POLITECNICO DI TORINO
Repository ISTITUZIONALE

Beyond the Black Box: Neuro-Symbolic Integration for Interpretable Video-Based Reinforcement Learning

Original
Beyond the Black Box: Neuro-Symbolic Integration for Interpretable Video-Based Reinforcement Learning / Cardone, L., Ghisolfo, G., Mongardi, G., Quer, S., Squillero, G.. - ELETTRONICO. - (2026), pp. 593-604. (21st International Conference on Software Technologies Porto (PRT) 16/07/2026 - 18/07/2026) [10.5220/0015174000004088].

Availability:
This version is available at: 11583/3014074 since: 2026-08-06T16:06:23Z

Publisher:
SCITEPRESS

Published
DOI:10.5220/0015174000004088

Terms of use:

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright

(Article begins on next page)

07 October 2026

Beyond the Black Box: Neuro-Symbolic Integration for Interpretable Video-Based Reinforcement Learning

Lorenzo Cardone[a], Giorgia Ghisolfo, Giorgio Mongardi[b], Stefano Quer[c] and Giovanni Squillero[d]

DAUIN Department of Control and Computer Engineering, Politecnico di Torino, Turin, Italy

Keywords: Neuro-Symbolic AI, Core Knowledge Theory, Interpretable Machine Learning, Intrinsic Motivation, Causal Reasoning, Deep Reinforcement Learning, Computer Vision.

Abstract: This paper introduces a cognitively inspired neuro-symbolic framework for extracting interpretable world models from raw video streams in dynamic 2D environments. While traditional end-to-end deep reinforcement learning systems often function as opaque "black boxes," our approach decouples visual perception from policy learning to enhance transparency. By leveraging Core Knowledge theory (specifically object persistence, physical causality, and agent representation), the system transforms visual patches into structured symbolic entities and governing interaction rules. The core of this architecture is a symbolic module that reconstructs persistent objects, infers parametric motion laws (such as velocity inversion upon contact), and incrementally consolidates class-specific behaviors into a compact knowledge base. This world model instantiates a domain-agnostic environment wrapper, enabling a Deep Q-Network to operate on symbolic state vectors rather than pixels. We further propose a domain-agnostic agent that develops complex behaviors driven by a composite intrinsic reward based on causal impact and event-driven curiosity. Experimental evaluations on Arkanoid and Pong demonstrate that this framework achieves near-optimal performance without task-specific external rewards. On Arkanoid, the agent matches the win rate of fully supervised models while remaining robust to structural modifications, such as changes in ball size or brick configuration. In Pong, the same mechanism transfers without architectural adjustments, consistently improving survival times. These results provide a transparent, generalizable alternative to dominant AI paradigms with good performance across varied environmental conditions.

1 INTRODUCTION

The field of Artificial Intelligence has seen a dramatic shift from the symbolic logic-based systems of the 20th century to the modern connectionist paradigm dominated by Deep Learning. While deep neural networks excel at high-dimensional pattern recognition, they often suffer from the "black box" problem, providing limited insight into their internal representations and offering no explicit handle for structured causal reasoning. These systems typically require vast amounts of labeled data and often fail to generalize to minor perturbations in their environment, particularly when the distribution shift affects relational or causal structure rather than low-level appearance (Marcus, 2018).

To address these limitations, researchers have increasingly turned to integrating neural and symbolic approaches (Neuro-symbolic AI) to combine the perceptual power of deep learning with the reasoning capabilities of symbolic logic. Furthermore, insights from developmental psychology, such as Core Knowledge theory, suggest that human intelligence is built upon innate systems for representing objects, space, and causality. These systems allow humans to learn with high sample efficiency and form "world models" that support counterfactual reasoning (Lake et al., 2017).

In this paper, we propose a framework that bridges the gap between raw visual perception and symbolic reasoning. Our system utilizes a U-Net architecture for semantic segmentation to ground symbolic entities in visual data. We process these entities

[a] https://orcid.org/0009-0008-7553-4839
[b] https://orcid.org/0009-0009-6175-2947
[c] https://orcid.org/0000-0001-6835-8277
[d] https://orcid.org/0000-0001-5784-6435

Cardone, L., Ghisolfo, G., Mongardi, G., Quer, S. and Squillero, G.
Beyond the Black Box: Neuro-Symbolic Integration for Interpretable Video-Based Reinforcement Learning.
DOI: 10.5220/0015174000004088
Paper published under CC license (CC BY-NC-ND 4.0)
In *Proceedings of the 21st International Conference on Software Technologies (ICSOFT 2026)*, pages 593-604
ISBN: 978-989-758-855-6; ISSN: 2184-2833
Proceedings Copyright © 2026 by SCITEPRESS – Science and Technology Publications, Lda.

by a symbolic reasoning module that infers interaction rules and maintains multiple competing hypotheses to manage perceptual ambiguity. Specifically, the system decomposes each frame into anonymous visual patches, tracks them over time using Core Knowledge-inspired heuristics (continuity, cohesion, and spatial proximity), and builds class-level rules that link discrete events (e.g., collisions) to parametric updates of object properties (such as velocity). Competing hypotheses about object identity and dynamics are scored for simplicity and explanatory power, and pruned when they fail to account for new observations. Finally, we demonstrate that a Deep Q-Network (DQN) agent operating on this structured state space can acquire complex skills through intrinsic motivation (specifically via "event-centered" rewards and causal impact bonuses), eliminating the need for handcrafted, task-specific reward functions.

To sum up, our main contributions are the following:

- A perceptually grounded symbolic world model for 2D arcade environments.
- An intrinsic reward scheme that operationalizes causal impact and event density over symbolic states.
- An empirical demonstration that this combination yields robust, near-optimal control and zero-shot generalization under structural perturbations.

