

X-VoE: Measuring eXplanatory Violation of Expectation in Physical Events

Bo Dai^{1,2}, Linge Wang³, Baoxiong Jia², Zeyu Zhang², Song-Chun Zhu^{1,2,3}, Chi Zhang^{2,✉}, Yixin Zhu^{4,✉}

<https://github.com/daibopku/X-VoE>

¹ School of Intelligence Science and Technology, Peking University

³ Department of Automation, Tsinghua University

✉ zhangchi@bigai.ai, yixin.zhu@pku.edu.cn

² Beijing Institute for General Artificial Intelligence

⁴ Institute for Artificial Intelligence, Peking University

Abstract

Intuitive physics is pivotal for human understanding of the physical world, enabling prediction and interpretation of events even in infancy. Nonetheless, replicating this level of intuitive physics in artificial intelligence (AI) remains a formidable challenge. This study introduces X-VoE, a comprehensive benchmark dataset, to assess AI agents' grasp of intuitive physics. Built on the developmental psychology-rooted Violation of Expectation (VoE) paradigm, X-VoE establishes a higher bar for the explanatory capacities of intuitive physics models. Each VoE scenario within X-VoE encompasses three distinct settings, probing models' comprehension of events and their underlying explanations. Beyond model evaluation, we present an explanation-based learning system that captures physics dynamics and infers occluded object states solely from visual sequences, without explicit occlusion labels. Experimental outcomes highlight our model's alignment with human commonsense when tested against X-VoE. A remarkable feature is our model's ability to visually expound VoE events by reconstructing concealed scenes. Concluding, we discuss the findings' implications and outline future research directions. Through X-VoE, we catalyze the advancement of AI endowed with human-like intuitive physics capabilities.

1. Introduction

Humans possess a profound understanding of the physical world, enabling them to predict the outcomes of physical interactions and events [6]. From infancy, humans demonstrate intuitive physics, comprehending actions and consequences even in unfamiliar scenarios. For the machine learning community, the challenge lies in emulating this level of intuitive physics understanding. This study introduces X-VoE, a comprehensive benchmark dataset designed to assess and push the limits of AI agents' intuitive physics comprehension.

The notion of intuitive physics, observed even in young infants, has been foundational in cognitive science and develop-

	Origin video	Surprise		Explaining result
		w-o explain	w. explain	
predictive (S1)		○	=	
		!	=	
hypothetical (S2)		!	≠	
		○	=	
explicative (S3)		!	≠	
		○	≠	

Figure 1: **Evaluation settings in the ball blocking exemplar scenario of X-VoE.** The explanation video illustrates potential hidden dynamics. Circles denote no surprise, and exclamation marks indicate surprise. In the predictive setup (S1), a solvable pair is presented without requiring explanation: predicting observed entities' dynamics suffices to reason about the outcome. In the hypothetical setup (S2), perceiving the direction of outgoing balls might lead to surprise, yet alternate explanations exist—e.g., a hidden blocker behind the wall causing ball rebound. However, a random agent's scores show negligible disparity, necessitating the explicative setup (S3) to discern surprises, demanding explanatory ability absent in predictive-only or random agents.

mental psychology [36]. Infants show surprise when physical events violate their expectations, indicating an understanding of fundamental physical principles [5]. Explanation-based learning has been proposed as a mechanism contributing to the development and refinement of intuitive physics understanding [4]. However, recent advances in this field have primarily resulted in predictive models, lacking the explanatory capacity and falling short of capturing even infant-level intuitive physics comprehension [30, 35].

Central to our work is the Violation of Expectation (VoE) paradigm, widely employed in psychological studies to evaluate infants' intuitive physics understanding [3, 5]. In this paradigm, participants exhibit surprise, indicated by prolonged attention, when exposed to events that either follow or violate intuitive physics laws. Inspired by the effective-

ness of this paradigm, we adopt it to evaluate AI agents' intuitive physics comprehension. In each trial, models encounter experiments adhering to or contravening intuitive physics laws. Models succeed in the VoE test if they display high surprise scores for physics-violating experiments and lower scores for compliant ones.

Existing works within the machine learning and computer vision community have embraced the VoE paradigm [10, 30, 32, 35, 43]. However, most of these efforts primarily focus on predictive abilities, disregarding the explanatory component [1, 29, 30, 32, 35, 37]. This perspective neglects the fundamental aspect of VoE—the act of explaining observed events. In psychological studies, human participants express surprise not at the moment a physics-violating event occurs, but upon learning of its outcome. This observation underscores the significance of explanation within VoE.

Motivated by these insights, we introduce X–VoE, an intuitive physics evaluation dataset designed specifically to incorporate explanation within VoE. Distinct from previous efforts that concentrated on predictive scenarios, our dataset encompasses setups that require explaining observed events in diverse VoE situations. We establish three VoE settings for each of the four scenarios: ball collision, blocking, object permanence, and continuity (see Fig. 2). Each scenario features predictive, hypothetical, and explicative setups. Notably, the three setups within the ball-blocking scenario distinguish explanatory agents from predictive and random ones.

Furthermore, we propose the eXplanation-based Physics Learner (XPL) model to emulate the explanation-based VoE process, inspired by findings in human studies [3, 4]. While XPL is adaptable to diverse deep architectures, we specifically build it upon PLATO [30] due to its robust performance. Our model incorporates three self-supervised modules: perception for image encoding, Transformer reasoning for occluded object prediction, and dynamic reasoning for simulating physical dynamics. Importantly, our model introduces a reasoning sub-component to update representations of occluded objects, akin to infants' explanation-based learning when confronted with unexpected outcomes [3].

In summary, our work makes three significant contributions:

- Introduction of X–VoE, a comprehensive intuitive physics evaluation dataset that challenges AI agents not only in predictive capabilities but also in their capacity to explain. The dataset covers four distinct scenarios, each with predictive, hypothetical, and explicative setups. This allows for a more comprehensive assessment of intuitive physics understanding within VoE.
- Proposition of the XPL model, enhancing existing approaches with an explanatory module that improves VoE evaluation. Our model comprises three modules—perception, reasoning, and dynamics learning—for holistic comprehension and simulation of physical dynamics.

- Experimental demonstration of XPL's enhanced performance in alignment with human commonsense compared to other baselines in X–VoE. Additionally, XPL offers insights into hidden factors, as depicted in Fig. 1.

2. Related work

Intuitive physics Intuitive physics forms a cornerstone of human cognition, enabling rapid and accurate predictions about moving object trajectories [19]. To evaluate machine understanding in this realm, benchmark datasets have emerged, often focusing on predicting future states [6, 8, 20, 45, 9] or inferring object properties [21, 22, 34]. These methods predominantly gauge model performance by comparing generated predictions to ground truth.

More recently, the Violation of Expectation (VoE) paradigm has garnered attention within the machine learning and computer vision community [10, 30, 32, 35, 43]. Rooted in developmental psychology, the VoE paradigm quantifies model surprise when presented with events that challenge intuitive physics laws. This perspective provides an alternative angle for assessing intuitive physics understanding. Notably, the IntPhys dataset [32] pioneered this VoE-based benchmarking approach. ADEPT [35] introduced a model combining re-rendering and object tracking. PLATO [30] decomposed the learning process into perception and dynamics prediction. Differing from conventional intuitive physics learning, the VoE paradigm does not rely on absolute ground truth. Instead, it hinges on relative measures of surprise, akin to developmental studies that assume higher responses indicate increased surprise. This emphasizes the role of explanation in VoE, as demonstrated in Fig. 1. In contrast to prior works that often neglected this vital component, our X–VoE includes scenarios that demand both traditional prediction-based understanding and explanation-based comprehension. Additionally, we propose an explanation-enhanced physics learner, XPL, which achieves improved performance and interpretability by incorporating explanations.

Video prediction The challenge of comprehending videos and making plausible predictions of future states from current observations has been a longstanding problem within computer vision [2, 27, 28], closely connected to the VoE paradigm. Solving VoE problems frequently involves predicting future frames for inference and evaluation. However, this prediction task is intricate due to the inherent complexity of modeling real-world dynamics and conditional image synthesis [38, 44]. Within the computer vision community, various architectures have been explored to address these challenges and enhance the quality of generated images [38, 44]. The task is further complicated by the need to model relationships between frames, leading to approaches that integrate spatial transformations over time [15, 25, 31]. Disentanglement of motion and content has also been pursued [11, 18, 24, 40]. More recent efforts involve learning physics-based dynam-

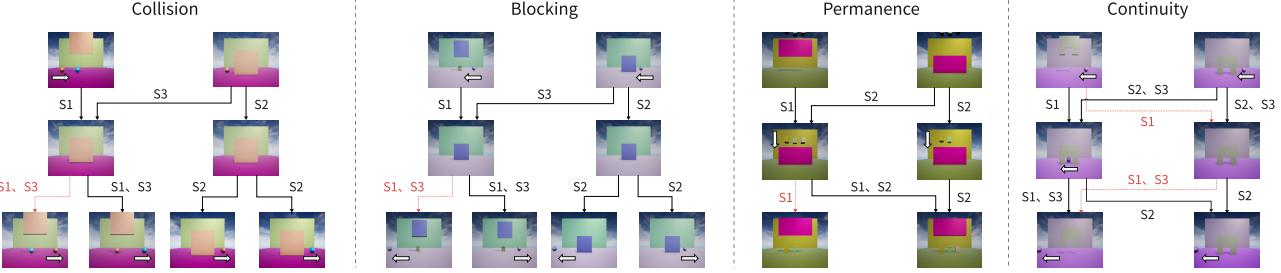


Figure 2: **Testing scenarios in X-VoE: ball collision, blocking, object permanence, and object continuity.** Within each scenario, frames in a testing video are linked by the same setup identification number (*e.g.*, S1). Black links denote non-surprising videos, while red links indicate surprising ones. Notably, certain videos require explanation to become non-surprising. For example, in the right S2 branch of the object permanence scenario, three cubes on the floor become non-surprising due to preceding observation of two cubes dropping, suggesting a hidden cube behind the wall.

ics from videos and reasoning about unknown factors [16]. Within X-VoE, we assess the performance of these video prediction models as baseline methods.

