NutriRL: A Benchmark for Nutritional Regulation under Delayed State Transitions

Aniket Khan, Charitha Palika, V. Srinivasa Chakravarthy

Keywords: benchmark,gymnasium environment,nutritional regulation,reinforcement learning

Summary

Designing adaptive diet and weight-control strategies requires coordinating multiple macronutrients under delayed physiological feedback. NutriRL is a Gymnasium-compatible benchmark environment developed to study reinforcement learning under delayed state transitions in this dietary regulation setting. The task models nutrition as a multi-dimensional control problem in which an agent sequentially selects foods to maintain hidden target levels of carbohydrates, fats, and proteins. Each action produces delayed and gradual changes to an internal nutritional state, and these effects overlap and accumulate over time, so rewards depend on past decisions across extended horizons. Unlike common RL benchmarks that reduce nutrition to a single calorie objective or introduce delay only at the reward level, NutriRL embeds heterogeneous nutrient delays directly into the state dynamics, making the problem partially observable and structurally delayed. The environment is minimal and computationally efficient, enabling systematic study of temporal credit assignment under multi-channel delayed dynamics. Empirically, the benchmark reveals distinct sensitivity patterns across representative value-based and policy-gradient methods as delay severity increases, highlighting the structural impact of delayed state evolution, and we further evaluate a GRU-based recurrent policy to examine whether temporal memory mitigates delay-induced credit assignment challenges. NutriRL thus provides a controlled testbed for adaptive, multi-macronutrient diet control under delayed physiological effects.

Contribution(s)

1. We introduce a Gymnasium-compatible benchmark that models multi-macronutrient dietary regulation as a partially observable control problem with delayed and irreversible state transitions embedded directly in the dynamics.
   Context: Unlike standard fully observed RL benchmarks in games and control (Mnih et al., 2015; Todorov et al., 2012) and prior reward- or observation-delay formulations (Liotet, 2023), this environment isolates structural transition delay within a minimal benchmarking setting.
2. We conduct a controlled delay-regime study that varies only the mean delay and the temporal spread of nutrient-specific absorption delays, while holding the reward structure, horizon, architectures, and training budgets fixed.
   Context: This design enables attribution of performance differences specifically to delayed transition dynamics (Dulac-Arnold et al., 2019).
3. We provide baseline evaluations across representative value-based and policy-gradient methods, to characterize algorithmic sensitivity to delayed transition dynamics, and evaluate a GRU-based recurrent policy to test whether temporal memory mitigates delay-induced credit assignment challenges.
   Context: The evaluated methods are widely used in deep RL and conclusions are limited to the tested configurations.
4. We propose a Gaussian-kernel absorption model that approximates temporally distributed digestion dynamics in a computationally efficient manner.
   Context: This phenomenological approximation avoids the complexity of full PBPK-style differential equation models (Peters, 2021) while retaining delayed and cumulative effects.

NutriRL: A Benchmark for Nutritional Regulation under Delayed State Transitions

Aniket Khan[1, †], **Charitha Palika**[2, †], **V. Srinivasa Chakravarthy**[3]
aniketkhan2003@gmail.com, charithapalika@gmail.com,
schakra@ee.iitm.ac.in

[1]Alumnus, Department of Mechanical Engineering, Indian Institute of Technology Madras
[2]PhD scholar, Department of Biotechnology, Indian Institute of Technology Madras
[3]Professor (Corresponding Author), Department of Biotechnology, Indian Institute of Technology Madras

[†]Equal contribution

Abstract

We introduce NutriRL, a Gymnasium-compatible benchmark for homeostatic reinforcement learning under delayed state transitions, where the objective is to maintain cumulative levels of carbohydrates, fats, and proteins near a target nutritional balance over time. The environment models dietary regulation as a partially observable control problem in which each action produces delayed and gradual changes to an underlying nutritional state. These changes overlap and accumulate irreversibly, so rewards depend on state updates induced by past decisions across extended horizons. NutriRL is motivated by biological regulation in healthcare, where food intake produces delayed internal responses and effective diet control requires coordinating multiple macronutrients rather than matching aggregate caloric intake. Unlike common reinforcement learning formulations that reduce nutrition to scalar energy balancing or introduce delay only at the reward level, NutriRL embeds heterogeneous nutrient delays within state evolution itself. The agent must continuously regulate intake toward a hidden macronutrient target while avoiding overshoot, reflecting safety-sensitive control. Across representative value-based and policy-gradient methods, the benchmark reveals distinct sensitivity patterns as delay severity increases, along with a GRU-based recurrent policy to examine whether temporal memory mitigates delay-induced credit assignment challenges, illustrating how delayed transition dynamics influence learning stability and credit propagation. By isolating structural delay within a minimal, standardized, and computationally efficient model, NutriRL provides a reproducible testbed for evaluating temporal credit assignment under multi-channel delayed transition dynamics.

1 Introduction

Reinforcement Learning (RL) has achieved transformative success in domains characterized by full observability and well-defined Markovian dynamics, such as deterministic board games including Go and Chess (Silver et al., 2016; 2017), as well as in simulated robotic control and physics-based continuous control environments (Kober et al., 2013; Todorov et al., 2012). In these environments, the current state typically contains all information necessary for optimal decision-making, and errors can often be corrected or "undone" through subsequent actions.

In contrast, physiological control and medical decision-making (Gottesman et al., 2019) violate these structural assumptions. The effects of actions unfold gradually through latent biological processes, so the observable state does not fully reflect recently executed decisions. For example, nu-

trient intake or medication initiates internal dynamics such as digestion, absorption, and metabolic conversion, which induce temporally distributed state transitions (Pires, 2017; Chudtong & De Gaetano, 2021; Santillán, 2025).

For example, after consuming a meal, the sensation of satiety does not arise immediately because nutrient absorption occurs gradually (Rakha et al., 2022; Tremblay & Bellisle, 2015; Garutti et al., 2025; Pino Santana et al., 2025). If additional food is consumed before the effects of the initial intake are reflected in the physiological state, cumulative intake may exceed the desired nutritional target and cannot be immediately corrected (Miquel-Kergoat et al., 2015).

Consequently, the system becomes partially observed and temporally coupled, and deviations cannot be immediately corrected once accumulation occurs. Physiological regulation therefore constitutes a predictive control problem in which actions must be selected based on anticipated future trajectories rather than instantaneous feedback.

Despite these properties, prior work in nutritional and health-related reinforcement learning often models such domains as standard Markov Decision Processes (Liu et al., 2022; Yu et al., 2021; Yang et al., 2017; Keramati & Gutkin, 2014; Chakravarthy et al., 2018), assuming fully observed states and near-instantaneous action effects. Many formulations reduce dietary optimization to aggregate energy balance, treating caloric intake as the primary control variable. This collapses heterogeneous nutrient dynamics into a single energy-based objective, whereas optimal diet and body weight regulation requires simultaneous control over multiple nutrient targets with distinct temporal profiles and interactions (Benton & Young, 2017; Hall et al., 2022; Theodorakis et al., 2024; Raubenheimer & Simpson, 2016). Consequently, the structural challenges introduced by delayed, multi-timescale nutrient dynamics are removed, and the learning problem is simplified to resemble fully observed, single-target control settings commonly used in reinforcement learning benchmarks (Dulac-Arnold et al., 2019).

To address this gap, we introduce **NutriRL**, a biologically inspired Gymnasium environment that embeds delayed absorption directly into the state transition dynamics. The objective is to optimize dietary decisions by regulating multiple macronutrient levels toward predefined nutritional targets under delayed and irreversible dynamics. The environment models three primary macronutrients, carbohydrates, fats, and proteins, whose effects unfold over heterogeneous time scales. Nutrient intake affects future states through delayed absorption, linking past actions to later physiological outcomes and requiring the agent to anticipate future effects.

