PKU-IAI Technical Report: PKU-IAI-TR-PKU-IAI-2023-0004

A Probabilistic Explanation for VoE-based Evaluation

Haowei Lin
Yuanpei College
Peking University
linhaowei@pku.edu.cn

Hang Ye
Yuanpei College
Peking University
yehang@pku.edu.cn

Hanyu Liu
Yuanpei College
Peking University
1900017711@pku.edu.cn

Abstract

Visual grounded *Violation of expectations* (VoE) paradigm is widely used to evaluate the physics learning capability of both humans and machines. It does this by measuring the prediction error, or *surprise*, of a physics learning model in a given scene. Despite intuitive formulation and perfect alignment with developmental psychology, the design of evaluation protocol based on *surprise* score is empirical. We point out the potential risks behind the traditional *surprise* score design and provide a probabilistic explanation of VoE paradigm based on *likelihood ratio theory*. Guided by the theoretical framework, we propose two novel and extensible surprise scores that are theoretically sounded. Furthermore, we implement a simple yet novel baseline based on PredRNN [29] that demonstrates the ability to perform physical reasoning through direct *pixel-level prediction*. Our model outperforms a strong *object-level prediction* baseline PLATO [20], achieving an overall accuracy of 90.0% on the `Probe` dataset, compared to 73.4% for PLATO. Additionally, we conduct experiments using our newly proposed metric.

1 Introduction

Research in cognitive science has provided extensive evidence of human cognitive ability in performing physical reasoning of objects from noisy perceptual inputs. Such cognitive ability is commonly known as intuitive physics [8]. As developed early in human life [11], intuitive physics is considered as the basic building block of human intelligence. Previous psychological works proposed intuitive physics as a *startup software* [9] and a sort of *core knowledge* [26]), demonstrating the importance of developing agents that display a basic understanding of the behavior of objects and forces. Research on intuitive physics is usually divided into two categories: one approach is to directly leverage physics knowledge (e.g. fully-fledged physics engine) into the agent architecture [12], and another approach is to learn physics from raw sensor data of the world [5]. The latter case raises a key challenge: how to evaluate the acquired knowledge during learning.

Recently, a series of work inspired by developmental psychology introduced complementary methods for probing physics knowledge in artificial systems. The evaluation methods are targeted to recognise specific physical concepts presented by developmental psychology over the past fifty years, including *solidity* (solid objects cannot interpenetrate), *continuity* (moving objects will follow smooth trajectories unless perturbed), *unchangeableness* (size, shape, pattern, color do not change spontaneously), to name a few. In addition to isolating these basic principles, developmental psychology has also invented and refined what is by now a widely accepted and replicated experimental technique for examining their acquisition, referred to as *violation of expectations* (VoE) paradigm [1]. In this paradigm, infants are presented with real or virtual animated 3D scenes, some of which violate basic

physical principles. When an infant stares longer at a physically impossible display with dilated pupils in comparison with a normal one, we can assume that the subject is surprised by the unusual sight. This experimental finding can provide clues for the infant's ability to perform physical reasoning with respect to the relevant principle.

This paradigm can be adapted to machine learning to evaluate the agent's acquisition of physical concepts [19, 20]. On top of that, Piloto et al. [20] has demonstrated that their designed deep learning system can learn a diverse set of physical concepts with a relatively small amount of visual experience (about 10 to 100 hours videos as learning material).

For clarity and simplicity, we only focus on visual grounded VoE in the following discussion. Note that in the aforementioned cognitive experiments, we investigate the ability of physical reasoning by examining if the infant shows surprise on the physically impossible display. To better evaluate the AI system under the VoE paradigm, we can quantify the extent of *surprise* correlated with a given input. In practice, the *surprise* score is computed as the model's prediction errors over the course of a test video. Let $\boldsymbol{x}_t$ denote the t-th frame of the test video, and the model generates prediction of the t-th frame $\hat{\boldsymbol{x}}_t$ based on the previous sequence $\boldsymbol{x}_{1:t-1}$, the *surprise* is computed as:[1]

$$\textit{surprise}(\boldsymbol{x}) = \sum_{t=2}^{T} \textit{surprise}(\boldsymbol{x}_t) = \sum_{t=2}^{T} ||\hat{\boldsymbol{x}}_t - \boldsymbol{x}_t||^2 \tag{1}$$

Here we measure the sum-squared errors of the pixel-level prediction. If the surprise score exceeds some pre-defined threshold γ (dependent on the dataset and model), the model decides the video sequence $\boldsymbol{x}_{1:T}$ as a VoE example. The formalization is simple and intuitive: when there is a huge discrepancy between prediction and perception, the model should be "surprised" just like the infant.

Despite intuitive formulation and perfect alignment with developmental psychology, the mathematical interpretation of current evaluation method using VoE paradigm receives little attention. The design of the surprise score is empirical, and a question is naturally raised: "Is it really true that VoE videos have higher surprise scores than non-VoE videos?" Actually, Piloto et al. [20] has proposed a counterexample: the imbalanced weight of multi-scale objects in pixel space, *i.e.* the prediction errors of small objects can be dominated by large objects. and the authors proposed a dataset of "probe scenario" that contains both VoE and non-VoE videos in the same scene (collect matched test videos that are tend to have comparable surprise score) to address the issue. This counterexample shows that empirical methods often have potential risks. Are there any other non-trivial counterexamples? To answer this question, we need a systematic theoretical examination of VoE paradigm.

In this paper, we carefully examine the VoE paradigm from a probabilistic perspective and provide discussions on previous works under our theoretical framework in Sec. 3. Additionally, we present two streams of novel surprise score designs under principled guidance derived from theoretical justification in Sec. 4. Furthermore, we implement a baseline model that learns intuitive physics from visual data and present the experimental results of our newly designed $\textit{surprise}_{KNN}$ in comparison with traditional pixel-error *surprise* in Sec. 5. Our contributions are as follows:

- We present a probabilistic explanation of VoE paradigm, which to the best of our knowledge is new to this field. This theoretical framework can provide insights on how to better assess the knowledge of physics learning from a novel perspective.
- We propose two novel surprise scorers based on theoretical derivation from likelihood ratio theory for evaluation under VoE paradigm: $\textit{surprise}_{naive}$ and $\textit{surprise}_{KNN}$. These two scorers have theoretical soundness and can be extended to various forms using different density estimation approaches.
- We implement a simple yet novel baseline based on PredRNN [29] to learn the physics concept. We explore the possibility of learning without explicit object modeling and discover that direct *pixel-level prediction* can equip model with physical reasoning capability. While previous models require hard pre-training to learn object features, our model is more data-efficient and more adaptive to real world settings.

[1] Here we make a simplification. The definition of surprise score is usually model-dependent, e.g. Piloto et al. [20] uses a ComponentVAE to decode predicted and input object codes and then computes the pixel error, but in essence, they basically take the form of Eq. (1).