Object-centric dynamics The “vision-as-inverse-graphics” framework and the versatility of physics simulation have led to models based on physics simulation, which offer notable advantages in terms of accuracy and generality [8, 33]. However, these models are often heavily reliant on specific physics engines, limiting their flexibility. In response, recent works have leveraged graph neural networks and object-centric representations to mitigate this dependence [30, 42]. By abstracting irrelevant signals and focusing on objects, these models establish a tighter mapping between visual inputs and physics engines. Further, some models can directly simulate real physics engines [6, 13, 45]. These object-centric dynamics models have demonstrated the ability to capture intricate dynamics. Our approach in X-VoE aligns with this framework, using object-centric representations for downstream computation and reasoning.

3. Generating X-VoE

Our X-VoE dataset encompasses four distinct scenarios, covering ball collision, ball blocking, object permanence, and object continuity. To evaluate various intuitive physics principles, each scenario, except object permanence, comprises three distinct settings: predictive, hypothetical, and explicative, as illustrated in Fig. 2. Within each setting, we create 1,000 procedurally generated scene pairs using Unreal Engine 4. Importantly, X-VoE primarily serves as a test suite for evaluating intuitive physics understanding, with no constraints on model training data.

3.1. Testing data

We generate testing videos that span four key aspects of object dynamics: ball collision, ball blocking, object permanence, and object continuity. Refer to Fig. 2 for a visual overview.

Collision In this scenario, a ball traverses the scene, while an occlusion wall is positioned centrally. In the predictive setting (S1), we design a scenario where a ball of differing color but identical mass stands behind a wall. The incoming ball collides with this hidden ball, resulting in the incoming ball coming to a halt and the concealed ball continuing its trajectory. To introduce VoE effects, we enable the incoming ball to pass through the hidden ball. In the hypothetical setting (S2), we create a scene featuring a central wall concealing objects behind it. An incoming ball enters the scene from the left and rolls behind the wall. In some cases, an additional ball appears to pass through the wall, while in others, the incoming ball does so. This distinction hinges on whether an unseen ball is situated behind the wall. The explicative setting (S3) closely mirrors the hypothetical setting, but we lift the wall to reveal the concealed scene’s contents.

Blocking The blocking scenario is conceptually similar to the collision scenario, substituting the hidden ball with a stationary cube. The impact of the incoming ball causes it to rebound upon collision with the cube.

Object permanence Drawing inspiration from developmental psychology literature, we recreate a scenario involving cubes falling to the ground and becoming occluded by a wall. In the predictive setting (S1), we devise a case where a wall descends to an initially vacant ground, followed by three cubes falling behind the wall. To elicit VoE effects, we raise the wall, revealing fewer than three objects. In the hypothetical setting (S2), the scenario begins with a wall positioned centrally, obscuring objects behind it. Three or two cubes fall behind the wall. When the wall is lifted, the scene consistently features three cubes, even when only two cubes initially fell. This reflects the possibility of one cube being hidden behind the wall from the outset.

Object continuity Motivated by psychology studies [1], we introduce a wall with a lower-half window. This setup allows a ball to traverse the scene from one side to the other. The ball becomes occluded when behind the wall, emerges

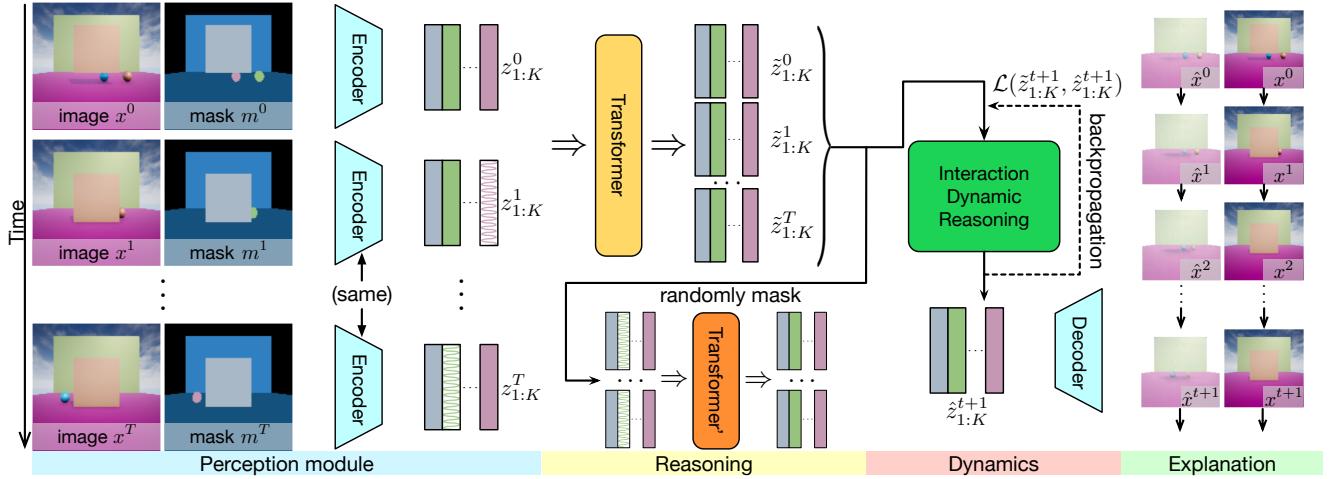


Figure 3: **Overview of the XPL model for explanation-based physics learning.** The model comprises three key modules: (i) the perception module, responsible for extracting object-centric representation from RGBD videos and segmentation masks; (ii) the reasoning module, utilizing two Transformer networks to infer representations of occluded objects; (iii) the dynamics module, which acquires intuitive physical knowledge and refines reasoning outcomes to align with intuitive physics. Additionally, the inferred object representation can be visualized using the decoder from the perception module, offering a **visual explanation** of events occurring behind the wall. Wavy curves indicate masking. Refer to the text for comprehensive details.

through the window, disappears, and subsequently reappears from the opposite end. The three distinct settings mirror the collision and blocking scenarios. The differentiation between plausible and implausible scenes revolves around whether the ball remains visible upon passing through the window. In the predictive setting (S1), all relevant information is presented at the video’s outset and conclusion, negating the presence of hidden objects. In the hypothetical setting (S2), information is deliberately withheld from the video’s start and finish, necessitating the model’s performance to align with infants [1], which involves explaining the existence of two balls. In the explicative setting (S3), the wall is lifted, verifying the absence of an additional ball behind the wall.

3.2. Training data

Though we do not impose constraints on the training data, for this study, we generate data adhering to the same structure as the test scenarios but without VoE effects. As shown in Fig. 4, the training set consists of 100,000 procedurally generated scenes, closely mirroring the scale used for training PLATO [30]. During training, we exclusively present videos following intuitive physics laws, raising the wall at the beginning and end of each video. This approach reduces reasoning complexity, simulating the developmental process where only non-surprising physical events are observed. Consequently, models must unsupervisedly learn from video sequences depicting ordinary scenes, developing intuitive physics understanding necessary for VoE. Furthermore, for the collision and blocking scenarios, we create videos depicting balls passing through walls without collision or obstruction, demonstrating the unimpeded path behind the wall as shown in Fig. 4(a). We also generate scenes

similar to the previously described settings but devoid of occlusion walls.

4. eXplanation-based Physics Learner (XPL)

4.1. Framework

Our proposed eXplanation-based Physics Learner (XPL) model draws inspiration from developmental psychology theories concerning infancy. As depicted in Fig. 3, the XPL model comprises three key components: (1) a perception module responsible for extracting object-centric representations to facilitate downstream processing, (2) a reasoning module tasked with inferring occluded object states by considering both spatial and temporal contexts, and (3) a dynamics module designed to acquire physical insights and evaluate inference outcomes for occluded objects.

Perception The perception module is designed to process input RGBD video sequences, represented as $\langle x^0, x^1, \dots, x^T \rangle$, alongside their corresponding segmentation masks, denoted as $\langle m^0, m^1, \dots, m^T \rangle$. The masks are generated using a pre-trained segmentation model. Notably, the simplicity of the scenes allows for direct use of ground truth segmentation, as observed in PLATO [30]. For each frame, the perception module employs a Component Variational Autoencoder (Component VAE) [7] to transform each input image into a concealed vector representation $\langle z_{1:K}^0, z_{1:K}^1, \dots, z_{1:K}^T \rangle$, where K represents the object count per frame.

