# Supplementary Material

## Contents

# 1 IWM vs. Other Physical Models.

Table S1: **Comparison of IWM with related physical models.**

| Models | Learning | | Functionality | |
|---|---|---|---|---|
| | **Based on observation** | **Symbolic properties** | **Explainability** | **Counterfactual** |
| ODDN[1] | ✓ | ✗ | ✗ | ✗ |
| SlotFormer[2] | ✓ | ✗ | ✗ | ✗ |
| IODINE-GNN[3] | ✓ | ✗ | ✗ | ✗ |
| CPL[4] | ✗ | ✓ | ✗ | ✓ |
| IWM | ✓ | ✓ | ✓ | ✓ |

In this section, we conduct a comparison of the learning manner and functionality between nine related physical frameworks to show how IWM advances upon existing work. The comparison is conducted based on five criteria:

1. **Based on observation**: suggesting whether the training is unsupervised, without relying on labels or built-in prior knowledge in the form of physical formulas or engines.
2. **Symbolic physical properties**: indicating if the framework is capable of learning symbolic representations of physics that are disentangled and low-dimensional, facilitating a more efficient and generalizable understanding of physical phenomena.
3. **Explainability** of the model: evaluating whether the model's predictions are explainable and can be easily understood by humans, which is crucial for further application of the model.
4. **Counterfactual** reasoning: testing whether the model has the ability to conduct counterfactual reasoning, that is, to predict what would happen in scenarios that deviate from the norm or from previously encountered situations, demonstrating a deeper level of causal understanding.

Grouping these criteria into categories of learning and functionality offers distinct perspectives. "Based on Observation" and "Symbolic Properties" focus on the frameworks' learning capabilities, while "Explainability" and "Counterfactual Reasoning" are concerned with their functional features. As indicated in the table, **(1.)** Both CPL and our method can learn symbolic property representations, but CPL's framework uses property labels for direct supervision during training. **(2.)** Only our method has explainability, as we explicitly model the force interactions between objects. **(3.)** Both CPL and our method have the capability of counterfactual prediction, as both methods include a prediction module that can forecast the future based on property inputs. The difference is that CPL uses GNN, which is an implicit way of prediction, and CPL utilizes labels to train the inference of physical properties.
Our approach, differs from our earlier version, PHYCINE[5], in that it operates independently of any pre-existing physical knowledge, making it completely data-driven. For example, PHYCINE's charge module, which simulated the interaction of Coulomb's force, included a built-in operation that multiplied the embeddings of the charge properties of two objects. IWM, on the other hand, autonomously learns to simulate physical interactions directly from the data, without the need for embedded physical equations. This marks a notable advancement in the design of data-driven physical frameworks.

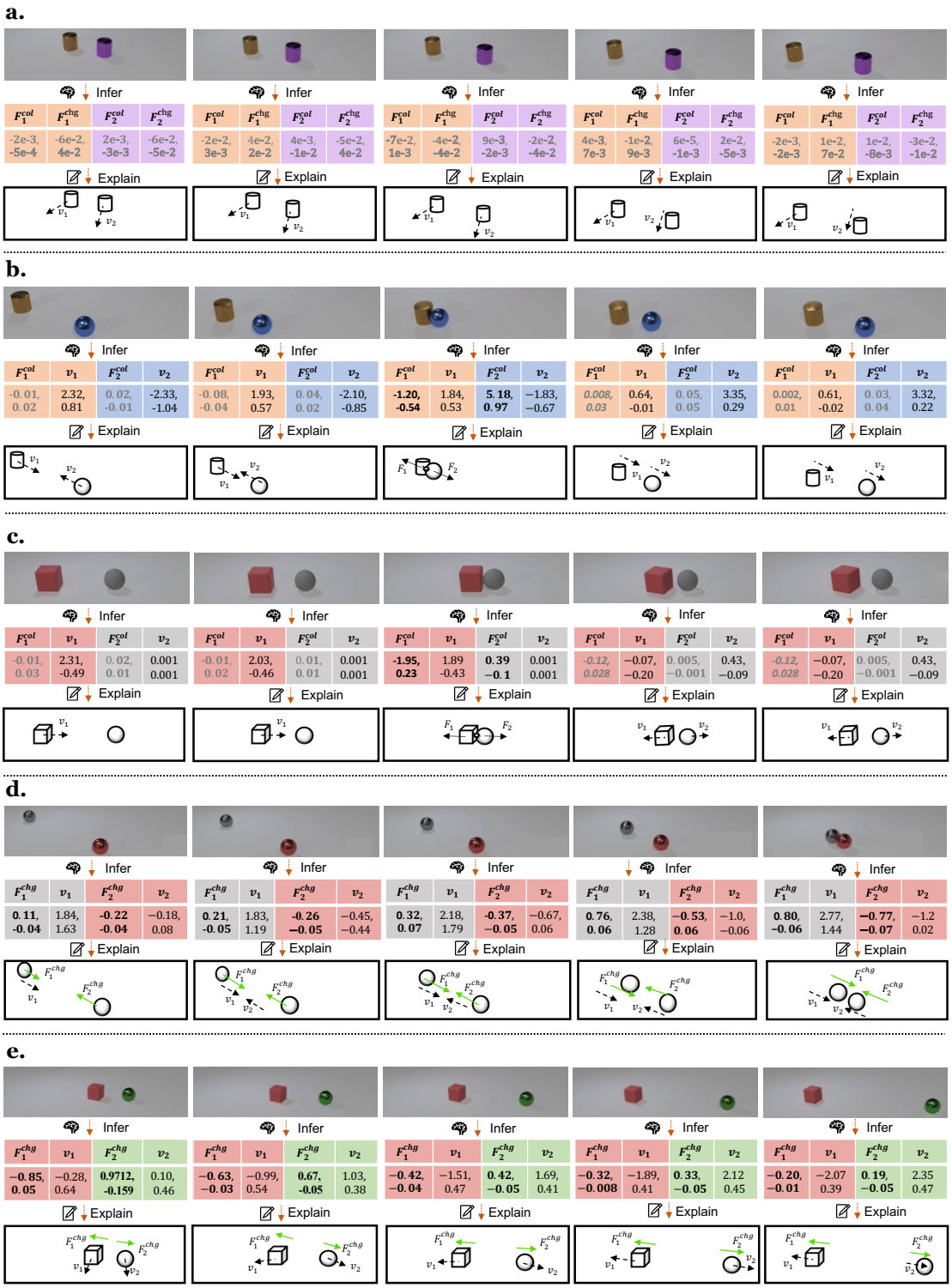


Figure S1: **The explanation ability of IWM. a**, displays two objects in linear trajectories with no interactive forces, illustrating independent motion. **b**, depicts an interaction where a heavier object collides with a lighter one while both are in motion. **c**, depicts a lighter moving object striking a heavier stationary one. **d**, reveals two objects converging under an attractive force. **e**, portrays two objects diverging due to a repulsive force.

## 2 Explanation examples in more scenarios.

More explanations of physical events are visualized in Fig. S1. The figure showcases IWM's analytical capabilities in interpreting physical events, offering insights that align with fundamental principles of physics and human intuition.

- Fig. S1.a depicts a scene featuring a cylinder and a sphere following linear trajectories without any interactive forces, illustrating the concept of independent motion. According to IWM's analysis of this scenario, there are no external forces affecting the objects, enabling them to continue along their respective paths without any alterations caused by external influences.