We evaluate our framework on Atari-like benchmarks, specifically Arkanoid[1] and Pong, showing that the system not only learns to play effectively but also maintains performance when physical parameters like object size or velocity are modified. This work represents a step toward autonomous agents that can construct a coherent, symbolic representation of their surroundings that is both effective and human-understandable.

This work extends a preliminary evolutionary proof-of-concept that focused solely on inferring objects, classes, and rules from Atari videos without a learning agent (Calabrese et al., 2024). In particular, we replace heuristic OpenCV-based segmentation with a U-Net perception module, introduce a symbolic world model with multi-hypothesis tracking and parametric motion laws, and connect it to a Deep Q-Network operating on symbolic states. The earlier paper focused on explaining observed videos; here, we additionally equip the system with intrinsic-motivation rewards (event-centered, causal impact, and curiosity) and demonstrate that a domain-agnostic agent can learn robust control policies in Arkanoid and Pong. Overall, this new work moves from offline rule discovery to an end-to-end agent that both constructs and exploits interpretable world models for reinforcement learning.

We organize the paper as follows. Section 2 reviews the state of the art and the theoretical foundations in Core Knowledge and neuro-symbolic AI. Section 3 details the proposed framework, including the U-Net perceptual grounding and symbolic rule inference. Section 4 presents the experimental evaluation on Atari benchmarks, while Section 5 reports some final considerations and directions for future work.

2 BACKGROUND

2.1 The Neuro-Symbolic Paradigm

A tension between two major paradigms has characterized the history of Artificial Intelligence: the "symbolic paradigm", which replicates human-like reasoning through the manipulation of logic-based rules, and the "connectionist paradigm", which learns representations directly from data through artificial neural networks. While symbolic systems excel at structured reasoning and provide inherent interpretability, they suffer from "brittleness" when faced with raw sensory data or uncertainty. Conversely, deep learning has achieved remarkable success in pattern recognition but lacks transparency, often resulting in "black-box" models that struggle with causal inference and factual consistency.

For several years, researchers have leveraged symbolic and formal techniques to enhance the validation power of testing within both hardware and software systems (Cabodi et al., 1994; Cabodi et al., 1999) Following this trend, Neuro-symbolic AI (Besold et al., 2021) emerges as a hybrid paradigm that seeks to bridge the gap between symbolic and connectionist approaches. By integrating the perceptual power of neural networks with the formal logic of symbolic systems, this approach aims to create agents that are both computationally powerful and explainable. Our framework aligns with this "third wave" of AI, using neural layers as perceptual front-ends to extract structured symbolic knowledge suitable for autonomous reasoning, with neural components serving strictly as perceptual front-ends and delegating world-model construction and decision-making to symbolic structures and intrinsic rewards (Liang et al., 2025).

[1] Arkanoid, released by Taito in 1986, is the more famous version of its spiritual "clone" Breakout, which was designed by Steve Wozniak and released by Atari a decade earlier in 1976.

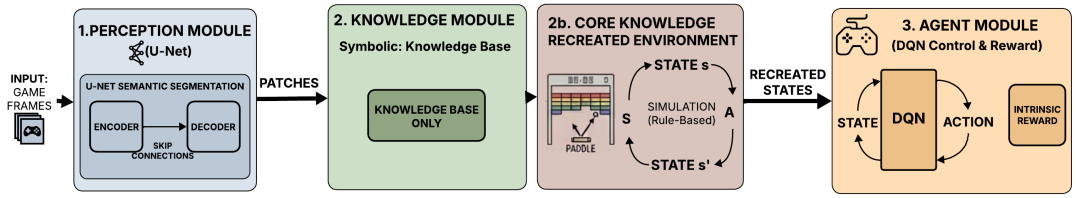

Figure 1: Overview of the proposed neuro-symbolic pipeline, from perceptual grounding to symbolic rule induction and DQN-based decision making. The system progressively transforms raw video frames into a structured, interpretable world model used for reinforcement learning.

2.2 Core Knowledge Theory

We use the Core Knowledge theory, from developmental psychology, to help our system identify important symbols in video streams. According to Spelke et al. (Spelke and Kinzler, 2007), human intelligence is built upon innate cognitive systems that represent fundamental aspects of the physical world, such as objects, space, and causality. In an AI context, these principles act as helpful assumptions that guide how the model learns. Indeed, rather than treating video as an undifferentiated stream of pixels, Core Knowledge suggests that a system should privilege object persistence (i.e., the understanding that entities endure over time), and physical causality (i.e., the ability to justify outcomes through cause-and-effect relations). By adopting these object-centric representations, an artificial system can construct internal "world models" that support not only prediction but also counterfactual reasoning and human-understandable explanations.

In our implementation, object persistence is enforced by tracking segmented patches across frames and discarding hypotheses that violate continuity or cohesion. We encode physical causality as parametric update rules that relate discrete events (such as contact with boundaries or bricks) to deterministic changes in velocity and object status. We also model agent representation via a special controllable entity and explicitly contrast their actions with a "null-action" baseline to quantify causal impact on other objects.

2.3 Deep Reinforcement Learning and the Atari Benchmark

The Atari 2600 benchmark has become the standard for evaluating the performance of autonomous agents. We align our symbolic world-model approach with (Mnih et al., 2015; Greff, Klaus and van Steenkiste, Sjoerd and Schmidhuber, Jürgen, 2020; Kassa, Yared and others, 2023). The landmark work by Mnih et al. (Mnih et al., 2015) demonstrates that DQN can achieve superhuman performance by learning directly from raw sensory inputs with minimal prior domain knowledge. Greff et al. (Greff, Klaus and van Steenkiste, Sjoerd and Schmidhuber, Jürgen, 2020) and Kassa et al. (Kassa, Yared and others, 2023) discuss object-centric representations and binding in neural networks.