Reasoning The reasoning module leverages the object embeddings obtained from the perception module as input and endeavors to enhance scene comprehension by inferring the attributes of occluded objects, whose masks remain va-

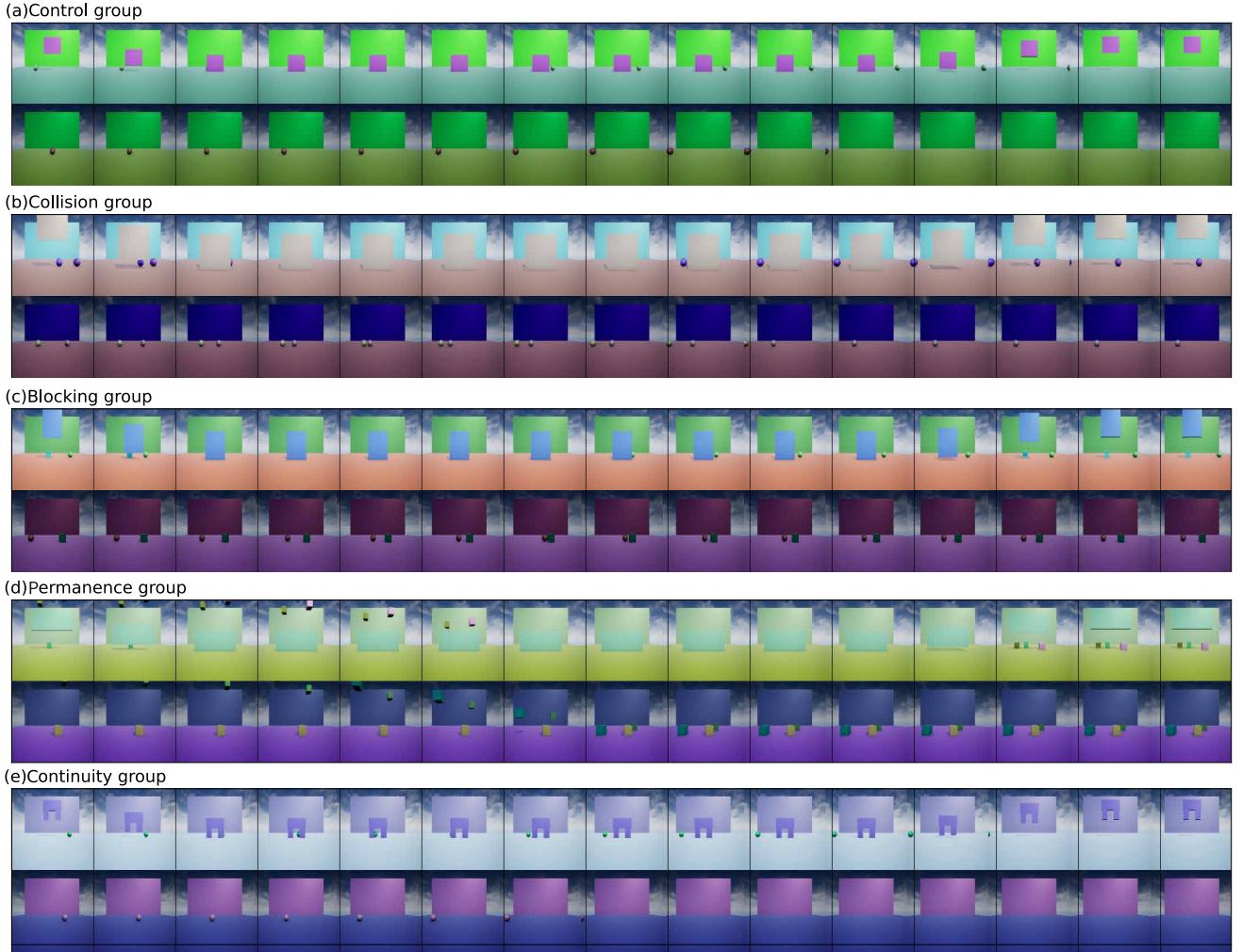


Figure 4: **Training scenarios for X-VoE.** The timeline progresses from left to right, where each row represents the control, collision, blocking, object permanence, and object continuity groups from top to bottom. Please refer to Appx. A.2 for additional details.

cant due to occlusion. This aspect employs two Transformer models to refine object embeddings and recover hidden objects. Both Transformers adopt flattened spatial-temporal embeddings and apply global attention mechanisms to contextualize information. The first Transformer refines input features of occluded objects to align with a learned dynamics module, producing \tilde{z} . The second Transformer is responsible for recuperating objects concealed within observation sequences of both original and refined features. It’s important to note that object recovery mirrors Masked Autoencoding [17], treating a random object as absent and necessitating reconstruction from contextual cues. Drawing from these observations, we train the second Transformer similarly to Masked Autoencoders (MAE).

Dynamics The dynamics module predicts object embeddings $\hat{z}_{1:K}^{t+1}$ in the succeeding frame based on the preceding frame’s refined object embeddings $\tilde{z}_{1:K}^{1:t}$. This involves employing the interaction dynamics module introduced in

PLATO [30], supplemented by a residual module. Unlike PLATO, we employ object embeddings subsequent to the reasoning module and jointly train the modules.

4.2. Model training

Initially, we pre-train the perception module to equip the system with foundational visual capabilities. Precisely, the perception module undergoes pre-training using RGBD images and segmentation masks. Throughout this phase, we segment objects and employ masked images for VAE training. During image reconstruction, depth information assists in calculating object mask details.

We then train one Transformer and the dynamics module, with latent codes frozen from the perception module, in an end-to-end manner employing the following loss:

$$\begin{aligned}\tilde{z} &= f_{\text{inf}}(z) \\ \mathcal{L} &= \|f_{\text{dyn}}(\tilde{z}_{1:K}^{0:t}) - \tilde{z}_{1:K}^{t+1}\|_2,\end{aligned}\quad (1)$$

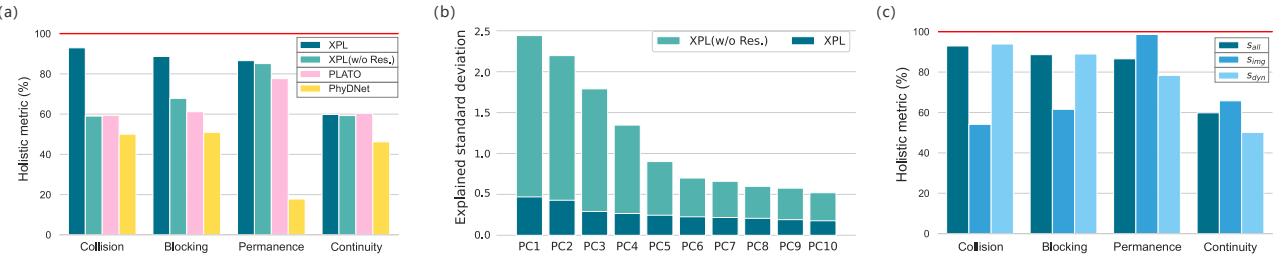


Figure 5: (a) **Performance of different models on X–VoE under the holistic metric.** The red line denotes the ideal performance. (b) **PCA with or without residual connection.** The first ten principal components are shown. (c) **Results from each score component.**

Here, the Transformer employs the architecture featured in Aloe [12] ($f_{\text{inf}}(\cdot)$), while the dynamics prediction module aligns with PLATO [30] ($f_{\text{dyn}}(\cdot)$). The second Transformer is trained independently using MAE.

5. Experiments

In this section, we thoroughly evaluate the performance of XPL using our X–VoE dataset across different experimental configurations: predicting future phenomena (predictive setup), interpreting existing phenomena (hypothetical setup), and understanding past occurrences given future conditions (explicative setup). We compare XPL against PhyDNet [16], a video prediction model, and PLATO [30] in our X–VoE dataset. These models are evaluated under two different metrics.

5.1. Defining accuracy and surprise

Before delving into different evaluative configurations, we first introduce how accuracy and surprise are formally defined.

In developmental psychology experiments on VoE, a surprise was defined by comparing infants' responses to normal scenes with those that violate expectations. Similar to existing works [35], we borrow the idea and define the model accuracy as the relative scores between two videos, one that violates intuitive physics laws and another that does not:

$$\text{Accuracy} = \frac{1}{N} \sum \mathbb{1}[s_{\text{nor}} < s_{\text{sur}}], \quad (2)$$

where N denotes the total number of such pairs, and s_{nor} and s_{sur} are scores of a normal physics video and one that violates physics, respectively. The scores are computed as the sum of the difference between the inferred results from the observation and that from the dynamics module's prediction, *i.e.*,

$$s = s_{\text{img}} + s_{\text{dyn}}, \quad (3)$$

where

$$s_{\text{img}} = \sum_{t=1}^T \ell(\mathbf{I}_t, \sum_i f_{\text{dec}}(\tilde{z}_i^t)), \quad (4)$$

and

$$s_{\text{dyn}} = \sum_{t=2}^T \ell(\sum_i f_{\text{dec}}(\tilde{z}_i^t), f_{\text{dec}}(f_{\text{dyn}}(\tilde{z}_{1:K}^{0:t-1}))). \quad (5)$$

Here, $f_{\text{dec}}(\cdot)$ denotes the learned decoder in our VAE, and we use MSE loss for $\ell(\cdot)$.

