

Neural Networks Discover Physical Principles from Observation with an Intuitive World Model

Xiangyu Zhu

`xiangyu.zhu@nlpr.ia.ac.cn`

Institute of Automation, Chinese Academy of Sciences

Qu Tang

Institute of Automation, Chinese Academy of Sciences

Haoyuan Zhang

Institute of Automation, Chinese Academy of Sciences

Zhen Lei

Institute of Automation, Chinese Academy of Sciences

Zhaoxiang Zhang

Institute of Automation, Chinese Academy of Science

Article

Keywords:

Posted Date: May 7th, 2025

DOI: <https://doi.org/10.21203/rs.3.rs-3744228/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: There is **NO** Competing Interest.

Neural Networks Discover Physical Principles from Observation with an Intuitive World Model

Qu Tang^{1,2†}, Xiangyu Zhu^{1,2†}, Haoyuan Zhang^{1,2}, Zhen Lei^{1,2,3*} and Zhaoxiang Zhang^{1,2,3*}

¹School of Artificial Intelligence, University of Chinese Academy of Sciences.

²MAIS, Institute of Automation, Chinese Academy of Sciences.

³Centre for Artificial Intelligence and Robotics, Hong Kong Institute of Science & Innovation, Chinese Academy of Sciences.

*Corresponding author(s). E-mail(s): zhaoxiang.zhang@ia.ac.cn; zlei@nlpr.ia.ac.cn;

[†]These authors contributed equally to this work.

Abstract

Humans readily grasp physical principles via visual observation, underpinning our ability to model the world and reason about events. In contrast, AI systems often fail to extract consistent physical rules from video, relying on black-box mechanisms that produce implausible outcomes. We introduce the Intuitive World Model (IWM), designed to learn physical principles directly from vision, mimicking human intuition. Drawing inspiration from developmental psychology, IWM utilizes a modular architecture to decompose interactions into force-based dynamics. Through observing objects collide and interact via charges, it identifies intrinsic properties like mass and charge and learns representations analogous to Newton's and Coulomb's laws. This paradigm uniquely allows neural networks to provide explicit explanations for physical phenomena and conduct counterfactual reasoning consistent with learned physical laws. Our work demonstrates that a developmentally-inspired approach empowers AI to acquire explicit, interpretable physical knowledge from observation, mirroring a core aspect of human intelligence.

From early childhood, infants begin to make sense of their surroundings through visual input, laying the foundation for their cognitive and intellectual development [1, 2, 3, 4, 5, 6]. They effortlessly develop the ability to predict the trajectory of a thrown ball and understand cause-and-effect relationships, such as the outcomes of collisions between objects [7]. These intuitive understandings of the physical world are deeply embedded in human cognition and play a crucial role in everyday interactions. Such insights highlight the necessity of enabling artificial intelligence (AI) systems to emulate human-like intuitions

and autonomously learn fundamental physical principles through visual perception [8, 9].

Cognitive research has revealed that the physics knowledge possessed by humans is intricately organized into structured intuitive theories [10, 11, 12, 13]. These theories encompass a comprehensive ontology of concepts and a coherent system of causal laws that govern the relationships between these concepts, exhibiting similarities to scientific theories. A crucial characteristic of intuitive theories lies in their explanatory capacity [10, 14]. Instead of merely depicting events, these theories interpret the

available evidence within their unique frameworks, providing insights and comprehension that extend beyond shallow observations. For example, in a physics scenario involving a collision between two objects, entailing an exchange of momentum and energy, a comprehensive collision theory should encompass not only observable factors such as position and velocity information but also fundamental concepts such as mass and force. By incorporating these underlying factors, the theory facilitates a deep analysis of the dynamic alterations taking place during the collision. Despite significant efforts, recent AI advancements have yet to produce a system that is capable of authentically replicating humans' ability to learn and apply intuitive theories. While deep learning offers advanced solutions for predicting future states in the physical domain, exemplified by its capability to forecast the stability of stacked blocks [15, 16] and the trajectories of objects [17, 18, 19], its ability to provide comprehensive explanations is typically limited, due to its inherent lack of interpretability [20]. On another note, while numerous models leverage built-in Newtonian principles to improve their explanatory abilities to a certain extent [21, 22, 23], establishing physical models purely through visual experiences, akin to the learning process employed by infants, remains challenging. Therefore, we advocate for the development of an AI system that is capable of discovering physical properties and principles from raw observations. This system should offer comprehensive explanations of physical phenomena that can be represented through, for example, a force diagram.

Our IWM is inspired by developmental psychology, reflecting key insights into how infants learn and adapt. First, when confronted with compelling evidence that challenges their existing understanding, infants are driven to develop new intuitive theories, which provide interpretable frameworks to explain their observations [24, 25]. Second, this process involves refining their predictions about future events, akin to how world models address uncertainty by learning latent variables [26, 27]. Third, instead of broadly generalizing across event types, infants acquire distinct theories for each category, building knowledge step-by-step [7]. Central to this is the idea that, through increasingly diverse visual experiences,

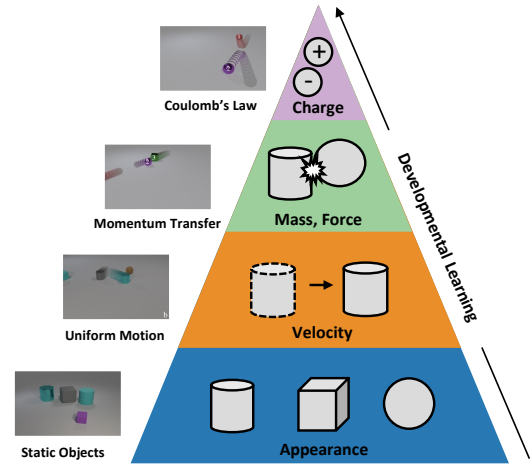


Fig. 1: Illustration of the data organization scheme in the hierarchical learning paradigm. From the bottom to the top: 1) static images showcasing the objects' appearances, 2) sequential frames depicting the linear motions of individual objects, 3) videos featuring impulsive collision forces between objects, and 4) videos involving continuous Coulomb forces between objects.

infants progressively discover novel intuitive physical principles [3, 28].

We assess our methodology within a context where multiple objects move and interact on a flat surface [29]. This dynamic environment includes two primary types of interactions. Firstly, there are collisions, the nature of which depends on the mass and velocity of the involved objects. Secondly, there are electrostatic interactions, influenced by the objects' electric charges. Our aim is to enable the model to discover fundamental properties of mass and charge, and their corresponding laws, including Newton's laws and Coulomb's law. Following this idea of hierarchical and category-specific learning from visual experiences, we propose a hierarchical organization of videos of physical events, encompassing progressive levels of object dynamics, as illustrated in Fig. 1. At each hierarchical level, IWM incrementally learns physics knowledge and develops an understanding of increasingly abstract concepts. At the lowest level, data consist of distinct static images, with a perception module trained to capture extrinsic object properties such as appearance, size, and spatial location. At the second level, data are organized as sequential frames, depicting uniform linear motion. A dynamic module is introduced to model how velocity influences extrinsic attributes,

