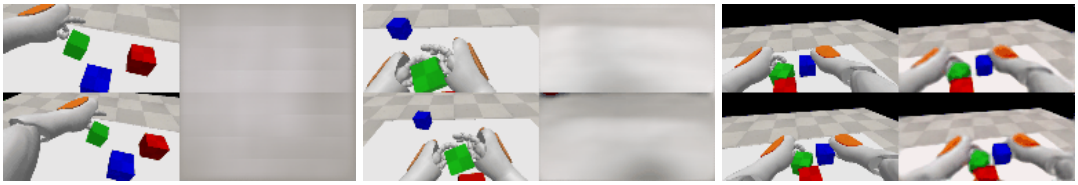
Os resultados demonstraram que o agente com previsões multimodais desenvolveu movimentos mais precisos e coordenados, sendo capaz de agarrar objetos e manipulá-los por vários passos consecutivos. A Figura 4 (c), mostra previsões visuais mais precisas da forma e do movimento dos dedos durante essas interações, algo desafiador para os agentes do experimento 1. Além disso, observamos um aumento significativo na intensidade e na diversidade das interações (Figura 3 (b) gráfico vermelho), com o agente explorando os cubos individualmente, realizando movimentos como esfregar e girar os objetos, e também em grupo, provocando múltiplas colisões e deslocamentos pela mesa. Essas ações demonstram claramente uma maior destreza e precisão nos dedos para manipular corretamente os objetos (Figura 4 (c)).

Nesse experimento, conseguimos demonstrar mais um passo do agente em direção a um desenvolvimento similar ao desenvolvimento infantil, em que a integração de múltiplas modalidades sensoriais contribui para o refinamento motor e para o surgimento de comportamentos exploratórios mais intencionais e coordenados.

5.3. Experimento 3: Desenvolvimento de Atenção Visual e Alinhamento Sensorial

Neste experimento, aumentamos a complexidade sensório-motora do agente. No vetor de observação $\mathbf{x_t}$, substituímos a visão em terceira pessoa pela visão estereoscópica em primeira pessoa, com câmeras nos olhos do robô. Também liberamos os dois atuadores do pescoço, permitindo movimentos autônomos da cabeça e ampliando o espaço de ações para controlar 28 juntas. Diferentemente dos experimentos anteriores, o robô agora tem visão limitada do ambiente, precisando girar e inclinar a cabeça para explorá-lo (Figura 5 (a)). Ele também não tem visão completa dos braços, ombros e torso simultaneamente.

Essa configuração torna as predições sensoriais especialmente desafiadoras, pois além de integrar visão, tato e propriocepção, o robô precisa decidir para onde olhar, prever



(a) Predições no passo 1×10^3 (b) Predições no passo 1×10^5 (c) Predições no passo 1×10^6

Figura 5. Visão e predições do robô imagens estereoscópicas em primeira pessoa e pescoço livre no experimento 3.

o que verá e inferir eventos fora do seu campo visual, como colisões não vistas ou efeitos de ações não observadas. Além disso, no início do treinamento, movimentos rápidos e descoordenados da cabeça causam desalinhamento sensorial e grandes variações entre *frames* consecutivos.

Nossa **hipótese** é que, ao enfrentar maior dificuldade para visualizar e prever o ambiente, o agente será mais estimulado pela recompensa curiosa, explorando o ambiente de forma mais profunda e ativa, e como consequência desenvolverá habilidades motoras e cognitivas mais sofisticadas que as observadas nos experimentos anteriores.