5.2. The holistic metric

Similar to Smith *et al.* [35], we adopt the holistic metric to evaluate VoE effects in all pairs of unexpected and normal event videos. Ideally, an intuitive physics model should produce higher surprise scores for unexpected events. Formally, the holistic metric is defined as such,

$$\frac{1}{n_s n_c} \sum_{i,j} \mathbb{1}[s(x_i^+) > s(x_j^-)], \quad (6)$$

where x_i^+ and x_j^- denote the unexpected and normal videos and n_s and n_c are the number of unexpected and normal videos. This metric aggregates results from all confounding factors, including interference from colors, shapes, scene complexity, *etc.* Therefore, it provides a holistic view of models' understanding of intuitive physics events; models need to judge the unexpectedness of outcomes from the intuitive physics perspective, disentangling all other confounding factors.

As shown in Fig. 5 (a), we measure the holistic value on different models on X–VoE. Both XPL and PLATO show better performance in all four testing scenarios, though with a notable gap from perfection. XPL is significantly better than PLATO in the collision, blocking, and permanence, but less so in continuity. We also compare different dynamic modules, with or without residual, in XPL. The results show that the residual connection in the dynamics module plays a critical role in our system, as evidenced by results for collision and blocking. An in-depth analysis from Principal Component Analysis (PCA) in Fig. 5 (b) shows that after adding the residual connection, the standard deviation in different principal components is particularly reduced, making learning easier.

To investigate the contribution of each of the two surprise components in Eq. (3), we compute the holistic metric from

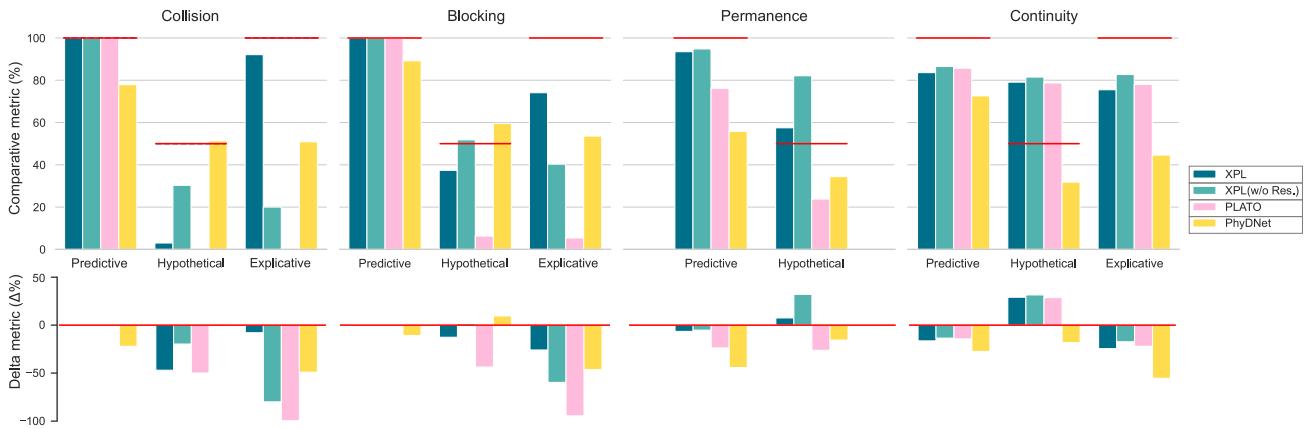


Figure 6: **Performance of different models on X-VoE under the comparative metric.** The red line denotes the ideal performance. The top part shows the absolute comparative values and the bottom part shows the difference from the ideal.

each of them separately. As shown in Fig. 5 (c), the performance of s_{dyn} is superior to that of s_{img} in the collision and blocking scenarios, whereas the performance of s_{img} is better in permanence and continuity. This result implies that the violation of physical knowledge plays a more important role in collision and blocking. In contrast, the mismatch from the observation is a more crucial factor for permanence and continuity. Thus, the residuals in XPL, explicitly taking earlier information into computation, could exert a greater influence on the dynamic module and its impact in the collision and blocking scenarios as shown in Fig. 5 (a).

The holistic metric only provides a global view of how a model understands intuitive physics. To paint a more complete landscape of a model, we look deeper into the comparative metric in the next section.

5.3. The comparative metric

The comparative metric, similar to ones proposed in literature [32, 43], is calculated in a pair of the unexpected and normal events within one specific setting in each scenario,

$$\frac{1}{n} \sum_i \mathbf{1}[s(x_i^+) > s(x_i^-)], \quad (7)$$

where x_i^+ and x_i^- are the two paired videos in each settings and n is the number of such pairs. The comparative metric is also most commonly used in evaluating infants' intuitive physics knowledge in developmental psychology [5, 23].

Whereas the holistic metric describes whether an observation sequence is absolutely surprising from a holistic perspective, the comparative metric assesses whether one observation sequence is more surprising than another from a comparative perspective. Although the holistic metric provides an overall perspective, it lacks the detailed results of the three specific cases the comparative metric provides; see Fig. 1. In each scenario in X-VoE, the two videos in the

hypothetical setting are likely to occur, while only one of the two videos in the predictive and explicative settings is likely to occur. Therefore, the comparative metric in the hypothetical setting should be ideally 50%, while the metric in the predictive and explicative settings should be ideally 100%.

Fig. 6 shows the comparative values of different models. The results in the predictive setting indicate that current AI systems, even as simple as general video prediction, can easily predict future outcomes accurately for such a simple task. However, when it comes to the setting that requires reasoning and explanation (*i.e.*, explicative), only XPL can consistently achieve over 50%. When common predictive models can only predict future occurrences based on past conditions, XPL can reason about the past conditions that lead to the observation, a critical ability necessary for successfully solving the explicative setting.

Of these, the hypothetical setting is where we notice the most performance volatility. For the hypothetical setting, both a random-answering human subject and an ideal human subject with perfect understanding would reach 50% accuracy. However, this is exactly why this problem is intriguing for psychologists. From this perspective, a model achieving 50% could mean it is either the worst or best. While in the hypothetical setup, PhyDNet achieves nearly 50%, it can only reach random-level performance in the explicative setting, showing that the model does not understand different possibilities behind the wall. This is why the explicative setting is so important. The explicative setting provides more new information in the video follow-up than the hypothetical setting. As shown in Fig. 1, the new information will change a possible scene to an impossible scene in the hypothetical setting. The metric gap between the hypothetical setting and explicative setting shows the power of the explanatory abilities. XPL demonstrates this property on both collision and blocking scenarios, especially on the collision scenario, where this gap reaches close to 90%.

Training data	Origin				⇒	PLATO				⇒	XPL			
Time	1	6	10	15		1	6	10	15		1	6	10	15
Collision														
Blocking														
Permanence														
Continuity														

Figure 7: Training: Visualization of the internal representation in PLATO and XPL during training.

Although the XPL with or without a residual module both have the reasoning module, they still have different explanatory abilities for hypothetical and explicative settings. In collision and blocking tasks, residuals' presence improves the explicative but not the hypothetical setting. The residual module enhances the connection between two consecutive frames, allowing the reasoning module to better infer the previous state based on the subsequent state. The main difference between the hypothetical and explicative setting is the inclusion of follow-up information. In the explicative setting, the presence of follow-up information enhances the performance of the reasoning module (with residual module) due to more subsequent state information. However, in the hypothetical setting, the absence of follow-up information negatively impacts the module's performance.

Overall, XPL improves over previous state-of-the-art but still fares worse on collision and continuity. While developmental psychology experiments have found the ability in infants [1], it remains a challenge for AI systems.

5.4. Visualization results

The challenge of visual occlusion persists in computer vision. Unless the ground-truth value is given directly, it is difficult to characterize occluded objects by vision alone, especially in the case of complete occlusion. However, humans can deduce occluded objects and corresponding physical phenomena intuitively, even under complete occlusion. We investigate whether XPL can reason about occluded objects through visualization.

We visualize occluded objects within the learned representation. Specifically, we mask the token associated with the wall and decode the resulting features to assess the model's ability to reconstruct hidden objects. Training visualization results are presented in Fig. 7. Notably, PLATO lacks a dedicated reasoning module for occluded objects, resulting in an inability to recover occluded factors. Conversely, XPL grad-

Test data	Origin				⇒	XPL			
Time	1	6	10	15		1	6	10	15
Predictive(S1)									
Hypothetical(S2)									
Explicative(S3)									

Figure 8: Testing: Visualization of the inferred internal representation in XPL during testing. This example corresponds to the settings in Figure 1.

ually learns to infer the presence of occluded objects behind the wall to explain observations. Crucially, we never provide ground-truth occluded object representations during training, emphasizing the importance of synchronized training of the inference and dynamic modules. This approach allows XPL to achieve improved occluded object restoration, though it still falls short of ground-truth results (Fig. 7).

For test visualization, detailed results corresponding to Fig. 1 are showcased in Fig. 8. The predictive setting demonstrates XPL's accurate reconstruction of observed objects. In the hypothetical setting, XPL provides coherent explanations involving hidden object interactions. In the explicative setting, the occluder is lifted toward the end of the videos, resulting in surprising outcomes.