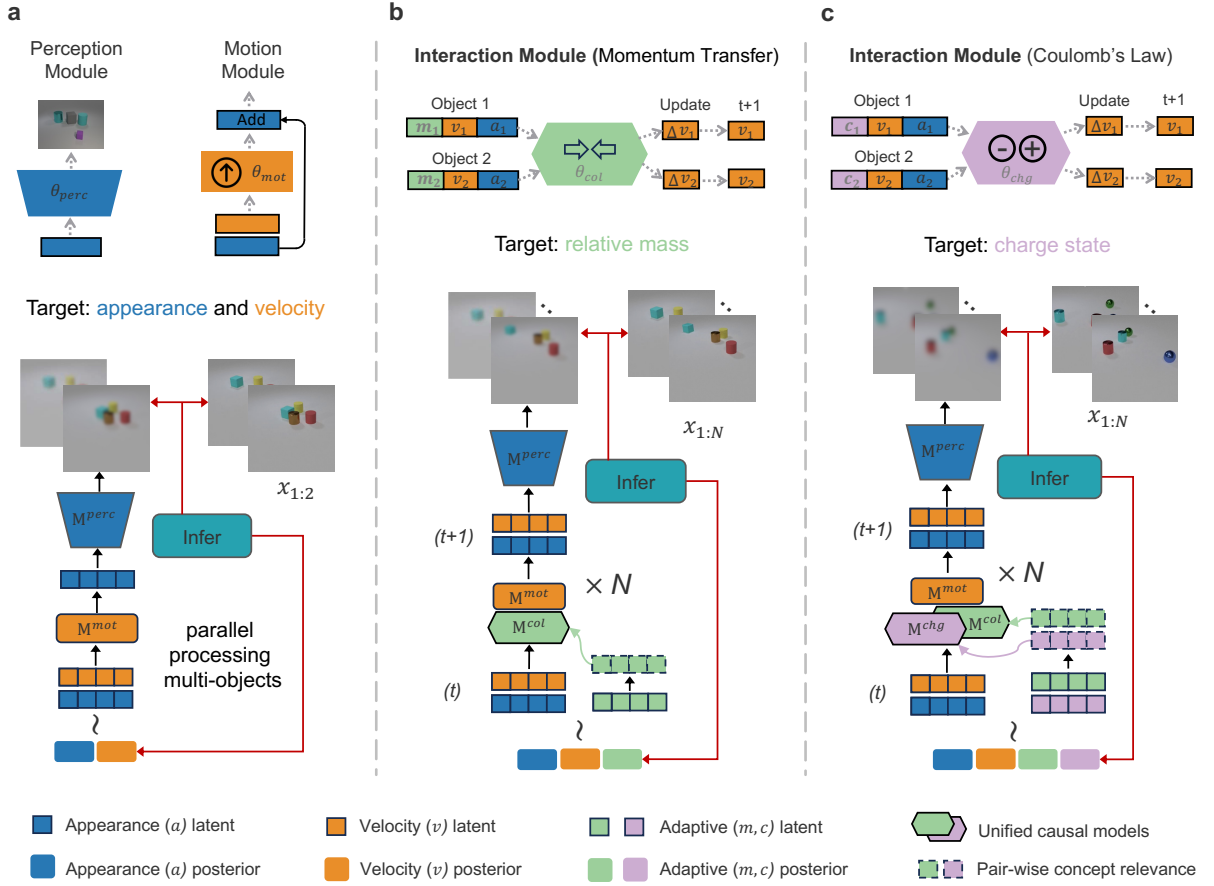


Fig. 2: An overview of the workflow. **a**, Depicts the perception and motion modules. The perception module (M^{perc}) infers appearance latents (a) from input images. The motion module (M^{mot}) takes appearance and velocity latents (v) to predict subsequent appearance. These modules operate in parallel on multiple objects. **b**, Illustrates the interaction module under momentum transfer (collision) dynamics. The module (M^{col}) takes as input the appearance, velocity, and pair-wise mass relevance to predict updated velocities after a collision. The target for inference in this scenario is relative mass. **c**, Illustrates the interaction module under Coulomb's (electrostatic force) dynamics. The module (M^{chg}) takes as input appearance, velocity, and pair-wise charge relevance to predict updated velocities. The target for inference is the charge state.

enabling short-term predictions. At the third level, videos illustrate dynamic changes driven by impulsive forces from object collisions. This requires a deeper understanding of interaction dynamics and longer-range predictions, enabling the model to infer intrinsic concepts of mass and force. At the final level, the physical events become increasingly complex, characterized by continuous Coulomb forces that vary dynamically based on the objects' charge and position states. The model is thus encouraged to comprehend a force that continuously changes, promoting an

understanding of the intrinsic concept of charge. This hierarchical paradigm enables the model to progressively acquire knowledge, moving from extrinsic to dynamic and ultimately to intrinsic physical properties. IWM is designed to interpret and reconstruct the video, ensuring that the physical properties and causal laws are uniquely incorporated within its four modular levels. Notably, IWM operates end-to-end, learning directly from raw videos without relying on predefined physics engines or Newtonian mechanics.

IWM’s hierarchical learning paradigm endows it with two fundamental capabilities concerning intuitive physics. First, IWM is able to explain the dynamic alterations resulting from the collaborative effects of collisions and Coulomb forces, reflecting its ability to infer underlying physical principles. This approach resonates with the viewpoint of developmental psychology, underscoring that infants organically formulate causal modules to explain why and how certain events happen [30, 31, 10]. Second, through the integration of physical properties and causal laws, IWM is able to imagine potential results by modifying physical variables that differ from those existing in the observed scenario. This capability demonstrates the model’s ability to learn and apply causal principles, enabling it to facilitate the explanation of how the model formulates causal judgments and explore the physical configurations required to produce specific events [10, 32].

Intuitive World Model

The IWM is designed to achieve two objectives for interpreting physical videos: generating object-centric and physically disentangled representations, and constructing causal modules to describe the progression of physical events. The model incorporates pair-wise concept relevance (denoted as r), which captures relationships between emergent concepts of interacting objects, such as relative mass or electrostatic repulsion/attraction. This allows for adaptive modeling of the physical interactions. Given an input video sequence $\{x^1, \dots, x^t\}$, the IWM generates an object-centric representation for each of the K objects involved in the physical events, denoted as $o^t = \{o_1^t, \dots, o_k^t\}$ in frame t . Each representation encapsulates the disentangled physical properties of a specific object, represented as $o = \{a, v, i\}$, where a , v , and i correspond to appearance, velocity, and adaptive intrinsic properties (like mass m or charge c), respectively. These adaptive properties are then used to derive the pair-wise concept relevance (r). Furthermore, based on these representations and the pair-wise relevance, the IWM constructs causal modules to explain the dynamics and interactions of objects, enabling future predictions as $o^{t+1} = M(o^t)$. The learning paradigm that

achieves these two objectives unfolds in a structured hierarchy with three stages, as illustrated in Fig. 2.

Perception and Motion. In the first stage (corresponding to Fig. 2a), the model learns to perceive the appearance and motion of objects. This involves two sub-modules:

- **Perception Module (M^{perc}):** Extracts object-level appearance features from images.

$$x = M^{perc}(\{a_1, \dots, a_k\}), \quad (1)$$

where M^{perc} searches for a set of object-level appearance features $a_1, \dots, a_k$ that can best reconstruct the original image x .

- **Motion Module (M^{mot}):** Predicts the next-frame appearance based on current appearance and velocity.

$$a_k^{t+1} = M^{mot}(a_k^t, v_k^t). \quad (2)$$

For the k th object, M^{mot} uses current appearance a_k^t and velocity v_k^t to predict the appearance in the next time step, a_k^{t+1} .

During training, the iterative amortized inference framework [33] is used to optimize the parameters of both M^{perc} (θ_{perc}) and M^{mot} (θ_{mot}). The motion module collaborates with the perception module, using the interpreted appearance features $\{a^t, a^{t+1}\}$ to infer the velocity v^t . At the end of this stage, the object representation is $o = \{a, v\}$. **Momentum Transfer Interaction.** In the second stage (corresponding to Fig. 2b), the model learns to simulate interactions based on momentum transfer (collisions).

$$v_{1,\dots,k}^{t+1} = M^{col}(a_{1,\dots,k}^t, v_{1,\dots,k}^t, r_{1,\dots,k}^m). \quad (3)$$

Here, the subscripts $1, \dots, k$ denote all k objects in the scene, and $r_{1,\dots,k}^m$ represents the pair-wise concept relevance related to mass for all object pairs. M^{col} learns the physics of collisions, utilizing the pair-wise mass relationships to predict the updated velocities $v_{1,\dots,k}^{t+1}$. The object representation can be considered as $o = \{a, v, m\}$.