NutriRL exposes four core challenges: **(1) latent digestion backlogs**, where absorbed nutrients are not directly observable; **(2) irreversible accumulation**, since nutrients cannot be actively removed once absorbed; **(3) heterogeneous multi-channel delays**, as different macronutrients evolve over distinct time scales; and **(4) temporal credit assignment under delayed transitions**, because actions influence the state only after a lag. Together, these properties define a delayed, multi-target control problem that differs structurally from standard reinforcement learning benchmarks.

To isolate the effects of delayed state transitions as cleanly as possible, NutriRL adopts a simplified binary action space in which the agent decides whether to consume or skip the currently presented food item. In addition, all episodes are generated using the same underlying nutrient absorption model and delay parameterization, rather than introducing inter-individual variability in metabolic or digestive processes. These design choices intentionally reduce confounding factors arising from combinatorial food selection and heterogeneous physiological responses, enabling controlled study of temporal credit assignment under delayed transition dynamics. Consequently, performance differences can be more directly attributed to the challenges introduced by delayed state evolution. Future extensions of the benchmark may incorporate richer action spaces, larger food repertoires, and personalized physiological parameters while retaining the delayed-transition framework.

Although NutriRL is not intended as a clinically validated dietary simulator, it abstracts key structural properties encountered in practical nutritional decision-making, including long-term diet planning, weight-management interventions, nutritional rehabilitation, and adaptive meal recommen-

dation systems. In these settings, physiological responses emerge gradually and are only partially observable, making delayed state evolution a central challenge for sequential decision-making.

2 Related Work

2.1 Temporal Credit Assignment under Delayed Transitions

Delayed feedback in reinforcement learning is commonly studied in the context of reward delay, where evaluative signals are withheld for several steps and must be attributed to earlier actions (Han et al., 2022). While these formulations emphasize temporal credit assignment, they typically assume that the underlying state transitions remain immediate and fully observed.

More challenging settings arise when delay is embedded directly in the transition dynamics or observation process, such that the consequences of an action are not immediately reflected in the perceived state (Karamzade et al., 2024; Liotet, 2023). In these cases, the agent must associate actions with outcomes that emerge only after intermediate transitions, increasing the horizon over which credit must be propagated. Although mechanisms such as eligibility traces (Sutton, 1988) can mitigate short-term delay, their effectiveness diminishes as the temporal gap between action and observable consequence grows.

Despite formal treatments of delayed rewards and observations, most benchmarks implement delay through fixed offsets or artificial reward withholding rather than modeling gradual, process-driven delay in state evolution. As a result, the structural impact of temporally distributed transitions on learning dynamics remains underexplored.

2.2 Reinforcement Learning in Healthcare and Nutrition

Applying RL to physiological domains requires navigating high-stakes environments with significant uncertainty. Existing literature in nutritional RL frequently simplifies the problem by optimizing for scalarized objectives, such as total caloric targets or generalized user preference scores (Tsolakidis et al., 2024; Mavrokotas et al., 2025). While these models incorporate recommendation logic, they often overlook the complex, multi-channel kinetics of macronutrient metabolism, which results in a temporally distributed nutrient absorption (Zimmerman et al., 2025; van Aken, 2024). In broader medical contexts, RL has been applied to blood glucose regulation and sepsis treatment (Emerson et al., 2023; Choudhary et al., 2024; Symeonidis et al., 2025); however, these approaches typically do not explicitly model temporal delays within the underlying state dynamics.

A related line of work considers physiological control tasks with prolonged action effects, where interventions influence future states over extended time horizons. Such settings have been explored in RL-based physiological regulation and highlight the challenges of temporal credit assignment when action consequences unfold gradually rather than instantaneously (Basu et al., 2023). While these environments exhibit delayed physiological responses, they are typically designed around specific clinical control objectives rather than as general-purpose benchmarks for systematically studying delayed state transitions.

Physiological simulators such as the UVA/Padova Type 1 Diabetes simulator have been widely used for evaluating control algorithms under delayed glucose-insulin dynamics (Man et al., 2014). However, these simulators primarily focus on single-channel glucose regulation within disease-specific physiological models. In contrast, NutriRL is designed as a lightweight benchmark environment for reinforcement learning research, emphasizing multi-channel nutrient regulation with heterogeneous delayed absorption dynamics and controlled scaling of delay parameters. The goal is not physiological fidelity comparable to clinical simulators, but rather reproducible evaluation of learning under delayed transition dynamics. However, many of these applications rely on domain-specific simulators that do not explicitly isolate the structural impact of delayed state evolution on learning behavior.

2.3 Abstraction of Physiological Dynamics

Detailed physiological modeling frequently employs Physiologically-Based Pharmacokinetic (PBPK) systems, which use systems of differential equations to track substance concentrations across organ compartments (Jones et al., 2015; Peters, 2021; Kuepfer et al., 2016). Although biologically grounded, PBPK models are computationally intensive and difficult to integrate into the large number of environment interactions required for deep reinforcement learning. At the opposite extreme, many reinforcement learning benchmarks introduce delay through fixed reward offsets without modifying the underlying transition process. Between these approaches, phenomenological models such as Gaussian convolution kernels provide a tractable alternative by approximating the temporal signature of absorption dynamics without full kinetic simulation (Chung & Tappenden, 2024; Rumessen & Gudmand-Høyer, 1986). Such models retain temporally distributed state evolution while remaining computationally efficient yet realistic with iterative learning.

2.4 Safety and Irreversibility

Safe reinforcement learning traditionally focuses on avoiding predefined constraint violations in state space (Garcıa & Fernández, 2015). Methods such as Constrained Policy Optimization (CPO) (Achiam et al., 2017) enforce safety by restricting policy updates to remain within feasible regions. A distinct challenge arises in monotonic systems where certain actions produce irreversible accumulation. In homeostatic regulation problems, excessive nutrient or medication intake cannot be actively reversed once initiated. Under delayed transitions, the risk lies not only in constraint violation but in cumulative over-commitment before consequences become observable. This highlights the link between safe RL and long-term credit assignment, where individually safe actions can accumulate over time and collectively lead to system failure (Dulac-Arnold et al., 2019).

3 Environment Design

NutriRL is formalized as a finite-horizon partially observable Markov decision process (POMDP) specified by the tuple $\mathcal{P} = (\mathcal{S}, \mathcal{A}, \mathcal{O}, \mathcal{T}, \Omega, r, \mu_0, \gamma, H)$, where $\mathcal{S}$ denotes the latent state space, $\mathcal{A}$ the action space, and $\mathcal{O}$ the observation space. The reward function is given by $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$. The transition operator $\mathcal{T} : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ defines a distribution over next states, where $\Delta(\mathcal{S})$ denotes the set of probability distributions over $\mathcal{S}$. The observation model $\Omega : \mathcal{S} \to \Delta(\mathcal{O})$ specifies a distribution over observations conditioned on the current latent state. The initial state distribution is given by $\mu_0 \in \Delta(\mathcal{S})$, the discount factor is $\gamma \in (0, 1)$, and $H \in \mathbb{N}$ denotes the finite planning horizon.

The objective is to learn a policy $\pi(a_t \mid h_t)$, where the interaction history is $h_t = (o_0, a_0, \ldots, o_t)$, that maximizes the expected discounted return $\mathbb{E}_\pi \left[\sum_{t=0}^{H-1} \gamma^t r(s_t, a_t)\right]$.

Latent state: The latent environment state at time t is defined as $s_t = (\mathbf{s}_t, \mathcal{D}_t, i_t)$, where $\mathbf{s}_t = [C_t, F_t, P_t]^T \in \mathbb{R}^3_{\geq 0}$ represents the cumulative absorbed carbohydrates, fats, and proteins, forming a continuous and monotonically increasing physiological state due to absorption-only dynamics. The component $\mathcal{D}_t$ is a finite multiset of active digestion processes initiated by prior consumption, each specifying a future nutrient absorption schedule that contributes to subsequent updates of $\mathbf{s}_t$, while $i_t \in \{1, \ldots, N\}$ denotes the currently presented food item, which is exogenous and not controlled by the agent.