However, despite their high scores, these agents remain profoundly fragile and breakable as they optimize heuristics based on raw pixel statistics rather than a structured understanding of the environment. Their performance often collapses when trivial changes (such as moving a paddle by a few pixels or changing a color) are introduced. This lack of an internal world model prevents true generalization. Our work addresses these limitations by replacing opaque pixel-to-action mapping with a structured state space, allowing the agent to maintain stable performance even under structural perturbations.

3 OUR FRAMEWORK

The proposed architecture is a modular pipeline that progressively increases the abstraction level of visual information, moving from raw pixels to semantic objects and finally to decision-making strategies. Figure 1 shows an overview of the proposed neuro-symbolic pipeline, made up of the following stages:

- Perceptual Grounding: The first module employs a U-Net architecture for pixel-wise semantic segmentation (Ronneberger et al., 2015). The segmentation masks are post-processed into a set of proto-objects with explicit attributes (position, size, and instantaneous velocity), which form the input alphabet for the symbolic module.
- Symbolic Rule Inference: Once the system identifies and tracks objects using Core Knowledge-inspired heuristics, the symbolic module extracts interaction rules. These rules capture conditional behaviors triggered by events (such as collisions or disappearances), providing a human-readable description of the environment's dynamics. Module (2b) analyzes a series of events to identify pat-

terns. It identifies which specific actions cause predictable changes to certain properties. Moreover, it incrementally groups objects into classes based on shared variability patterns and rule sets, yielding a compact, human-readable description of the environment's dynamics.

- Domain-Agnostic RL Wrapper: The third and last module integrates the inferred symbolic model with DQN. Unlike standard RL, the agent operates on a structured state space rather than raw pixels, guided by a composite intrinsic reward based on causal impact and event density. At each time step, the DQN observes a flattened vector encoding object-class identifiers (such as positions, velocities, and recent event flags), making the policy invariant to purely visual changes such as color or texture.

The following sections elaborate on the core ideas of each step.

3.1 Training Procedure

The framework follows a modular training strategy rather than an end-to-end one. First, we train the U-Net perception component offline on procedurally generated frames with pixel-level annotations. Once trained, the perception module is frozen and used to produce semantic segmentation masks. The symbolic reconstruction module does not rely on gradient-based optimization. On the contrary, it incrementally reconstructs object representations, infers interaction rules, and groups objects into symbolic classes from the segmented observations. Finally, the reinforcement learning component learns from symbolic state vectors produced by the symbolic subsystem. Consequently, only the DQN parameters are updated during reinforcement learning, while perception and symbolic reasoning provide a stable abstraction layer. This design improves interpretability and allows each component to be analyzed independently.

3.2 Perceptual Grounding

The first stage of the pipeline addresses the grounding problem by transforming raw video frames into semantic segmentation masks. We use a U-Net architecture, which is particularly effective for this task due to its symmetrical encoder-decoder structure and skip connections (Ronneberger et al., 2015). We organize the network as follows:

- Encoder and Bottleneck: The encoder progressively reduces spatial dimensions while increasing feature depth, thereby capturing increasingly abstract representations of the scene through convolutional blocks. It consists of three convolutional blocks, each composed of two consecutive 3×3 convolutional layers, followed by ReLU activation. These blocks progressively increase the number of feature channels while reducing the spatial resolution via max pooling with a stride of two. This design allows the network to capture increasingly abstract representations of the input image, as the spatial resolution decreases, while the number of feature channels increases. At the bottleneck, a deeper convolutional block increases the feature dimensionality, enabling the network to learn the most abstract and global image features.
- Decoder and Skip Connections: The decoder path reconstructs the spatial resolution and a dense segmentation map. Crucially, skip connections concatenate high-resolution features from the encoder to the upsampled features, preserving fine-grained spatial details and ensuring precise localization of object boundaries. The decoder mirrors the encoder structure and employs transposed convolutions for upsampling. At each decoding stage, the upsampled feature maps are concatenated with the corresponding encoder outputs through skip connections. This mechanism preserves spatial information and improves boundary localization in the predicted segmentation masks.

Overall, the network produces a pixel-wise classification map. This step is fundamental as it discretizes the continuous visual field into proto-objects, which serve as the raw material for the symbolic abstraction. For Arkanoid and Pong, we use a lightweight U-Net variant trained on procedurally generated frames with pixel-wise annotations for balls, paddles, bricks, and the background. In the experiments reported in Section 4, segmentation achieves sufficient accuracy to sustain stable object tracking. The multi-hypothesis symbolic reconstruction compensates for occasional mis-segmentations by discarding interpretations that repeatedly contradict the observed motion.

3.3 Symbolic System as Knowledge Construction Module

We decompose the environment into "patches," i.e., anonymous and minimal visual units that represent potential objects or their parts. In our implementation, we generate these patches by post-processing the U-Net semantic segmentation module's output.

The symbolic subsystem operates on the hierarchical representation shown in Figure 2. It first converts segmentation masks into anonymous patches

Figure 2: Hierarchical symbolic metamodel. The system associates raw segmented patches across frames to reconstruct persistent objects and infers rules from their transitions. It then groups objects with compatible properties and behaviors into generalized classes.

characterized by position, shape, and size. Next, it associates patches across consecutive frames to reconstruct persistent objects and analyzes their transitions to infer or update symbolic rules that capture behavior and interactions. When objects exhibit compatible property-variance patterns and behavioral rules, the subsystem groups them into symbolic classes. As it processes new frames, it repeats this procedure, allowing the symbolic knowledge base to evolve incrementally over time.

Guided by Core Knowledge principles, the object formation module reconstructs persistent entities by associating patches across consecutive frames. The object formation process acts as a temporal glue, enforcing the principle of "object permanence" even when visual evidence is intermittent or ambiguous. Since the system lacks prior information on object identity, it utilizes interpretable heuristics (such as spatial proximity, shape similarity, and behavioral consistency) to establish continuity. Rather than seeking a single "ground truth" immediately, the system adopts a probabilistic yet symbolic approach, entertaining multiple world models that compete to explain the incoming sensory evidence. This multi-hypothesis tracking ensures that the system remains robust against temporary occlusions or sudden environmental shifts that would otherwise lead to a loss of identity.