Coulomb’s Interaction. In the third stage (corresponding to Fig. 2c), the model learns to simulate interactions based on Coulomb’s Law (electrostatic forces).

$$v_{1,\dots,k}^{t+1} = M^{chg}(a_{1,\dots,k}^t, v_{1,\dots,k}^t, r_{1,\dots,k}^c). \quad (4)$$

Here, $r_{1,\dots,k}^c$ represents the pair-wise concept relevance related to charge for all object pairs. M^{chg} learns how like charges repel and opposite charges attract, predicting the updated velocities $v_{1,\dots,k}^{t+1}$ based on the pair-wise charge relationships and initial states. The object representation can be considered $o = \{a, v, c\}$. A more general representation for the object is $o = \{a, v, i\}$, where i represents the adaptive properties.

Results

In this section, we evaluate IWM using the ComPhy dataset [29]. This dataset consists of videos with objects such as cylinders, spheres, and cubes interacting on a flat surface. These videos capture various dynamics, including collisions and Coulomb attractions and repulsions, typically featuring 3-5 objects per video. ComPhy distinguishes itself through its challenge of understanding intrinsic properties such as mass and charge, which cannot be observed from static appearances. This complex comprehension task, especially when addressing the continuous nature of Coulomb forces, makes it an ideal candidate for evaluating our learning paradigm.

Casual Reasoning Performance

We first test IWM in the causal reasoning task, which encompasses video question answering related to physical events; this is a common approach for evaluating a model’s understanding of intuitive physics [41, 42]. This task requires the model to respond to questions related to the physical events, typically in a multiple-choice format. The ComPhy dataset contains three types of causal reasoning questions related to mass and charge: factual, predictive, and counterfactual questions. Factual questions assess the relative masses or charges of two objects in a video. Predictive questions forecast upcoming events, such as potential collisions, while counterfactual questions explore what might have happened under altered initial conditions. As an example, consider a scenario where an initial collision occurs between a sphere and a cylinder, subsequently resulting in the cylinder colliding with a cube. In this context, the posed question could be: "What would have happened if the sphere were heavier?", and the model should choose from three answers,

including "A: The cube would have collided with the sphere.", "B: The red object would have collided with the cylinder.", and "C: The sphere would have collided with the metal object."

We compare IWM with a series of state-of-the-art methods. Among them, CNN-LSTM [34], HCRN [35], and MAC [36] are multimodal approaches that enable holistic reasoning. They utilize CNN backbones to extract visual features, creating video embeddings. However, they do not investigate object-centric representations or physical properties. On the other hand, the object-centric models focus solely on encoding visual scenes, including SAVi [37], SlotFormer [19], MoNet [38], and IODINE [39]. They process single images into K slots, serving as object-centric representations for question answering. We further enhance IODINE by incorporating graph neural networks (GNNs) [40] to enable the modeling of interactions and the inference of intrinsic physical properties in an implicit manner; this variant is denoted as IODINE-GNN. To achieve question answering, questions and answer choices are converted into word embeddings, and the given video is transformed into an embedding using these candidate models. Subsequently, a cross-modal transformer [43] obtains its inputs from these three embeddings and employs a classification head to generate the answer. Another noteworthy physics-based method is CPL [29], a high-performance symbolic model. However, it is distinctive in that its physical properties, i.e., mass and charge, are trained with ground truth labels, setting it apart from the comparison methods discussed here. Furthermore, to evaluate the contribution of IWM, we assess both the full models and submodels trained at various stages, with different learned physical properties.

In the factual questions evaluation, we find that both CNN-LSTM and HCRN exhibit sub-optimal performance, while MAC demonstrates better performance due to its compositional attention mechanism. IWM stands out as the most effective model, surpassing the performance of other methods. For the predictive questions, which require an understanding of dynamics for accurate forecasting, IWM achieves the highest accuracy among all state-of-the-art models. IWM also shows notable advancements in counterfactual tasks, where embodying human-like intuition to

Table 1: Causal reasoning accuracy comparison conducted on ComPhy with the state-of-the-art models.

Models	Physics	Object	Factual	Predictive		Counterfactual	
				per-opt	per-que	per-opt	per-que
CNN-LSTM [34]	$\times$	$\times$	41.9	59.6	29.4	57.2	12.8
HCRN [35]	$\times$	$\times$	47.3	62.7	32.7	58.6	14.2
MAC [36]	$\times$	$\times$	65.8	60.2	30.7	60.3	14.3
SAVi [37]	$\times$	$\checkmark$	51.2	59.4	26.1	65.0	22.2
SlotFormer [19]	$\times$	$\checkmark$	50.0	60.5	31.1	63.9	21.7
MoNet [38]	$\times$	$\checkmark$	57.7	67.9	37.1	67.9	22.2
IODINE [39]	$\times$	$\checkmark$	58.9	64.8	35.6	67.8	25.0
IODINE-GNN [39, 40]	$\checkmark$	$\checkmark$	56.6	66.1	36.3	71.7	32.1
IWM ^a	$\checkmark$	$\checkmark$	59.0	64.6	35.0	67.8	24.9
IWM ^{av}	$\checkmark$	$\checkmark$	60.2	67.1	38.8	69.6	28.4
IWM ^{avm}	$\checkmark$	$\checkmark$	66.0	70.5	43.1	73.2	33.0
IWM ^{avmc(all)}	$\checkmark$	$\checkmark$	69.2	72.6	47.5	73.6	33.3

The evaluation metric used is the question answering accuracy, with the best results highlighted in bold. The ‘Physics’ column indicates whether the model incorporates physical information, while the ‘Object’ column signifies whether the scene is represented in an object-centric manner. The following columns are three question types: ‘Factual’, ‘Predictive’, and ‘Counterfactual’. Here, ‘Predictive’ and ‘Counterfactual’ are multiple-choice questions. For the multiple-choice questions, accuracy is measured both per-option (per-opt) and per-question (per-que). In the case of our method, IWM, the superscripts denote the interpreted properties used for reasoning, which include appearance (a), velocity (v), mass (m), and charge (c).

represent physical scenes is essential. In this scenario, the CNN-based models demonstrate limited performance (60.3), and the object-centric models perform better, with a notable improvement from IODINE to IODINE-GNN (67.8 to 71.7). The contrast between the CNN-based and object-centric models highlights the benefits of object-level representations, with the incorporation of object dynamics into IODINE-GNN yielding further improved accuracy. IWM achieves the best performance value of 73.6, which validates that factorizing the entity-specific and rule-based properties of a , v , m , and c is crucial for physical reasoning.

An ablation analysis illuminates the beneficial effects of acquiring individual physical properties at distinct stages. This study involves four properties, each corresponding to one of the four training stages. We assess the performance achieved with individual properties as well as their combined effect. It is evident that with each successive learning stage, our model acquires a deeper understanding of physics, resulting in consistent performance enhancements. The peak performance is reached when all properties are seamlessly integrated.

The results reveal that the acquisition of the physical representations enhances the performance achieved in downstream causal reasoning tasks. Though IWM delivers state-of-the-art performance, it is crucial to underscore that the main focus of IWM is not question answering. In the upcoming experiments, we will assess IWM’s ability to offer explicit physical explanations and conduct counterfactual inference, which are capabilities that the current methods lack.

Learned Physical Principles

IWM’s main purpose is to uncover physical principles from raw observations. It achieves this by employing causal modules that capture the interactions among objects, accounting for physical properties like appearance a , velocity v , mass m , and charge c . The model predicts acceleration Δv and updates velocity v (both two-dimensional vectors). Given a video, IWM optimizes its internal interpretation of physical events until it accurately reproduces the input. We can then use the derived v and Δv values to create a force diagram, illustrating the physical event.