- Fig. S1.b depicts a collision scenario featuring a brown cylinder that possesses greater mass than a blue sphere. The collision occurs in the third frame, and we observe a significant difference in acceleration due to their differing masses. When we examine the perceived mass values of these objects in IWM, we find that the brown cylinder has a mass value of 2.41, while the blue sphere has a mass value of -1.25. IWM's force analysis reveals that the gray cylinder's acceleration (-1.20, 0.54) is considerably smaller than that of the blue sphere (5.18, 0.97), highlighting the fundamental principle that, in a collision, a heavier object tends to experience a smaller change in velocity compared to a lighter one. Furthermore, the forces acting on both objects are in opposite directions, in accordance with Newton's third law of motion, which states that for every action, there is an equal and opposite reaction.

- Fig. S1.c depicts a collision event between a red cube with a smaller mass and an initial velocity and a heavier, stationary gray sphere. This collision, occurring in the third frame, serves as an illustration of how both mass and the initial motion state influence the outcome of such interactions. IWM's analysis reveals that the red cube, which has a mass value of -0.95, experiences a more significant force acceleration of (-1.95, 0.23) compared to the gray sphere with a mass value of 2.30, which experiences an acceleration of (0.39, -0.1). This observation aligns with the expectation that a lighter, moving object would undergo a more substantial acceleration upon colliding with a heavier, stationary object. Furthermore, the interaction also is in accordance with the principles of conservation and Newton's third law of motion. Notably, in contrast to the collision depicted in Fig. S1.b, the velocity of the lighter object, the red cube, does not undergo as drastic a change. This can be attributed to the stationary state of the heavier object, the gray sphere, prior to the collision, resulting in a lack of momentum transfer capable of significantly altering the velocity of the lighter object. IWM's analysis captures this subtle aspect of the collision.

- Fig. S1.d depicts a scenario involving Coulomb interaction, where two objects with opposite charges are depicted as being drawn together. This scenario successfully captures the development of an attractive force in line with Coulomb's law. Based on data from IWM, the charges of the two objects are 0.71 and -4.09, respectively. IWM's analysis indicates that as the objects move closer, the force magnitudes increase, as seen with the red sphere where the force escalates from (-0.22, -0.04) to (-0.77, -0.07). This analysis by IWM effectively demonstrates the inverse relationship between force magnitude and the distance between two charged objects, a fundamental aspect of Coulomb's law.

- Fig. S1.e depicts a contrasting scenario to the previous one, displaying two objects with similar charges and their resultant repulsive interaction. In this case, as the objects move farther apart, the repulsive force between them decreases. This phenomenon is accurately captured and analyzed by IWM through both graphical and numerical representations. The charges of these objects are identified as 0.86 and 1.20, respectively. Notably, the force acting on the green sphere diminishes from (0.97, -0.16) to (0.19, -0.05), showcasing the system's adeptness in interpreting how the force lessens as the distance between like-charged objects increases.

Collectively, these scenarios demonstrate IWM's advanced capability to simulate, analyze, and elucidate observed physical events. Through detailed force analyses and clear visual representations, IWM establishes itself as an intuitive and intelligent framework for understanding the intricate interplay of physical forces.

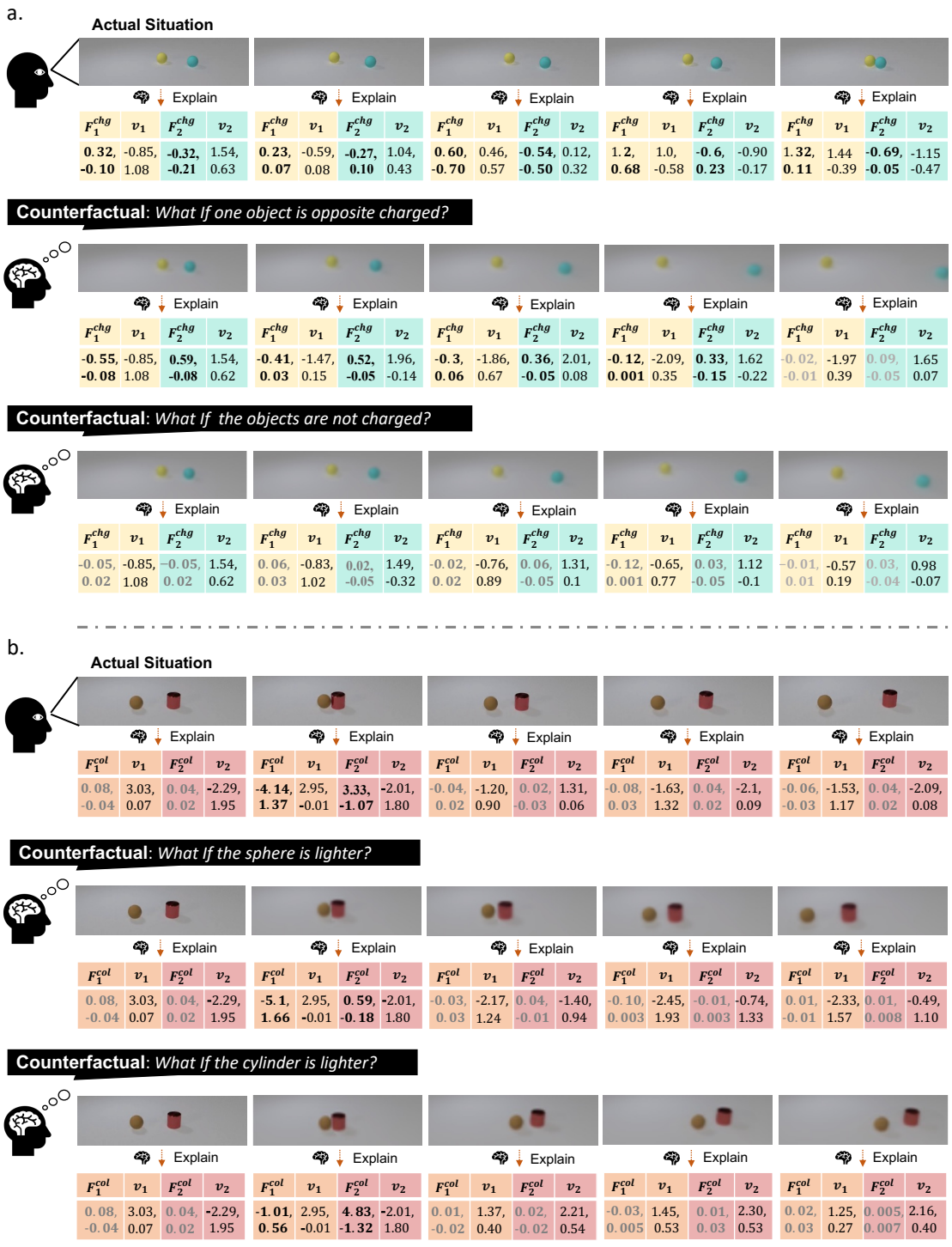


Figure S2: **IWM can make counterfactual reasoning with accompanying explanations.** In scenario (a), two spheres initially move apart but eventually converge due to electrostatic attraction, considering two counterfactual situations: one where both objects carry like charges, prompting an inquiry into the resultant repulsive force dynamics; another posits a neutral charge state, prompting speculation on how their trajectories would remain unaltered by electrical forces. In scenario (b), a collision between a sphere and a cylinder of equal mass with two counterfactual situations: if the sphere's mass were reduced; if the cylinder's mass were reduced.