To handle the inherent ambiguity of visual streams, the system maintains multiple competing interpretations (hypotheses) in parallel and evaluates them based on their explanatory power. Each hypothesis defines both a mapping between visual patches and object identities, and a coherent set of class-level rules governing their behavior. Building on this, the symbolic reasoning module analyzes transitions between object states to infer the environment's underlying "laws of physics." By identifying recurring correlations between observed events (i.e., collisions and resulting changes in object properties), the system abstracts general, human-readable rules and stores them in an evolving knowledge base.

A pruning mechanism, implemented via a scoring system, ensures that simpler, more coherent explanations are preferred over more complex ones. In particular, the score-based system minimizes unexplained changes and maintains consistent object–patch associations. For instance, the system more naturally explains a ball moving at constant speed and bouncing off an object by a single change in velocity rather than by multiple unexplained shifts in position.

The system further evaluates each hypothesis based on how well it adheres to object-class constraints and how accurately it can predict its behavior from the environmental rules. To balance flexibility with computational efficiency, overly complex explanations are penalized; however, lower-scoring hypotheses are not discarded immediately. Instead, they are granted a "grace period," allowing them to accumulate additional evidence over time. This strategy prevents the premature elimination of potentially correct interpretations that require more temporal context to be validated.

Overall, this evaluation process is essential for scalability, as it filters out noise-driven hypotheses that lack consistency or predictive power, ultimately yielding a compact set of human-readable rules. For example, the speed of the ball along the x and y axis can be expressed in different instants (i and $i+1$) of the game as:

$$\begin{array}{lcl} \mathit{game_start} & \rightarrow & vx[i+1] = 0 \cdot vx[i] + 1 \\ \mathit{contact_bottom} & \rightarrow & vy[i+1] = -1 \cdot vy[i] + 0 \end{array} \quad (1)$$

These rules mirror the intuitive physics of the environment and form the basis for model-based predictions used in the causal impact computation described in Section 3.4.2.

Notice that the inference module does not merely record transitions; it seeks to extract the "latent grammar" of the environment and formulate it as a set of human-interpretable causal laws. These rules serve as a predictive forward model, allowing the system to project the future state of an object and, consequently, to detect anomalies when observations deviate from the learned physics. By encoding behaviors as variable transformations, the system achieves a level of abstraction that allows for the transfer of knowledge across different but functionally similar game scenarios.

3.4 Intrinsic Motivation and Learning

The decision-making core is a DQN, but with a fundamental change: it does not process image tensors.

Its input is a flattened vector of symbolic object properties. As a result, the network is significantly smaller and faster to train than standard pixel-based DQNs. Since the input is symbolic, the learned policy may exhibit improved robustness. If the environment's appearance changes, the symbolic representation remains unchanged, allowing for zero-shot generalization to new visual skins or modified game parameters.

Since no extrinsic goal is defined, we do not evaluate performance through traditional game scores. Instead, we assess interaction coherence, that is, the agent's ability to maintain behavioral stability, survive for extended periods, and sustain a high density of causal interactions. By combining the discount factor, the replay buffer, and the forward dynamics model, the agent develops temporal reasoning and long-term strategies. It learns to anticipate the consequences of its actions and handle delayed rewards, effectively "making sense" of the environment by continuously pursuing causal impact and informational gain.

In addition, the agent's behavior is governed by an internal drive to understand and manipulate its environment, rather than a pursuit of an external game score. This logical behavior follows the broader view of intrinsic motivation in autonomous agents, where internally generated signals guide exploration and skill acquisition in the absence of explicit external rewards (Oudeyer et al., 2007).

3.4.1 Event-Centered Rewards

To foster autonomous skill acquisition, we implement Event-Centered Rewards. Instead of traditional sparse rewards, the agent receives a positive signal whenever it triggers a change in the world's symbolic state. Concretely, we counted an event whenever at least one object property in the symbolic state vector (position, velocity, or existence flag) differs from its predicted null-action value, and the reward is proportional to the number or magnitude of such changes within a time step. This process encourages the agent to interact with the environment. The agent is rewarded for hitting an element because it triggers a sensor event in the symbolic state, even if it has no concept of winning or losing. This behavior aligns with the developmental psychology view that infants learn by trying to cause "interesting" effects in their surroundings.

3.4.2 Causal Impact Bonus

To distinguish between accidental events and those directly caused by the agent's agency, we introduce the Causal Impact Bonus. The system compares the actual state transition S_{t+1} resulting from an action a_t with a predicted "null-action" state (what would have happened if the agent had done nothing), in line with causal reasoning frameworks that emphasize counterfactual comparisons (Pearl, 2009). The null-action dynamics are provided by the current symbolic model, which predicts how objects would move in the absence of control. The causal impact bonus is then computed as a distance between the actual and null-action states in the symbolic space, emphasizing actions that meaningfully perturb other entities, akin to counterfactually guided policy search methods that evaluate actions against hypothetical alternatives (Buesing et al., 2018). If the agent's action substantially changes the trajectory or status of other objects, it receives a higher reward. This mechanism prioritizes actions that exert control over the environment.

3.4.3 Curiosity-Driven Bonus

Intrinsic motivation is further reinforced by a curiosity mechanism based on a forward predictive model. In the spirit of curiosity-driven exploration, which rewards prediction error over future observations (Pathak, Deepak and Agrawal, Pulkit and Efros, Alexei A. and Darrell, Trevor, 2017), this model attempts to predict the next symbolic state given the current state and action. The prediction error serves as an additional reward, pushing the agent to explore regions of the state space that are novel, complex, or poorly understood. This curiosity term is normalized over a sliding window to avoid unbounded growth and to prevent the agent from being trapped in stochastic or noisy configurations.