2 Related Work

Intuitive physics modeling in machine learning In recent years, researchers have made attempts to endow artificial intelligence systems with human-level capability of physical reasoning. A straightforward approach is to incorporate a fully-fledged physics engine in the framework, which supports Bayesian inference via simulation [4, 28]. Galileo [30] further incorporated a physics engine with representation learning to infer physical properties directly. Another line of works acquire physical concepts implicitly from training data without actual engines. Specifically, sensory inputs are fed into deep-learning models to generate predictions for various reasoning tasks, including physical property inference, dynamic prediction [10], puzzle-solving [3, 12] and visual question answering [31].

VoE paradigm In developmental psychology, researchers often assess the possession of a physical concept under the VoE paradigm, where the surprise of human viewers is measured via gaze duration. A pioneering work [19] first applied this technique to artificial learning systems and computed KL-divergence between prior expectations and posterior beliefs given perceptual inputs to estimate the "surprise". Follow-up works further explored this idea by proposing more evaluation benchmarks, which consist of matched videos of physically possible versus impossible events [22, 20].

Temporal modeling in computer vision Inspired by remarkable success of language models in natural language processing, the community of computer vision has borrowed similar model architectures to process video sequences. For example, temporal visual features can be extracted via LSTM [29] and transformers [2, 14]. In addition to pixel-level prediction, previous works often resort to intermediate representations to boost the performance. In complex visual reasoning, a series of works focus on object-centric representation to characterize the video frame, such as RPIN [21], Aloe [6] and ADEPT [25].

Likelihood ratio theory In statistics, the likelihood-ratio test assesses the goodness of fit of two competing statistical models based on the ratio of their likelihoods. When both models have no unknown parameters, use of the likelihood-ratio test can be justified by the Neyman–Pearson lemma. In our theoretical analysis, we compute the likelihood ratio in binary hypothesis test to decide whether a video clip violates expectation or vice versa.

3 A Theoretical Understanding of VoE

As mentioned in Sec. 1, our work falls into the second type of strategy, focusing on learning physics from raw sensor data of the world. Our final goal is to build a machine that is capable of physical reasoning, and the first step towards the goal is to design proper evaluation protocols. Here we focus on the evaluation methods based on visual grounded VoE paradigm. Then we can formalize the relevant component of this problem as follows.

3.1 Formalization of Physics Learning Model

We let $\mathbf{\Theta}$ denote the physics learning model (typically a RNN-like generative model), which takes a video clip $\boldsymbol{x}_{1:t}$ as input, and outputs the prediction of the next frame $\hat{\boldsymbol{x}}_{t+1}$. In essence, we aim to learn a joint probability distribution which can be factorized as follows:

$$\mathbf{P}(\boldsymbol{x}_{1:T}) = \mathbf{P}(\boldsymbol{x}_1) \cdot \mathbf{P}(\boldsymbol{x}_2|\boldsymbol{x}_1) \cdot \mathbf{P}(\boldsymbol{x}_3|\boldsymbol{x}_{1:2}) \cdots \mathbf{P}(\boldsymbol{x}_T|\boldsymbol{x}_{1:T-1})$$

Here $\mathbf{\Theta}$ can be viewed as a conditional generative model $\hat{\boldsymbol{x}}_t = \mathbf{\Theta}(\boldsymbol{x}_{1:t-1}) = \arg\max_{\hat{\boldsymbol{x}}_t} q(\hat{\boldsymbol{x}}_t|\boldsymbol{x}_{1:t-1})$. Recall that we use *surprise* to measure the physical improbability of a given video and it takes the form of square errors. Now we examine the probabilistic context of surprise from the view of regression analysis. Suppose we utilize a regression model that satisfies $\boldsymbol{x}_t = \mathbf{\Theta}(\boldsymbol{x}_{1:t-1}) + \boldsymbol{\epsilon}$, where $\boldsymbol{x}_t$ is the response variable, $\boldsymbol{x}_{1:t-1}$ is the co-variate, and $\boldsymbol{\epsilon}$ is a random error with zero mean. If we introduce an extra ass suppose that $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma_0^2 \boldsymbol{I})$, then the square error *surprise* can be related to the probability density of $\boldsymbol{\epsilon}$:

$$\mathbf{P}(\boldsymbol{\epsilon}|\boldsymbol{x}_{1:t-1}) = \mathbf{P}(\hat{\boldsymbol{x}}_t - \boldsymbol{x}_t|\boldsymbol{x}_{1:t-1}) = \frac{1}{\sqrt{(2\pi\sigma_0)^n}} e^{-\frac{1}{2\sigma_0{}^{2n}}||\hat{\boldsymbol{x}}_t - \boldsymbol{x}_t||^2} \propto -||\hat{\boldsymbol{x}}_t - \boldsymbol{x}_t||^2$$

In test time (the model $\boldsymbol{\Theta}$ is fixed), we can omit $\hat{\boldsymbol{x}}_t$ since it's deterministic given the input $\boldsymbol{x}_{1:t-1}$ (i.e. $\hat{\boldsymbol{x}}_t = \boldsymbol{\Theta}(\mathbf{x_{1:t-1}})$), and the last term is exactly the opposite number of *surprise*$(\boldsymbol{x}_t)$. Thus *surprise* measures the negative log likelihood of $\boldsymbol{x}_t$, i.e. $q(\boldsymbol{x}_t|\boldsymbol{x}_{1:t-1})$. So using *surprise*$(\boldsymbol{x}_t)$ to decide whether the model should be surprised is to decide whether the current perception has a small probability density.

3.2 Formalization of VoE

Note that we still have not formalized the problem of VoE paradigm yet. In our task, the model is asked to make a binary decision whether $\boldsymbol{x}_t|\boldsymbol{x}_{1:t-1}$ is surprising or not. It can be interprted from two probabilistic perspectives. From the perspective of frequentist, the binary decision is basically a binary hypothesis test:

$$\mathcal{H}_0 : \boldsymbol{x} \sim \mathcal{P}_{\textit{VoE}} \quad v.s. \quad \mathcal{H}_1 : \boldsymbol{x} \sim \mathcal{P}_{\textit{non-VoE}} \tag{2}$$

Here the null hypothesis $\mathcal{H}_0$ denotes that $\boldsymbol{x}$ is drawn from some VoE distribution $\mathcal{P}_{\textit{VoE}}$ while the alternative hypothesis $\mathcal{H}_1$ denotes $\boldsymbol{x}$ is drawn from some non-VoE distribution $\mathcal{P}_{\textit{non-VoE}}$. Since our training data are obviously non-VoE, our model actually models the probability distribution $\mathcal{P}_{\textit{non-VoE}}$ (i.e. $\mathbf{P}(\cdot)$ mentioned in Sec. 3.1 relates to the non-VoE distribution). Suppose we have $\mathcal{R}$ as the reject region of Eq. (2), then the Type I error β_1 and Type II error β_2 are defined as:

$$\beta_1(\mathcal{R}) = \int_{\mathcal{R}} \mathbf{P}_{\textit{VoE}}(\boldsymbol{x}) d\boldsymbol{x}$$
$$\beta_2(\mathcal{R}) = 1 - \int_{\mathcal{R}} \mathbf{P}_{\textit{non-VoE}}(\boldsymbol{x}) d\boldsymbol{x},$$