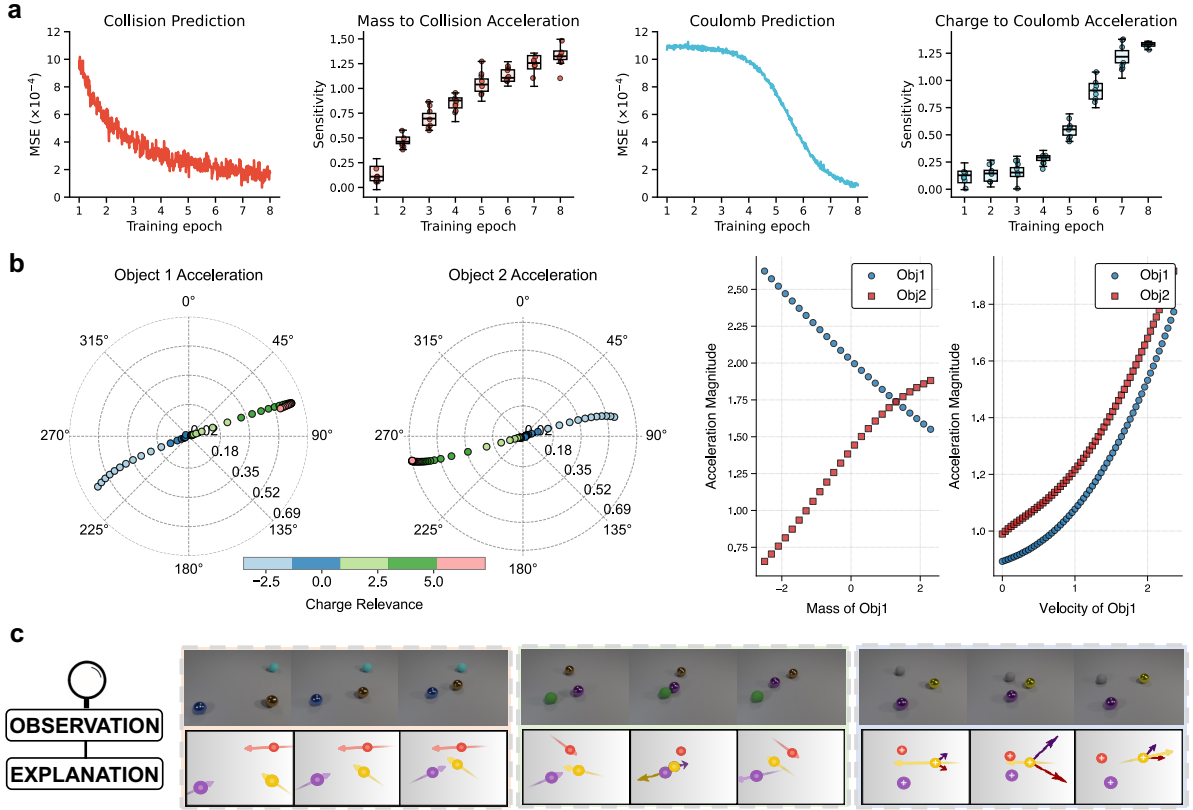


Fig. 3: IWM's learned physical principles and explanatory capabilities. (a) Learning progression of physical principles for predicting collision and Coulomb interactions, illustrated by a decreasing mean squared error (MSE) and an increasing sensitivity of acceleration to changes in mass (left) and charge (right). (b) Analysis of learned physical principles. Left: Polar plots showing the influence of "Charge Relevance" on the acceleration (direction and magnitude) of two interacting objects. Right: Acceleration magnitude changes as a function of mass and velocity of one object during a collision. (c) Visual explanations of physical events. Each column shows an "Observation" (video frame), and a simplified "Explanation" derived from the IWM. The examples illustrate understanding of (left to right) inertia (Newton's First Law), momentum transfer during collisions, and Coulombic repulsion.

In Fig. 3.a, we explore the progression of acquired physical knowledge during training. For mass m , at each training snapshot, we examine the causal module by manipulating m and observing the resulting acceleration (Δv). We then compute the ratio between the change in m and the change in Δv , quantifying mass's influence on the collision acceleration. The training MSE is plotted concurrently, showing convergence. Initially, mass alterations have minimal effect on force, as the causal module hasn't yet captured the relationships between the adaptive latent variables and the collision force. As training advances and reconstruction loss decreases, a connection between the adaptive latent variables

and collision is established, and the influence becomes more pronounced. Eventually, the model effectively reproduces the observed data in its intuitive world (converged training loss). A similar pattern is observed between the charge latent variable and Coulomb's acceleration. These findings demonstrate that IWM continually refines its physical laws by better reproducing observed data throughout training.

We manually adjust physical variable values fed into the causal modules and observe the predicted acceleration. Fig. 3.b (left) uses two polar plots to illustrate how "Charge Relevance" affects the acceleration of two interacting objects

(Obj1 and Obj2). The angle in each plot represents acceleration direction, while radial distance signifies magnitude. As Charge Relevance increases from -2.5 to 5.0, both the direction and magnitude of acceleration change for each object. This demonstrates the model has learned the relationship between the charge relevance and repulsion/attraction, analogous to Coulomb's Law, where changes in charge influence both the strength and direction of interactions. Fig. 3.b (right) analyzes the collision module. By keeping velocity constant and varying the mass of one object from small to large, we observe that the acceleration of that object decreases, while the acceleration of the other object increases. Similarly, with constant masses and varying input velocity of one object from low to high, both objects experience heightened interaction forces. This indicates the causal module grasps the fundamental principles of momentum transfer.

IWM's ability to explain physical events stems from the intuitive theories acquired by the causal modules concerning collisions and charges, which approximate intuitive physical principles. Fig. 3.c visually demonstrates this explanatory ability by comparing observed video frames ("Observation") with a simplified explanation ("Explanation") derived from the IWM. The first column depicts uniform linear motion. The 'Observation' frame shows two objects and the "Explanation" reinforces this with straight, equal-length arrows, demonstrating understanding of inertia (Newton's First Law). The second illustrates a collision between objects of different masses: "Observation" shows objects mid-collision; "Explanation" highlights these velocity changes, showcasing a grasp of momentum transfer. The third column illustrates a scenario involving three positively charged objects experiencing mutual Coulombic repulsion. The "Observation" panel displays the objects. The IWM represents this interaction with acceleration vectors emanating outwards from each object, indicating repulsion; the magnitude of these vectors (represented by arrow length) is inversely related to the inter-object distances. The "Explanation" panel depicts outward-pointing arrows on each object, confirming that the model understands Coulombic repulsion and that the net acceleration on each object results from the superposition of multiple repulsive forces. These examples effectively bridge the gap between raw

visual input and the underlying physical principles IWM has learned.

Counterfactual Inference

Consistent with the findings concerning the intuitive theories of human infants, IWM can also perform counterfactual inference based on its acquired theories. In this section, we explore IWM's ability to address counterfactual situations. Fig. 4 illustrates the model's capability for conducting explainable counterfactual reasoning. It demonstrates how physical events unfold when the initial conditions deviate from real observations (depicting "what if..."). This includes predicting the likelihood of events, providing rationales for why certain outcomes are feasible, and delineating strategies for achieving particular results.

We begin by evaluating IWM's capacity for counterfactual imagination by presenting it with "what if" questions and examining the generated counterfactual videos. In Fig. 4.a, a gray sphere collides with a heavier green cylinder, causing the sphere to rebound. Based on this observed input, IWM infers that the mass relevances of the two objects are $r_1^m = -0.5$ and $r_2^m = 1.8$, respectively. For this scenario, we present the first question, "What if the mass relevances of the two objects were swapped?", instructing the model to exchange the interpreted mass relevances and regenerate the physical event as a video. Specifically, we manually set $r_1^m = 1.8$ and $r_2^m = -0.5$ while keeping the other physical properties unchanged, and then we make predictions from the first frame and generate the resultant video with the decoder. The counterfactual video in the second row of Fig. 4.a shows that the brown sphere undergoes a decreased displacement after the collision due to an increase in its mass relevance, while the green cylinder changes its trajectory and moves downward to the right due to its lower mass relevance. This imaginative outcome aligns with our intuition. Next, we examine a scenario that goes even further. We ask "What would happen if the cylinder becomes even lighter?", and manually set $r_1^{m'} = 1.8$, $r_2^{m'} = -2$. The imagination process indicates that the cylinder would shift further in the downward and to the right, which also aligns with our intuitive theory. Significantly, in this scenario, the tested