To conclude, XPL proficiently reconstructs occluded objects and provides visual explanations for various events, underscoring its capacity to reason about hidden factors in the context of intuitive physics.

6. Conclusion and discussion

In this paper, we introduced X-VoE, a novel explanation-based Violation of Expectation (VoE) dataset consisting of four distinct scenarios, each encompassing three unique settings: predictive, hypothetical, and explicative. While the predictive setting aligns with conventional VoE tasks, the other two settings focus on evaluating a model's explanatory capacity. Our proposed XPL combines reasoning and explanation processes to address occluded objects, offering enhanced performance within the X-VoE settings. Our experiments revealed that XPL excels in scenarios requiring explicit explanations for occluded objects, positioning it ahead of other methodologies. Notably, the decoded representation from XPL offers visual explanations for occluded events, highlighting its ability to reason about hidden factors.

Our work underscores the pivotal role of explanations in VoE tasks, particularly concerning occluded objects and their contribution to video comprehension. Even when objects are obscured by walls, the possibility of underlying physical events remains, and a model equipped with explanation capabilities performs more adeptly in such situations. The capacity to reason about occluded objects extends the model's scope beyond mere video prediction, enabling it to capture intuitive physics principles more effectively.

However, certain challenges persist. Notably, XPL encounters difficulties in scenarios that demand high-level explanations, such as the hypothetical setting in collision or continuity (Fig. 6). These limitations underscore the need for further advancements in the reasoning aspect of our model, paving the way for future research. The ability to handle complex interactions and provide meaningful explanations remains a challenging aspect that requires careful consideration in model design.

In conclusion, while our model's reasoning capabilities are still a work in progress, our study sheds light on the integration of explanations into VoE tasks, aiming to develop models with a level of intuitive physics comprehension akin to infants. The focus on occluded objects and their explanatory potential broadens the scope of VoE tasks and encourages the development of AI systems with deeper understanding.

6.1. Limitations

Method Despite its strengths, XPL faces certain limitations. It struggles in some experiments, particularly the hypothetical setting in collision or continuity (Fig. 6), where its performance falls short of human-like comprehension. Furthermore, our explanation process employs a basic Transformer module, lacking physics-related inductive biases that could enhance performance. A promising direction for future research lies in incorporating domain-specific inductive biases that exploit physical principles to improve reasoning and explanatory capabilities.

Accuracy metric Although our accuracy metrics draw inspiration from developmental psychology experiments and prior works, they rely on video comparisons to evaluate violations of intuitive physics. This approach, while effective, assumes that one of the videos violates intuitive physics laws, even if the difference in surprise values is marginal. As a result, the method might struggle to achieve the desired metrics in scenarios like the hypothetical setting. Exploring metrics that focus on higher-level concepts and the detection of fundamental violations could yield insights into the underlying mechanisms that drive these evaluations.

Dataset X-VoE pioneers the evaluation of physical explanatory abilities in VoE tasks. However, our test scenarios could be more diverse and comprehensive. Future efforts will expand and diversify these scenarios to create a more robust framework for testing intuitive physics understanding in VoE. By incorporating a wider range of physical phenomena and interactions, future datasets can challenge AI systems with greater complexity.

6.2. Future Directions

Future research should focus on refining XPL's reasoning capabilities, enhancing its performance in scenarios demanding higher-order explanations. Introducing more sophisticated physics-based inductive biases could contribute to better occluded object reasoning. Additionally, exploring hybrid approaches that combine neural networks with symbolic reasoning could lead to more advanced models with enhanced explanatory capabilities.

Additionally, X-VoE can serve as a stepping stone for designing more intricate and varied VoE scenarios. Incorporating more complex physical interactions, occlusions, and multiple objects would lead to a richer and more challenging testbed for evaluating AI systems' intuitive physics comprehension. Diverse scenarios can provide comprehensive evaluation of models' understanding across a wide range of intuitive physics principles.

In summary, our study provides insights into the integration of explanations in VoE tasks and sets the stage for future advancements in both model design and dataset development. The intersection of explanations and intuitive physics comprehension holds promise for creating AI systems that not only predict events but also understand the underlying physical principles that govern them.

Acknowledgment

The authors would like to thank four anonymous reviews for constructive feedback, Huiyin Li (BIGAI) for designing the figures, and NVIDIA for their generous support of GPUs and hardware. This work is supported in part by the National Key R&D Program of China (2022ZD0114900) and the Beijing Nova Program.

References

- [1] Andréa Aguiar and Renée Baillargeon. Developments in young infants' reasoning about occluded objects. *Cognitive Psychology*, 45(2):267–336, 2002. 2, 3, 4, 8
- [2] Mohammad Babaeizadeh, Chelsea Finn, Dumitru Erhan, Roy H Campbell, and Sergey Levine. Stochastic variational video prediction. In *International Conference on Learning Representations (ICLR)*, 2018. 2
- [3] Renée Baillargeon. Physical reasoning in young infants: Seeking explanations for impossible events. *British Journal of Developmental Psychology*, 12(1):9–33, 1994. 1, 2
- [4] Renée Baillargeon and Gerald F DeJong. Explanation-based learning in infancy. *Psychonomic bulletin & review*, 24(5):1511–1526, 2017. 1, 2
- [5] Renée Baillargeon, Elizabeth S Spelke, and Stanley Wasserman. Object permanence in five-month-old infants. *Cognition*, 20(3):191–208, 1985. 1, 7
- [6] Peter W Battaglia, Jessica B Hamrick, and Joshua B Tenenbaum. Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences (PNAS)*, 110(45):18327–18332, 2013. 1, 2, 3
- [7] Christopher P Burgess, Loic Matthey, Nicholas Watters, Rishabh Kabra, Irina Higgins, Matt Botvinick, and Alexander Lerchner. Monet: Unsupervised scene decomposition and representation. *arXiv preprint arXiv:1901.11390*, 2019. 4, 2
- [8] Michael B Chang, Tomer Ullman, Antonio Torralba, and Joshua B Tenenbaum. A compositional object-based approach to learning physical dynamics. *arXiv preprint arXiv:1612.00341*, 2016. 2, 3
- [9] Zhenfang Chen, Jiayuan Mao, Jiajun Wu, Kwan-Yee Kenneth Wong, Joshua B Tenenbaum, and Chuang Gan. Grounding physical concepts of objects and events through dynamic visual reasoning. In *International Conference on Learning Representations (ICLR)*, 2020. 2
- [10] Arijit Dasgupta, Jiafei Duan, Marcelo H Ang Jr, Yi Lin, Su-hua Wang, Renée Baillargeon, and Cheston Tan. A benchmark for modeling violation-of-expectation in physical reasoning across event categories. *arXiv preprint arXiv:2111.08826*, 2021. 2
- [11] Emily L Denton et al. Unsupervised learning of disentangled representations from video. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 2
- [12] David Ding, Felix Hill, Adam Santoro, Malcolm Reynolds, and Matt Botvinick. Attention over learned object embeddings enables complex visual reasoning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 6, 2
- [13] Mingyu Ding, Zhenfang Chen, Tao Du, Ping Luo, Josh Tenenbaum, and Chuang Gan. Dynamic visual reasoning by learning differentiable physics models from video and language. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 3
- [14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Trans-formers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2
- [15] Chelsea Finn, Ian Goodfellow, and Sergey Levine. Unsupervised learning for physical interaction through video prediction. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016. 2
- [16] Vincent Le Guen and Nicolas Thome. Disentangling physical dynamics from unknown factors for unsupervised video prediction. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 3, 6
- [17] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16000–16009, 2022. 5
- [18] Jun-Ting Hsieh, Bingbin Liu, De-An Huang, Li Fei-Fei, and Juan Carlos Niebles. Learning to decompose and disentangle representations for video prediction. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018. 2
- [19] James R Kubricht, Keith J Holyoak, and Hongjing Lu. Intuitive physics: Current research and controversies. *Trends in Cognitive Sciences*, 21(10):749–759, 2017. 2
- [20] Adam Lerer, Sam Gross, and Rob Fergus. Learning physical intuition of block towers by example. In *International Conference on Machine Learning (ICML)*, 2016. 2
- [21] Wei Liang, Yibiao Zhao, Yixin Zhu, and Song-Chun Zhu. What is where: Inferring containment relations from videos. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2016. 2
- [22] Wei Liang, Yixin Zhu, and Song-Chun Zhu. Tracking occluded objects and recovering incomplete trajectories by reasoning about containment relations and human actions. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2018. 2
- [23] Yi Lin, Maayan Stavans, and Renée Baillargeon. *Infants' physical reasoning and the cognitive architecture that supports it*. Cambridge University Press, 2020. 7
- [24] Runtao Liu, Zhirong Wu, Stella Yu, and Stephen Lin. The emergence of objectness: Learning zero-shot segmentation from videos. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 2
- [25] Ziwei Liu, Raymond A Yeh, Xiaoou Tang, Yiming Liu, and Aseem Agarwala. Video frame synthesis using deep voxel flow. In *International Conference on Computer Vision (ICCV)*, 2017. 2
- [26] Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 2
- [27] William Lotter, Gabriel Kreiman, and David Cox. Deep predictive coding networks for video prediction and unsupervised learning. In *International Conference on Learning Representations (ICLR)*, 2016. 2
- [28] Michael Mathieu, Camille Couprie, and Yann LeCun. Deep multi-scale video prediction beyond mean square error. In *International Conference on Learning Representations (ICLR)*, 2016. 2