Likelihood ratio is a principled way to decide VoE. In a hypothesis test, our ultimate goal is to find a *uniformly most powerful* (UMP) test (a test function $\varphi(\boldsymbol{x})$ or reject region $\mathcal{R}$), i.e. minimizing β_2 with β_1 kept under a certain level α. Following the standard protocols in statistics, we first define the likelihood ratio $\textit{LR}(\boldsymbol{x}) := \mathcal{P}_{\textit{VoE}}(\boldsymbol{x})/\mathcal{P}_{\textit{non-VoE}}(\boldsymbol{x})$. With the Neyman-Pearson lemma [16], a simple theorem can be derived to show the principled role of likelihood ratio in VoE paradigm:

Theorem 1. *A binary test with test function $\varphi(\boldsymbol{x}) = \mathbf{1}_{\boldsymbol{x}\in\mathcal{R}}(\boldsymbol{x})$ and rejection region $\mathcal{R}$ defined as follows is a unique UMP test for Eq.* (2)*.*

$$\mathcal{R} := \{\boldsymbol{x} : \textit{LR}(\boldsymbol{x}) < \lambda_0\}$$

Here $\mathbf{1}_{\boldsymbol{x}\in\mathcal{R}}(\cdot)$ is an indicator function and λ_0 is a threshold that can be chosen to obtain a specified significance level.

Proof. The proof is presented in Appendix A. □

In sharp contrast, from a Bayesian perspective, to decide a given $\boldsymbol{x}$ is VoE is to learn a binary classifier $\mathbf{P}(\textit{VoE}|\boldsymbol{x}) = 1 - \mathbf{P}(\textit{non-VoE}|\boldsymbol{x})$. We introduce an auxiliary variable $C \sim \mathcal{B}(p)$, where $\mathcal{B}(p)$ denotes a Bernoulli distribution with parameter $p \in [0, 1]$. And let $\boldsymbol{x} \sim \mathcal{P}_{\textit{VoE}}$ if $C = 1$ and $\boldsymbol{x} \sim \mathcal{P}_{\textit{non-VoE}}$ if $C = 0$. Given an input $\boldsymbol{x}$,

$$\begin{aligned}
\mathbf{P}(\textit{VoE}|\boldsymbol{x}) &= \mathbf{P}(C = 1|\boldsymbol{x}) \\
&= \frac{\mathbf{P}(\boldsymbol{x}|C = 1) \cdot \mathbf{P}(C = 1)}{\mathbf{P}(\boldsymbol{x}|C = 1) \cdot \mathbf{P}(C = 1) + \mathbf{P}(\boldsymbol{x}|C = 0) \cdot \mathbf{P}(C = 0)} \\
&= \frac{p}{p + (1 - p)\textit{LR}(\boldsymbol{x})^{-1}}
\end{aligned}$$

The second equation is simply the Bayes' theorem. The A similar result still holds: likelihood ratio is a principled way to decide VoE under the Bayesian perspective. When $\textit{LR}(\boldsymbol{x})$ is small, $\boldsymbol{P}(\textit{VoE}|\boldsymbol{x})$ is small, thus the model considers $\boldsymbol{x}$ as a non-VoE scene.

Above all, we introduce the probabilistic formalization of VoE. Basically we introduce a concept of $\mathcal{P}_{\textit{VoE}}$ and $\mathcal{P}_{\textit{non-VoE}}$. Under this framework, we have shown that the likelihood ratio $\textit{LR}(\boldsymbol{x})$ is an optimal choice from both frequentist and Bayesian perspectives. Then we will examine the validity of using $\textit{surprise}(\boldsymbol{x})$ to decide VoE under this framework.

3.3 Problems with *surprise*

As revelaed by Sec. 3.1 and Sec. 3.2, $\textit{surprise}(\boldsymbol{x})$ measures the non-VoE probability density of $\boldsymbol{x}$. More precisely, the relationship can be expressed as a Boltzmann distribution:

$$\boldsymbol{P}_{\textit{non-VoE}}(\boldsymbol{x}) = \frac{1}{Z} e^{-\textit{surprise}(\boldsymbol{x})/T},$$

where $Z > 0$ is a normalizing constant and $T > 0$ is a temperature scaling term. However, we have pointed out that the principled way to decide VoE is through likelihood ratio $\textit{LR}(\boldsymbol{x}) := \mathcal{P}_{\textit{VoE}}(\boldsymbol{x})/\mathcal{P}_{\textit{non-VoE}}(\boldsymbol{x})$. The use of $\textit{surprise}(\boldsymbol{x})$ may suffer from some problems.

Here we give a intuitive counterexample of using $\textit{surprise}(\boldsymbol{x})$ to decide VoE. Suppose $\mathcal{P}_{\textit{VoE}} = \mathcal{N}(0, 0.001)$ and $\mathcal{P}_{\textit{non-VoE}} = \mathcal{N}(0, 1)$, then the surprise score for $x = 1$ and $x = 0$ are calculated as $\textit{surprise}(0) = -T \cdot \log(\frac{1}{\sqrt{2\pi}} \cdot Z) < \textit{surprise}(1) = -T \cdot \log(\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}} \cdot Z)$. Thus 0 is less surprising than 1 in this case. However, considering the likelihood test, it is obvious that $\textit{LR}(0) > \textit{LR}(1)$ since density of $\mathcal{P}_{\textit{VoE}}$ is highly concentrated at 0.

The failure of *surprise* here is not a paradox but a expected result. Some may think that if we draw our training dataset from $\mathcal{N}(0, 1)$, most samples will look more like 0 instead of 1, but why should we surprise at 0 rather than at 1? This intuitive idea in fact makes two hypothesises:

- $\mathcal{P}_{\textit{VoE}}$ is a uniform distribution.
- the supporting set of $\boldsymbol{x}$ lies in a low dimensional space.

For the first hypothesis, we usually tend to model $\mathcal{P}_{\textit{VoE}}$ as a uniform distribution since we have no training data, thus it remains the same as prior distribution (suppose we hold a Bayesian learning view). But in real life the forming of $\mathcal{P}_{\textit{VoE}}$ may be complex: our cognitive ideas towards VoE scenes may be shaped by many factors such as evolution, gene, experiences etc. We should be cautious to make the assumption that VoE scenes are uniformly distributed.