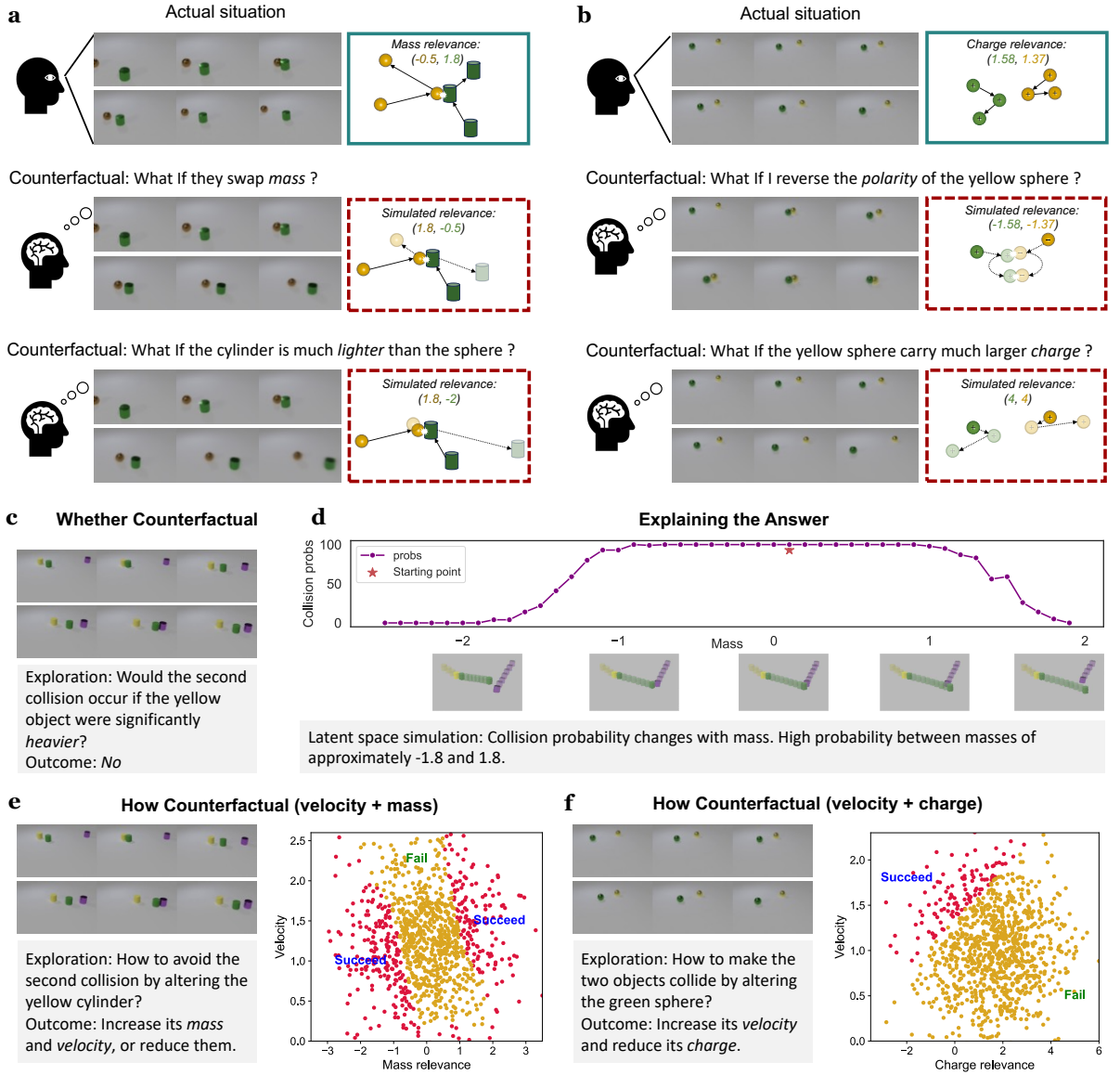


Fig. 4: Counterfactual inference and explanation using IWM. (a) Illustration of counterfactual imagination by manually adjusting inferred mass values in a collision scenario, demonstrating the impact on predicted outcomes. (b) Counterfactual imagination applied to charge and Coulombic interactions (attraction and repulsion), showing the effects of altering the charge relevance of objects. (c) Example of a “whether” counterfactual query, exploring the possibility of a specific future event under a modified configuration. (d) Explanation process for the “whether” judgment in (c), using IWM as an intuitive physics engine to simulate outcomes under different configurations. (e) Left: Example of a “how” counterfactual query, investigating solutions to achieve a desired outcome in a collision. Right: IWM’s search results, showing possible configurations satisfying the query’s constraints. (f) Example of a “how” counterfactual query in a scenario combining Coulombic forces and potential collisions.

mass relevance ($r_2^{m'} = -2$) exceeds the value distribution of the training dataset, showcasing the model’s strong generalization capabilities.

In Fig. 4.b, we examine a repulsion event in which two spheres, both carrying like charges, approach each other and are pushed apart by

Coulomb’s force before colliding. For this event, IWM infers that $r_1^c = 1.58$ and $r_2^c = 1.37$. We present an initial question: “What would happen if we reversed the polarity of the yellow sphere?” As depicted in the second row of Fig. 4.b, by manually setting $r_1^c = -1.58$ and $r_2^c = -1.37$, IWM imagines that the repulsion would turn into attraction, leading to a collision. After the collision, IWM predicts that the objects would move apart and be subsequently attracted back to each other due to the continuous attractive force. This outcome illustrates the model’s solid understanding of charges and the combination of Coulomb and collision forces. Moreover, the model can also make out-of-distribution predictions for the charge scenarios. We increase the charge relevance of the two objects to 4.0, which is beyond the range observed in the training data. The result presented in the third row of Fig. 4.b shows that the objects are pushed further apart, which also aligns with human intuition.

Another application of counterfactual imagination is to provide an explanation for a counterfactual judgment, moving beyond simple multiple-choice question answering in previous methods [34, 35, 36, 29]. As an illustration, consider Fig. 4.c, which presents a classic counterfactual judgment question: “If the yellow object were heavier, would the green cylinder collide with the purple cylinder?” A traditional answer in previous methods is “Yes”. IWM goes beyond this by not only predicting the probability of the suggested event but also providing an explanation for the underlying reasons behind this judgment. To achieve this aim, we treat IWM’s causal modules as a noisy physical engine. We introduce noise from a normal distribution to the inferred representations of physical properties and utilize the causal modules to predict collisions in counterfactual scenarios. This process is repeated multiple times (100 times for each mass relevance value), allowing us to approximate the collision probability. In addition, the video of each trial can also be provided as an explanation for this judgment. As shown in Fig. 4.d, in the actual situation (marked as ‘starting point’), the perceived $r^m = 0.1$. When the mass relevance is increased, IWM finds the collision does not occur at a high probability when $r^m > 1.8$. This corresponds to a scenario where the velocity of the green object increases due to the heavier yellow cylinder, passing the purple

object without a collision. Moreover, leveraging this feature, IWM can expand the question by examining the outcome when the mass relevance of the yellow cylinder varies from small to large. As illustrated in Fig. 4.d, it is evident that at a small mass relevance, the green object gains a small amount of kinetic energy from the yellow object, consequently missing a further collision. As the mass relevance increases, the green object attains a greater amount of velocity from the yellow object, enabling it to reach the purple object. However, as the mass relevance continues to increase, the green object accelerates to a point where it moves too fast, ultimately causing it to bypass the purple object. This “miss-collision-miss” judgment offers a broader perspective on the correlation between mass relevance and collisions, surpassing the initial scope of the posed question.

The generative causal module enables IWM to solve the challenging “how counterfactual” problem, determining the initial configurations required to ensure a desired future event, which remains impossible for black-box models. Continuing with the scenario introduced in Fig. 4.c, we now pose the following question: “How can we prevent the collision between the green cylinder and the purple cylinder by altering the yellow object?” The original event occurs in the latter half of the sequence. To solve this question, IWM explores a range of values for the velocity and mass relevance of the yellow cylinder, generating counterfactual scenarios. The answer to this question is presented in the right part of Fig. 4.e, where any configuration corresponding to purple points can successively avoid the collision. Two possible approaches are evident: either increasing its mass relevance and velocity or reducing them. Fig. 4.f depicts an event in which two spheres with like charges are initially repelled from each other before colliding. To make the two objects collide, IWM suggests either increasing their initial velocities or decreasing their charge relevance, which is illustrated by the successful trials in the upper-left portion of the scatter plot.