- [29] Luis Piloto, Ari Weinstein, Dhruva TB, Arun Ahuja, Mehdi Mirza, Greg Wayne, David Amos, Chia-chun Hung, and Matt Botvinick. Probing physics knowledge using tools from developmental psychology. *arXiv preprint arXiv:1804.01128*, 2018. 2
- [30] Luis S Piloto, Ari Weinstein, Peter Battaglia, and Matthew Botvinick. Intuitive physics learning in a deep-learning model inspired by developmental psychology. *Nature Human Behaviour*, 6(9):1257–1267, 2022. 1, 2, 3, 4, 5, 6
- [31] Fitsum A Reda, Guilin Liu, Kevin J Shih, Robert Kirby, Jon Barker, David Tarjan, Andrew Tao, and Bryan Catanzaro. Sdcnnet: Video prediction using spatially-displaced convolution. In *European Conference on Computer Vision (ECCV)*, 2018. 2
- [32] Ronan Riocet, Mario Ynocente Castro, Mathieu Bernard, Adam Lerer, Rob Fergus, Véronique Izard, and Emmanuel Dupoux. Intphys: A framework and benchmark for visual intuitive physics reasoning. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2020. 2, 7
- [33] Ronan Riocet, Josef Sivic, Ivan Laptev, and Emmanuel Dupoux. Occlusion resistant learning of intuitive physics from videos. *arXiv preprint arXiv:2005.00069*, 2020. 3
- [34] Adam N Sanborn, Vikash K Mansinghka, and Thomas L Griffiths. Reconciling intuitive physics and newtonian mechanics for colliding objects. *Psychological Review*, 120(2):411, 2013. 2
- [35] Kevin Smith, Lingjie Mei, Shunyu Yao, Jiajun Wu, Elizabeth Spelke, Josh Tenenbaum, and Tomer Ullman. Modeling expectation violation in intuitive physics with coarse probabilistic object representations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 1, 2, 6
- [36] Elizabeth S Spelke and Katherine D Kinzler. Core knowledge. *Developmental Science*, 10(1):89–96, 2007. 1
- [37] Aimee E Stahl and Lisa Feigenson. Observing the unexpected enhances infants’ learning and exploration. *Science*, 348(6230):91–94, 2015. 2
- [38] Yu Tian, Jian Ren, Menglei Chai, Kyle Olszewski, Xi Peng, Dimitris N Metaxas, and Sergey Tulyakov. A good image generator is what you need for high-resolution video synthesis. In *International Conference on Learning Representations (ICLR)*, 2021. 2
- [39] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 2
- [40] Ruben Villegas, Jimei Yang, Seunghoon Hong, Xunyu Lin, and Honglak Lee. Decomposing motion and content for natural video sequence prediction. In *International Conference on Learning Representations (ICLR)*, 2017. 2
- [41] Nicholas Watters, Loic Matthey, Christopher P Burgess, and Alexander Lerchner. Spatial broadcast decoder: A simple architecture for learning disentangled representations in vaes. *arXiv preprint arXiv:1901.07017*, 2019. 2
- [42] Nicholas Watters, Daniel Zoran, Theophane Weber, Peter Battaglia, Razvan Pascanu, and Andrea Tacchetti. Visual interaction networks: Learning a physics simulator from video. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 3
- [43] Luca Weihs, Amanda Rose Yuile, Renée Baillargeon, Cynthia L Fisher, Gary Marcus, Roozbeh Mottaghi, and Aniruddha Kembhavi. Benchmarking progress to infant-level physical reasoning in ai. *Manuscript under review*, 2022. 2, 7
- [44] Dirk Weissenborn, Oscar Täckström, and Jakob Uszkoreit. Scaling autoregressive video models. In *International Conference on Learning Representations (ICLR)*, 2020. 2
- [45] Jiajun Wu, Erika Lu, Pushmeet Kohli, Bill Freeman, and Josh Tenenbaum. Learning to see physics via visual de-animation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 2, 3

A. Dataset

A.1. Test data

For the VoE task, we divided the four scenarios into 11 groups, each with two comparison cases. The setups in the testing data are very similar to the ones in the training data except for the behavior of the wall. All scenarios except Permanence contain predictive, hypothetical, and explicative settings. The predictive and explicative settings contain both plausible and implausible events, while the hypothetical setting contains two plausible events. In the predictive setting, the wall is moved away at the beginning and end of the video, so all information is shown at the beginning and end of the video. In the hypothetical setting, the wall always stays in the middle of the scene. In the explicative setting, the wall is moved away only at the end of the video, so new information is shown to the model at the end of the video.

Collision The Collision scenario is shown in Fig. A1. Collision contains predictive, hypothetical, and explicative settings. In the predictive setting, the wall is moved away at the beginning and end of the video, so two balls are visible to the model. We can easily tell from intuitive physics that the case in the first row is possible while the case in the second row is not, because the red ball cannot pass through the blue ball without collision. In the hypothetical setting, the wall always stays in the middle of the scene, so we can not tell how many balls there are in the scene. As we can not infer if a blue ball is hidden behind the wall at the beginning of the video, both cases in the setting are possible. In the explicative setting, the wall is moved away at the end of the video, so additional information is given. We can infer that a blue ball must be hidden behind the wall, so the case in the first row is possible, while the case in the second row is not.

Blocking The Blocking scenario is shown in Fig. A2. The Blocking scenarios are similar to the Collision scenarios, except that the ball hidden behind the wall is replaced by a fixed cube. In the predictive setting, the wall is moved away at the beginning and end of the video, so the cube is visible to the model. Similar to Collision, we can easily tell that the case in the first row is possible while the case in the second row is not, because the blue ball can not pass through the green cube without collision. In the hypothetical setting, the wall always stays in the middle of the scene, so we can not tell if there is a cube behind the wall. Therefore, both cases in the setting are possible. In the explicative setting, the wall is moved away at the end of the video, so we can infer that a cube must be hidden behind the wall. Furthermore, we can tell that the case in the first row is possible while the case in the second row is not.

Permanence The Permanence scenario is shown in Fig. A3. In the Permanence scenarios, three cubes are randomly divided into two groups (allowing empty groups), where cubes in the first group are dropped to the ground and

the second rest on the floor. We do not have an explicative setting for this scenario, as there is no new evidence at the end of the video. In the predictive setting, the wall is moved away at the beginning of the video, so we can infer that there is no object on the ground at the beginning. So the case in the second row is impossible, while the case in the first row is possible. In the hypothetical setting, the wall stays in the middle of the scene at the beginning, so we can not tell if there are cubes on the ground at the beginning, so both cases are possible.

Continuity The Continuity scenario is shown in Fig. A4. In the Continuity scenarios, we create a window on the lower half of the wall. In the case of the wall, the ball rolls across the scene. When the ball passes through the wall, it can be seen going from one side to the other. In the predictive setting, the wall is moved away at the beginning of the video, so we can infer that only one ball is in the scene. We can tell that the case in the second row is impossible while the case in the first row is possible. In the hypothetical setting, the wall always stays in the middle of the scene, and we can easily infer that the case in the first row is possible. Considering the case in the second row, we can not tell if there are two balls with the same appearance in the scene, one of which is visible at the beginning and the other one is hidden by the right part of the wall. If that is true, the case in the second row is also possible. So both cases are possible. In the explicative setting, the wall is moved away at the end of the video, so we can infer that there is only one ball in the scene. Thus we can tell that the case in the first row is possible while the case in the second row is not.

A.2. Train data

For four scenarios, we created 5 groups for training. Each of Permanence and Continuity contains 1 group, while Collision and Blocking in total contain 3 groups. Each group contains 2 kinds of cases: cases with a wall and ones without a wall. In the case with a wall, a movable wall stands in the middle of the scene and will be moved away at the beginning and the end of the video. In the case without the wall, everything stays the same except that the wall does not exist, showing that the wall won't interact with other objects physically. Each row in the Fig. 4 corresponds to one sampled video in a specific case. See Fig. 4 for all training groups.

Control group In the control group, a ball rolls across the scene without interacting with other objects, indicating that the environment follows basic physics.

Collision group A ball rolls across the scene in the Collision scenario with the wall. Another ball with the same mass but a different color is hidden behind the wall and will collide with the incoming ball, causing the first ball to stop and itself to pass through. In a setting without a wall, the second ball will always be visible.

Blocking group The Blocking scenarios are similar to the Collision scenario, except that the ball hidden behind the wall is replaced by a fixed cube. A ball rolls across the scene in the blocking setting with the wall. A fixed cube is hidden behind the wall and will collide with the incoming ball, causing the incoming ball to turn around. In the setting without a wall, everything stays the same except that the wall doesn't exist, and the cube will always be visible.

Permanence group In the Permanence scenario, three cubes are randomly divided into two groups (allowing empty groups), where cubes in the first group are dropped to the ground and the second rest on the floor. In the setting with the wall, the wall will be moved away at the end of the video, showing that all of the cubes still exist. In the setting without the wall, the cubes will always be visible.