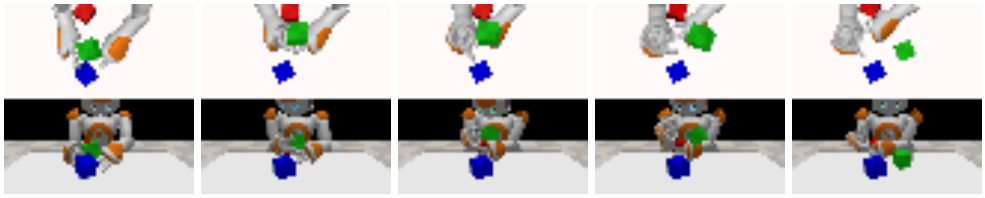
Os resultados dessa configuração desafiadora foram muito promissores, representando um avanço importante na construção de robôs humanoides autônomos inspirados no desenvolvimento humano. No início do treinamento, a liberação do pescoço gerou movimentos caóticos da cabeça e dos braços. No entanto, após apenas 80 mil passos, emergiu espontaneamente um controle coordenado da cabeça, permitindo a **atenção sustentada** sobre os objetos, como mostrado na Figura 6 (a).

Essa habilidade, típica do desenvolvimento infantil, é amplamente reconhecida como essencial para o aprendizado motor e cognitivo, mas raramente surge de forma natural em sistemas artificiais. Na literatura atual, comportamentos desse tipo geralmente exigem supervisão externa, modelagem explícita ou recompensas específicas para serem induzidos. O fato de termos obtido essa capacidade apenas com motivação intrínseca demonstra a robustez e o potencial do nosso *framework* para promover o surgimento de habilidades cognitivas complexas de maneira autônoma.

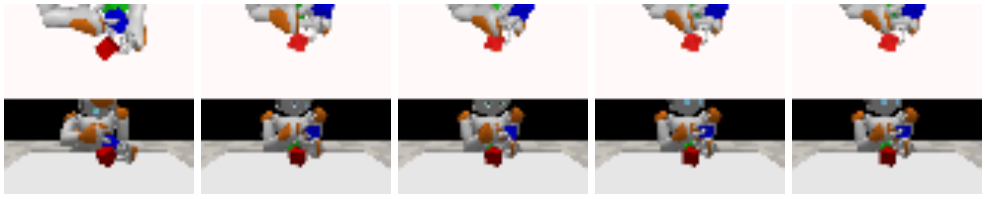
A atenção emergiu como resposta à dificuldade do agente em prever o futuro sensorial diante do seu comportamento aleatório inicial, levando-o a reduzir a complexidade perceptiva de forma autônoma. Após desenvolver a atenção, o agente passou a apresentar alinhamento sensorial, mantendo simultaneamente olhar e mãos direcionados aos mesmos pontos do ambiente (Figura 6 (a) e (c)). Esse comportamento também é raro em sistemas sem regras explícitas e comprova teorias da cognição corporificada que afirmam que humanos aprendem o alinhamento sensorial para aliviar a carga cognitiva do cérebro.

A combinação entre atenção sustentada nos objetos e o alinhamento sensorial permitiram movimentos de manipulação mais naturais, como levantar os cubos com as duas mãos enquanto mantém o foco visual, a capacidade de rastrear objetos em movimento, explorar o próprio corpo e interagir deliberadamente com o ambiente demonstram o avanço do desenvolvimento autônomo do nosso agente, que agora atua de forma mais próxima a uma criança de até três anos de idade (Figura 6). Esses resultados validam a nossa

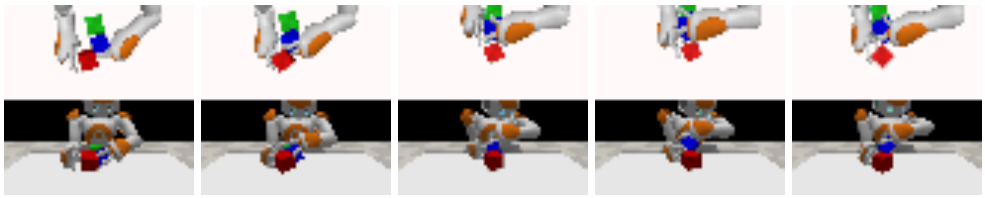
(a) Trajetória A: Robô levantando o cubo da mesa com as duas mãos e olhar fixo sobre o objeto.



(b) Trajetória B: Robô rastreando o movimento do cubo verde.



(c) Trajetória C: Robô levantando o cubo com máxima atenção visual sobre o cubo azul.