Observation: At each time step, the agent observes $o_t = (\mathbf{s}_t, \mathbf{e}(i_t))$, where $\mathbf{e}(i_t) \in \mathbb{R}^{d_e}$ is a fixed embedding of the presented food item. The digestion backlog $\mathcal{D}_t$ and the target nutritional state are unobserved, resulting in partial observability, since a given observation may correspond to multiple underlying digestion histories. Consequently, observations alone do not uniquely determine future state transitions, leading to distinct future nutrient absorption trajectories.

Action space: The action space is binary, $\mathcal{A} = \{0, 1\}$, corresponding to skipping or consuming the currently presented food item. Food availability is exogenous and uncontrollable, reflecting dietary settings in which action selection is constrained by externally presented options. Food items are drawn from a fixed catalog, where each item is associated with a fixed nutrient composition and a consistent identity across episodes. In the current benchmark implementation, the food catalog consists of 10 distinct food items selected to provide diverse macronutrient compositions while maintaining a controlled experimental setting. The environment itself is fully parameterized and can accommodate larger food catalogs without modification. Food embeddings and item indices remain fixed throughout training and evaluation. At each time step, a food item is sampled uniformly at random from the catalog and presented to the agent, generating diverse nutritional trajectories while preserving the underlying food definitions. Consequently, the agent must learn a policy that generalizes across varying food presentations rather than memorizing fixed food sequences. This design isolates the effects of delayed state transitions while avoiding additional confounding factors arising from combinatorial food selection.

Environmental Dynamics: Consider a food item i consumed at time t_c (i.e., action $a_{t_c} = 1$). Each food item is associated with its own nutrient absorption profile, reflecting physiological differences in how nutrients are digested and absorbed over time. In particular, some foods exhibit rapid absorption with short delays and narrow temporal spread, while others are digested more slowly, resulting in longer delays and broader absorption windows. Moreover, different macronutrients exhibit distinct intrinsic absorption dynamics: carbohydrates are typically absorbed rapidly with relatively short delays, whereas fats and proteins are digested and absorbed more gradually, leading to broader temporal spreads. For example, simple carbohydrates such as glucose-rich foods induce sharp and early absorption peaks, while high-fat meals delay gastric emptying and extend nutrient absorption over longer periods. Similarly, protein digestion involves breakdown prior to absorption, resulting in intermediate, yet temporally distributed, uptake. Consequently, the overall absorption profile of a food item depends jointly on its macronutrient composition and the nutrient-specific digestion kinetics, motivating food- and nutrient-specific delayed absorption kernels.

To capture these food- and nutrient-specific absorption dynamics, each macronutrient $n \in \{C, F, P\}$ in food item i is assumed to contribute to the physiological state through a delayed and temporally distributed absorption process. This process is modeled using a Gaussian parameterization, where the mean delay $\tau_n^{(i)}$ represents the characteristic onset time of digestion following consumption, and the variance $(\sigma_n^{(i)})^2$ controls the temporal spread of nutrient uptake around this onset. This choice provides a smooth and unimodal approximation of nutrient digestion dynamics while remaining analytically simple and computationally tractable. Formally, the absorption schedule for nutrient n from food item i is given by

$$K_n^{(i)}(t) = \frac{\exp\left(-\frac{1}{2}\left(\frac{t-(t_c+\tau_n^{(i)})}{\sigma_n^{(i)}}\right)^2\right)}{\sum_{u=t_c+1}^{t_c+H_n^{(i)}} \exp\left(-\frac{1}{2}\left(\frac{u-(t_c+\tau_n^{(i)})}{\sigma_n^{(i)}}\right)^2\right)}, \quad \Delta s_n(t) = m_n(i)\, K_n^{(i)}(t). \tag{1}$$

where $m_n(i)$ denotes the total amount of nutrient n contained in food item i, and the normalization ensures that $\sum_{t=t_c+1}^{t_c+H_n^{(i)}} K_n^{(i)}(t) = 1$. The discrete truncation horizon is given by

$$H_n^{(i)} = \left\lceil 6\,\sigma_n^{(i)} \right\rceil.$$

The continuous absorption schedules are discretized over environment time steps, and the physiological state evolves as the superposition of all active digestion processes as follows

$$\mathbf{s}_{t+1} = \mathbf{s}_t + \sum_{j \in \mathcal{D}_t} \begin{bmatrix} \Delta s_C^{(j)}(t) \\ \Delta s_F^{(j)}(t) \\ \Delta s_P^{(j)}(t) \end{bmatrix}, \tag{2}$$

where each $j \in \mathcal{D}_t$ corresponds to a previously consumed food item whose absorption process remains active at time t.

Note that the physiological state $\mathbf{s}_t$ evolves irreversibly, as cumulative nutrient absorption is non-decreasing and any excess intake beyond the target nutritional state cannot be undone by future actions.

Reward: The agent is rewarded for making progress toward a predefined nutritional target, which is fixed by the environment but hidden from the agent. At each time step, the reward is primarily a function of the distance between the current physiological state $\mathbf{s}_t$ and the target state $\mathbf{T}$, encouraging actions that reduce this distance over time. Specifically, the agent receives reward when it moves closer to the target, Let the distance to the target be defined as $d_t = \|\mathbf{s}_t - \mathbf{T}\|_1$. Progress is measured by the reduction in this distance, $\Delta d_t = d_{t-1} - d_t$. In addition to progress-based rewards, a comfort term penalizes deviation from the target using a smooth, saturating function of the distance d_t, encouraging the agent to remain near the target once reached and smoothly diminishing as the agent moves farther away. Finally, an overshoot penalty discourages excessive intake beyond the target nutritional state, penalizing elementwise positive deviations beyond the target, defined as $\mathbf{e}_t = \max(\mathbf{0}, \mathbf{s}_t - \mathbf{T})$.

The relative importance of progress, comfort, and overshoot avoidance is controlled by the coefficients λ_p, λ_c, and λ_o, respectively.

$$d_t = \|\mathbf{s}_t - \mathbf{T}\|_1, \tag{3}$$

$$r_t = \lambda_p(d_{t-1} - d_t) - \lambda_c \tanh\left(\frac{d_t}{\tau_c}\right) - \lambda_o \tanh\left(\left\|\max(\mathbf{0}, \mathbf{s}_t - \mathbf{T})\right\|_2^2\right). \tag{4}$$

The complete step-by-step environment dynamics are provided in Algorithm 1 in the supplementary material, detailing the delayed absorption updates, digestion backlog management, and reward computation.

3.1 Benchmark Design Principles

NutriRL is constructed according to four design criteria: (1) structural isolation of delayed state transitions; (2) parametric controllability of delay mean and variance; (3) computational tractability for large-scale training; and (4) compatibility with standard Gymnasium interfaces. These principles ensure reproducibility and facilitate fair cross-algorithm comparison under delayed transition dynamics. To preserve interpretability of benchmark outcomes, the current version intentionally employs a simplified action space and a single nominal physiological configuration, allowing delayed transition dynamics to be isolated from additional sources of variability.

4 Standardized Delay Regime Protocol

To evaluate RL algorithms under delayed physiological consequences, we conduct controlled experiments in which only the statistical properties of nutrient absorption delays are varied, while all other environmental and training components remain fixed. Specifically, we analyze performance under

progressively increasing delay magnitudes and under two distinct levels of temporal dispersion. By isolating delay in this manner, performance differences can be attributed directly to delayed transition dynamics, enabling controlled cross-algorithm comparison under a standardized evaluation protocol. The detailed description of delays and temporal spreads used is described in the following section.