# 3 Counterfactual imagination with explanation.

Fig. S2 highlights IWM's capability in explaining imagined scenarios and engaging in counterfactual reasoning, a key feature for exploring hypothetical situations and discerning potential outcomes that diverge from actual events.

- In Fig. S2.a, the original scenario describes two spheres initially moving apart but eventually converging due to attractive force. IWM's analysis (Charge: 0.84 and -4.01) reveals the dynamics of this attraction. In the first counterfactual situation, both objects are given like charges (0.84 and 1). IWM then predicts the dynamics under this new condition, showing how the repulsive forces would alter their paths. Another counterfactual scenario posits both objects with neutral charges (0.84 and -1), leading to an absence of electrostatic interaction. IWM predicts unchanged trajectories, emphasizing the influence of charge on motion. Comparing the force analysis of IWM in these three scenarios, in the second counterfactual scenario, the Coulomb force between objects is not activated. In the original scenario and the first counterfactual scenario, the electric charges are opposite. Taking the yellow sphere in the first frame as an example, the acceleration from Coulomb forces it experiences in these two scenarios are respectively (0.32, -0.1) and (-0.55, -0.08).
- In Fig. S2.b, the original scenario depicts a collision between a sphere and a cylinder of equal mass. IWM's analysis (Mass: 2.33 and 2.17) shows the accelerations during the collision. The counterfactual situation explores two possibilities: first, reducing the sphere's mass (2.33 to -1), and second, reducing the cylinder's mass (2.17 to -1). IWM predicts how these mass changes would affect the collision dynamics, demonstrating the impact of mass on collision outcomes. Specifically, we compare the force analysis in the collision frame (second frame) by IWM for the three scenarios. In the original scenario, the acceleration of the gray sphere and the red cylinder are respectively (-4.14, 1.37) and (3.33, -1.07), which are close in magnitude. In the first counterfactual scenario, where the gray sphere becomes lighter, their accelerations change to (-5.1, 1.66) and (0.59, -0.18). In the second counterfactual scenario, where the red cylinder becomes lighter, their accelerations become (-1.01, 0.56) and (4.83, -1.32). These counterfactual imaginations and corresponding explanations are consistent with our physical intuition.

Overall, Fig. S2 illustrates IWM's advanced capability to not only analyze actual physical interactions but also to simulate and extrapolate outcomes in a range of hypothetical scenarios, providing deeper insights into the intuitive principles of physics.

# 4 Bounding Box Supervision for Complex Scene Adaptation.

In IWM, pixel-level supervision enables the generation of counterfactual videos by leveraging an object-centric model that mimics human-like prior understanding of object dynamics. This approach does not rely on external detectors or trackers, as the model itself learns to segment objects directly from the visual context. However, in natural scenes, predicting future frames based on object-centric features is a challenging task due to factors such as texture variations, occlusions, and changing lighting conditions. One promising solution is to incorporate bounding boxes (bbox) as a supervisory signal, offering a robust alternative to pixel-level reconstruction in the IWM framework. By leveraging bbox annotations, IWM can effectively focus on learning the physical dynamics of object trajectories and interactions, rather than pixel-wise details.
Let's consider a set of object bounding boxes at time $t$, denoted as $\mathcal{B}_t = \{\mathbf{b}_t^1, \mathbf{b}_t^2, \ldots, \mathbf{b}_t^K\}$, where each bounding box $\mathbf{b}_t^k = [x_t^k, y_t^k, w_t^k, h_t^k]$ consists of the center coordinates $(x_t^k, y_t^k)$, width $w_t^k$, and height $h_t^k$ for object $k$. The loss function for predicting future bounding boxes is computed as:

$$\mathcal{L}_{\text{bbox}} = \sum_{k=1}^{K} \left\| \mathbf{b}_{t+1}^k - \hat{\mathbf{b}}_{t+1}^k \right\|_1$$

where $\hat{\mathbf{b}}_{t+1}^k$ is the predicted bounding box at time $t+1$, expressed as:

$$\hat{\mathbf{b}}_{t+1}^k = \mathbf{b}_t^k + \Delta\mathbf{b}_k$$

Here, $\Delta\mathbf{b}_k$ represents the change inferred by the model, capturing the object's velocity and acceleration. The use of the $L_1$ norm ensures stable gradient propagation, even in noisy or occluded environments.
In comparison, pixel-level supervision, commonly employed in IWM, typically involves a reconstruction loss:

$$\mathcal{L}_{\text{pixel}} = -\log(p_\theta(x_{t+1}|z_{t+1}^{app}))$$

where $x_{t+1}$ is the ground-truth frame, and $p_\theta(x_{t+1}|z_{t+1}^{app})$ is the probability of reconstructing $x_{t+1}$ from the appearance encoding $z_{t+1}^{app}$.
Bounding box supervision offers several distinct advantages over pixel-based methods:

- It is more robust to changes in lighting and occlusions, as the focus is on the object's movement rather than the entire scene's visual appearance.
- It enables the model to concentrate on learning the physical principles governing object dynamics (e.g., velocity and acceleration) rather than attempting to reconstruct detailed pixel-level scene information.
- It is more scalable and allows the model to generalize across diverse environments more effectively.
- Predicting bounding boxes requires fewer parameters and less memory compared to reconstructing entire frames, making it a more resource-efficient approach in terms of both computation and memory usage.

Figure S3 compares the computational time and memory usage between bounding box supervision and pixel-level supervision, clearly showing that bounding box supervision significantly reduces both time and memory requirements, making it a more efficient solution. Figure S4 presents a selection of objects from the GSO dataset [6], which consists of 3D-scanned household items. This dataset is designed for robotics and computer vision applications, enabling large-scale interactive 3D simulations. Its diversity and realism make it ideal for evaluating the generalization capabilities of IWM. Figure S5 illustrates how the mass concept captured by IWM affects the colliding objects, in accordance with Newton's laws, suggesting that bounding box supervision can be effectively used for conducting physics-based learning in more realistic and complex environments.

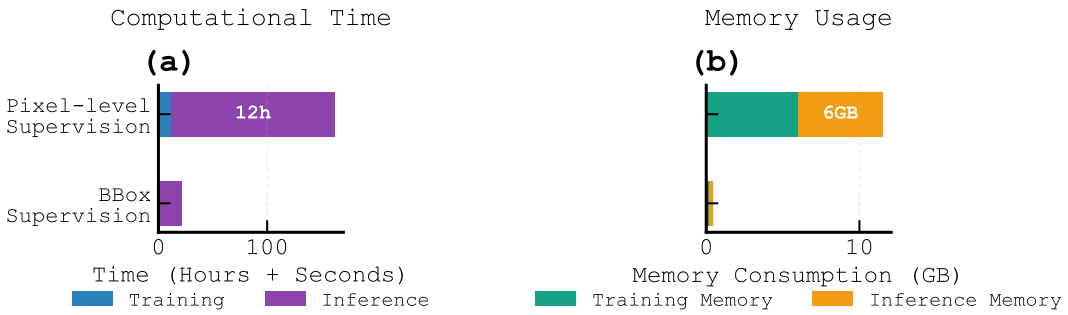


Figure S3: **Comparison of computational time and memory usage between bounding box supervision and pixel-level supervision in the IWM framework.**