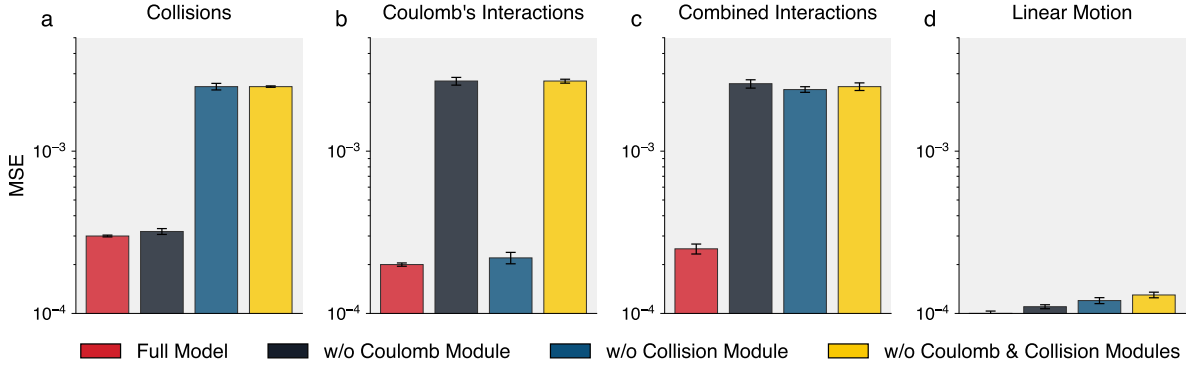


Fig. 5: Correspondence between causal modules and specific physical events. The test data are categorized into four types of physical events, evaluating the reproduction precision (MSE) of IWM and its two variants with specific causal modules disabled. 'w/' and 'w/o' denote 'with' and 'without', respectively. 'col' and 'chg' represent 'collision' and 'charge', respectively. Error bars represent the SEM of the MSE, calculated across $N = 3$ independent training and evaluation runs for each model configuration.

Disentanglement of Physical Concepts in IWM

Developmental psychology suggests that humans learn distinct concepts for different classes of physical events [7]. This ability to separate and understand individual physical factors is also crucial for achieving explicit and disentangled physical properties, which are necessary for explanation and counterfactual reasoning [10]. To validate the disentanglement of physical properties in IWM, we present a detailed analysis in Fig. 5, focusing on the causal modules tied to collisions and Coulombic interactions, and providing insights into their distinct functionalities. Specifically, we evaluate IWM's performance by selectively deactivating the collision and charge modules, individually and simultaneously, thus preventing the calculation of their respective forces.

In Fig. 5, we quantitatively evaluate IWM's accuracy in reconstructing input videos. The test data are divided into four groups based on the presence or absence of collisions and charged interactions. We systematically deactivate each causal module and evaluate the model on these four data categories. Disabling a causal module demonstrably degrades the interpretation of the corresponding physical events. For instance, in the first group (collision-only events), disabling the charge module has a marginal impact on performance; the MSE remains comparable to

the full model (3×10^{-4} vs. 2×10^{-4}). Conversely, deactivating the collision module induces a substantial performance decrease, resulting in an MSE of 2.5×10^{-3} . In the second group (charged objects only), disabling the charge module leads to significant deterioration, with the MSE increasing from 2×10^{-4} to 2.7×10^{-3} . Disabling the collision module, however, has little influence. In the third group (both collisions and charged interactions), disabling either the charge or the collision module significantly degrades performance. In the fourth group (neither collisions nor charged interactions), disabling either module has minimal impact, as linear motion does not necessitate force modeling.

These results demonstrate that IWM successfully establishes direct and disentangled connections between its causal modules and specific physical concepts. A model lacking the charge module can still accurately model collision forces but fails to capture Coulombic forces. Further removing the collision module restricts the model to capturing only linear motion. This disentanglement serves as the foundation for the explanation and counterfactual reasoning capabilities demonstrated previously.

Conclusion

The relations and interactions among objects play fundamental roles in our understanding of the physical world. These relationships are governed by a set of intuitive theories that

describe how objects behave and interact with each other. However, when it comes to analyzing and interpreting visual data, such as videos, these underlying physical laws often remain hidden in what can be considered "dark matter" within our everyday observations. In recent years, deep learning has become a powerful tool for identifying patterns and extracting significant insights from visual data. Nonetheless, a primary challenge encountered by deep learning methods is their restricted capability to offer clear explanations for their recognition and decision-making processes. To construct a system that is capable of acquiring intuitive physics akin to humans, we introduce IWM, a novel framework that merges the learning capabilities of deep neural networks with the foundational principles of intuitive theories to achieve explainability.

IWM is trained in a hierarchical learning paradigm. Initially, IWM decomposes a scene into object-centric representations. Subsequently, it trains causal modules to handle increasingly complex physical scenarios. Its progression from appearance to velocity and then to mass and charge, serves as a naive yet illustrative example of a self-evolving learning model that continually enhances its understanding of the physical world through the data it processes. IWM does not rely on built-in Newtonian laws or physical engines but is rather designed to autonomously extract and learn abstract physical principles directly from raw observational data. The constructed IWM approach has two significant abilities: 1) explaining what happened within a physical event through a series of force diagrams and 2) imagining counterfactual scenarios to predict what would have happened if the initial state had differed. Furthermore, our experiments demonstrate that distinct physical concepts are embedded within separate causal modules, enabling a thorough probing of the learned intuitive physical laws.

IWM possesses the potential to cover a broader range of physical concepts. Its self-supervised learning paradigm empowers it to adaptively delve into and discover intuitive physical concepts from observations. IWM has the potential to construct explainable world models grounded in explicit physical properties and causal laws, thereby enhancing its understanding of causality and the fundamental mechanisms underlying events.

Thus, it holds the promise of reinforcing the reliability of decision-making and planning.

Method

In this section, we provide the mathematical and computational details of IWM, an explanatory system that interprets physical events with intuitive theories. IWM has the capacity to infer both extrinsic and intrinsic physical properties as well as learn their corresponding physical laws based on visual inputs. It employs an explicit factorization approach to represent the physical properties of objects by composing latent variables o with disentangled vectors, including appearance a , velocities v , mass m , and charge c properties. In this representation, a is responsible for capturing extrinsic properties, encompassing essential concepts such as colors, shapes, materials, locations, and sizes. These properties can be inferred by reconstructing the original image. v encodes velocity, which determines the transition of the latent variable a to its state in the subsequent time step. v can be inferred through short-term predictions. In scenarios where objects exert forces on each other, such as attraction, repulsion, or collision, the velocities of the objects are influenced by these forces. To model these interactions and update v , IWM incorporates the intrinsic properties of mass m and charge c with the guidance of longer-term predictions. The acquisition of increasingly abstract properties follows a hierarchical scheme.

Learning Object Appearances via Reconstruction

In the first stage illustrated in Fig. 2, the data are structured as individual images. During this stage, IWM acquires object-centric appearance representations, which are denoted as $a_1, a_2, \dots, a_k$. IWM utilizes a multislot framework, which leverages a spatial Gaussian mixture model to break down a static image into K slots, with each slot representing a distinct object within the image. This framework allows IWM to identify different objects in the visual input, paving the way for further analyses and interpretations of physical events. By utilizing a decoder, each latent vector a_k can be decoded to determine the pixel-wise mean μ_k^i and the pixel-wise mask χ_k^i . These components provide information about the assignment

probability and appearance of pixel i for object k . The generative mechanism is shared, meaning that different orderings of the K slots yield the same image. To ensure that the sum of the K mask logits for each pixel is equal to 1, the K mask logits are normalized using a softmax function. The generative process can be described as follows:

$$p(x | \{a_1, \dots, a_k\}) = \prod_{i=1}^D \sum_{k=1}^K \chi_k^i \mathcal{N}(x_i; \mu_k^i, \sigma^2). \quad (5)$$

In this equation, μ and χ parameterize the final spatial Gaussian mixture distribution, while σ remains fixed and retains the same value for all i and k . This formulation represents the probability distribution over the pixel values x given the latent appearance variables a .