4 EXPERIMENTAL RESULTS

We evaluate our framework on Atari-inspired benchmarks, specifically Arkanoid and Pong as represented in Figure 3. We conduct a series of experiments to assess the performance of our symbolic strategy on Arkanoid (on the left-hand side of Figure 3), which serves as the base task. Once we have done this, we evaluate its generalization capabilities within the same environment under visual perturbations. Finally, we assess its cross-domain generalization on Pong (on the right-hand side of Figure 3).

The experimental results highlight several key findings regarding both the learned representations and the reinforcement learning dynamics enabled by the proposed framework. In particular, this section presents a comprehensive evaluation of the synergy among perceptual grounding, symbolic representa-

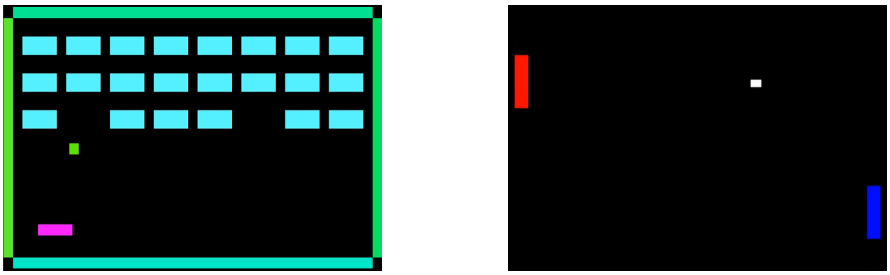

Figure 3: Example of perceptual grounding and symbolic reconstruction of the Arkanoid and Pong games.

tion, and reinforcement learning driven by intrinsic causal rewards. The primary objective is to determine whether a domain-agnostic agent can achieve goal-directed behavior solely by maximizing structured interactions, without access to extrinsic task-specific rewards or handcrafted game scores.

4.1 Experimental Setup and Metrics

In Arkanoid, the game terminated either when the player destroyed all the bricks or when they lost the ball. In Pong, an episode ended when the agent conceded a fixed number of points. We trained all models with the same replay buffer size, discount factor, and optimizer, differing only in the reward configuration and input representation (pixels versus symbols). We averaged the training curves across several independent seeds. We tested our framework on procedurally generated, deterministic arcade environments: Arkanoid (featuring a paddle, a ball, and 24 destructible bricks) and Pong.

We evaluate our methodology in terms of:

- Task Performance, analyzing the win rate, average bricks destroyed, and episode length (survival time).
- Structural Coherence, assessing the stability of induced symbolic rules and object classes under parameter changes.
- Generalization power, evaluating the trained models under structural modifications (e.g., altered object sizes, colors, or brick number) without any retraining.

As an upper baseline, we trained a supervised reference DQN directly on raw pixels with a dense reward derived from the game score. This model achieved a 100% win rate, destroyed 24 bricks in Arkanoid, and serves as the performance ceiling rather than a competing approach in terms of interpretability or sample efficiency.

4.2 Perceptual Grounding and Symbolic Reconstruction

The preliminary challenge we address is the transition from raw semantic masks to stable symbolic logs. This phase follows previous vision-based traffic-tracking systems that combine low-level segmentation with higher-level tracking and structural priors (Bottino et al., 2016). Raw U-Net segmentation often contains noise (e.g., intermittent missing detections). To mitigate this problem, we implemented a Refinement Pipeline that leverages a structural prior (the known geometric properties of game entities) to regularize malformed contours and ensure temporal continuity.

Overall, this phase transforms noisy masks into structured dictionaries (encoding position, velocity, and hitboxes), providing the high-quality input necessary for the symbolic rule induction. Figure 3 provides a visual representation intended for human inspection. Internally, the framework stores each entity as a symbolic record containing object identifier, class membership, position, velocity, size, and event information. The graphical visualization is generated from these symbolic records and does not constitute the representation used by the reinforcement learning module. Instead, the DQN receives a flattened symbolic state vector derived from the underlying object descriptions and inferred causal events.

4.3 Symbolic Representation and Rule Induction

In exploratory tests, the symbolic module successfully reconstructs persistent objects, infers transition rules, and groups objects into consistent symbolic classes. By analyzing the temporal evolution of patches, the system moves beyond simple tracking to achieve a structured understanding of the environment's physics. In practice, a stable set of ball rules–wall interactions emerges after only a few hundred environment steps. These rules remain valid even when we alter the ball size and the brick lay-

out. When rule violations occur (for example, due to transient segmentation errors), the corresponding hypotheses are downweighted and eventually pruned, while the dominant rule set persists across training runs.

For example, in the Pong game (see right-hand side of Figure 3), the system must interpret a simple scenario with a ball, four walls, and two paddles. The ball is initialized with a randomly selected direction and changes its velocity upon colliding with a wall or a paddle. In this environment, the system easily distinguishes between static environmental boundaries (walls) and dynamic entities (the ball and the paddles). We formalize this distinction by inducing Object Classes that aggregate objects based on shared property variances and behavioral signatures. Overall, our system describes the game using three classes.

Objects belonging to the first class (*Class 1*, represented in Figure 4) have positions that remain constant across all frames. Although each object preserves its shape over time, objects within the class may differ in shape. Thus, the system correctly inserts all structural, non-reactive elements of the environment into this class, such as walls.

Property Variance:
- Pos_x: constant
- Pos_y: constant
- Shape_x: constant
- Shape_y: constant

Rules:
- No Rules

same_shape: False
Assigned Objects: [0, 1, 2, 3]

Figure 4: Main characteristics of Class 1, representing the walls of the Pong puzzle.

The system identifies a second class, i.e., *Class 2*, represented in Figure 5, including objects whose position and speed are variable. In this class, the system specifically represents the ball, capturing the underlying dynamics of its environment. Indeed, the system learns that the ball begins moving at the start of the game and reverses its velocity upon collision with any boundary present on the scene.