Figure S4: **Sampled objects from the GSO [6] dataset**

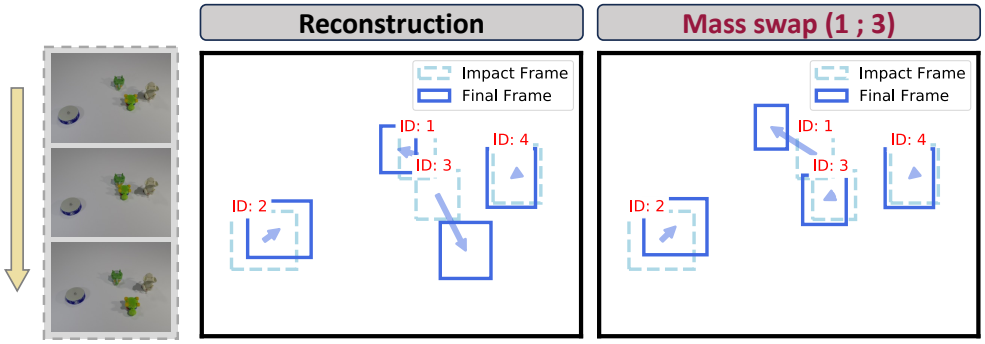


Figure S5: **BBox IoU between IWM-interpretation and GT at the final frame of colliding objects across different dataset splits.** 'rm' (random mass) represents replacing the object mass inferred by IWM with a randomly sampled mass within training distribution.

# 5 Nonlinear Activation in Causal Modules.

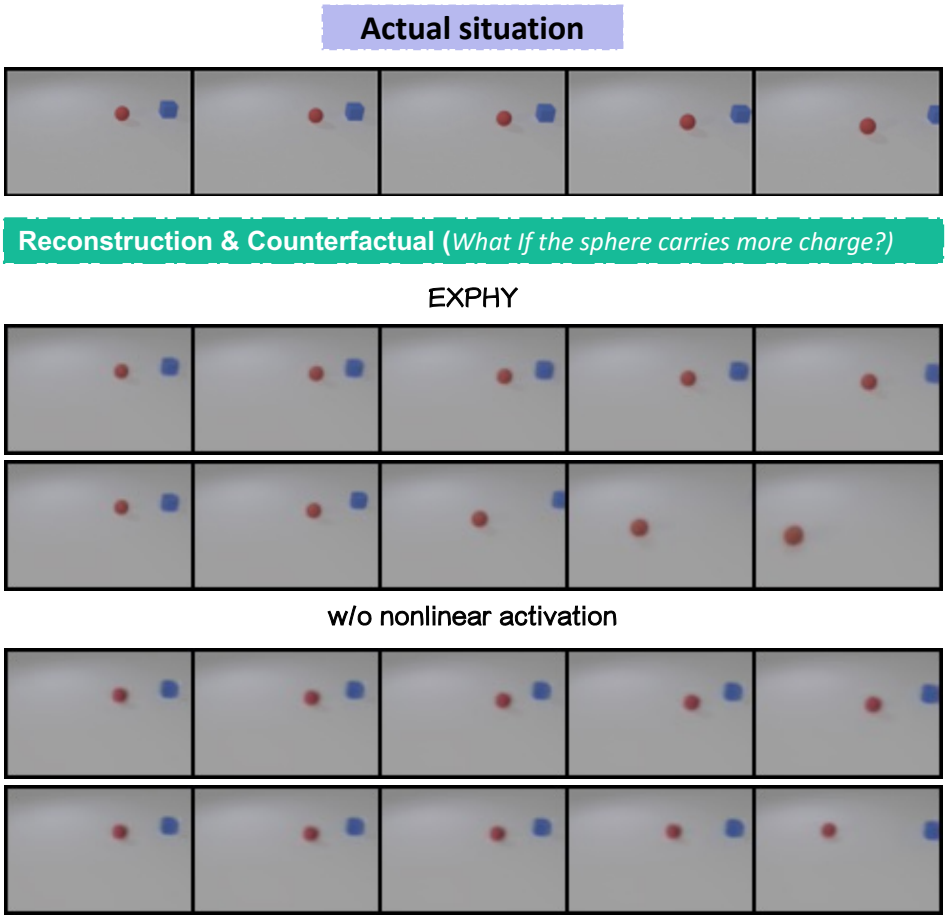


Figure S6: **The counterfactual visualization comparison between modeling interaction forces with and without nonlinear activation functions.**

In causal modeling, the forces between objects are modeled as functions of the objects' physical properties. These functions usually have nonlinear characteristics because physical forces in the real world, such as Coulomb's force, depend on nonlinear relationships, like the inverse square law of distance. Nonlinear activation functions in these models allow the network to capture such nonlinear dependencies.

If the introduction of nonlinear activation functions is reduced when modeling the forces between objects, maintaining only the necessary activation functions such as the absolute value function to ensure the force magnitude remains positive, the model would lose its ability to capture and express complex physical interactions. Take Coulomb interaction modeling as an example:

$$E_m = \cancel{ReLU}(FC(a; v; c)); f_m = \cancel{ReLU}^{abs}(FC(E)), \tag{1}$$

where $f_m$ denotes the force magnitude, and $E_m$ denotes the embedding function that embeds all properties of objects (a, v, and c) in a pairwise manner. As illustrated in Fig. S6, in the input video, two objects carrying like charges exhibit electrostatic repulsion. In IWM's counterfactual experiment, we assume that the red object has a larger charge. The counterfactual visualization shows an increase in the repulsive force, causing the objects to be repelled further apart. In the comparative experiment with a model lacking nonlinear activation, it is observed that only the position of the red object changes significantly, while the blue object remains almost unaffected. This suggests that the interaction module of the model has not adequately modeled the interaction between the objects. In such a case, the model might be unable to accurately describe the dynamic interactions that occur when the properties of objects change and would only be able to describe the effect of these changes on the motion state of individual objects.

# 6 Addressing the Binding Problem.

The "binding problem" in cognitive science refers to the challenge of how the brain integrates information from different sensory attributes and parts to produce a unified, coherent perception of an object. For example, when we see a red, round object bouncing, we don't perceive the color, shape, and motion separately; we see a single coherent event – a red ball bouncing. This problem extends to understanding how the brain links related properties (such as color and shape) to a single object, rather than to different objects. Understanding the binding problem is not only crucial for cognitive science and neuroscience but is also relevant for the development of artificial intelligence and machine learning, particularly in creating systems that can integrate various types of information into a cohesive whole[7].

IWM constructs an upper-level framework for the discovery of implicit physical properties based on an object-centric representation approach. We hope that the form of representation for the physical properties of objects inferred by the model has a one-to-one binding relationship with the object itself. Specifically, the physical property representation of one object corresponds only to that object's slot and is not distributed among other slots. Through the use of counterfactual imagination, we aim to illustrate the binding of representations of physical properties inferred by IWM to object slots. Specifically, we adjust the physical property values corresponding to individual slots, and IWM expands the corresponding counterfactual prediction visualizations to observe the associated changes.