The representation is derived using amortized variational inference, which can be described by the following equations:

$$a \sim q_{\lambda^r}(a | x), \quad (6)$$

$$\lambda^{r+1} \leftarrow \lambda^r + f(a, x, \mu, \chi, \nabla_{\mu} p(x | a), \nabla_{\chi} p(x | a)), \quad (7)$$

where $q(a | x)$ represents the posterior probability of a given the image x , with the parameters denoted by λ^r . The function f is a network responsible for refining these parameters. Initially, the posterior parameters are set to an arbitrary guesses, and the latent representations a are sampled from q_{λ} . In each refinement step, the input image x along with a set of variables (means μ , masks χ , mean gradient $\nabla_{\mu} p(x | a)$, and mask gradient $\nabla_{\chi} p(x | a)$) computed from z and x are fed into the refinement network f to obtain the updated posterior λ^{r+1} .

Learning Object Dynamics via Short-Term Prediction

In the second stage, the data are organized as sequential frames. This extension involves applying the concept of amortized variational inference to videos. The extension is built upon the idea that object velocities influence object appearance alterations, such as location and depth changes. To accomplish this, IWM learns a transition function that takes both the appearance at time t , denoted as a^t , and the velocity at time t , denoted as v^t , as inputs. This transition function then

generates the state at the next step, which is represented as a^{t+1} :

$$a^{t+1} = W_{pred}(v^t) + a^t, \quad (8)$$

where W_{pred} takes v^t as input and outputs the offset of a^t . Initially, the predicted a^{t+1} may have arbitrary values, but as the training process progresses, the transition function is trained to incorporate dynamics into the location information, while excluding other irrelevant properties such as shape and color. a^{t+1} is required to generate the next frame x^{t+1} with the learned decoder, as employed in the first stage, and the precise inference of v^t is assured through this guidance, guaranteeing accurate predictions for future image frames. This process can be accomplished through one-step predictions, as described by:

$$v^t \sim q_{\lambda}(v^t | x^{t+1}, a^t). \quad (9)$$

We note that a can be directly inferred by observing the image, while the discovery of v relies on accurately predicting the transition of a . Therefore, we think v is at a higher level of abstraction level than a .

Learning Masses and Collisions via Long-Term Prediction

In the third stage, the training dataset consists of videos containing object collisions. IWM incorporates intrinsic properties to effectively capture the outcomes of object interactions. With the interpretation of the training data, the model gradually internalizes these properties and understands the corresponding causal laws. IWM establishes a uniform framework for discovering intrinsic physical properties, which is accomplished by representing the interaction forces among objects through the utilization of three fundamental components: attention f_a , force directions f_d , and force magnitudes f_m . Herein, f_d is normalized, and f_m is constrained to be positive.

IWM's collision module computes acceleration directly from object appearance (a), velocity (v), and mass relevance (r^m). The collision module is described as follows:

$$r_{ij}^m = E_{r^m}(m_i, m_j), \quad (10)$$

$$f_a = \text{Sigmoid}(W_a(E_a(a, v, r^m))), \quad (11)$$

$$f_d = \text{L2-norm}(W_d(E_d(a, v, r^m))), \quad (12)$$

$$f_m = \text{ReLU}(W_m(E_m(a, v, r^m))), \quad (13)$$

where W_a , W_d , and W_m are weight matrices, Sigmoid and ReLU are activation functions, and E_a , E_d , and E_m denote embedding functions that process the object properties (a , v , and r^m) in a pairwise manner. The collision force is then calculated as:

$$\Delta v = f_a \cdot f_d \cdot f_m. \quad (14)$$

The collision force updates v and subsequently updates a for the next time step. The mass inference process can be formulated as:

$$m^t \sim q_\lambda(m^t \mid x^{\leq t+N}, a^t, v^t). \quad (15)$$

Learning Charge and Coulomb interactions via Long-Term Prediction

In the final stage, IWM learns to interpret Coulomb's force with video including charged objects, where the learned collision module is fixed. Assuming that the system has already grasped the concepts of object appearance a and motion velocity v , we now introduce a novel variable c to encode the charge quantity of the object. In this context, guided by the existing physical concepts, the force module can be represented in the following form:

$$r_{ij}^c = E_{rc}(c_i, c_j), \quad (16)$$

$$f_a = \text{Sigmoid}(W_a(E_a(a; v; r^c))), \quad (17)$$

$$f_d = \text{L2-norm}(W_d(E_d(a; v; r^c))), \quad (18)$$

$$f_m = \text{ReLU}(W_m(E_m(a; v; r^c))). \quad (19)$$

Finally, the resultant force is calculated as the product of these components:

$$\Delta v = f_a \cdot f_d \cdot f_m, \quad (20)$$

and the mass and charge inference calculations can be formulated as:

$$m^t, c^t \sim q_\lambda(m^t, c^t \mid x^{\leq t+N}, a^t, v^t). \quad (21)$$

Training and Inference

Throughout the training process, optimization is performed for the initial posterior λ and the parameters of three key models: the decoder p_θ that generates images from a ; the inference model

q_ϕ that updates the posterior parameters using the ELBO [44] gradient information obtained by the current parameter estimates; and the causal modules r_δ that roll out a sequence of $a^{2:N}$. We train all the parameters by gradient descent through the reconstruction and prediction objective. The training loss $\mathcal{L}$ is the negative ELBO, which can be described as:

$$\mathcal{L} = \sum_{t=0}^N (-\log(p_\theta(x^t \mid a^t)) + \beta \text{KL}(q_\phi(o^t \mid x^t, o^{<t}) \parallel p(o))). \quad (22)$$

The first term is the negative log-likelihood of the decoder given our mixture of decoder distributions, as discussed above. The second term is the Kullback–Leibler divergence (KL) divergence of the latent posterior with the latent prior, which is weighted with a hyperparameter β .

During the testing phase, IWM first iteratively refines the prior distribution based on the input video, ultimately obtaining λ . At this point, we can sample from it to obtain the physical representations of the objects within the scene:

$$a^1, v^1, m, c \sim \lambda. \quad (23)$$

Based on the inferred scene configuration, IWM can replay the physical event with its causal modules:

$$a^{2:N}, v^{2:N}, \Delta v^{1:N-1} = r_\delta(a^1, v^1, m, c). \quad (24)$$

This process corresponds to the reconstruction and explanation experiments depicted in Fig. 3 and Fig. 5, respectively. IWM further alters the inferred representations to conduct counterfactual reasoning:

$$a'^{2:N}, v'^{2:N}, \Delta v'^{1:N-1} = r_\delta(a^1, v^1, m', c'). \quad (25)$$

This corresponds to the characterization of causal laws in Fig. 3 and the counterfactual experiments in Fig. 4.

Data availability

The ComPhy dataset used in this paper is publicly available and can be accessed at <https://comphyreasoning.github.io/>.

Code availability

The source code can be accessed at <https://github.com/tqace/EXPHY>.

Author contributions

Conceptualization, Q.T., X.Z., Z.L., and Z.Z.; Methodology, Q.T.; Formal analysis, Q.T. and X.Z.; Investigation, Q.T. and X.Z.; Validation, Q.T. and H.Z.; Data Curation, Q.T. and H.Z.; Visualization, Q.T. and H.Z.; Writing, Q.T. and X.Z.; Supervision, Z.L. and Z.Z.; Project Administration, Z.L. and Z.Z.; Funding Acquisition, Z.Z.