For the second hypothesis, it is basically a confusion between *lack of samples* and *low probability density*. Considering high dimensional data, which is a practical representation of visual scene features, a sample with higher probability density doesn't mean that it's more likely to appear in sampling procedure. For example, we have *Gaussian Annulus Theorem*, which states that nearly all the probability of a spherical Gaussian with unit variance is concentrated in a thin annulus of width $O(1)$ at radius $\sqrt{d}$:

Theorem 2 (Gaussian Annulus Theorem)**.** *For a d-dimensional spherical Gaussian with unit variance in each direction, for any $\beta \leq \sqrt{d}$, all but at most $3e^{-c\beta^2}$ of the probability mass lies within the annulus $\sqrt{d} - \beta \leq |\boldsymbol{x}| \leq \sqrt{d} + \beta$, where c is a fixed positive constant.*

Proof. The proof is presented in this notes.[2] □

Therefore, in spite of the highest density at the origin, the $\mathbf{0} \in \mathbb{R}^d$ is very unlikely to appear when sampling from high dimensional standard Gaussian. This falsehood also demonstrates that our intuition usually fails on high dimensional space, which reminds us of the importance to find theoretical principle in research problems rather than purely relying on empirical observations.

4 Methodology

In this section, we will introduce several methods that follows the principled like-likelihood ratio theory as a refinement to traditional surprise score design.

[2]Proseminar Theoretical Computer Science notes by Wolfgang Mulzer, see `https://www.inf.fu-berlin.de/lehre/WS17/SemAlg/notes/02_highdim2.pdf`.

4.1 A Naive Scorer

Firstly, we will present a naive scorer $\textit{surprise}_{\textit{naive}}(\boldsymbol{x})$ based on likelihood ratio. Since we train a model $\boldsymbol{\Theta}$ to model the non-VoE probability $\mathcal{P}_{\textit{non-VoE}}$, to obtain the full estimation of $\textit{LR}(\boldsymbol{x})$, we can still train a generative model $\tilde{\boldsymbol{\Theta}}$ to model the VoE probability $\mathcal{P}_{\textit{VoE}}$. The training procedure is just the same as $\boldsymbol{\Theta}$: First we need to collect a set of VoE videos $\{\boldsymbol{x}^{(i)}\}_{i=1}^{N}$ as our training dataset, and using training strategies like *next frame prediction* or object-based prediction methods as in [20]. Then for a given test video, we compute the surprise score $\textit{surprise}_{\tilde{\boldsymbol{\Theta}}}(\boldsymbol{x})$, too. The final score is computed using likelihood ratio:

$$\textit{surprise}_{\textit{naive}}(\boldsymbol{x}) \propto C_1 \cdot \exp\{-(\gamma \cdot \textit{surprise}_{\tilde{\boldsymbol{\Theta}}}(\boldsymbol{x}) - \textit{surprise}_{\boldsymbol{\Theta}}(\boldsymbol{x})) \cdot C_2\},$$

where the R.H.S. is derived from the Boltzmann distribution formalization, γ is a hyper-parameter, and C_1 and C_2 are constants, thus our naive scorer can be simplified as:

$$\textit{surprise}_{\textit{naive}}(\boldsymbol{x}) = \textit{surprise}_{\boldsymbol{\Theta}}(\boldsymbol{x}) - \gamma \cdot \textit{surprise}_{\tilde{\boldsymbol{\Theta}}}(\boldsymbol{x})$$

When we assume that $\textit{surprise}_{\tilde{\boldsymbol{\Theta}}}(\boldsymbol{x})$ is a constant mapping (i.e. the VoE distribution is uniform), $\textit{surprise}_{\textit{naive}}(\boldsymbol{x})$ will degrade to the vanilla $\textit{surprise}(\boldsymbol{x})$. This scorer is theoretically excellent, however, it requires collection of VoE data which is usually unavailable in the real world. Also, the collection of VoE data may be hard, expensive, and biased. Under this circumstance, we need to find an alternative way to estimate VoE distribution from only a small part of collected data. A possible way may be transfer learning and few-shot fune-tuning, but generally it's hard to analyse pre-trained models under a theoretical framework, here we only assume that we are only accessible to $\boldsymbol{\Theta}$ trained on non-VoE data as a simulation of human physics learning in the real world.

4.2 Scorer Based on Density Estimation

Note that contrary to $\boldsymbol{\Theta}$ as a physics learning agent model, $\tilde{\boldsymbol{\Theta}}$ is simply a statistical model that aims to capture the distribution of VoE scenes. We can consider using traditional estimation methods from statistics to estimate $\mathcal{P}_{\textit{VoE}}$.

Density estimation using K-nearest neighbor There are many probability estimation methods such as IForest [13], OCSVM [23], PCA [24], Mahalanobis distance [15]. Among them the simplest method is K-nearest Neighbor (KNN) [18], which will be used as an example to illustrate our surprise scorer design based on density estimation. The KNN estimation method is performed as follows. Suppose we want to estimate the underlying distribution $\mathcal{P}$ of an observation set $\mathcal{Z} = \{\boldsymbol{z}_i\}_{i=1}^{n}$ where $||\boldsymbol{z}||_2 = 1$ and $\boldsymbol{z} \in \mathbb{R}^m$, we calculate the Euclidean distance $||\boldsymbol{z}_i - \boldsymbol{z}^\star||_2$ for a given test sample $\boldsymbol{z}^\star$. Then we reorder $\mathcal{Z}$ according to the increasing distance $||\boldsymbol{z}_i - \boldsymbol{z}^\star||_2$. The reordered sequence is denoted as $\mathcal{Z}' = (\boldsymbol{z}_{(1)}, \boldsymbol{z}_{(2)}, \cdots, \boldsymbol{z}_{(n)})$, and the density estimation of $\boldsymbol{z}^\star$ is given by:

$$\hat{p}(\boldsymbol{z}^\star) \propto -r_k(\boldsymbol{z}^\star) := ||\boldsymbol{z}^\star - \boldsymbol{z}_{(k)}||_2$$

We have a convergence guarantee for this estimator as the following theorem:

Theorem 3. *Suppose we have a KNN estimator with hyper-parameters k and n, where k denotes the index of element in the reordered sequence $\mathcal{Z}'$, n denotes the cardinality of the observation set $\mathcal{Z}$, the convergence holds:*

$$\lim_{\frac{k}{n} \to 0} \hat{p}(\boldsymbol{z}^\star; k, n) = p(\boldsymbol{z}^\star)$$

Specifically,

$$\mathbb{E}[|\hat{p}(\boldsymbol{z}^\star; k, n) - p(\boldsymbol{z})|] = o(\sqrt[m-1]{\frac{k}{n}} + \sqrt{\frac{1}{k}})$$