Finally, the system inserts the two paddles into the third and last class, i.e., *Class 3*, as shown in Figure 6. In this case, a user can actively control the paddle, and the class represents objects whose vertical position is variable and dependent on inputs, while their horizontal position remains constant. Despite this increased complexity, the system converges on a single, consistent hypothesis, identifies the two involved objects, and infers a structured set of behavioral rules introducing a dependency on external inputs. The paddles

Property Variance:
- Pos_x: variable
- Pos_y: variable
- Shape_x: constant
- Shape_y: constant
- Speed_x: variable
- Speed_y: variable

Rules:
- game_start $\rightarrow$ vx[i+1] = 0 $\cdot$ vx[i] + 1
- game_start $\rightarrow$ vy[i+1] = 0 $\cdot$ vy[i] + 1
- Contact_Bottom $\rightarrow$ vy[i+1] = -1 $\cdot$ vy[i] + 0
- Contact_Right $\rightarrow$ vx[i+1] = -1 $\cdot$ vx[i] + 0
- Contact_Top $\rightarrow$ vy[i+1] = -1 $\cdot$ vy[i] + 0
- Contact_Left $\rightarrow$ vx[i+1] = -1 $\cdot$ vx[i] + 0

same_shape: True
Assigned Objects: [4]

Figure 5: Main characteristics of Class 2, representing the ball of the Pong puzzle.

change their positions not autonomously but in response to control signals. When these inputs are considered, the system correctly infers rules linking them to paddle movement. However, if such signals are absent, the representation remains structurally similar but lacks any behavioral rules for the paddle. This process highlights an important limitation: the system can only infer causal relationships when the relevant variables are observable. At the same time, these inferred rules play a crucial role. When available, they provide actionable knowledge that could guide an agent using this representation as an internal model. For instance, understanding that activating the up arrow (i.e., up_arrow_pressed) results in an upward paddle movement allows the agent to predict the consequences of its actions and make informed decisions based on the expected evolution of the environment.

Property Variance:
- Pos_x: constant
- Pos_y: variable
- Shape_x: constant
- Shape_y: constant

Rules:
- up_arrow_pressed $\rightarrow$ pos_y[i+1] = 1 $\cdot$ pos_y[i] - 2
- down_arrow_pressed $\rightarrow$ pos_y[i+1] = 1 $\cdot$ pos_y[i] + 2

same_shape: True
Assigned Objects: [5, 6]

Figure 6: Main characteristics of Class 3, representing the paddles of the Pong puzzle.

In our second experiment, conducted with the Arkanoid environment, the system proceeds as in the previous case, by generalizing the representation and grouping entities into Object Classes based on shared properties and dynamics. The walls and the ball retain abstractions similar to those derived in the Pong

experiment. In contrast, the paddle follows analogous control-dependent rules along the horizontal axis rather than the vertical axis. However, an additional object class emerges to account for the bricks that are initially 24 and disappear after contact with the ball. Figure 7 represents the physical rules for the bricks.

Property Variance:
Pos_x: constant
Pos_y: constant
Shape_x: constant
Shape_y: constant
Rules:
Contact(ball,brick) → Disappearance(brick)
same_shape: True
Assigned Objects: [5, 6, 7, 8, 9, 10, 11, 12, 13, 14
15, 16, 17, 19, 20, 21, 22, 23, 24]

Figure 7: Main characteristics of Class 4, for the Arkanoid game, representing the 24 bricks initially present on the game pitch.

4.4 Reinforcement Learning and Intrinsic Reward Analysis

The core contribution is tested by training a DQN agent under four reward configurations in Arkanoid and comparing it against a fully supervised Reference Model (with a 100% win rate). We compare four intrinsic reward schemes:

- Pure Causal, which relies solely on the causal impact bonus.
- Pure Causal plus Curiosity, which adds the prediction-error term.
- Shaping Only, which uses handcrafted shaping based on paddle–ball alignment.
- Causal plus Density, which augments causal impact with an event-density component.

All symbolic agents shared the same architecture and training hyperparameters.

Table 1 reports our results.

Table 1: Performance comparison across all model configurations.

Model	Win Rate	Avg Bricks
Reference Model (supervised)	100%	24.00
Pure Causal	83%	21.17
Pure Causal plus Curiosity	99%	23.98
Shaping Only	53%	20.21
Causal plus Density	**100%**	24.00

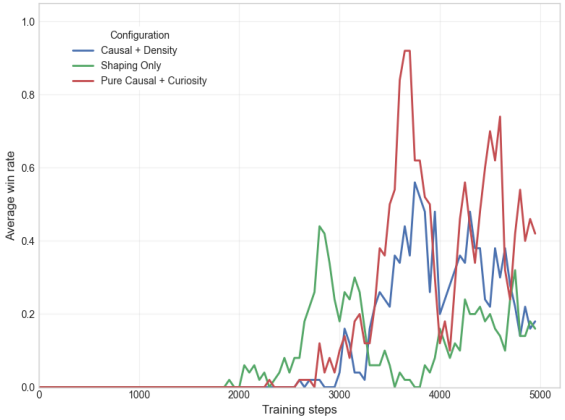

Figure 8: Training dynamics of different reward configurations on Arkanoid. Purely structural intrinsic rewards (causal, causal plus density, and causal plus curiosity) converge to near-optimal performance, while shaping-only rewards lead to convergent but suboptimal policies.

The data show that purely structural signals are sufficient to achieve performance comparable to that of the supervised reference. In contrast, naively designed shaping can be exploited by the agent, leading to suboptimal but reward-maximizing behaviors. The Pure Causal plus Curiosity configuration achieved a 99% win rate without any external score signal. However, compared to a standard pixel-based DQN, the proposed approach does not improve sample efficiency. In our experiments, the baseline DQN reached comparable performance in approximately 100 episodes, whereas the intrinsic-reward agent required around 2500 episodes. Additionally, the overall computational cost could be higher due to the overhead introduced by state extraction, event detection, causal reasoning, and the forward dynamics model, which we have not yet extensively optimized. Nevertheless, the main contribution of this approach lies in its interpretability. The agent's behavior can be traced through explicit event transitions and causal chains derived from the symbolic representation, providing a structured, explainable decision process. In contrast, standard DQN methods rely on a largely statistical mapping from observations to actions, which remains inherently opaque.