Across all experiments, the agent performs multi-nutrient regulation over carbohydrates, fats, and proteins. The toxicity penalty remains active ($\lambda_o > 0$), introducing irreversible cost for excessive intake and preserving asymmetric consequences of overconsumption. Reward coefficients are fixed across all configurations to maintain identical optimization objectives. The episode horizon is defined as $H = 50$, representing a 24-hour duration where each discrete time step $\Delta t \approx 30$ minutes, and the stochastic food presentation distribution is identical across experiments. Furthermore, neural network architectures and training budgets are matched across algorithms to ensure fair comparison. By holding these factors constant, we guarantee that the only experimental variation arises from delay statistics.

4.1 Delay Regimes

To systematically vary nutrient absorption dynamics, we construct seven distinct datasets corresponding to Run 0 through Run 6. Run 0 serves as a baseline configuration with no absorption delay. For the delayed settings, we introduce progressively increasing mean absorption delays for carbohydrates, proteins, and fats, organized into three regimes: Low, Medium, and Long. These regimes represent increasing temporal scales of nutrient absorption. Within each delay regime, we define two temporal spread configurations, yielding a total of six delayed runs. The temporal spread is specified relative to the corresponding mean delay. In the low spread configuration, the standard deviation is defined as 20 to 30 $\%$ of the mean delay, resulting in more concentrated absorption profiles. In the high spread configuration, the standard deviation is defined as 50 to 70 $\%$ of the mean delay, producing broader absorption windows and increased temporal overlap across nutrients. Delay values are expressed in environment steps, where one step corresponds to approximately 30 minutes. The complete parameter specification for each run is provided in Table 1.

Table 1: Experimental Design: Mean Delays and Temporal Spread Across Runs

Run	**Description**	τ_C	τ_P	τ_F	σ_C	σ_P	σ_F
Run 0	Baseline (no delay)	0	0	0	0	0	0
Run 1	Low delay, low temporal spread	[4,5]	[5,6]	[6,7]	[0.8, 1.5]	[1.0, 1.8]	[1.2, 2.1]
Run 2	Low delay, high temporal spread	[4,5]	[5,6]	[6,7]	[2.0, 3.5]	[2.5, 4.2]	[3.0, 4.9]
Run 3	Medium delay, low temporal spread	[6,8]	[8,10]	[10,12]	[1.2, 2.4]	[1.6, 3.0]	[2.0, 3.6]
Run 4	Medium delay, high temporal spread	[6,8]	[8,10]	[10,12]	[3.0, 5.6]	[4.0, 7.0]	[5.0, 8.4]
Run 5	Long delay, low temporal spread	[10,12]	[14,16]	[18,20]	[2.0, 3.6]	[2.8, 4.8]	[3.6, 6.0]
Run 6	Long delay, high temporal spread	[10,12]	[14,16]	[18,20]	[5.0, 8.4]	[7.0, 11.2]	[9.0, 14.0]

5 Benchmark Algorithms

To evaluate NutriRL and how delayed and partially irreversible dynamics influence learning, we benchmark representative RL algorithms that span key design dimensions: value-based vs. policy-based optimization and on-policy vs. off-policy learning. This selection provides representative baselines spanning key algorithmic design axes, enabling comparative evaluation under delayed absorption and irreversible state evolution.

We evaluate the following algorithms:

- **DDQN** (Van Hasselt et al., 2016): An off-policy, value-based method relying on temporal-difference bootstrapping and target networks for stable value estimation. Its dependence on bootstrapped value targets makes it sensitive to delayed reward propagation.
- **MC** (Schulman et al., 2015): An on-policy Monte Carlo policy-gradient method using generalized advantage estimation (GAE). By reducing bootstrapping bias, it provides a contrast in how delayed effects are integrated into policy updates.
- **PPO** (Schulman et al., 2017): An on-policy policy-gradient algorithm with clipped surrogate updates for stability, augmented with GAE. PPO represents a widely adopted baseline for continuous control and stochastic environments.
- **SAC** (Haarnoja et al., 2018): An off-policy actor–critic algorithm incorporating entropy regularization to encourage exploration and robustness under stochastic dynamics.

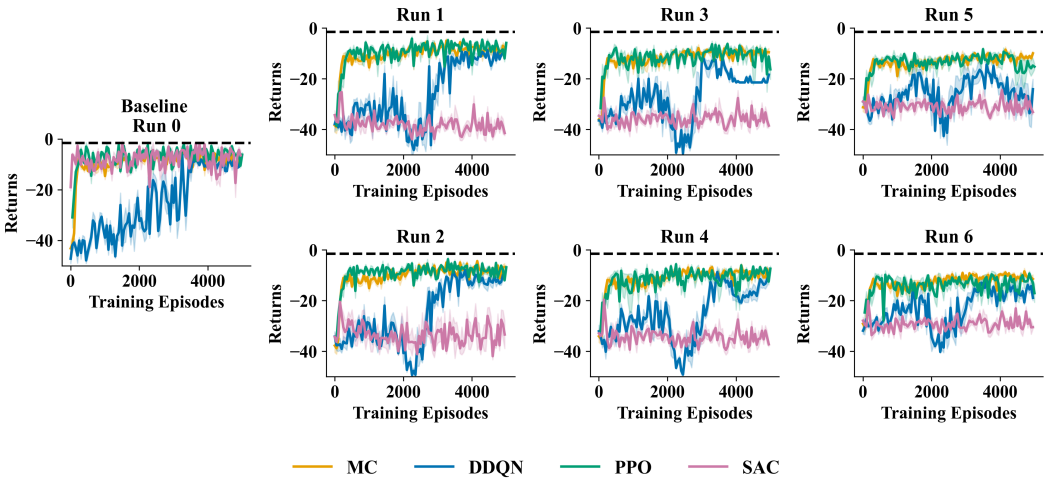

Figure 1: The figure shows the returns per episode during training across all experimental runs. MC and PPO maintain relatively stable adaptation, DDQN shows increased variability, and SAC exhibits strong sensitivity to delayed dynamics.

6 Results

All algorithms are trained under identical architectural and computational budgets to ensure that performance differences reflect algorithmic design rather than implementation asymmetries, and hyperparameters are selected to achieve the best possible results under these controlled conditions. For each baseline, hyperparameters were selected through preliminary tuning aimed at achieving stable convergence while maintaining comparable training budgets across algorithms. The final configurations were then fixed and used for all delay-regime experiments to ensure consistency of evaluation. All algorithms are trained for 5000 episodes with a fixed horizon of 50 steps per episode. Each experimental run is evaluated across five independent random seeds to account for variability in training dynamics. To evaluate the learned policies, five episodes per seed are executed, and the

results are averaged across seeds. We record four primary metrics: the average episodic return, the average number of foods consumed per episode, the final distance to the target nutritional state at the end of the episode, and the frequency of overshoot events reflecting excessive nutrient intake. Run 0 corresponds to the no-delay baseline. Runs 1–2, 3–4, and 5–6 correspond to increasing delay mean (Levels 1–3), with each pair differing only in temporal spread (low versus high variance).

6.1 Training Dynamics Under Increasing Delay

Training returns across all runs are shown in Fig. 1, while the corresponding food consumption curves are shown in Fig. 2. The dashed horizontal line in the return plots denotes the best empirical return achieved. The theoretical upper bound is $r_{\max} = 0$, achieved when the agent perfectly minimizes the distance-to-target objective, eliminating all deviation penalties. In practice, this bound is unattainable. Under a finite horizon ($H = 50$), cumulative regret cannot be zero because delayed absorption requires time to influence the state. In addition, food items correspond to fixed, discrete nutrient vectors in a three-dimensional space, and the agent can only move via additive, irreversible transitions, making exact alignment with the target generally infeasible.

