Predictive Inference Model of the Physical Environment that emulates Predictive Coding

Eri Kuroda[1][0000−0001−6248−5056] and Ichiro Kobayasi[1][0000−0001−7789−475X]

Ochanomizu University, Tokyo, Japan
{kuroda.eri, koba}@is.ocha.ac.jp

Abstract. In recent years, the significance of artificial intelligence in comprehending the real-world has increased, by leveraging the inherent ability of humans to process intuitive physics on a computer. Prior investigations on real-world understanding have mainly relied on image inference to recognize the physical environment. In contrast, we propose an inference model that can predict the observed environment using both visual and physical features, emulating the predictive coding hypothesized to occur in the human brain, and detects change points in response to predictive events. Additionally, the model verifies the correctness of the timing of important physical events of objects, such as object collisions and disappearances. Furthermore, the results of the physical information prediction are also described as natural language sentences to confirm whether the model accurately recognizes the real-world and predicts the next behavior based on the physical information.

Keywords: physical characteristics · latent hierarchical structure of physical relationships · prediction

1 Introduction

When faced with a specific circumstance, humans possess the innate capacity to swiftly comprehend environmental cues, predominantly through visual perception. This capability is believed to rely on the mental construction and simulation of the environment within the brain, contingent upon perceived stimuli [9]. Concurrently, humans are able to apprehend and anticipate the actions of objects in the environment, founded on the environmental framework constructed within their brain. At this point, humans generate predictions concerning both the physical and visual aspects of the perceived objects. It is believed that physical prediction pertains to significant events in the object, rather than forecasting all possible states of the environment. Considerable research has been conducted to achieve the human capacity to identify and forecast environmental information on a computer [9, 30, 22, 26, 18, 6, 13, 1, 8]. Nonetheless, the majority of real-world prediction studies have produced results based on either visual predictions via pixel alterations, or physical predictions via numerical variations in simulators, and no prediction model that can simultaneously generate both visual and physical predictions has been put forward, as humans are capable of doing. In this

investigation, we present a novel model capable of producing both visual and physical predictions regarding objects in the environment, whilst simultaneously extracting the timing of important events amongst the predicted events. The model is constructed through a combination of PredNet [20], a prediction model that replicates the top-down and bottom-up hierarchical information processing in the human brain, and the Variational Temporal Abstraction (VTA) [14] mechanism, which retrieves change points within the observed environment based on the visual information's image characteristics. The proposed model is rigorously evaluated to confirm its efficacy, wherein the timing of predicted object collisions within the event is ascertained using CLEVRER [33], representing physical phenomena such as object collisions, and the model's accuracy is verified by computing the correct timing. Furthermore, the physical prediction results are generated as sentences to facilitate interpretation and validate whether the model accurately forecasts the next action based on physical information.

2 Related Work

Real-world cognition. Real-world cognition refers to the study of machine learning and artificial intelligence for recognizing and interpreting the real-world. Ha et al [9] proposed the concept of world models as a mechanism by which humans perceive and understand the environment. When humans visually observe the environment, they can quickly recognize the objects and their behavior in the environment. This is made possible by modeling and simulating the environment in the brain based on the sensory input. LeCun [16] identified one of the three challenges that AI research must address in the future: "How can machines learn to represent, predict, and act on the world from observation?" Humans and animals can gain insight into how the world works and acquires background knowledge through limited interaction and observation. This is considered the basis of common sense, which not only predicts future outcomes but also fills in information gaps in time and space. Common sense consists of models of the world that inform us about what is probable and what is improbable. This allows humans and animals to predict, reason, plan, explore sequences of actions, and imagine novel solutions to problems. The study of real-world cognition is therefore crucial.

Prediction. Research on real-world cognition often focuses on visual prediction and commonly employs Recurrent Neural Networks (RNN) and Long Short Term Memory (LSTM) methods [11]. However, while Chang et al. [2] proposed a new model, STIP, to address the problem of generating high-resolution predictions due to a loss function based on information loss and mean squared error, the emphasis of many studies is solely on capturing temporal dependencies between frames, with little discussion of the spatial features within frames. To rectify this, Wang et al. introduced a spatio-temporal LSTM (ST-LSTM) structure to predict high-quality videos and proposed novel prediction models such as PredRNN++ [28] and PredRNN [29]. Additionally, to enable long-term prediction, Lin et al. [19] integrated a self-attention mechanism into ST-LSTM to

store long-range spatial features, while Lee et al. [17] introduced memory alignment learning to store long-term temporal dependencies. Other proposed models include Iso-Dream [23], an improved version of Dreamer [10], which separately learns controllable and uncontrollable state transitions and combines them with prediction, and Gao et al.'s SimVP [7], a prediction model that merges image recognition with Transformer technology and uses Vision Transformer [5]. These studies aim to produce highly accurate prediction results and expand research on models that can make long-term predictions, such as for humans.