(d) Trajetória D: Robô inspecionando o buraco da mesa.

Figura 6. Amostras das habilidades adquiridas autonomamente pelo robô no experimento 3, evidenciando o desenvolvimento de atenção visual focada nos objetos e o alinhamento sensorial entre múltiplos sentidos. Esses avanços possibilitaram um comportamento de manipulação mais natural, semelhante ao observado em crianças de até três anos. As imagens estão com qualidade reduzida devido a limitações de memória durante o treinamento.

hipótese de que predições mais desafiadoras contribuem significativamente para avanços no desenvolvimento autônomo do agente, e demonstram que é possível replicar aspectos centrais do desenvolvimento infantil através do nosso *framework*.

5.4. Experimento 4: Transferência de Habilidades para uma nova Tarefa

Neste experimento, testamos a **hipótese** de que habilidades motoras e cognitivas adquiridas de forma autônoma durante o treinamento intrínseco podem ser transferidas para uma tarefa específica. A tarefa para testar essa hipótese foi a *capturingBall*, que consiste em capturar uma bola em movimento sobre a mesa, com posição inicial e velocidade variando aleatoriamente a cada episódio. Essa tarefa exige atenção, coordenação motora e predição de movimento.

Para validar nossa hipótese, colocamos o *framework* no modo 2, realizamos o *fine-tuning* do agente autônomo da Seção 6.3 para a tarefa e comparamos seu desempenho

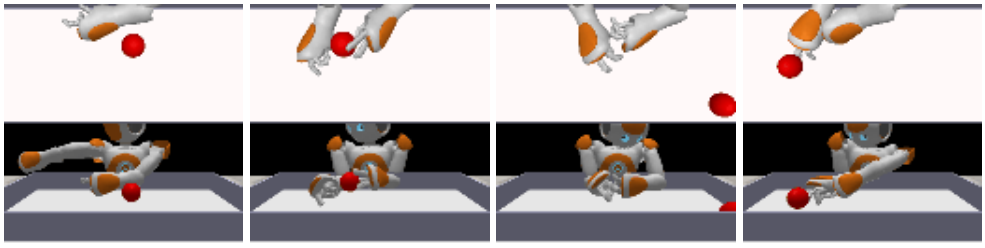


Figura 7. Comparação entre comportamentos aprendidos pelo agente treinado do zero e pelo agente autônomo adaptado a tarefa. em (a) temos o comportamento do agente treinado do zero, com dificuldade para controlar o pescoço e nas demais alternativas o comportamento do agente autônomo adaptado a tarefa.

com o de um agente idêntico treinado do zero apenas com a recompensa da tarefa atual. A recompensa da tarefa foi definida como $r_{\text{ext}} = \sum_{i=1}^{M} c_i$, onde $c_i = 1$ quando há colisão entre a falange i e a bola e 0 caso contrário, com M sendo o número total de falanges. O desempenho foi avaliado pela curva de recompensa média por episódio acumulada em treinamento e por 16 testes independentes com configurações já vistas, analisando média e desvio padrão das recompensas obtidas.

Os resultados mostram que o agente autônomo que foi adaptado para a tarefa conseguiu se adaptar rapidamente e atingiu picos maiores de recompensa (Figura 3 (c)), validando a nossa hipótese. Qualitativamente, esse agente também foi muito superior, exibindo estratégias complexas como antecipar trajetórias da bola, posicionar a mão por vários passos e empurrar a bola contra a parede, enquanto o agente treinado do zero na tarefa não desenvolveu habilidades básicas, como manter o olhar sobre a mesa, e portanto não conseguiu executar bem a tarefa (Figura 7). Nos testes, o agente adaptado à tarefa também se destacou, com recompensa de $1,35 \pm 1,26$ e falha em tocar a bola em 12% dos casos. Já o agente treinado do zero obteve $0,84 \pm 0,82$, falhando em 21% dos testes, evidenciando a superioridade da nossa abordagem.