In the no-delay baseline (Run 0), all algorithms learn the task effectively and converge toward a stable optimum, as the state transitions immediately reflect the consequences of each action. MC and PPO converge rapidly to near-optimal returns, with learning curves stabilizing early and exhibiting low variance across seeds. SAC also achieves competitive final returns under this regime, and its policy entropy decreases steadily during training, indicating stable convergence toward a consistent action strategy. DDQN improves more gradually and exhibits larger fluctuations, but ultimately learns a functional policy. The corresponding food consumption curves show a sharp decrease during early training for all methods, followed by stabilization at approximately 3–5 food selections per episode. This intake range is sufficient to maintain minimal homeostatic distance while avoiding overshoot. The convergence of both return and consumption to stable levels indicates that, in the absence of delayed state updates, the credit assignment problem is straightforward and all algorithms are able to learn an effective control policy.

Introducing a small delay reveals distinct policy adaptation patterns across algorithms. At smaller delays (Runs 1–2), MC and PPO remain near-optimal, although their asymptotic return decreases relative to Run 0. Food consumption increases slightly compared to baseline, indicating compensation for delayed nutrient realization, though the magnitude of this increase is modest and does not represent a significant deviation from baseline behavior. In contrast, SAC exhibits strong sensitivity to delayed state realization even at this smallest delay. Returns remain substantially below baseline and do not recover during training. Notably, entropy remains high under delayed conditions, suggesting that the policy fails to reduce exploration and does not converge to a stable control strategy. DDQN exhibits increased volatility, with several runs displaying substantial mid-training performance drops; however, despite these instabilities, the agent eventually recovers and converges to a reasonable policy, albeit with higher variance than the policy-gradient methods.

At medium delays (Runs 3–4), the control problem becomes more anticipatory relative to Runs 1–2. For both MC and PPO, the total returns achieved are lower than those obtained under small delays, indicating a further but modest degradation. DDQN behaves similarly to Runs 1–2: it continues to show fluctuations during training, yet it is still able to learn a workable policy. In contrast, SAC does not show meaningful learning under these conditions and remains at low return levels throughout training.

At large delays (Runs 5–6), the environment requires increasingly long-horizon anticipation across algorithms. Both MC and PPO show slightly higher variability during training and achieve lower final returns compared to the small and medium delay settings. Food consumption increases for MC, contributing to higher overshoot, while PPO also experiences reduced reward alongside more unstable behavior. DDQN does not recover stable performance and remains inconsistent throughout training. SAC maintains low performance across both runs without clear recovery. Across all delay

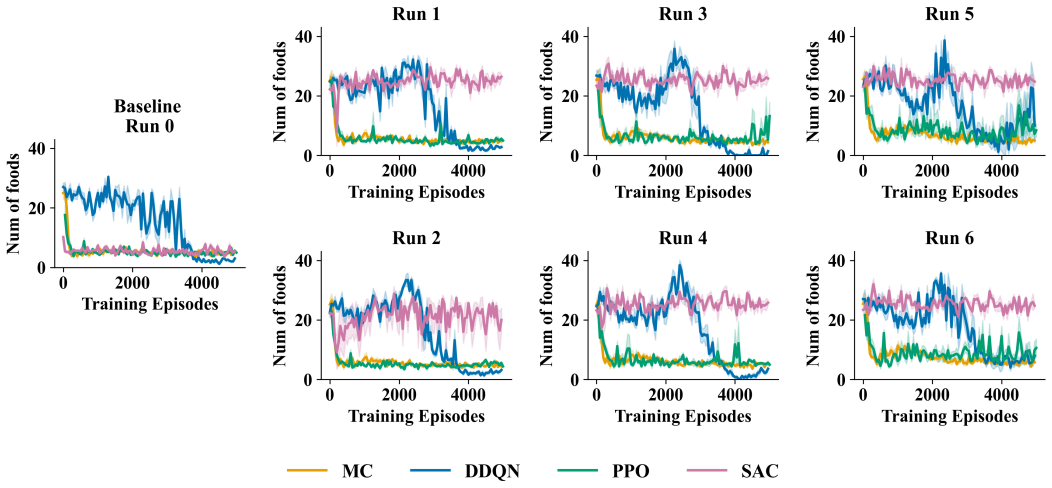

Figure 2: Average number of foods consumed per episode during training across delay regimes (Run 0–6). Shaded regions denote standard error across five seeds. Increased delay leads to higher consumption variability and greater overshoot for most algorithms. MC and PPO exhibit gradual increases in intake under larger delays, while SAC converges to persistently high consumption under all delayed settings.

regimes, differences between low and high temporal spread within each pair remain small, indicating that the delay mean is the primary factor influencing performance.

Overall, increasing delay shifts the task toward longer-horizon predictive regulation, producing distinct learning dynamics across algorithms; notably, SAC demonstrates pronounced sensitivity even at small delays, with limited recovery as delay magnitude increases.

6.2 Fixed-Policy Inference Across Delay Regimes

We evaluate the learned policies across all delay regimes. The aggregated inference metrics are shown in Fig. 3, including reward, distance-to-target, foods consumed, and overshoot ratio.

The MC agent shows a gradual degradation in performance as the delay mean increases. In Run 0, it achieves the highest reward and the lowest distance-to-target. As the delay increases from small to large, food consumption rises almost monotonically. This increased intake results in greater overshooting, which in turn leads to higher distance-to-target values.

PPO exhibits similar baseline behavior as MC but greater sensitivity to regime transitions. In Runs 0–2, distance-to-target remains relatively low; however, from medium delays onward (Runs 3–6), performance decreases further. Distance-to-target increases and the overshoot ratio becomes noticeably higher, exceeding that of MC at comparable delay levels. As the delay increases, particularly under medium and large delays, PPO tends to consume more food than required relative to the target, leading to increased overshooting. Although returns decrease with increasing delay, the degradation remains progressive rather than abrupt.

SAC displays a consistent high-consumption regime under delay. While competitive in the baseline regime, it converges to a near-constant high food consumption level (approximately 25 foods per episode) across all delayed conditions. The overshoot ratio approaches 1.0 and remains effectively independent of delay mean or variance. Distance-to-target remains elevated and reward remains low. This suggests that SAC does not effectively adapt its policy to delayed credit assignment under the tested configuration.

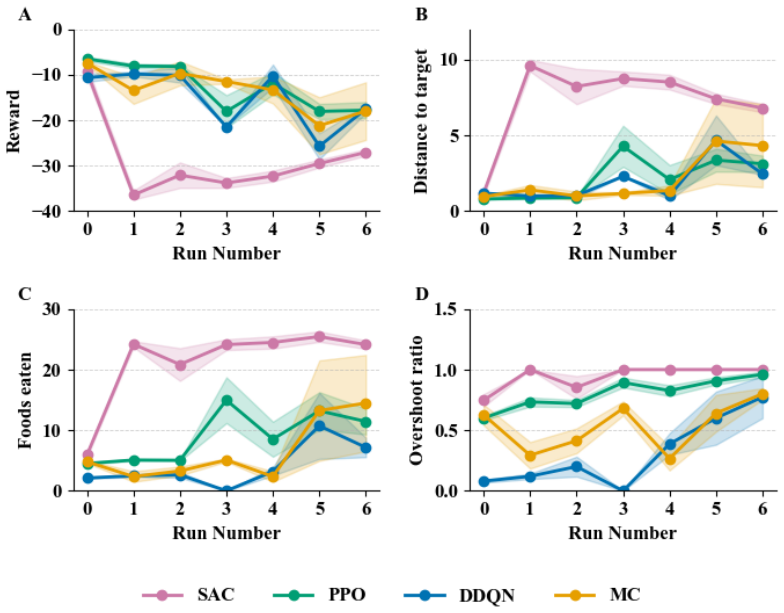

Figure 3: Fixed-policy inference performance across delay regimes. Panels show (A) episodic reward, (B) final distance-to-target, (C) foods consumed per episode, and (D) overshoot ratio. Results are averaged across five seeds and five evaluation episodes per seed, with shaded regions indicating standard error. Increasing delay mean progressively shifts the benchmark toward longer-horizon regulation regimes, while temporal spread within each regime has comparatively minor impact.