In Fig. S7, the blocks of each slot that represent velocity are adjusted in a counterfactual manner, showing the impact of these counterfactual changes on the progression of physical events. The conclusions that can be drawn are: 1) Adjustments to the slots that do not encode objects do not cause any alterations in the temporal progression of physical events. 2) Adjusting the slot that corresponds to a particular object will only affect the changes in that object's own movement velocity and will not affect other objects. This effectively demonstrates that the representation of the velocity property is bound to the object. Similarly, we continue to explore the representation of mass and charge properties and their binding to slots. By holding other properties constant and only adjusting the variable values of mass or charge, the changes in each slot and the counterfactual visualizations they induce are presented in Fig. S8 and Fig. S9. The conclusions are: 1) For slots that do not encode objects, adjusting the corresponding variable values does not change the physical event. 2) Adjusting slots that encode objects brings about changes in the motion of that object and other objects by affecting the forces of interaction between them. 3) Adjustments to the representations of objects that are involved in interactions lead to symmetric changes (indicating that the forces are mutual).

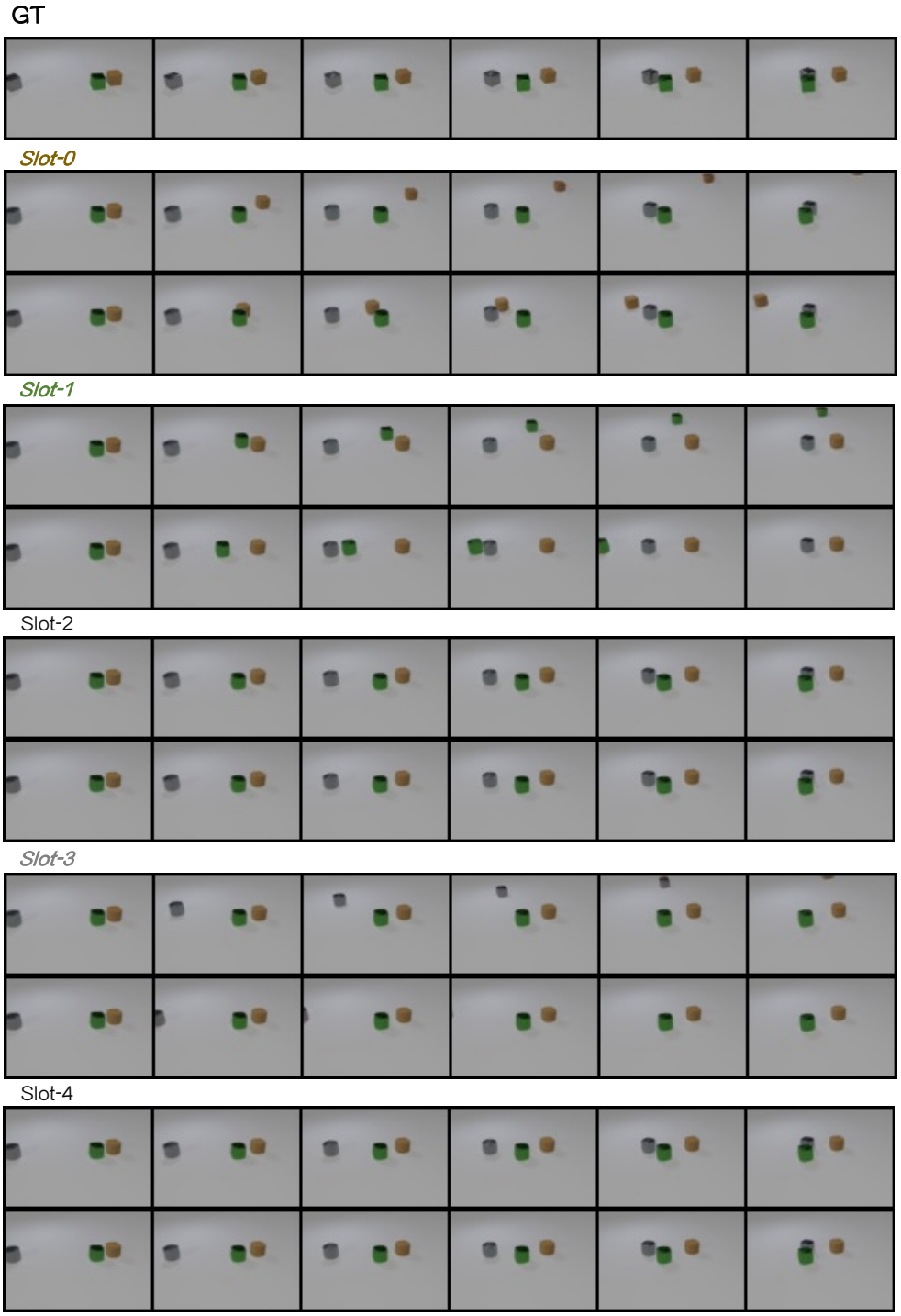


Figure S7: **Illustration of the slot-wise counterfactual visualization related to the velocity property.**


Figure S8: **Illustration of the slot-wise counterfactual visualization related to the mass property.**

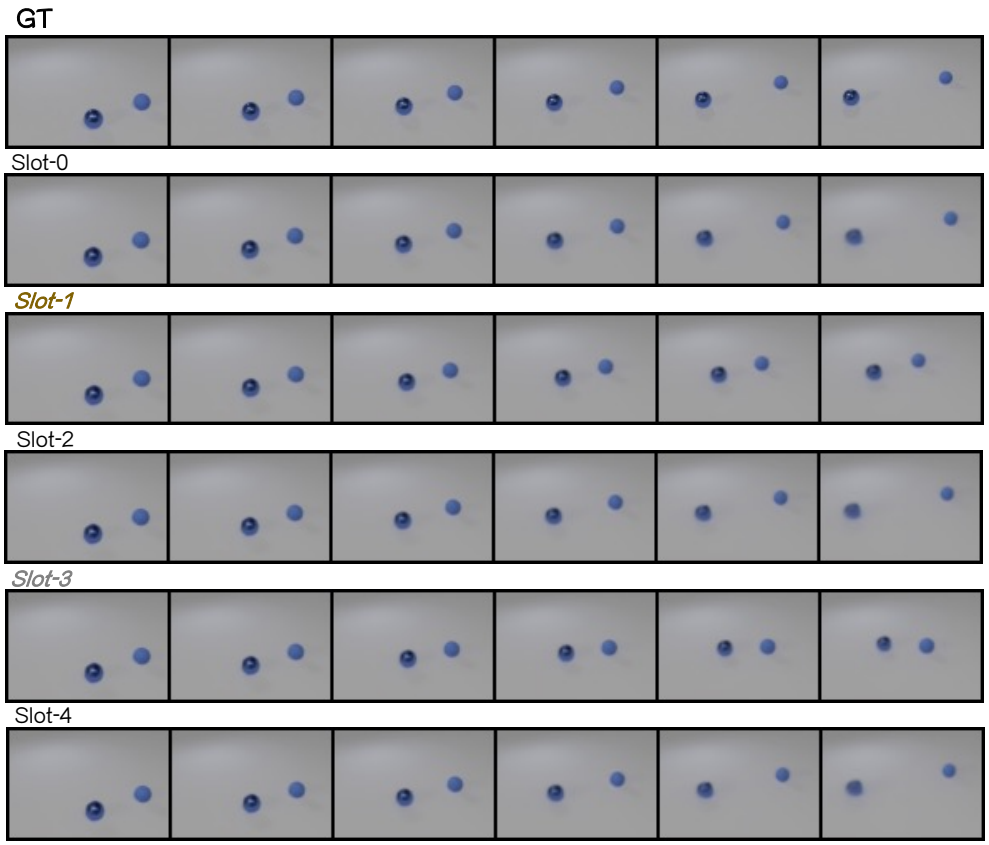


Figure S9: **Illustration of the slot-wise counterfactual visualization related to the charge property.**