Proof. The proof is presented in Appendix A. □

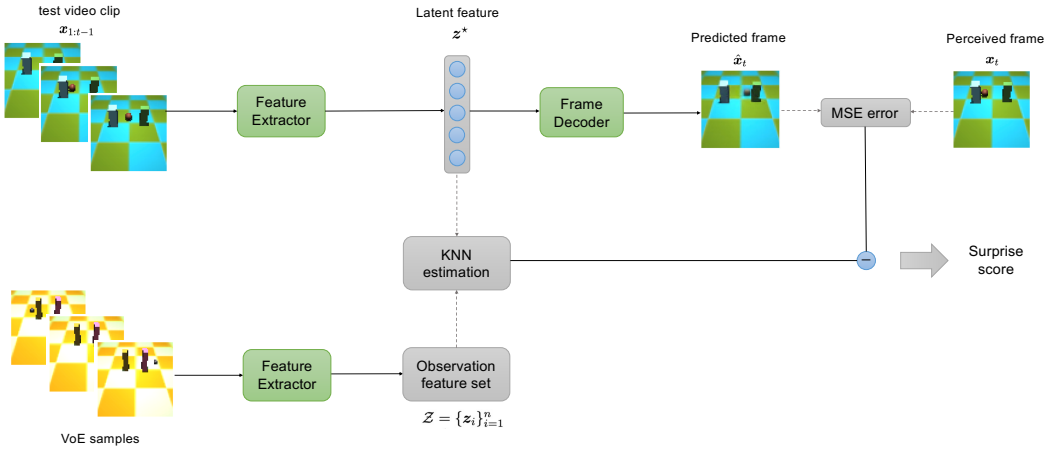

Figure 1: **Illustration of our proposed surprise scorer.** In this pipeline, the upper part is identical to the standard protocol, where we compute the naive surprise by summing the square errors between predicted frames and perceived frames. Our key design lies in the estimation of the probability density of VoE distribution. We split the test set and generate feature embeddings of some VoE samples as an observation set. For a test video clip, we additionally compute the distance of its dense representation to the k-nearest neighbor in the observation set. We subtract the scaled distance from the naive surprise to obtain the final score.

Design of $surprise_{KNN}$ Based on KNN estimation method, we can design a novel scorer $surprise_{KNN}$ as follows:

$$surprise_{KNN}(\boldsymbol{x}) = surprise_{\boldsymbol{\Theta}}(\boldsymbol{x}) - \gamma \cdot r_k(\boldsymbol{z})$$

Here $surprise_{\boldsymbol{\Theta}}(\boldsymbol{x})$ is the same as previously defined traditional surprise score that estimate $\mathcal{P}_{non\text{-}VoE}$ using model $\boldsymbol{\Theta}$, and $r_k(\boldsymbol{z})$ is the distance of $\boldsymbol{z}$ to the k-nearest neighbor. $\boldsymbol{z}$ is a dense representation vector of $\boldsymbol{x}$ in the high dimensional space, which can be obtained from the hidden states of the $\boldsymbol{\Theta}$ model (e.g. the final states of RNN, average pooling of each frame's feature). We need some VoE samples to prepare the observation set $\mathcal{Z}$ and the choice of k and n can be guided by Theorem 3. Note that KNN here is only an example, and we can substitute it with any other density estimation method. This scorer is also based on likelihood ratio, which indicates its theoretical soundness. The pseudo code for computing $surprise_{KNN}(\boldsymbol{x})$ is presented in Appendix B. To this end, the overall pipeline is illustrated in Fig. 1.

Requirement of VoE data Notice that we still need VoE data as our observation set to compute $surprise_{KNN}(\boldsymbol{x})$. Here we want to highlight the difference of data requirement between $surprise_{KNN}(\boldsymbol{x})$ and $surprise_{naive}(\boldsymbol{x})$. The data we need in $surprise_{naive}(\boldsymbol{x})$ is used to train a generative model from scratch, whose amount is generally huge (e.g. PLATO model uses 4,500,000 $3 \times 64 \times 64$ images in the training phase), but KNN estimator usually requires small amount of data to converge (e.g. Sun et al. [27] use 500 dense vectors $\boldsymbol{z} \in \mathbb{R}^{128}$ as observation set and set $k = 50$). We have to emphasize that the only two ways to decide $\mathcal{P}_{VoE}$ are (1) using collected data to estimate. (2) using proxy VoE distribution $\mathcal{P}_{proxy}$ to substitute $\mathcal{P}_{VoE}$ (e.g. uniform distribution). A small amount of VoE data will benefit our evaluation quality.

5 Experiment

In this section, we design a set of experiments to find some properties of our theory in real practice. First, we implement a simple yet novel baseline model based on PredRNN [29] to learn the physics concept, and we also explore its effectiveness under the traditional framework of VoE paradigm based evaluation. Second, we examine the $\mathcal{P}_{VoE}$ distribution using KNN estimator and compare the estimated $\mathcal{P}_{VoE}$ with uniform proxy distribution. Third, we test our model on `Probe` dataset [20] using our newly designed $surprise_{KNN}$. The relevant details are as follows.

Dataset. Following Piloto et al. [20], we conducted our experiment on the Physical Concepts dataset. The `freeform` training set consists of 300,000 scenes which encompass a wide range of complex physical interactions. And the validation set contains 5,000 video clips capturing common physical

Method	Average Accuracy(%)					
	Continuity	Directional Inertia	Object Persistence	Solidity	Unchangeableness	**Overall**
Piloto et al. [20]	89.1	72.7	67.8	71.9	65.6	73.4
Ours	**99.9**	**81.7**	**94.1**	**77.6**	**96.3**	**90.0**

Table 1: **Comparison of average accuracy for each physical concept on the probe test set.** Without intermediate object-centric representation, our PredRNN outperforms Piloto et al. [20] by a large margin.

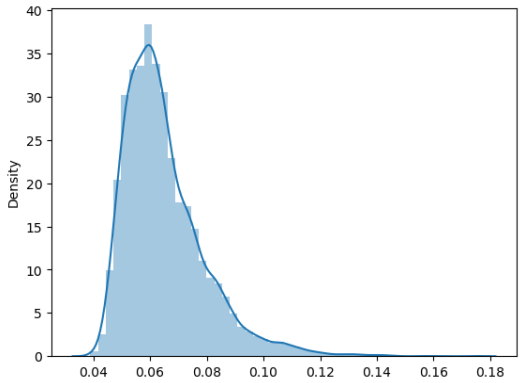

Figure 2: **Distribution plot for $r_k(\boldsymbol{z})$ on test dataset.**

events, which can be used to select the best prediction model. To leverage the VoE paradigm, five physical concepts are targeted for evaluation: continuity, directional inertia, object persistence, solidity and unchangeableness. For a single physical concept, the `probe` test set contains 5,000 tuples, each comprising two physically possible probes and two physically impossible probes. The latter were constructed by splicing together frames from the possible probes in a way that violates physics. Note that each video was restricted to 15 frames at 64×64 resolution in RGB channels.