To better contextualize the trade-offs of the proposed approach, Table 2 summarizes the main differences between a conventional pixel-based DQN and the symbolic DQN adopted in this work. As shown in the table, the proposed framework primarily targets interpretability and explicit world modeling rather than sample efficiency or raw performance. Instead, its main contribution lies in replacing opaque pixel-level representations with an explicit symbolic world model. This process enables causal reasoning,

human-readable explanations, and robustness to visual perturbations, at the cost of increased computational overhead and slower learning.

Table 2: Qualitative comparison between a conventional pixel-based DQN and the proposed symbolic DQN framework.

Property	Pixel DQN	Symbolic DQN
Input	Pixels	Symbols
Reward	Extrinsic	Intrinsic
Interpretability	Low	High
Explicit world model	No	Yes
Visual invariance	Limited	High

This interpretability is complemented by empirical evidence showing that maximizing "interaction richness," operationalized as causal event chains, constitutes a viable proxy for goal-directed behavior. In particular, handcrafted rewards (e.g., paddle–ball alignment) often degraded performance. Indeed, in the Shaping Only setup, the agent learned to exploit the reward by remaining close to the ball without actually clearing bricks, highlighting the limitations of manually designed heuristics. Nevertheless, intrinsic structural signals proved more robust. Furthermore, augmenting the reward with an event-density component—capturing the magnitude of state changes—enabled the agent to match the efficiency of a supervised reference model, at the cost of increased variance across training checkpoints.

Figure 8 shows the average win rate (or average bricks destroyed) versus the training steps for the three models previously analyzed, i.e., Pure Causal plus Curiosity, Shaping Only, and Causal plus Density. The average win rate exhibits a rich and non-trivial learning dynamic, with performance peaks indicating the agent's ability to discover effective strategies. The observed oscillations should not be interpreted as a limitation, but rather as evidence of active exploration within the policy space. In particular, different configurations that leverage intrinsic reward signals achieve significantly higher peak performance. The combination of causal signals, curiosity, and event density enables the agent to learn richer representations of the environment dynamics. The non-stationary nature of intrinsic rewards promotes sustained exploration, preventing premature convergence to suboptimal policies. Curiosity-driven exploration proves especially effective in guiding the agent toward informative and high-value states. Overall, these results highlight a fundamental trade-off between exploration and stability. Richer reward signals favor the discovery of effective strategies, but introduce noise into the learning process. This behavior is consistent with the reinforcement learning literature, which shows that complex intrinsic rewards can compromise the stability of the value function. The observed instability suggests the need for techniques to normalize or balance the different contributions to the reward. The "Pure Causal plus Curiosity" configuration achieves the highest performance values and emerges as the most promising strategy, even if it also exhibits the greatest volatility. In contrast, the "Shaping Only" approach produces more stable learning, though it is limited in terms of maximum performance. The "Causal plus Density" combination represents a compromise between stability and exploratory capacity, maintaining moderate variance.

4.5 Generalization and Cross-Domain Validation

To test whether the agent acquired a coherent symbolic representation of the environment, we conduct zero-shot generalization tests:

- Structural Robustness: When the ball size is increased by 5×, and the brick count is altered, the agent maintains near-perfect performance. Since the policy relies on relational dynamics (alignment and collisions) rather than fixed-pixel layouts, it remains scale-invariant. Importantly, the symbolic agent is evaluated in these modified environments without any additional training, using the policy learned on the original configuration.
- Visual Invariance: Altering object colors has zero impact on performance, as the symbolic extractor operates on physical attributes rather than raw RGB values.
- Cross-Domain Applicability: The same framework is applied to Pong. While learning is slower due to sparser interaction events, the agent shows a steady increase in survival time, confirming that the causal reward mechanism is indeed domain-agnostic. Over the course of a single training run, the agent's average survival time increases from 1.2 seconds to 13.3 seconds (computed from an environment working at 60 frame per second), including over 5000 episodes. This survival time demonstrates that the intrinsic reward signal is sufficient to drive the emergence of stable and progressively improving behaviors, even without external supervision or handcrafted objectives.
- Knowledge Transfer: In a preliminary experiment, initializing the framework with the object classes learned in the Pong domain enables it to reuse the abstractions for the walls and the ball in

the Arkanoid environment. The system preserves the compatible classes and their associated behavioral rules, while creating new classes for domain-specific entities, such as the bricks and the paddle. A distinct class represents the latter because its movement follows different control-dependent rules. This qualitative result suggests that the symbolic representation supports the reuse of previously acquired knowledge across structurally related domains.

The results highlight a "revealing paradox" in reverse: while standard DRL is superhuman yet brittle, our neuro-symbolic approach is competent and robust. By grounding reinforcement learning in a symbolic world model, we achieve an agent that does not just "react" to pixels but captures explicit causal relationships between its actions and their consequences. This achievement demonstrates that structured intrinsic motivation is sufficient for mastering complex tasks and offers a clear path toward more transparent and generalizable AI.

4.6 Limitations

Our preliminary results suggest that our structured approach improves interpretability, robustness, and explainability in dynamic environments. However, there are a few areas where the current framework can be further refined.