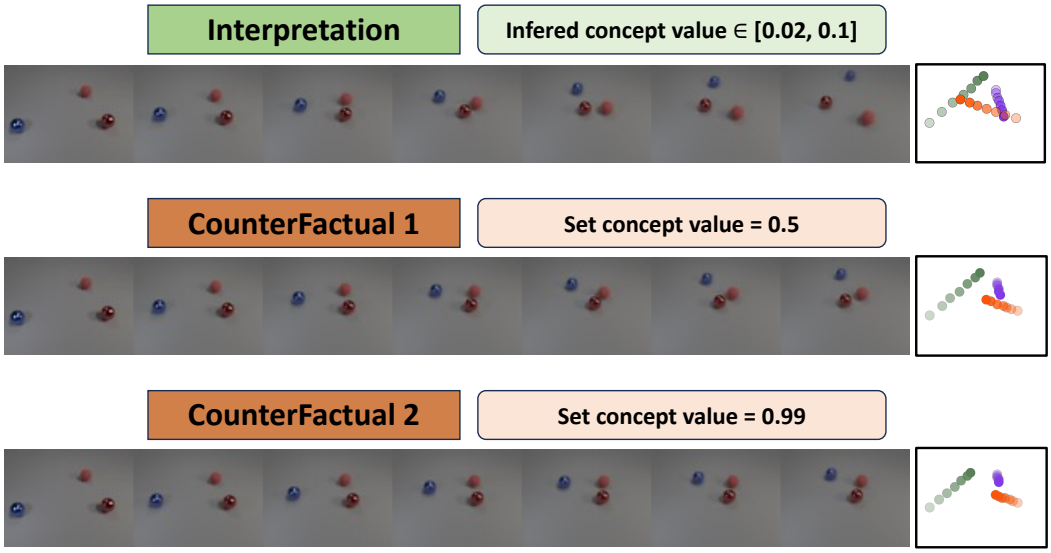


Figure S10: **Counterfactual Demonstration for the Inferred Friction Coefficient Property.** This figure demonstrates counterfactual generation examples based on the learned friction coefficient concept. The top row shows the initial scenario. The second and third rows display counterfactuals.

# 7 Generalizing IWM to Friction Force Modeling.

IWM, initially designed to model interactions like collision forces (which involve learning about mass) and Coulomb forces (which involve learning about electric charge states), offers a promising framework for understanding physical events. However, the potential of this framework extends beyond these interactions, allowing for generalization to other physical scenarios. In this section, we focus on expanding IWM to friction force modeling, a key interaction that plays a significant role in real-world physical systems. We demonstrate that IWM can be extended to successfully model frictional forces and discover the fundamental concept of the friction coefficient. The learning process for frictional forces follows the two-stage training approach of IWM. First, the model learns the object's velocity in a uniform linear motion scenario. Then, an adaptive concept latent (ac) is introduced to model the frictional forces. The input-output flow is as follows:

$$\text{Input: } [a, v, ac] \quad \rightarrow \quad \text{FFN}_{\text{direction}} \quad \rightarrow \quad \text{L2 norm} \quad \rightarrow \quad \delta v_{\text{direction}}$$

$$\text{Input: } [a, v, ac] \quad \rightarrow \quad \text{FFN}_{\text{magnitude}} \quad \rightarrow \quad \text{ReLU} \quad \rightarrow \quad \delta v_{\text{magnitude}}$$

Where $a$ is the appearance, $v$ is the velocity, and $ac$ is the adaptive concept latent. The first FFN learns the direction of acceleration, with the output $\delta v_{\text{direction}}$ calculated via the L2 norm. The second FFN learns the magnitude of acceleration, with the output $\delta v_{\text{magnitude}}$ calculated using ReLU. These two components together allow the model to learn the friction coefficient and update the object's velocity accordingly. In this experiment, the friction coefficient of the object is set to 0 or 1. In addition, we apply sigmoid normalization to the adaptive concept latent, constraining this feature to a range between 0 and 1.

Figure S10 demonstrates counterfactual generation examples based on the inferred friction coefficient property. In the top row, the reconstruction of the original scene is shown, where the friction coefficient of the objects in the scene is set to 0, and the inferred concept value lies within a certain range near 0. The second and third rows display counterfactual scenarios, where the friction coefficient of all objects is set to fixed values (0.5 and 0.99, respectively). These visualizations illustrate how changes in the friction coefficient influence the system's behavior, providing insight into the model's ability to adapt and understand the underlying frictional dynamics. Figure S11 further delves into the physical principles governing the learned frictional forces. Panel (a) illustrates the distribution of the adaptive concept latent values in smooth and frictional motion scenarios, explicitly demonstrating that the adaptive mechanism successfully learns the concept of friction coefficient, as evidenced by the distinct latent value distributions

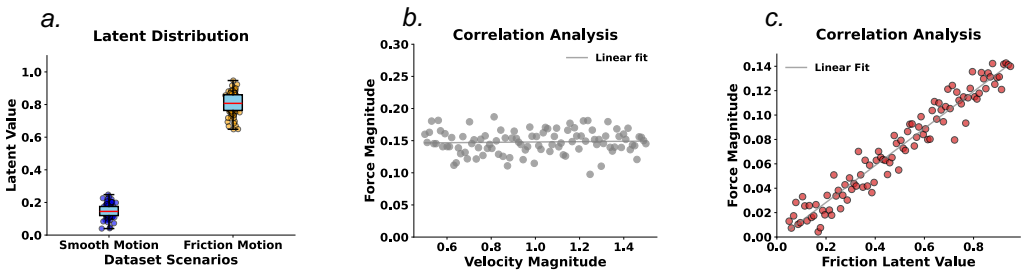


Figure S11: **Analysis of Physical Principles of Frictional Force.** Panel (a) presents the latent value distribution of the adaptive concept for smooth and friction motion scenarios. Panels (b) and (c) display the correlation analyses between frictional force magnitude and the learned physical properties, where the results are obtained by manually adjusting the dependent variable during acceleration calculation and observing the changes in the output acceleration.

corresponding to different physical behaviors. Panels (b) and (c) analyze the correlation between the learned frictional properties and the actual force magnitude. Panel (b) highlights the weak correlation between frictional force and velocity magnitude, suggesting that the model has understood that there is no causal relationship between the object's speed and the frictional force. In contrast, panel (c) demonstrates a strong linear correlation between the friction latent value and the force magnitude, reinforcing the model's ability to capture the fundamental physical principle that frictional force is directly related to the friction coefficient. This analysis confirms the effectiveness of IWM in modeling the friction force and its underlying principles.

# 8 Rationale and Insights.

## 8.1 The Violation of Expectation Paradigm

The concept of Violation of Expectation (VoE) originates in developmental psychology and plays a critical role in understanding how infants develop an intuitive understanding of the physical world. In VoE experiments, unexpected events tend to capture infants' attention, leading to longer looking times. This phenomenon indicates that infants are sensitive to discrepancies between their expectations and observations. Building on this foundational concept, recent works have begun incorporating VoE-inspired ideas into AI. PLATO [8] introduced a deep learning framework that mimics human-like intuitive physics learning, focusing on the observation and prediction of physical events. The X-VoE benchmark [9] expanded this concept by evaluating AI models on their ability to explain observed interactions, with an emphasis on causal reasoning. Human learning under the VoE framework, however, extends beyond mere detection and reaction. Key aspects of this process include:

- Infants, when observing physical events, construct causal models that enable them to interpret their observations and make predictions [10].
- Infants' causal models evolve through VoE experiences, as they incrementally learn to identify the variables that are critical for refining their predictions [11].
- Infants do not generalize variables across event categories. Instead, they learn separately about each category of interaction [12].