Model architecture. The architrecture of PredRNN [29] is enlightened by the idea that a predictive learning system should memorize both spatial appearances and temporal variations in a unified memory pool. The core of this network is a new Spatio-temporal LSTM (ST-LSTM) unit that extracts and memorizes spatial and temporal representations simultaneously. In our experiment, the PredRNN model consists of four ST-LSTM layers with 128 hidden states each. The convolution filters inside ST-LSTMs are set to 3×3.

Implementation details. We train the PredRNN [29] model to generate pixel-level frame prediction from scratch for 50000 epochs, where the batch size is set to 64. We use the Adam [7] optimizer with $\beta_1, \beta_2 = 0.9, 0.95$, and adjust the learning rate to 3×10^{-4}. To implement our proposed novel metric, we further split the `probe` dataset into observation set (20%) and test set (80%). The former consists of $N = 2000$ VoE video clips, which provide useful information for estimating $\mathcal{P}_{VoE}$. Note that we estimate the probability density of $\mathcal{P}_{VoE}$ for each physical concept separately. As mentioned in Sec. 4.2, the dense latent feature $\boldsymbol{z}$ is obtained from the final hidden states of our PredRNN after global max-pooling across each channel, with a dimension of $d = 128$. And we choose $k = 50$ and $\gamma = 0.01$ for KNN-based density estimation.

Comparison with PLATO. Following Piloto et al. [20], we compute a binary accuracy score associate with each physical concept, where a probe tuple is correctly classified when the relative surprise is greater than zero. As illustrated in Tab. 1, our model attains higher average accuracy and greatly surpasses Piloto et al. [20], demonstrating that direct pixel-level supervision may suffice to endow the model with strong capability of physical reasoning.

x**Uniform assumption on $\mathcal{P}_{VoE}$**. Note that the main focus of this paper is to propose a novel metric for evaluation under VoE paradigm. The key insight is that vanilla *surprise* only uses $\mathcal{P}_{non\text{-}VoE}$ but we highlight the information provided by $\mathcal{P}_{VoE}$. The difference is that vanilla *surprise* makes an assumption of $\mathcal{P}_{VoE}$ to be a uniform distribution. In this experiment, we test this assumption using statistical methods. Specifically, we compute the distance of dense feature z to the k-nearest neighbor (i.e. $r_k(\boldsymbol{z}) \propto -\mathbf{P}_{VoE}(\boldsymbol{z})$) for each test video clip. As shown in Fig.2, the $\{r_k(\boldsymbol{z})\}$ is apparently not a

Metric	Continuity	Directional Inertia	Object Persistence	Solidity	Unchangeableness	**Overall**
Relative Surprise ($\times 10^{-4}$)	8.8	8.0	77.5	1.6	21.2	23.4
Average Accuracy (%)	99.9	81.8	94.0	77.4	96.4	89.9

Table 2: **Evaluation results on the probe test set with our newly designed surprise scorer.**

uniform distribution and the p-value is significantly larger than 0.05. We can draw the conclusion that the uniform assumption on $\mathcal{P}_{VoE}$ is incorrect.

Evaluation with *surprise*$_{KNN}$. Again, we test our model on the `Probe` test set using the proposed *surprise*$_{KNN}(\boldsymbol{x})$. The experimental results are presented in Tab. 2. The results are overall consistent with vanilla *surprise*, which indicates that the effect of the uniform assumption on $\mathcal{P}_{VoE}$ is not huge **on this dataset**. Using *surprise* to evaluate physics learning is reasonable in this case, but it doesn't mean that $\mathcal{P}_{VoE}$ can be overlooked in other datasets either.

Discussion

To prevent misunderstanding, we make a clarification on our setting (VoE-based evaluation) in this section. We want to clarify that the VoE paradigm is used to evaluate a physics learning model, but the vanilla surprise score design is empirical and ignore $\mathcal{P}_{VoE}$ (assumes it as a uniform distribution). This work theoretically justified the principled role of using the likelihood ratio in this evaluation paradigm, and propose a KNN-based surprise score. This score is derived from the likelihood ratio theory, which proves its superiority. Remenber that we are doing **evaluation**, and the ultimate goal of VoE-based evaluation is to reveal the physics learning ability of a certain learning method (or a certain model), thus a better score does not necessarily improve the VoE detection accuracy. In our experiment, the KNN surprise score achieves nearly the same results as the vanilla surprise score on the Probe dataset. This only means that it's feasible to use the vanilla surprise score on this dataset. Given the same model evaluated, higher VoE accuracy doesn't mean the score is better or "more correct".

6 Conclusion

In this work, we provide a potential probabilistic explanation of physics learning evaluation methods based on VoE paradigm. The key insight behind our theoretical framework is the likelihood ratio between two probability distribution: $\mathcal{P}_{VoE}$ and $\mathcal{P}_{non\text{-}VoE}$. Our theory shows that traditional *surprise* overlooks the existence of $\mathcal{P}_{VoE}$ and assumes it to be a uniform distribution. We propose two novel and extensible surprise scores (*surprise*$_{naive}$ and *surprise*$_{KNN}$) to correct the traditional *surprise*. Experimental results also show that the uniform assumption of $\mathcal{P}_{VoE}$ is inaccurate. The second contribution of this work is the implementation of an *pixel-level prediction* physics learning model based on PredRNN. The model outperforms a strong object-level prediction model PLATO [20] with a large margin, which indicates that using *pixel-level prediction* is promising for models to learn physics.

Author Contribution

In this course project, Haowei proposed the theoretical framework and wrote the main body of this report. Hang conducted the experiments and did the main coding of this project. Three of them surveyed the related work and reviewed literature.

References

[1] Andréa Aguiar and Renée Baillargeon. 2.5-month-old infants' reasoning about when objects should and should not be occluded. *Cognitive psychology*, 39(2):116–157, 1999. 1

[2] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6836–6846, 2021. 3

[3] Anton Bakhtin, Laurens van der Maaten, Justin Johnson, Laura Gustafson, and Ross Girshick. Phyre: A new benchmark for physical reasoning. *Advances in Neural Information Processing Systems*, 32, 2019. 3

[4] 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*, 110(45): 18327–18332, 2013. 3

[5] Misha Denil, Pulkit Agrawal, Tejas D Kulkarni, Tom Erez, Peter Battaglia, and Nando De Freitas. Learning to perform physics experiments via deep reinforcement learning. *arXiv preprint arXiv:1611.01843*, 2016. 1

[6] David Ding, Felix Hill, Adam Santoro, Malcolm Reynolds, and Matt Botvinick. Attention over learned object embeddings enables complex visual reasoning. *Advances in neural information processing systems*, 34:9112–9124, 2021. 3