DDQN shows a clear decline in performance once delay is introduced. Similar to MC and PPO, increasing delay leads to higher distance-to-target and reduced reward, with unstable food consumption patterns across runs. The value-based estimator appears unable to propagate credit effectively once reward is temporally displaced.

6.3 Benchmark Sensitivity Profile

Across both training and inference analyses, the benchmark exhibits three consistent sensitivity patterns. First, increasing the delay mean systematically shifts the control problem toward longer-horizon anticipation across all algorithms. Second, temporal spread has a comparatively minor influence relative to the delay mean. Third, different algorithm classes exhibit distinct adaptation patterns under moderate delay, whereas entropy-regularized actor–critic (SAC) and bootstrapped value-based methods (DDQN) exhibit pronounced instability or invariant failure once temporal decoupling becomes significant.

These results characterize how the proposed environment stresses temporal credit assignment under controlled delay scaling. Algorithms capable of stable return propagation and conservative policy updates degrade gracefully, whereas methods relying heavily on bootstrapped value estimation or persistent entropy regularization exhibit increased sensitivity as delay increases.

6.4 Recurrent Policy Evaluation

NutriRL is naturally modeled as a POMDP; the digestion backlog and nutritional targets remain partially hidden. Because delayed absorption introduces latent state evolution not captured by instantaneous observations, we evaluate whether recurrent architectures can help mitigate this temporal decoupling. To test this, we compared the feedforward PPO agent with a GRU-PPO variant, keeping all other hyperparameters and training budgets identical.

Figure 4 illustrates the performance across delay regimes. In the no-delay baseline (Run 0), both architectures achieve near-optimal performance, indicating no inherent advantage to memory when transitions are immediate. As delay increases (Runs 3–6), GRU-PPO consistently outperforms feedforward PPO, achieving higher returns, lower distance-to-target, and reduced overshoot ratios. This suggests the recurrent agent can better anticipate delayed absorption by integrating historical intake history.

Despite these improvements, increasing delay regimes remain substantially more challenging. While the GRU-PPO reduces overconsumption, its efficiency remains below the baseline regime, where the agent typically consumes only 3–4 food items to reach targets, providing a useful dietary efficiency benchmark. Even with temporal memory, the agent's performance degrades as delay scales, highlighting that structural digestion delays introduce increasingly complex temporal credit assignment, making NutriRL a useful benchmark for evaluating RL methods under delayed physiological feedback.

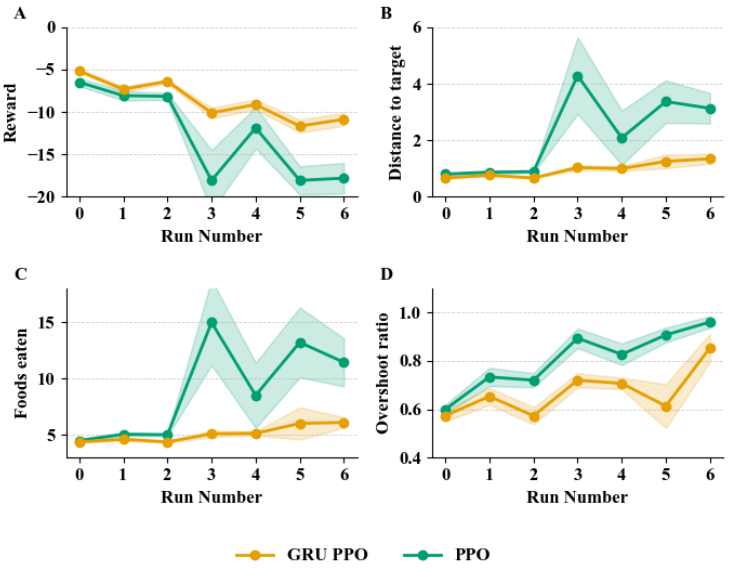

Figure 4: Inference performance comparison between feedforward PPO and GRU-PPO across delay regimes. Panels show (A) episodic reward, (B) distance-to-target, (C) foods consumed per episode, and (D) overshoot ratio. Shaded regions denote standard error across seeds. While both architectures degrade as delay increases, GRU-PPO consistently achieves higher reward, lower distance-to-target, and reduced overshoot under medium and long delays.

7 Discussion and Conclusion

NutriRL was designed to capture a structural property central to physiological decision-making but largely abstracted away in standard reinforcement learning benchmarks: delayed state evolution under irreversible accumulation. Our empirical characterization shows that embedding delay directly within transition dynamics fundamentally alters the structure of the learning problem. When transitions are immediate, all evaluated algorithms converge reliably to stable optima. Once delay is introduced, the task requires anticipatory regulation: returns decrease, distance-to-target increases, and transient overshoot becomes more frequent. Importantly, even small delays measurably influence policy dynamics, while larger delays amplify variability and reduce achievable reward across methods. In particular, MC and PPO exhibit gradual sensitivity to increasing delay, DDQN shows increasing instability, and SAC demonstrates high sensitivity even under small delays. Collectively, these findings indicate that predictive regulation under delayed accumulation introduces measurable differences relative to standard immediate-transition benchmarks.

The importance of this benchmark lies in its structural fidelity to physiological regulation. In real dietary systems, intake decisions produce gradual and overlapping internal effects that cannot be instantly reversed. Effective control, therefore, requires anticipating cumulative consequences rather than reacting to immediate feedback. NutriRL captures this property in a minimal, computationally efficient, Gymnasium-compatible framework. By embedding heterogeneous delays directly into state evolution, the environment preserves multi-channel nutrient coupling and irreversible accumulation, providing a principled testbed for studying reinforcement learning under temporally extended dynamics.

This work focuses on isolating delayed transition dynamics within a minimal and computationally efficient setting; accordingly, we employ phenomenological absorption models and standard feedforward RL algorithms to study structural delay effects in a controlled manner. The current benchmark also adopts a simplified binary consume/skip action space and a shared nutrient absorption model across all episodes, intentionally reducing variability from combinatorial food selection and inter-individual physiological differences. Preliminary evaluation of a recurrent PPO variant indicates that memory improves stability under delayed transitions. While recurrence improves stability and reduces overshoot in medium and long delay regimes, the task remains challenging under severe temporal decoupling. More expressive recurrent, belief-based, or model-based approaches, therefore, remain promising directions for future work. Future benchmark extensions may incorporate richer action spaces, personalized physiological parameterizations, and longer-term regulatory dynamics while preserving the benchmark's focus on delayed transition effects.

NutriRL is intended as a standardized evaluation environment for studying reinforcement learning under delayed transition dynamics. All delay regimes, reward coefficients, environmental parameters, and baseline configurations are fixed and documented to support reproducible cross-study comparison. The benchmark implementation, environment configurations, and baseline code are publicly available on GitHub to facilitate reproducibility and community adoption.

8 Future Scope

While NutriRL intentionally employs a simplified action space and a single nominal physiology to isolate delayed transition dynamics, future benchmark versions may incorporate richer food-selection actions, larger food repertoires, personalized physiological models, and multi-day regulation tasks. Extensions may also include continuous action spaces that allow agents to determine both food choice and consumption quantity, enabling finer-grained nutritional control. Additional developments could incorporate stochastic digestion dynamics, demographic-specific nutritional targets, and clinically informed absorption models. Such enhancements would increase ecological validity while preserving the benchmark's primary role as a controlled testbed for studying temporal credit assignment and decision-making under delayed state evolution.

References

Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization. In *International conference on machine learning*, pp. 22–31. Pmlr, 2017.