First, the perceptual module targets simplified 2D arcade environments with limited visual complexity. Extending this approach to high-resolution and cluttered scenes requires a more robust segmentation and tracking system and may challenge the current multi-hypothesis mechanism. Second, the symbolic rule language focuses on affine updates of a small set of kinematic properties, which is sufficient for Atari-like physics but may be inadequate for richer domains involving continuous forces, partial observability, or stochastic dynamics. Third, the intrinsic reward design, while domain-agnostic in principle, depends on a careful balance between causal impact, event density, and curiosity. Our findings suggest that certain shaping configurations may still lead to unintended reward-hacking behaviors. Finally, the experiments primarily target short-horizon control. The current system lacks context-based reasoning, relies on a still immature knowledge base, and may not scale well as the number of hypotheses grows. Moreover, integrating explicit causal reasoning and explicit planning into the learned symbolic model remains an open research question.

5 CONCLUSIONS AND FUTURE WORKS

This work demonstrates that "interpretability" and "performance" are not mutually exclusive, but rather complementary pillars of robust AI. By actively constructing explicit, human-readable world models from raw video, our framework achieves a degree of robustness and generalization that often eludes standard "black-box" models. Through the lens of Core Knowledge theory, we have shown that grounding reinforcement learning in symbolic entities and causal rules allows agents to build an explicit symbolic world model of their environments rather than merely reacting to pixel-level patterns.

Our experimental evaluations on Arkanoid and Pong underscore several key strengths of the neuro-symbolic approach:

- Interpretability: The decision process is transparent, allowing behaviors to be traced through explicit event transitions and causal chains.
- Structural Robustness: The agent maintains near-optimal performance even under significant environmental perturbations, such as changes in object size or brick configuration, without requiring additional training.
- Zero-Shot Generalization: Because the system operates on symbolic physical attributes rather than raw RGB values, it remains entirely invariant to visual changes, such as color or texture shifts.
- Autonomous Learning: By leveraging a composite intrinsic reward—based on causal impact and event density—the agent acquires complex skills without the need for task-specific external rewards or human-designed game scores.

While the results are encouraging, the current framework serves as a foundational step with several areas identified for refinement and already described in Section 4.6:

- Perception: The current U-Net-based module is targeted at simplified 2D environments. Future work will focus on enhancing the perception pipeline to handle the high-resolution, cluttered, and occluded scenes typical of more complex domains.
- Rule Representations: The current symbolic language focuses on affine updates to kinematic properties. We plan to explore richer representations that capture continuous forces, stochastic dynamics, and partial observability.
- Model-Based Planning: While the current agent uses the symbolic model to inform its state space,

we aim to transform the system into a fully model-based agent by integrating explicit, long-horizon planning over the learned symbolic world model.

- Knowledge Transfer: We intend to develop meta-learning mechanisms that enable the agent to align entities and reuse its established knowledge base (e.g., object classes and causal laws) across entirely new, yet functionally similar, domains.
- Normalization techniques: Further improvements could be achieved by using reward normalization, preserving the approach's benefits while improving stability.

Ultimately, this research provides a generalizable alternative to dominant AI paradigms, suggesting that structured intrinsic motivation and symbolic reasoning are sufficient for mastering complex tasks while ensuring the resulting systems remain human-understandable and stable under pressure.

REFERENCES

Besold, T. R., d'Avila Garcez, A., Bader, S., Bowman, H., Domingos, P., Hitzler, P., Kühnberger, K., Lamb, L. C., Lima, P. M. V., de Penning, L., Pinkas, G., Poon, H., and Zaverucha, G. (2021). Neural-Symbolic Learning and Reasoning: A Survey and Interpretation. In *Neuro-Symbolic Artificial Intelligence: The State of the Art*, volume 342 of *Frontiers in Artificial Intelligence and Applications*, pages 1–51. IOS Press.

Bottino, A., Garbo, A., Loiacono, C., and Quer, S. (2016). Street viewer: An autonomous vision based traffic tracking system. *Sensors*, 16(6).

Buesing, L., Weber, T., Racaniere, S., et al. (2018). Woulda, Coulda, Shoulda: Counterfactually-Guided Policy Search. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Cabodi, G., Camurati, P., and Quer, S. (1994). Full Symbolic ATPG for large Circuits. In *ITC'94: IEEE International Test Conference*, pages 980–988, Washington, District of Columbia.

Cabodi, G., Camurati, P., and Quer, S. (1999). Improving symbolic traversals by means of activity profiles. In *Proceedings 1999 Design Automation Conference*, pages 306–311.

Calabrese, A., Quer, S., Squillero, G., and Tonda, A. (2024). Towards an Evolutionary Approach for Exploting Core Knowledge in Artificial Intelligence. In *GECCO '24 Companion: Proceedings of the Genetic and Evolutionary Computation Conference Companion*, pages 259–262, Melbourne, VIC, Australia.

Greff, Klaus and van Steenkiste, Sjoerd and Schmidhuber, Jürgen (2020). On the Binding Problem in Artificial Neural Networks. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Kassa, Yared and others (2023). Object-Centric Learning with Deep Reinforcement Learning. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., and Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40:e253.

Liang, B., Wang, Y., and Tong, C. (2025). Ai reasoning in deep learning era: From symbolic ai to neural-symbolic ai. *Mathematics*, 13(11):1707.

Marcus, G. (2018). Deep learning: A critical appraisal. *arXiv preprint arXiv:1801.00631*.

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

Oudeyer, P.-Y., Kaplan, F., and Hafner, V. V. (2007). Intrinsic Motivation Systems for Autonomous Mental Development. *IEEE Transactions on Evolutionary Computation*, 11(2):265–286.

Pathak, Deepak and Agrawal, Pulkit and Efros, Alexei A. and Darrell, Trevor (2017). Curiosity-driven Exploration by Self-supervised Prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

Pearl, J. (2009). Causality: Models, reasoning and inference. *Cambridge University Press*.

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, volume 9351 of *Lecture Notes in Computer Science*, pages 234–241. Springer.

Spelke, E. S. and Kinzler, K. D. (2007). Core knowledge: Cognitive principles and psychological foundations. *Developmental Science*, 10(1):89–96.