Inspired by these developmental insights, IWM and IWM incorporate similar principles to mirror human-like learning and reasoning. IWM organizes events into distinct categories, paralleling how infants construct causal models for each type of interaction. IWM further builds on this by introducing an adaptive concept mechanism, which infers latent properties such as mass or charge to explain specific interactions, and an iterative inference process that refines predictions based on observed outcomes. Together, these components align with the nuanced learning processes observed in human development under the VoE paradigm.

## 8.2 The World Model Perspective

As highlighted in HJEPA [13], building hierarchical and predictive representations is essential for intelligent reasoning. A world model serves two primary roles: estimating missing information about the environment and predicting plausible future states. These principles align closely with the design objectives of IWM, which extends them into physical reasoning.

**Hierarchical and Predictive Learning.** HJEPA emphasizes that hierarchical structures in world models enable the progressive acquisition of higher-order abstractions, where higher-level concepts build upon lower-level representations. Short-term predictions, such as immediate trajectories, lay the groundwork for learning mid-level concepts like stability and momentum. Over longer prediction horizons, abstract principles such as gravity or conservation laws naturally emerge. These multi-level abstractions allow world models to reason across different time scales and levels of complexity.

Similarly, IWM employs IWM's structured event categorization to facilitate hierarchical learning. Each event category (e.g., collisions, electrostatics) focuses on specific domains of causal reasoning, where lower-level abstractions, such as objects and velocities, are first inferred. These foundational representations serve as the basis for discovering higher-level abstractions, including latent properties (e.g., mass) and physical laws (e.g., Newton's law), ensuring robust generalization. Both JEPA and IWM emphasize predictive modeling as a fundamental mechanism driving learning. As predictions extend to longer time horizons, systems uncover deeper patterns and abstractions. IWM builds on this through iterative refinement, where predictive feedback continuously adjusts its latent properties. For instance, IWM infers and refines mass during repeated collisions, illustrating the evolution of causal understanding. By bridging short-term accuracy with interpretable, long-term abstractions, IWM aligns with the hierarchical principles of JEPA while extending them through structured reasoning and domain-specific physical explanations.

**Latent Variable Modeling for Uncertainty.** JEPA emphasizes the role of latent variables in capturing hidden or uncertain aspects of the environment. These variables enable models

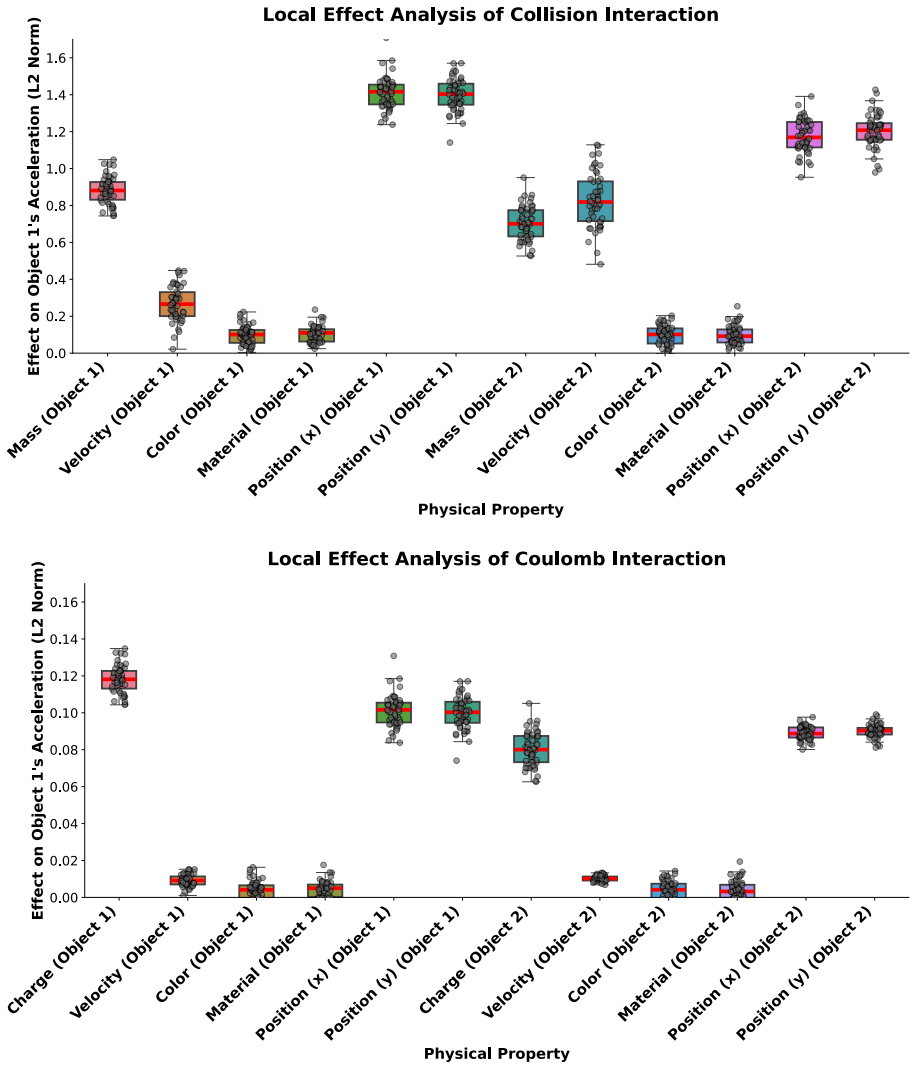


Figure S12: **Local Effect Analysis,** quantifying the influence of various physical properties on the L2 norm of the change in Object 1's acceleration.

to represent multi-modal outcomes and handle ambiguity effectively. IWM adopts a similar philosophy, using its adaptive concept mechanism to introduce latent properties, which explain and predict complex interactions. For example, in electrostatic scenarios, IWM models the uncertainty in interaction forces by refining the inferred charge values, ensuring both predictive accuracy and interpretability.

**Regularization to Prevent Collapse.** JEPA employs regularization to ensure meaningful latent representations by constraining the information content of latent variables. This prevents collapse, where the latent space becomes uninformative or redundant. By limiting the capacity of latent variables to encode only task-relevant information, JEPA ensures that the representation space remains compact and focused, enabling robust learning and multi-modal predictions.

IWM adopts a similar approach to prevent collapse through two complementary strategies tailored to physical reasoning. First, it employs low-dimensional latent variables, such as 2D velocity or 1D concept relevance, ensuring that each dimension captures essential and interpretable physical properties. This natural dimensionality constraint regularizes the latent space, focusing on key factors relevant to specific interactions. Second, IWM employs the interaction module to decompose forces into three interpretable components: attention, direction, and magnitude. By explicitly separating these elements, The interaction module ensures sparsity and prevents the collapse of any single component, maintaining an interpretable and task-relevant representation.