Sumana Basu, Marc-André Legault, Adriana Romero-Soriano, and Doina Precup. On the challenges of using reinforcement learning in precision drug dosing: Delay and prolongedness of action effects. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 14102–14109, 2023.

David Benton and Hayley A Young. Reducing calorie intake may not help you lose body weight. *Perspectives on Psychological Science*, 12(5):703–714, 2017.

Srinivasa Chakravarthy, Pragathi Priyadharsini Balasubramani, Alekhya Mandali, Marjan Jahanshahi, and Ahmed A Moustafa. The many facets of dopamine: Toward an integrative theory of

the role of dopamine in managing the body's energy resources. *Physiology & behavior*, 195: 128–141, 2018.

Kartik Choudhary, Dhawal Gupta, and Philip S Thomas. Icu-sepsis: A benchmark mdp built from real medical data. *arXiv preprint arXiv:2406.05646*, 2024.

Mantana Chudtong and Andrea De Gaetano. A mathematical model of food intake. *Mathematical Biosciences and Engineering*, 18(2):1238–1279, 2021.

Brian M Chung and Kelly A Tappenden. Macronutrient digestion, absorption, and metabolism. In *Clinical nutrition in gastrointestinal disease*, pp. 77–96. CRC Press, 2024.

Gabriel Dulac-Arnold, Daniel Mankowitz, and Todd Hester. Challenges of real-world reinforcement learning. *arXiv preprint arXiv:1904.12901*, 2019.

Harry Emerson, Matthew Guy, and Ryan McConville. Offline reinforcement learning for safer blood glucose control in people with type 1 diabetes. *Journal of Biomedical Informatics*, 142:104376, 2023.

Javier Garcıa and Fernando Fernández. A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research*, 16(1):1437–1480, 2015.

Mattia Garutti, Marianna Sirico, Claudia Noto, Lorenzo Foffano, Mark Hopkins, and Fabio Puglisi. Hallmarks of appetite: a comprehensive review of hunger, appetite, satiation, and satiety. *Current Obesity Reports*, 14(1):12, 2025.

Omer Gottesman, Fredrik Johansson, Matthieu Komorowski, Aldo Faisal, David Sontag, Finale Doshi-Velez, and Leo Anthony Celi. Guidelines for reinforcement learning in healthcare. *Nature medicine*, 25(1):16–18, 2019.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, pp. 1861–1870. Pmlr, 2018.

Kevin D Hall, I Sadaf Farooqi, Jeffery M Friedman, Samuel Klein, Ruth JF Loos, David J Mangelsdorf, Stephen O'Rahilly, Eric Ravussin, Leanne M Redman, Donna H Ryan, et al. The energy balance model of obesity: beyond calories in, calories out. *The American journal of clinical nutrition*, 115(5):1243–1254, 2022.

Beining Han, Zhizhou Ren, Zuofan Wu, Yuan Zhou, and Jian Peng. Off-policy reinforcement learning with delayed rewards. In *International conference on machine learning*, pp. 8280–8303. PMLR, 2022.

HM Jones, Yuan Chen, Christopher Gibson, Tycho Heimbach, Neil Parrott, SA Peters, Jan Snoeys, VV Upreti, Ming Zheng, and SD Hall. Physiologically based pharmacokinetic modeling in drug discovery and development: a pharmaceutical industry perspective. *Clinical Pharmacology & Therapeutics*, 97(3):247–262, 2015.

Armin Karamzade, Kyungmin Kim, Montek Kalsi, and Roy Fox. Reinforcement learning from delayed observations via world models. *arXiv preprint arXiv:2403.12309*, 2024. URL https://arxiv.org/abs/2403.12309.

Mehdi Keramati and Boris Gutkin. Homeostatic reinforcement learning for integrating reward collection and physiological stability. *Elife*, 3:e04811, 2014.

Jens Kober, J Andrew Bagnell, and Jan Peters. Reinforcement learning in robotics: A survey. *The International Journal of Robotics Research*, 32(11):1238–1274, 2013.

L Kuepfer, C Niederalt, T Wendl, J-F Schlender, S Willmann, J Lippert, M Block, T Eissing, and D Teutonico. Applied concepts in pbpk modeling: how to build a pbpk/pd model. *CPT: pharmacometrics & systems pharmacology*, 5(10):516–531, 2016.

Pierre Liotet. Delays in reinforcement learning. *arXiv preprint arXiv:2309.11096*, 2023.

Mingyang Liu, Xiaotong Shen, and Wei Pan. Deep reinforcement learning for personalized treatment recommendation. *Statistics in medicine*, 41(20):4034–4056, 2022.

Chiara Dalla Man, Francesco Micheletto, Dayu Lv, Marc Breton, Boris Kovatchev, and Claudio Cobelli. The uva/padova type 1 diabetes simulator: new features. *Journal of diabetes science and technology*, 8(1):26–34, 2014.

Konstantinos I Mavrokotas, Eleni I Georga, Costas Papaloukas, and Dimitrios I Fotiadis. A nutrition recommendation system based on reinforcement learning. In *2025 47th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 1–4. IEEE, 2025.

Sophie Miquel-Kergoat, Veronique Azais-Braesco, Britt Burton-Freeman, and Marion M Hetherington. Effects of chewing on appetite, food intake and gut hormones: a systematic review and meta-analysis. *Physiology & behavior*, 151:88–96, 2015.

Volodymyr Mnih et al. Human-level control through deep reinforcement learning. *Nature*, 518 (7540):529–533, 2015.

Sheila Annie Peters. *Physiologically based pharmacokinetic (PBPK) modeling and simulations: principles, methods, and applications in the pharmaceutical industry*. John Wiley & Sons, 2021.

Andrés Pino Santana, Santiago Garcia-Gen, Laurent Dewasme, and Alain Vande Wouwer. Model predictive control of anaerobic digestion processes using a long short-term memory network predictor. *Water Science & Technology*, 92(7):1063–1076, 2025.

Jorge Guerra Pires. *Mathematical modeling in energy homeostasis, appetite control and food intake with a special attention to ghrelin*. Jorge Guerra Pires, 2017.

Allah Rakha, Fakiha Mehak, Muhammad Asim Shabbir, Muhammad Arslan, Muhammad Modassar Ali Nawaz Ranjha, Waqar Ahmed, Claudia Terezia Socol, Alexandru Vasile Rusu, Abdo Hassoun, and Rana Muhammad Aadil. Insights into the constellating drivers of satiety impacting dietary patterns and lifestyle. *Frontiers in Nutrition*, 9:1002619, 2022.

David Raubenheimer and Stephen J Simpson. Nutritional ecology and human health. *Annual review of nutrition*, 36:603–626, 2016.

JJ Rumessen and E Gudmand-Høyer. Absorption capacity of fructose in healthy adults. comparison with sucrose and its constituent monosaccharides. *Gut*, 27(10):1161–1168, 1986.

Moisés Santillán. Quantitative insights into glucose regulation: A review of mathematical modeling efforts. *Dynamics of physiological control: Contributions in honor of Michael C. Mackey*, pp. 125–148, 2025.

John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. *arXiv preprint arXiv:1506.02438*, 2015.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.

David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, et al. Mastering chess and shogi by self-play with a general reinforcement learning algorithm. *arXiv preprint arXiv:1712.01815*, 2017.

Richard S Sutton. Learning to predict by the methods of temporal differences. *Machine learning*, 3 (1):9–44, 1988.

Panagiotis Symeonidis, Evangelos Rizos, Christos Andras, Stergios Hairistanidis, Yannis Manolopoulos, and Markus Zanker. Deep reinforcement learning for personalized insulin dosing and glucose control of hospitalized in icu patients. *International Journal of Data Science and Analytics*, 20(7):6841–6854, 2025.