References

- [1] R. Baillargeon, “How do infants learn about the physical world?” *Current Directions in Psychological Science*, vol. 3, no. 5, pp. 133–140, 1994.
- [2] A. E. Stahl and L. Feigenson, “Observing the unexpected enhances infants’ learning and exploration,” *Science*, vol. 348, no. 6230, pp. 91–94, 2015.
- [3] S. J. Hespos and R. Baillargeon, “Infants’ knowledge about occlusion and containment events: A surprising discrepancy,” *Psychological Science*, vol. 12, no. 2, pp. 141–147, 2001.
- [4] R. Baillargeon, “8. object permanence in 31/2-and 41/2-month-old infants,” *An Introductory Reader in Developmental Psychology*, vol. 23, no. 5, p. 103, 1995.
- [5] R. Baillargeon, E. S. Spelke, and S. Wasserman, “Object permanence in five-month-old infants,” *Cognition*, vol. 20, no. 3, pp. 191–208, 1985.
- [6] E. S. Spelke, R. Kestenbaum, D. J. Simons, and D. Wein, “Spatiotemporal continuity, smoothness of motion and object identity in infancy,” *British journal of developmental psychology*, vol. 13, no. 2, pp. 113–142, 1995.
- [7] R. Baillargeon, “Infants’ physical world,” *Current directions in psychological science*, vol. 13, no. 3, pp. 89–94, 2004.
- [8] B. M. Lake, T. D. Ullman, J. B. Tenenbaum, and S. J. Gershman, “Building machines that learn and think like people,” *Behavioral and brain sciences*, vol. 40, p. e253, 2017.
- [9] P. W. Battaglia, J. B. Hamrick, and J. B. Tenenbaum, “Simulation as an engine of physical scene understanding,” *Proceedings of the National Academy of Sciences*, vol. 110, no. 45, pp. 18 327–18 332, 2013.
- [10] T. Gerstenberg and J. B. Tenenbaum, “Intuitive theories,” 2017.
- [11] A. N. Sanborn, V. K. Mansinghka, and T. L. Griffiths, “Reconciling intuitive physics and newtonian mechanics for colliding objects.” *Psychological review*, vol. 120, no. 2, p. 411, 2013.
- [12] K. A. Smith and E. Vul, “Sources of uncertainty in intuitive physics,” *Topics in cognitive science*, vol. 5, no. 1, pp. 185–199, 2013.
- [13] M. McCloskey, “Intuitive physics,” *Scientific american*, vol. 248, no. 4, pp. 122–131, 1983.
- [14] T. Lombrozo, “Explanation and abductive inference.” 2012.
- [15] W. Li, S. Azimi, A. Leonardis, and M. Fritz, “To fall or not to fall: A visual approach to physical stability prediction,” *arXiv preprint arXiv:1604.00066*, 2016.
- [16] A. Lerer, S. Gross, and R. Fergus, “Learning physical intuition of block towers by example,” in *International conference on machine learning*. PMLR, 2016, pp. 430–438.
- [17] M. B. Chang, T. Ullman, A. Torralba, and J. B. Tenenbaum, “A compositional object-based approach to learning physical dynamics,” *arXiv preprint arXiv:1612.00341*, 2016.
- [18] N. Watters, D. Zoran, T. Weber, P. Battaglia, R. Pascanu, and A. Tacchetti, “Visual interaction networks: Learning a physics simulator from video,” *Advances in neural information processing systems*, vol. 30, 2017.
- [19] Z. Wu, N. Dvornik, K. Greff, T. Kipf, and A. Garg, “Slotformer: Unsupervised visual dynamics simulation with object-centric models,” *arXiv preprint arXiv:2210.05861*, 2022.
- [20] A. G. Alias Parth Goyal, A. Didolkar, N. R. Ke, C. Blundell, P. Beaudoin, N. Heess, M. C. Mozer, and Y. Bengio, “Neural production systems,” *Advances in Neural Information Processing Systems*, vol. 34, pp. 25 673–25 687, 2021.
- [21] J. Wu, E. Lu, P. Kohli, B. Freeman, and J. Tenenbaum, “Learning to see physics

via visual de-animation,” *Advances in Neural Information Processing Systems*, vol. 30, 2017.

- [22] J. Wu, I. Yildirim, J. J. Lim, B. Freeman, and J. Tenenbaum, “Galileo: Perceiving physical object properties by integrating a physics engine with deep learning,” *Advances in neural information processing systems*, vol. 28, 2015.
- [23] J. Wu, J. J. Lim, H. Zhang, J. B. Tenenbaum, and W. T. Freeman, “Physics 101: Learning physical object properties from unlabeled videos.” in *BMVC*, vol. 2, no. 6, 2016, p. 7.
- [24] D. Shapere, “The structure of scientific revolutions,” *The Philosophical Review*, vol. 73, no. 3, pp. 383–394, 1964.
- [25] E. S. Spelke, K. Breinlinger, J. Macomber, and K. Jacobson, “Origins of knowledge.” *Psychological review*, vol. 99, no. 4, p. 605, 1992.
- [26] S.-h. Wang and R. Baillargeon, “Inducing infants to detect a physical violation in a single trial,” *Psychological science*, vol. 16, no. 7, pp. 542–549, 2005.
- [27] Y. LeCun, “A path towards autonomous machine intelligence,” *preprint posted on openreview*, 2022.
- [28] N. D. Goodman, J. B. Tenenbaum, and T. Gerstenberg, “Concepts in a probabilistic language of thought,” Center for Brains, Minds and Machines (CBMM), Tech. Rep., 2014.
- [29] Z. Chen, K. Yi, Y. Li, M. Ding, A. Torralba, J. B. Tenenbaum, and C. Gan, “Comphy: Compositional physical reasoning of objects and events from videos,” *arXiv preprint arXiv:2205.01089*, 2022.
- [30] S.-h. Wang and R. Baillargeon, “Can infants be “taught” to attend to a new physical variable in an event category? the case of height in covering events,” *Cognitive Psychology*, vol. 56, no. 4, pp. 284–326, 2008.
- [31] W.-k. Ahn and C. W. Kalish, “The role of mechanism beliefs in causal reasoning,” *Explanation and cognition*, pp. 199–225, 2000.
- [32] T. Gerstenberg, N. D. Goodman, D. A. Lagnado, and J. B. Tenenbaum, “How, whether, why: Causal judgments as counterfactual contrasts.” in *CogSci*, 2015.

- [33] J. Marino, Y. Yue, and S. Mandt, “Iterative amortized inference,” in *International Conference on Machine Learning*. PMLR, 2018, pp. 3403–3412.
- [34] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, “Vqa: Visual question answering,” in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2425–2433.
- [35] T. M. Le, V. Le, S. Venkatesh, and T. Tran, “Hierarchical conditional relation networks for video question answering,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 9972–9981.
- [36] D. A. Hudson and C. D. Manning, “Compositional attention networks for machine reasoning,” *arXiv preprint arXiv:1803.03067*, 2018.
- [37] T. Kipf, G. F. Elsayed, A. Mahendran, A. Stone, S. Sabour, G. Heigold, R. Jonschkowski, A. Dosovitskiy, and K. Greff, “Conditional object-centric learning from video,” *arXiv preprint arXiv:2111.12594*, 2021.
- [38] C. P. Burgess, L. Matthey, N. Watters, R. Kabra, I. Higgins, M. Botvinick, and A. Lerchner, “Monet: Unsupervised scene decomposition and representation,” *arXiv preprint arXiv:1901.11390*, 2019.
- [39] K. Greff, R. L. Kaufman, R. Kabra, N. Watters, C. Burgess, D. Zoran, L. Matthey, M. Botvinick, and A. Lerchner, “Multi-object representation learning with iterative variational inference,” in *International Conference on Machine Learning*. PMLR, 2019, pp. 2424–2433.
- [40] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini, “The graph neural network model,” *IEEE transactions on neural networks*, vol. 20, no. 1, pp. 61–80, 2008.
- [41] K. Yi, C. Gan, Y. Li, P. Kohli, J. Wu, A. Torralba, and J. B. Tenenbaum, “Clevrer: Collision events for video representation and reasoning,” *arXiv preprint arXiv:1910.01442*, 2019.
- [42] R. Girdhar and D. Ramanan, “Cater: A diagnostic dataset for compositional actions and temporal reasoning,” *arXiv preprint arXiv:1910.04744*, 2019.

- [43] D. Ding, F. Hill, A. Santoro, M. Reynolds, and M. Botvinick, “Attention over learned object embeddings enables complex visual reasoning,” *Advances in neural information processing systems*, vol. 34, pp. 9112–9124, 2021.
- [44] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe, “Variational inference: A review for statisticians,” *Journal of the American statistical Association*, vol. 112, no. 518, pp. 859–877, 2017.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [IWMNCEsup.pdf](#)