Continuity group In the Continuity scenario, we create a window on the lower half of the wall. In the setting with the wall, the ball rolls across the scene. When the ball passes through the wall, it can be seen going from one side to the other, especially visible from the window. In the setting without the wall, the ball will always be visible.

A.3. Environment

Our X-VoE dataset comprises 22K+100K procedurally generated scenes using Unreal Engine 4. In addition to the floors and the backgrounds, there are four different object types: balls, cubes, walls, and windowed walls. In all videos, the size of the ball and the cube are the same, while the size of the wall with or without windows are randomly different. The positions of objects are randomly set in the videos, except for the walls in the permanent scenes in which the wall is placed in the middle. All objects, including the floor and the background, are randomly set in different colors.

B. Model

B.1. Perception

The perception module in XPL is similar to that of Component Variational Autoencoder (ComponentVAE) in the PLATO model [30]. For each object k in an image, we take as input a 128×128 RGBD (0-255 for each channel) image x_k that is masked except around the object. Then we use a Vision Transformer [14] encoder Φ to encode the image with only one object into a 32-dimensional Gaussian posterior distribution $q_\Phi(z_k|x_k)$. The sample from this distribution, z_k , is decoded by a spatial broadcast decoder [41] to an RGBD image. To address occlusion, we use the depth of the decoder image to combine all objects in the image by multiplying them with softmaxed depth values. We first pre-trained the perception module by optimizing the variational objective defined in [7]. We set σ to 0.05, β to 0.5, and γ to 0 to ensure that the model reconstructs object masks without segmentation information in the loss function.

ViT encoder We first reshape the $128 \times 128 \times 4$ images into a sequence of flattened $16 \times 16 \times 256$ patches, followed by a linear layer with 256 dimensions. Next, we add 2D position embeddings and learnable embeddings, flatten, and send them to a Transformer. We use 8 multi-head, 32 key dimensions, 1024 MLP layer dimensions, and 6 Transformer layers for the Transformer model [39]. Finally, we use an MLP layer with size [512, 64] and a leaky-ReLU activation function to the Transformer output and obtain 32-dimensional Gaussian posterior distributions for each object.

Spatial broadcast decoder Our spatial broadcast decoder is similar to that in [26]. As shown in Tab. A1, we use position embeddings and CNN model to decode the object embeddings, where the parameter θ in the softmax layer is learnable, thus representing the mask in terms of depth.

B.2. Reasoning

In the reasoning module, we use two Transformer modules to reason the hidden object which is occluded in some or all of the frames. All objects in a video can be reshaped as $F \times N \times D$ embeddings, where F is 15 frames, N is 8 objects, and D is 32 dimensions in our work. As shown in Tab. A2, we use a Transformer model to reason the masked objects in video, similar to the self-supervised learning module in Aloe [12]; the parameter [M] in the Mask (1) part is learnable.

First Transformer We set the mask to 0 for objects that are temporally occluded in some frames, and 1 for others. As shown in Tab. A2, we can use the Transformer model to reason the new object embeddings whose mask equals 0. We use it in both the training and testing steps to have better object embedding for the whole video.

Second Transformer In our test dataset, there may be cases where an object is obscured in all frames. So in the training step, we set the mask to 0 for one random object (including empty object) in all frames. Then we can train the second Transformer model in a self-supervised manner. In the test step, we set the mask to 0 for one object that is not visible in all frames. Then we can reason about the occluded object to explain the whole video.

B.3. Dynamics

In fact, the occluded objects are never directly seen for the Transformer model. After the first reasoning module, we obtain reasonable video object embeddings based on experience. In the dynamics module, we predict the value of the incremental change of the object embeddings in the time step by using the same dynamics module from PLATO [30] with the only difference in object dimension used (from 16 to 32). We refer the readers to [30] for architectural details.

Table A1: Spatial broadcast decoder architecture (from top to down).

Type	Size	Activation	Comment
Spatial Broadcast	8×8	-	-
Position Embedding	-	-	-
Conv 5×5	64	ReLU	stride: 2
Conv 5×5	64	ReLU	stride: 2
Conv 5×5	64	ReLU	stride: 2
Conv 5×5	64	ReLU	stride: 2
Conv 5×5	64	ReLU	stride: 1
Conv 3×3	4	-	stride: 1
Channels	RGBD(4)	Softmax (on depth channel)	$\text{softmax}(\text{depth} \times \text{abs}(\theta) \times -1000.0)$

Table A2: The Transformer architecture (from top to down). The [M] is a learnable mask token for Transformer.

Type	Size	Activation	Comment
LP (1)	256	-	-
Mask (1)	-	$\times \text{mask} + [\text{M}] \times (1-\text{mask})$	mask : (size $F \times N \times 1$), (value 0 or 1)
Position Embedding	-	-	-
Transformer	256, 256 (MLP)	ReLU (MLP)	head=8,key=32, layers=6
LP (2)	256	-	-
Mask (2)	-	$\times (1-\text{mask}) + \text{inputs} \times \text{mask}$	mask : (size $F \times N \times 1$), (value 0 or 1)

Table A3: Training parameters. The pre-processed video features are calculated by the Perception module, which is pre-trained.

Model	batch size	training step	optimizer	learning rate	warm step	delay step
Perception module (in XPL, PLATO)	300 (images)	472000	Adam	0.0004	2000	100000
XPL	500 (pre-processed video features)	32000	Adam	0.0004	1000	10000
PLATO	500 (pre-processed video features)	32000	Adam	0.0004	1000	10000
PhyDNet	100 (videos)	70000	Adam	0.001	-	-

C. Training

C.1. Training detail

In a scene with occlusion, we cannot get the representation of the occluded object directly by observation. Therefore, we first use the dynamics loss on the object embeddings after the first Transformer to train our first Transformer and dynamics model. Then, we use the object embeddings after the first Transformer to train our second Transformer model. We randomly mask an object throughout the video frame and use the model to predict representations of the objects throughout the video, enabling the model to infer whether there is a fully hidden object in the test task.

C.2. Training parameters

We first pre-train the perception module and use it for both PLATO and XPL. Then we train our model XPL, PLATO, and PhyDNet with the parameters shown in Tab. A3.

C.3. Training steps

During the development of the model, we explored how the size of the training dataset impacted the pixel loss of the dynamics module. We use the expected video in the predictive setting of all scenarios as the test dataset to calculate

the average pixel loss. Fig. A5 shows that more training data will improve the performance of the dynamics module.

D. Visualize supplementary

In the main text, we visualize the reasoning results by our XPL model in the Blocking scenario. Here, we visualize the reasoning results for the rest of the scenarios.

D.1. Collision

As shown in Fig. A6, in the predictive setting, XPL has no problem accurately reconstructing the objects, and the surprise video can be found directly. In the hypothetical setting, the possible explanation for the first video is that another ball collides with the incoming ball. In contrast, no such ball is in the second video, explaining both cases. This result also shows the limitation of our XPL as the incoming ball did not stop behind the wall. In the explicative setting, the occluder is only moved away at the end of the videos. Unlike the hypothetical, when showing a hidden ball behind it, it is impossible for the ball to pass through, causing surprise.

D.2. Permanence

As shown in Fig. A7, in the predictive setting, XPL can reconstruct the objects behind the wall, and the surprise

video can be found by comparing it with the origin image. The visual effect of the reconstructed objects does not seem to be very well, which is still a limitation of our XPL. In the hypothetical setting, the possible explanation for the second video is that there exists another object behind the wall, and our XPL can reason about the object.

D.3. Continuity

As shown in Fig. A8, the visualization results of our XPL are the same in all settings. Even though the visualization results can show surprise in predictive and explicative settings by comparing with the origin videos, our XPL still can not deal with the hypothetical setting due to the limitation discussed in the main text. Our XPL requires given masks and identification of objects. Therefore, it can not reason about the hypothetical setting in continuity by changing the identification of objects and suggesting that there are two same objects as infants do [1].