Nikolaos Theodorakis, Magdalini Kreouzi, Andreas Pappas, and Maria Nikolaou. Beyond calories: individual metabolic and hormonal adaptations driving variability in weight management—a state-of-the-art narrative review. *International journal of molecular sciences*, 25(24):13438, 2024.

Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ international conference on intelligent robots and systems*, pp. 5026–5033. IEEE, 2012.

Angelo Tremblay and France Bellisle. Nutrients, satiety, and control of energy intake. *Applied Physiology, Nutrition, and Metabolism*, 40(10):971–979, 2015.

Dimitris Tsolakidis, Lazaros P Gymnopoulos, and Kosmas Dimitropoulos. Artificial intelligence and machine learning technologies for personalized nutrition: a review. In *Informatics*, volume 11, pp. 62. MDPI, 2024.

George A van Aken. Computer modeling of digestive processes in the alimentary tract and their physiological regulation mechanisms: closing the gap between digestion models and in vivo behavior. *Frontiers in Nutrition*, 11:1339711, 2024.

Hado Van Hasselt, Arthur Guez, and David Silver. Deep reinforcement learning with double q-learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30, 2016.

Longqi Yang, Cheng-Kang Hsieh, Hongjian Yang, John P Pollak, Nicola Dell, Serge Belongie, Curtis Cole, and Deborah Estrin. Yum-me: a personalized nutrient-based meal recommender system. *ACM Transactions on Information Systems (TOIS)*, 36(1):1–31, 2017.

Chao Yu, Jiming Liu, Shamim Nemati, and Guosheng Yin. Reinforcement learning in healthcare: A survey. *ACM Computing Surveys (CSUR)*, 55(1):1–36, 2021.

Christopher A Zimmerman, Scott S Bolkan, Alejandro Pan-Vazquez, Bichan Wu, Emma F Keppler, Jordan B Meares-Garcia, Eartha Mae Guthman, Robert N Fetcho, Brenna McMannon, Junuk Lee, et al. A neural mechanism for learning from delayed postingestive feedback. *Nature*, 642(8068): 700–709, 2025.

Supplementary Materials

The following content was not necessarily subject to peer review.

Algorithm 1 NutriRL Environment Dynamics (Finite-Horizon POMDP)

Require: Planning horizon H, discount factor γ; food stream $\{i_t\}_{t=1}^{H}$;
nutrient parameters $\{m_n(i), \tau_n^{(i)}, \sigma_n^{(i)}\}$;
hidden target $\mathbf{t}$; reward weights $(\lambda_p, \lambda_c, \lambda_o)$
Ensure: Trajectory $\{(o_t, a_t, r_t)\}_{t=1}^{H}$

$\mathbf{s}_0 \leftarrow \mathbf{0}$
$\mathcal{D}_0 \leftarrow \emptyset$
Sample initial food item i_1
$o_1 \leftarrow (\mathbf{s}_0, \mathbf{e}(i_1))$
for $t = 1$ to H **do**
  Select action $a_t \in \{0, 1\}$ given observation o_t
  if $a_t = 1$ **then**
    for $n \in \{C, F, P\}$ **do**
      Compute mean $\mu \leftarrow t + \tau_n^{(i_t)}$
      Compute horizon $H_n \leftarrow \lceil 6\,\sigma_n^{(i_t)} \rceil$
      Define kernel
      $K_n^{(i_t)}(u) \leftarrow \dfrac{\exp\left(-\frac{1}{2}\left(\frac{u-\mu}{\sigma_n^{(i_t)}}\right)^2\right)}{\sum_{v=t+1}^{t+H_n} \exp\left(-\frac{1}{2}\left(\frac{v-\mu}{\sigma_n^{(i_t)}}\right)^2\right)}$
      Create digestion vector
      $\Delta s_n(u) \leftarrow m_n(i_t)\, K_n^{(i_t)}(u)$
    end for
    Append digestion process to backlog $\mathcal{D}_t$
  end if
  $\mathbf{s}_t \leftarrow \mathbf{s}_{t-1} + \sum_{j \in \mathcal{D}_t} \begin{bmatrix} \Delta s_C^{(j)}(t) \\ \Delta s_F^{(j)}(t) \\ \Delta s_P^{(j)}(t) \end{bmatrix}$
  Remove expired digestion entries from $\mathcal{D}_t$
  $d_t \leftarrow \|\mathbf{s}_t - \mathbf{t}\|_1$
  $r_t \leftarrow \lambda_p(d_{t-1} - d_t) - \lambda_c \tanh\left(\frac{d_t}{\tau_c}\right) - \lambda_o \tanh\left(\|\max(\mathbf{0}, \mathbf{s}_t - \mathbf{t})\|_2^2\right)$
  Sample next food item i_{t+1}
  $o_{t+1} \leftarrow (\mathbf{s}_t, \mathbf{e}(i_{t+1}))$
end for

Table 2: Common Experimental Configuration (All Algorithms)

State representation	3 physiological + 10-dim one-hot food embedding
Episode horizon	50 steps
Normalization	Enabled
Delay regimes	6 dataset variants
Seeds per configuration	5 seeds (seed values: 0, 21, 9, 2003, 1997)
Evaluation episodes	5 per seed

Table 3: Architectural and Algorithmic Comparison: PPO vs DDQN

Component	**PPO**	**DDQN**
Learning type	On-policy	Off-policy
Network structure	Shared Actor–Critic	Action-conditioned Q-network
Backbone MLP	$256 \to 128$	$256 \to 128$
Value networks	1 Value head	1 Q-network + 1 target network
Replay buffer	No	Yes (300k capacity)
Return estimator	GAE ($\gamma = 0.99$, $\lambda = 0.95$)	5-step return ($n = 5$)
Batch size	64 (minibatch)	128
Gradient clipping	$\|\nabla\|_2 \leq 0.5$	$\|\nabla\|_2 \leq 10.0$
Learning rate	1×10^{-3}	$1 \times 10^{-3} \to 1 \times 10^{-4}$ (linear decay)

Table 4: Architectural and Algorithmic Comparison: MC Agent vs SAC

Component	**MC Agent**	**SAC**
Learning type	On-policy Monte Carlo	Off-policy entropy-regularized
Network structure	Shared Actor–Critic	Separate Actor + Twin Q (Q1,Q2)
Backbone MLP	$256 \to 128$	$256 \to 128$
Value networks	1 Value head	2 Q-networks + 2 targets
Replay buffer	No	Yes (100k capacity)
Return estimator	Full-episode GAE	1-step soft Bellman backup
Entropy regularization	No	Yes ($\alpha = 0.2$ fixed)
τ	—	0.005
Batch size	Episode-level	128
Actor LR	1×10^{-3}	1×10^{-4}
Critic LR	1×10^{-3}	1×10^{-3}

Table 5: Optimization and Update Schedule (All Algorithms)

	MC	DDQN	PPO	SAC
Episodes	5000	5000	5000	5005
Horizon		50 steps		
Optimizer	Adam	Adam	Adam	Adam
Replay buffer	No	300k	No	100k
Min replay size	—	10k	—	Batch size
LR schedule	None	Linear decay	None	None

Table 6: GRU-PPO Architectural Modification

Component	**Specification**
Base architecture	Shared Actor–Critic (PPO backbone)
Physiological encoder	1-layer GRU (input size = 3, hidden size = 128)
Food embedding encoder	1-layer GRU (input size = 10, hidden size = 128)
Hidden state combination	Concatenation of final GRU hidden states (256-dim total)
Actor head	Linear($256 \to 128$) $\to$ ReLU $\to$ Linear($128 \to 2$)
Critic head	Linear($256 \to 128$) $\to$ ReLU $\to$ Linear($128 \to 1$)
Weight initialization	Orthogonal (linear + GRU weights), bias = 0
Sequence handling	Hidden states reset at episode start
Other hyperparameters	Identical to feedforward PPO