[7] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 8

[8] James R Kubricht, Keith J Holyoak, and Hongjing Lu. Intuitive physics: Current research and controversies. *Trends in cognitive sciences*, 21(10):749–759, 2017. 1

[9] Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. Building machines that learn and think like people. *Behavioral and brain sciences*, 40, 2017. 1

[10] Adam Lerer, Sam Gross, and Rob Fergus. Learning physical intuition of block towers by example. In *International conference on machine learning*, pages 430–438. PMLR, 2016. 3

[11] Alan M Leslie. Spatiotemporal continuity and the perception of causality in infants. *Perception*, 13(3):287–305, 1984. 1

[12] Shiqian Li, Kewen Wu, Chi Zhang, and Yixin Zhu. On the learning mechanisms in physical reasoning. *arXiv preprint arXiv:2210.02075*, 2022. 1, 3

[13] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation forest. In *2008 eighth ieee international conference on data mining*, pages 413–422. IEEE, 2008. 6, 13

[14] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3202–3211, 2022. 3

[15] Goeffrey J McLachlan. Mahalanobis distance. *Resonance*, 4(6):20–26, 1999. 6, 13

[16] Jerzy Neyman and Egon Sharpe Pearson. Ix. on the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231(694-706):289–337, 1933. 4, 12

[17] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. 13

[18] Leif E Peterson. K-nearest neighbor. *Scholarpedia*, 4(2):1883, 2009. 6

[19] 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, 3

[20] 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, 6, 7, 8, 9

[21] Haozhi Qi, Xiaolong Wang, Deepak Pathak, Yi Ma, and Jitendra Malik. Learning long-term visual dynamics with region proposal interaction networks. *arXiv preprint arXiv:2008.02265*, 2020. 3

[22] Ronan Riochet, Mario Ynocente Castro, Mathieu Bernard, Adam Lerer, Rob Fergus, Véronique Izard, and Emmanuel Dupoux. Intphys: A framework and benchmark for visual intuitive physics reasoning. *arXiv preprint arXiv:1803.07616*, 2018. 3

[23] Bernhard Schölkopf, John C Platt, John Shawe-Taylor, Alex J Smola, and Robert C Williamson. Estimating the support of a high-dimensional distribution. *Neural computation*, 13(7):1443–1471, 2001. 6, 13

[24] Mei-Ling Shyu, Shu-Ching Chen, Kanoksri Sarinnapakorn, and LiWu Chang. A novel anomaly detection scheme based on principal component classifier. Technical report, Miami Univ Coral Gables Fl Dept of Electrical and Computer Engineering, 2003. 6

[25] 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. *Advances in neural information processing systems*, 32, 2019. 3

[26] Elizabeth S Spelke and Katherine D Kinzler. Core knowledge. *Developmental science*, 10(1): 89–96, 2007. 1

[27] Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-distribution detection with deep nearest neighbors. *arXiv preprint arXiv:2204.06507*, 2022. 7, 12

[28] Tomer D Ullman, Elizabeth Spelke, Peter Battaglia, and Joshua B Tenenbaum. Mind games: Game engines as an architecture for intuitive physics. *Trends in cognitive sciences*, 21(9): 649–665, 2017. 3

[29] Yunbo Wang, Mingsheng Long, Jianmin Wang, Zhifeng Gao, and Philip S Yu. Predrnn: Recurrent neural networks for predictive learning using spatiotemporal lstms. *Advances in neural information processing systems*, 30, 2017. 1, 2, 3, 7, 8

[30] Jiajun Wu, Ilker Yildirim, Joseph J Lim, Bill Freeman, and Josh Tenenbaum. Galileo: Perceiving physical object properties by integrating a physics engine with deep learning. *Advances in neural information processing systems*, 28, 2015. 3

[31] Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B Tenenbaum. Clevrer: Collision events for video representation and reasoning. *arXiv preprint arXiv:1910.01442*, 2019. 3

[32] Puning Zhao and Lifeng Lai. Analysis of knn density estimation. *IEEE Transactions on Information Theory*, 68(12):7971–7995, 2022. 12

A Theoretical Proofs

Proof of Theorem 1 For any test method towards Eq. (2) with rejection set $\mathcal{R}$, and any $\alpha \in [0,1]$, we say that it satisfies condition P_α if (1) $\alpha = \Pr(\boldsymbol{x} \in \mathcal{R}|\textit{VoE})$. (2) $\exists \eta \geq 0$ such that

$$\boldsymbol{x} \in \mathcal{R} \setminus \mathcal{A} \Rightarrow \textit{LR}(\boldsymbol{x})^{-1} > \eta$$
$$\boldsymbol{x} \in \mathcal{R}^c \setminus \mathcal{A} \Rightarrow \textit{LR}(\boldsymbol{x})^{-1} < \eta,$$

where $\mathcal{A}$ is a set ignorable in both $\mathcal{P}_{\textit{VoE}}$ and $\mathcal{P}_{\textit{non-VoE}}$, i.e. $\Pr(\boldsymbol{x} \in \mathcal{A}|\textit{VoE}) = \Pr(\boldsymbol{x} \in \mathcal{A}|\textit{non-VoE}) = 0$. For any $\alpha \in [0,1]$, let the set of level α tests be the set of all tests with size (the probability of falsely rejecting the null hypothesis) at most α. That is, letting its rejection set be $\mathcal{R}'$, we have $\Pr(\boldsymbol{x} \in \mathcal{R}') \leq \alpha$.

Apparently, we know that the rejection region $\mathcal{R} = \{\boldsymbol{x} : \textit{LR}(\boldsymbol{x}) < \lambda_0\}$ satisfies condition P_α, where $\alpha = \Pr(\boldsymbol{x} \in \mathcal{R}|\textit{VoE})$. When $\boldsymbol{x} \in \mathcal{R} \setminus \mathcal{A}$, $\textit{LR}(\boldsymbol{x})^{-1} > \frac{1}{\lambda_0}$, and when $\boldsymbol{x} \in \mathcal{R}^c \setminus \mathcal{A}, \textit{LR}(\boldsymbol{x})^{-1} < \frac{1}{\lambda_0}$. Here $\mathcal{A} = \{\boldsymbol{x} : \textit{LR}(\boldsymbol{x}) = \lambda_0\}$ if we assume that $\mathcal{P}_{\textit{VoE}}$ and $\mathcal{P}_{\textit{non-VoE}}$ are continuous.

From Neyman-Pearson lemma [16], we know that $\mathcal{R}$ is a uniformly most powerful test in the set of level α tests (**existence**), and every UMP test $\mathcal{R}'$ in the set of level α and $\mathcal{R}$ will agree with probability 1 whether measured by $\mathcal{P}_{\textit{VoE}}$ or $\mathcal{P}_{\textit{non-VoE}}$ (**uniqueness**).