6. Conclusão e Trabalhos Futuros

Este trabalho investigou como o aprendizado preditivo com motivação intrínseca baseada na curiosidade pode levar robôs humanoides a desenvolver, de forma autônoma, habilidades motoras e cognitivas complexas sem a necessidade de intensa engenharia humana. Ao prever o funcionamento do mundo a partir de múltiplas modalidades sensoriais, o robô humanoide NAO construiu representações internas ricas e comportamentos com trajetória semelhante ao desenvolvimento infantil. Mostramos também que essas habilidades podem ser transferidas para novas tarefas por meio de um *fine-tuning* simples e com mínima intervenção humana. Como trabalhos futuros, o DreamerRL possibilita investigar o uso de imaginação ativa, incorporar linguagem e elementos simbólicos, realizar experimentos domésticos mais complexos, testar adaptação a diferentes tarefas e explorar novas formas de recompensa intrínseca para avançar estudos em cognição artificial no nosso grupo.

Referências

Bengio, Y., Louradour, J., Collobert, R., and Weston, J. (2009). Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning -*

ICML '09, pages 1–8, Montreal, Quebec, Canada. ACM Press.

Cleveston, I., Santana, A. C., Costa, P. D., Gudwin, R. R., Simões, A. S., and Colombini, E. L. (2025). Instructrobot: A model-free framework for mapping natural language instructions into robot motion. *arXiv preprint arXiv:2502.12861*.

Colas, C., Karch, T., Lair, N., Dussoux, J.-M., Moulin-Frier, C., Dominey, P., and Oudeyer, P.-Y. (2020). Language as a cognitive tool to imagine goals in curiosity driven exploration. *Advances in Neural Information Processing Systems*, 33:3761–3774.

Craik, K. J. W. (1967). *The nature of explanation*, volume 445. CUP Archive.

de Santana Correia, A. and Colombini, E. L. (2022). Attention, please! a survey of neural attention models in deep learning. *Artificial Intelligence Review*, 55(8):6037–6124.

Gottlieb, J., Oudeyer, P.-Y., Lopes, M., and Baranes, A. (2013). Information-seeking, curiosity, and attention: computational and neural mechanisms. *Trends in cognitive sciences*, 17(11):585–593.

Hawkins, J., Ahmad, S., and Cui, Y. (2017). A theory of how columns in the neocortex enable learning the structure of the world. *Frontiers in neural circuits*, 11:295079.

LeCun, Y. (2022). A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review*, 62(1):1–62.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *nature*, 521(7553):436–444.

Rohmer, E., Singh, S. P., and Freese, M. (2013). V-rep: A versatile and scalable robot simulation framework. In *2013 IEEE/RSJ international conference on intelligent robots and systems*, pages 1321–1326. IEEE.

Santana, A. and Colombini, E. (2021). Neural attention models in deep learning: Survey and taxonomy. *arXiv preprint arXiv:2112.05909*.

Santana, A., Costa, P. P., and Colombini, E. L. (2025). Learning to explore with predictive world model via self-supervised learning. *arXiv preprint arXiv:2502.13200*.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. (2017). Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Shah, R. and Kumar, V. (2021). Rrl: Resnet as representation for reinforcement learning. *arXiv preprint arXiv:2107.03380*.

Shamsuddin, S., Ismail, L. I., Yussof, H., Zahari, N. I., Bahari, S., Hashim, H., and Jaffar, A. (2011). Humanoid robot nao: Review of control and motion exploration. In *2011 IEEE international conference on Control System, Computing and Engineering*, pages 511–516. IEEE.

Sun, Z., Pang, B., Yuan, X., Xu, X., Song, Y., Song, R., and Li, Y. (2025). Hierarchical reinforcement learning with curriculum demonstrations and goal-guided policies for sequential robotic manipulation. *Engineering Applications of Artificial Intelligence*, 153:110866.

Wang, Z., Bovik, A. C., and Sheikh, H. R. (2017). Structural similarity based image quality assessment. In *Digital Video image quality and perceptual coding*, pages 225–242. CRC Press.