# 9 Correlation Analysis of Latent Physical Concepts

In this section, we investigate the correlation between the latent variables learned by the Intuitive World Model (IWM) and various physical outcomes, such as accelerations. The goal of this analysis is to explore how IWM captures both the correlations and independencies between physical concepts, offering insights into the model's ability to learn realistic physical principles. Specifically, we examine the relationships between causal interactions and different levels of abstraction in physical properties, including appearance, velocity, mass, and charge states. To perform this analysis, we manually adjust the physical latent inputs during IWM's interaction modeling in the interactive frames and observe their impact on the outputs.

Figure S12 provides a multifaceted view of the IWM's internal workings, specifically focusing on how different physical properties influence the outcome of two distinct interaction types: collisions and Coulomb interactions. The figure presents a local effect analysis, quantifying the influence of various physical attributes on the L2 norm of the change in a specific object's acceleration.

The analysis, visualized in Figure S12, reveals several key insights about the IWM's learned representations:

- **Quantitative Assessment with L2 Norm and Controlled Experiments:** The effect of each physical property is quantified using the L2 norm of the change in acceleration of the object of interest (Object 1 in this case). This provides a standardized measure for comparison. To obtain these results, we performed controlled experiments where we:

  1. *Discretized and varied physical properties:* We systematically and independently varied each physical latent input (representing properties like mass, velocity, etc.) during the IWM's interaction modeling. To ensure comprehensive coverage of each property's influence, we discretized the range of possible values for each property into a set of distinct bins. This allowed us to test the IWM's response to a representative set of values across the entire plausible range for each property, rather than relying on a potentially biased or limited sample of continuous values.
  2. *Observed acceleration changes:* For each discrete value (bin) of a latent variable, we recorded the resulting change in the object's acceleration.
  3. *Calculated the L2 norm:* We then calculated the L2 norm (Euclidean norm) of this change in acceleration, providing a single scalar value representing the magnitude of the effect.

- **Context-Dependent Feature Importance:** The two subplots in Figure S12 demonstrate a crucial capability of the IWM: the ability to adapt its sensitivity to different physical properties depending on the interaction context. In the collision scenario (top subplot), properties directly related to momentum and impact (mass, velocity, position) have a strong influence on Object 1's acceleration, while appearance-related properties (color, material) have minimal impact. This selective attention to relevant features is further supported by the findings in Figure S13 (left), which shows that the IWM focuses its computational resources on the frames immediately surrounding the collision event. In contrast, the Coulomb interaction scenario (bottom subplot of Figure S12) highlights the dominance of charge and position, consistent with the principles of electromagnetism. The IWM's ability to capture these nuanced relationships, as also qualitatively illustrated in the Coulomb force dynamics visualization (Figure S13 right), suggests that it is not simply learning a fixed set of correlations but rather adapting its internal representations to the specific physics governing each interaction type.

- **Prioritization of Physically Relevant Properties:** Across both interaction types, the IWM consistently prioritizes properties that are physically relevant to the outcome. This aligns with the core principles of intuitive physics, where factors like mass, velocity, and position are crucial for predicting collision dynamics, and charge and position are paramount for understanding electrostatic interactions. The model's ability to disregard or downplay the influence of irrelevant features like color and material further strengthens this observation.

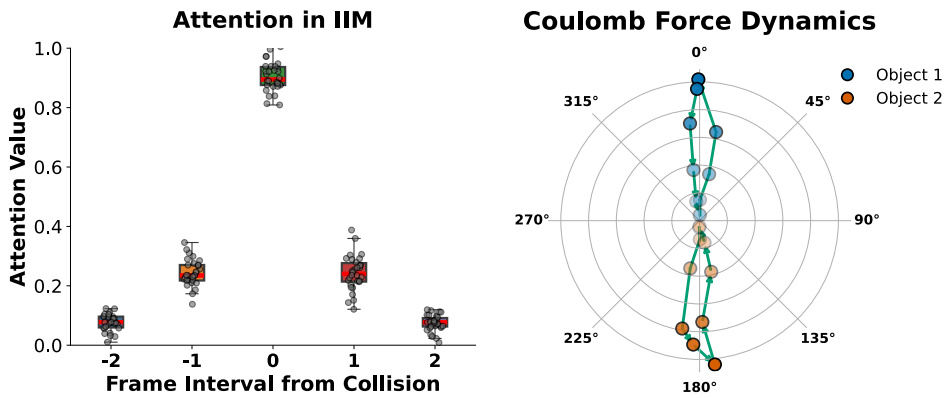


Figure S13: **Analysis of IWM's attention and interaction dynamics.** (Left) Distribution of attention values within the IWM as a function of frame interval from collision, showing highest attention at the collision frame (interval 0). (Right) Visualization of Coulomb force dynamics between two interacting objects, with radial distance representing force magnitude and angle representing object orientation; arrows indicate the trajectory over time.

- **Learning Realistic Physical Principles:** The overall pattern of results strongly suggests that the IWM is learning representations that reflect realistic physical principles. The model's sensitivity to the appropriate properties in each context (collision vs. Coulomb interaction) indicates an understanding of the underlying causal mechanisms at play.

In conclusion, the local effect analysis presented in Figure S12 provides compelling evidence that the IWM learns to identify and prioritize physically relevant features in a context-dependent manner. This ability to capture both the correlations and independencies between physical concepts, and to adapt its representations to different interaction types, highlights the IWM's potential as a model of intuitive physics understanding. Further research could explore the model's performance on a wider range of physical scenarios and with more complex interactions to assess the generalizability of these findings.

# References


[1] Q. Tang, X. Zhu, Z. Lei, and Z. Zhang, "OBJECT DYNAMICS DISTILLATION FOR SCENE DECOMPOSITION AND REPRESENTATION," in *International Conference on Learning Representations*, 2022. [Online]. Available: https://openreview.net/forum?id=oJGDYQFKL3i

[2] 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.

[3] 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.

[4] 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.

[5] Q. Tang, X. Zhu, Z. Lei, and Z. Zhang, "Intrinsic physical concepts discovery with object-centric predictive models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023, pp. 23 252–23 261.

[6] L. Downs, A. Francis, N. Koenig, B. Kinman, R. Hickman, K. Reymann, T. B. McHugh, and V. Vanhoucke, "Google scanned objects: A high-quality dataset of 3d scanned household items," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 2553–2560.

[7] K. Greff, S. Van Steenkiste, and J. Schmidhuber, "On the binding problem in artificial neural networks," *arXiv preprint arXiv:2012.05208*, 2020.

[8] L. S. Piloto, A. Weinstein, P. Battaglia, and M. Botvinick, "Intuitive physics learning in a deep-learning model inspired by developmental psychology," *Nature human behaviour*, vol. 6, no. 9, pp. 1257–1267, 2022.

[9] B. Dai, L. Wang, B. Jia, Z. Zhang, S.-C. Zhu, C. Zhang, and Y. Zhu, "X-voe: Measuring explanatory violation of expectation in physical events," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 3992–4002.

[10] T. Gerstenberg and J. B. Tenenbaum, "Intuitive theories," 2017.

[11] R. Baillargeon, "Innate ideas revisited: For a principle of persistence in infants' physical reasoning," *Perspectives on Psychological Science*, vol. 3, no. 1, pp. 2–13, 2008.

[12] ——, "Infants' physical world," *Current directions in psychological science*, vol. 13, no. 3, pp. 89–94, 2004.

[13] Y. LeCun, "A path towards autonomous machine intelligence," *preprint posted on openreview*, 2022.