Proof of Theorem 3 The idea of this proof is based on [27]. We consider the convergence when estimating VoE distribution $\mathcal{P}_{\textit{VoE}}$. Note that $\boldsymbol{z}$ is a normalized feature vector in $\mathbb{R}^m$, which means $\boldsymbol{z}$ locates on the surface of a m-dimensional unit sphere. We denote $B(\boldsymbol{z}, r) = \{\boldsymbol{z}' : ||\boldsymbol{z}' - \boldsymbol{z}||_2 \leq r\} \cap \{\boldsymbol{z}' : ||\boldsymbol{z}'||_2 = 1\}$, which is a set of data points on the unit hyper-shpere and are at most r Euclidean distance away from the center $\boldsymbol{z}$. The local dimension of $B(\boldsymbol{z}, r)$ is $m-1$. Assuming the density satisfies Lebesgue's differentiation theorem, the probability density function can be attained by:

$$\mathbf{P}_{\textit{VoE}}(\boldsymbol{z}^\star) = \lim_{r \to 0} \frac{\Pr(\boldsymbol{z} \in B(z^\star, r)|\boldsymbol{z} \sim \mathcal{P}_{\textit{VoE}})}{|B(\boldsymbol{z}^\star, r)|}$$

We denote our observation set as $\mathbb{Z}_n = \{\boldsymbol{z}_1, \boldsymbol{z}_2, \cdots, \boldsymbol{z}_n\}$. We assume each sample $\boldsymbol{z}_i$ is $i.i.d.$ drawn from $\mathcal{P}_{VoE}$, then the empirical point-density for the test data can be estimated by KNN distance as:

$$\hat{\mathbf{P}}_{\textit{VoE}}(\boldsymbol{z}; k, n) = \frac{\Pr(\boldsymbol{z}_i \in B(\boldsymbol{z}^\star, r_k(\boldsymbol{z}^\star))|\boldsymbol{z}_i \in \mathbf{Z}_n)}{|B(\boldsymbol{z}^\star, r_k(\boldsymbol{z}^\star))|} = \frac{k}{cn(r_k(\boldsymbol{z}^\star))^{m-1}},$$

where c is a constant. Using the ℓ_α bound results in [32] can prove the convergence and establish the convergence rate of the estimator.

B Pseudo Code for Computing *surprise*$_{KNN}$

Algorithm 1 Compute *surprise*$_{KNN}$

Input: A VoE video dataset $\{\boldsymbol{x}^{(i)}\}_{i=1}^N$, a trained physics learning model $\boldsymbol{\Theta}$, hyper-parameters n, k, γ. A test video $\boldsymbol{x}$. traditional *surprise*$_{\boldsymbol{\Theta}}(\boldsymbol{x})$
Return: *surprise*$_{KNN}(\boldsymbol{x})$

Get the hidden representation $\{\boldsymbol{z}^{(i)}\}_{i=1}^N$of $\{\boldsymbol{x}^{(i)}\}_{i=1}^N$ using $\boldsymbol{\Theta}$
Get the hidden representation $\boldsymbol{z}$ of $\boldsymbol{x}$ using $\boldsymbol{\Theta}$
$\boldsymbol{z}^{(i)} \leftarrow \boldsymbol{z}^{(i)}/||\boldsymbol{z}^{(i)}||_2$ ▷ Normalize each feature vector into unit norm
$\boldsymbol{z} \leftarrow \boldsymbol{z}/||\boldsymbol{z}||_2$
compute distance $||\boldsymbol{z} - \boldsymbol{z}^{(i)}||_2$
Reorder $\{\boldsymbol{z}^{(i)}\}_{i=1}^N$ to $\{\boldsymbol{z}^{(s_i)}\}_{i=1}^N$ with increasing distance of $||\boldsymbol{z} - \boldsymbol{z}^{(i)}||_2$
$r_k(\boldsymbol{z}) \leftarrow ||\boldsymbol{z} - \boldsymbol{z}^{(s_k)}||_2$ ▷ distance to $\boldsymbol{z}$'s k-th nearest neighbor
surprise$_{KNN}(\boldsymbol{x}) \leftarrow$ *surprise*$_{\boldsymbol{\Theta}}(\boldsymbol{x}) - \gamma \cdot r_k(\boldsymbol{z})$

C Density Estimation Methods

Apart from KNN, we can also estimate $\mathcal{P}_{VoE}$ using the following three density estimators in the experiment. The implementations are based on `sklearn` [17]. And the code is attached in the supplemental material.

IForest [13] The Isolation Forest (IForest) 'isolates' observations by randomly selecting a feature and then randomly selecting a split value between the maximum and minimum values of the selected feature. Since recursive partitioning can be represented by a tree structure, the number of splittings required to isolate a sample is equivalent to the path length from the root node to the terminating node. This path length, averaged over a forest of such random trees, is a measure of normality and our decision function. Random partitioning produces noticeably shorter paths for anomalies. Hence, when a forest of random trees collectively produce shorter path lengths for particular samples, they are highly likely to have large distance to the data distribution. We use 100 base estimators in the ensemble.

OCSVM [23] One-Class SVM (OCSVM) is an unsupervised learning technique to learn the ability to differentiate the test samples of a particular class from other classes. OCSVM works on the basic idea of minimizing the hypersphere of the single class of examples in training data and considers all the other samples outside the hypersphere to be outliers or "out of observation data distribution". OCSVM estimates the support of a high-dimensional distribution, and we adapt it as a non-parametric density estimation method.

Mahalanobis distance [15] The Mahalanobis distance is a measure of the distance between a point P and a distribution $\mathcal{D}$, introduced by P. C. Mahalanobis in 1936. Mahalanobis's definition was prompted by the problem of identifying the similarities of skulls based on measurements. It is a multi-dimensional generalization of the idea of measuring how many standard deviations away P is from the mean of $\mathcal{D}$. This distance is zero for P at the mean of $\mathcal{D}$ and grows as P moves away from the mean along each principal component axis. If each of these axes is re-scaled to have unit variance, then the Mahalanobis distance corresponds to standard Euclidean distance in the transformed space. The Mahalanobis distance is thus unitless, scale-invariant, and takes into account the correlations of the data set. The equation is as follows:

$$r_{maha}(P) = diag[(P - \mu_{\mathcal{D}})\Sigma_{\mathcal{D}}^{-1}(P - \mu_{\mathcal{D}})^T] \tag{3}$$

Here Σ^{-1}, μ_D is the empirical precision matrix and empirical mean of observed data drawn from $\mathcal{D}$. Note that $P \in \mathbb{R}^d$, $\mu_{\mathcal{D}} \in \mathbb{R}^d$, $\Sigma^{-1} \in \mathbb{R}^{d\times d}$, and $diag(\cdot)$ denotes the extraction of diagonal of a matrix.