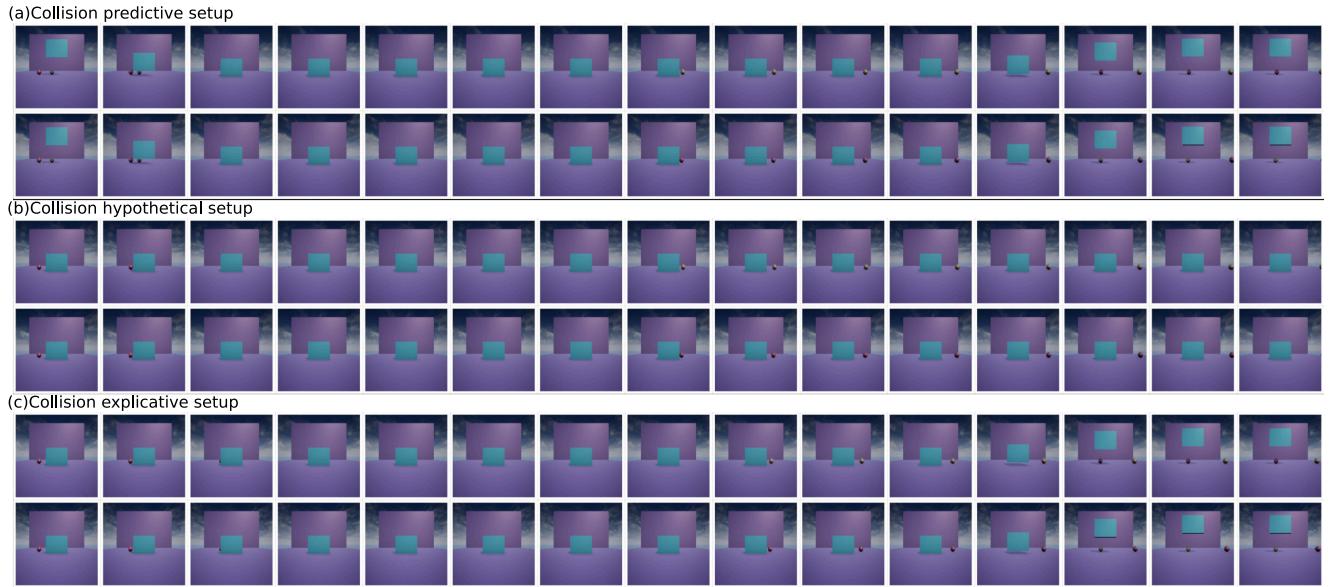
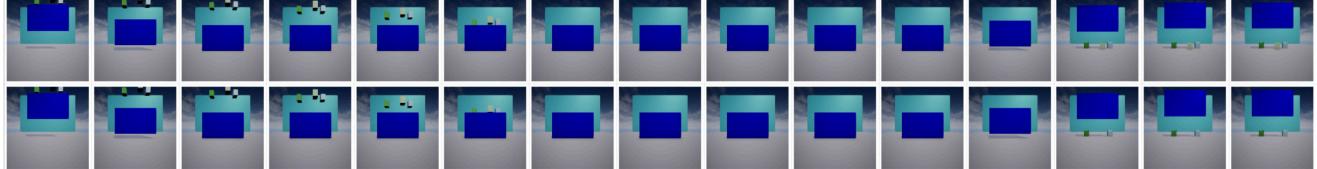


Figure A1: Collision test groups.



Figure A2: Blocking test groups.

(a) permanence predictive setup



(b) permanence hypothetical setup

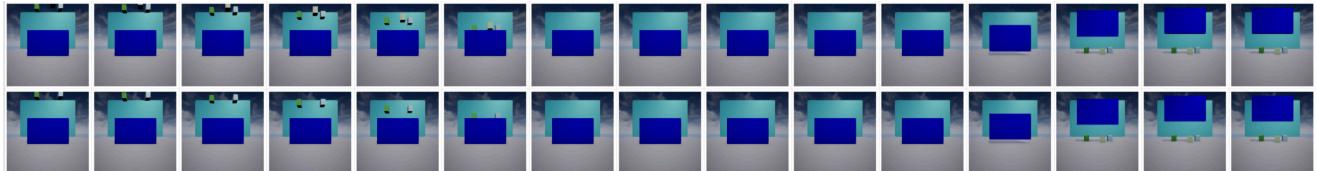


Figure A3: Permanence test groups.

(a) continuity predictive setup



(b) continuity hypothetical setup



(c) continuity explicative setup

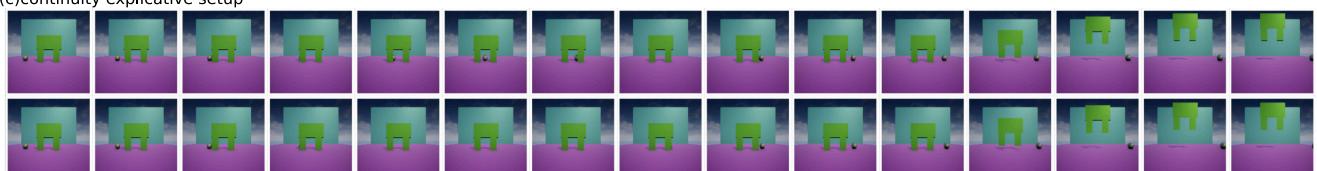


Figure A4: Continuity test groups.

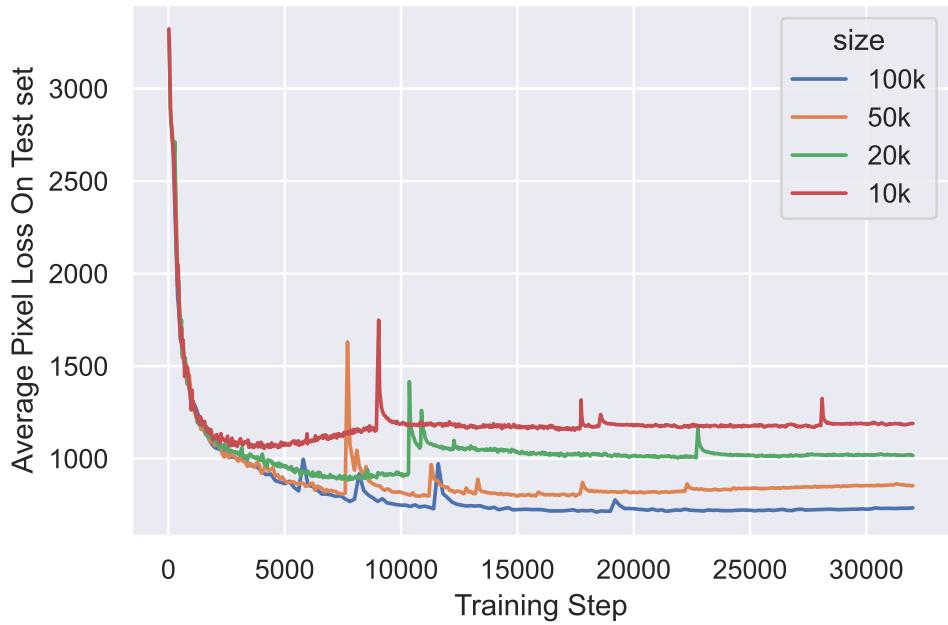


Figure A5: Average pixel loss of test data for different sizes of training data.

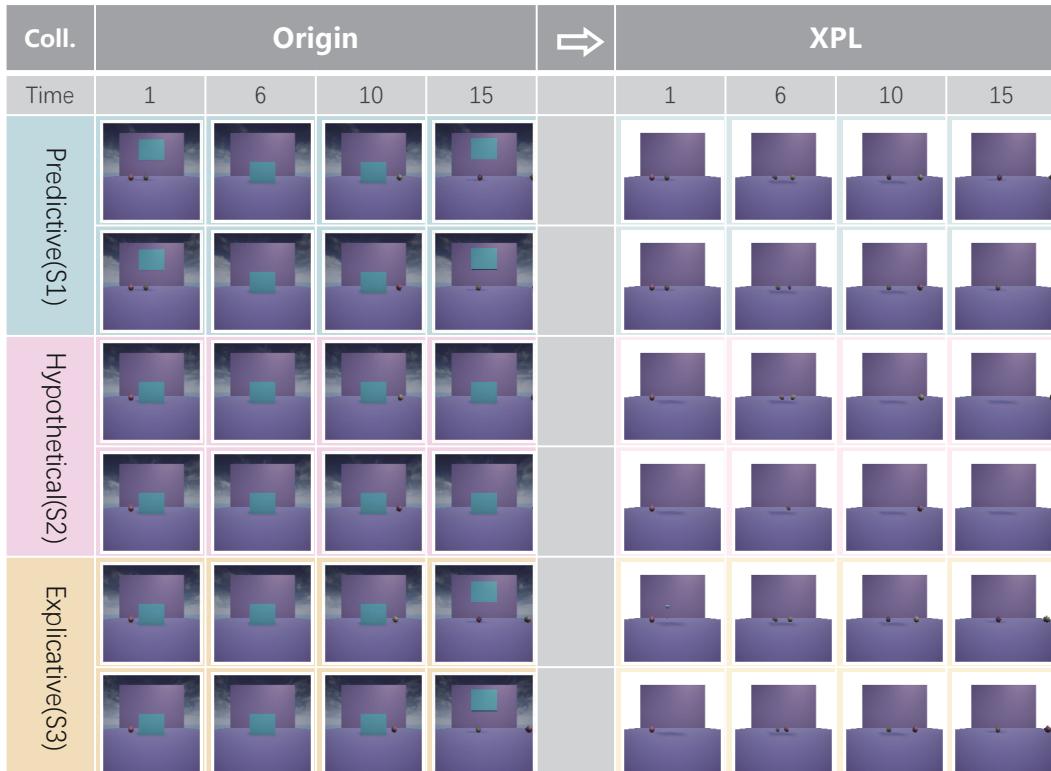


Figure A6: Visualization of the inferred internal representation in XPL during testing in collision scenarios.

Perm.	Origin				↑	XPL			
	1	6	10	15		1	6	10	15
Time	1	6	10	15		1	6	10	15
Predictive(S1)									
Hypothetical(S2)									

Figure A7: Visualization of the inferred internal representation in XPL during testing in permanence scenarios.

Cont.	Origin				↑	XPL			
	1	6	10	15		1	6	10	15
Time	1	6	10	15		1	6	10	15
Predictive(S1)									
Hypothetical(S2)									
Explicative(S3)									

Figure A8: Visualization of the inferred internal representation in XPL during testing in continuity scenarios.