Physical Reasoning, Intuitive Physics. The field of common sense or intuitive physics, which involves computational understanding of the physical world, has been studied extensively in recent years. Representing intuitive physics is crucial for modeling object interactions and predicting their dynamics, and has received considerable attention [4, 3, 26]. Tang et al. [26] proposed PHYCINE, a hierarchical prediction model that focuses not only on first-order features such as object position and shape, but also on hidden behaviors of objects such as mass and charge, by discovering physical concepts of objects from low-level (color, shape) to high-level abstract (mass, charge) from video images. Ye et al. [31] and Piloto et al. [25] focus on learning intuitive physical properties that can be interpreted. In addition, many studies have attempted to learn intuitive physical properties from a few frames of a video image. Yi et al. [32] have focused on the complex temporal and causal structures underlying object interactions, using the image reasoning dataset CLEVR (Compositional Language and Elementary Visual Reasoning diagnostics dataset) [12] with CLEVRER (CoLlision Events for Video REpresentation and Reasoning) [33]. They also extended CLEVRER and proposed CLEVRER-Humans as a video inference dataset for human-labeled causality inference [22].

3 PredNet

PredNet [21] is a deep prediction neural network construct to mimic the concept of predictive coding. An overview of the model is shown in Figure 1. Each module has four internal components: an input convolutional layer (x_{t_k}), a recurrent convolutional representation layer (R), a convolutional prediction layer (A), and an error representation (E). The representation layer in each module captures the state for prediction, while the input layer processes the input information. The prediction layer generates the internal prediction state, and the error layer outputs the error representation by taking the difference between the prediction state and the input. PredNet utilizes a bidirectional process to generate predictions, where predictions made in the upper layers of the network are conveyed to the lower layers via the representation module, and errors detected in the lower layers are transmitted to the upper layers. This mechanism mimics the operation of a generalized state equation, which enables accurate predictions to be made.

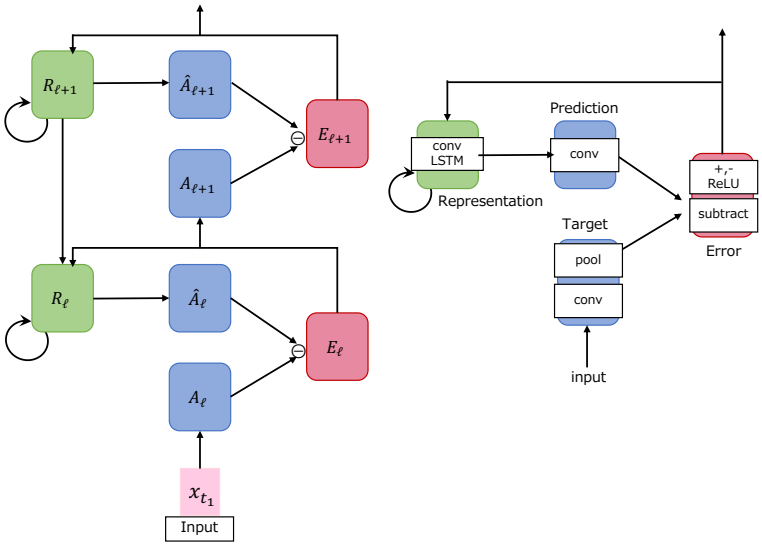

Fig. 1. Schematic diagram of PredNet.

4 Variational Temporal Abstraction (VTA)

In Variational Temporal Abstraction (VTA) [14], a state-space model is proposed to extract hierarchical abstractions from series data and detect change points. Figure 2 (right) is the graphical model of the hierarchical state-space model obtained by VTA. In this figure, X is the input, S is the observation abstraction, and Z is the temporal abstraction. X is the lowest layer closest to the input, and S and Z are the upper layers in that order. By processing the input series information and obtaining the hierarchical structure, VTA enables the acquisition of the upper Z representation, which indicates the transition of the environment. However, as for the state space models that handle sequential series information, in general, it is difficult to determine when to transition to the upper layer Z, taking into account temporal transitions, as depicted in Figure 2 (left) to (right). To address this issue, VTA introduces a binary latent variable m that determines the timing, as shown in Figure 2(left). The boundary indicator $M = m_{1:T}$ takes the value 0 or 1. When the change in the observed or temporal abstraction is significant, m becomes 1, and the upper layer Z transitions accordingly.

5 Proposed Model

5.1 Mechanism of the change point prediction model

PredNet [20] and VTA mechanisms are integrated to construct an change point prediction model. This model mimics predictive coding which is a hypothesized function in the human brain. The model architecture is presented in Figure 3.

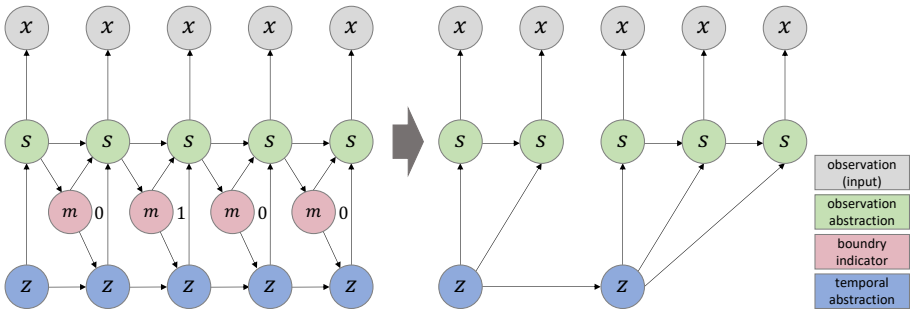

Fig. 2. Schematic diagram of VTA. (left) Model with boundary index $M = \{0, 1, 0, 0\}$. (right) Model with time structure obtained from the boundary index M.

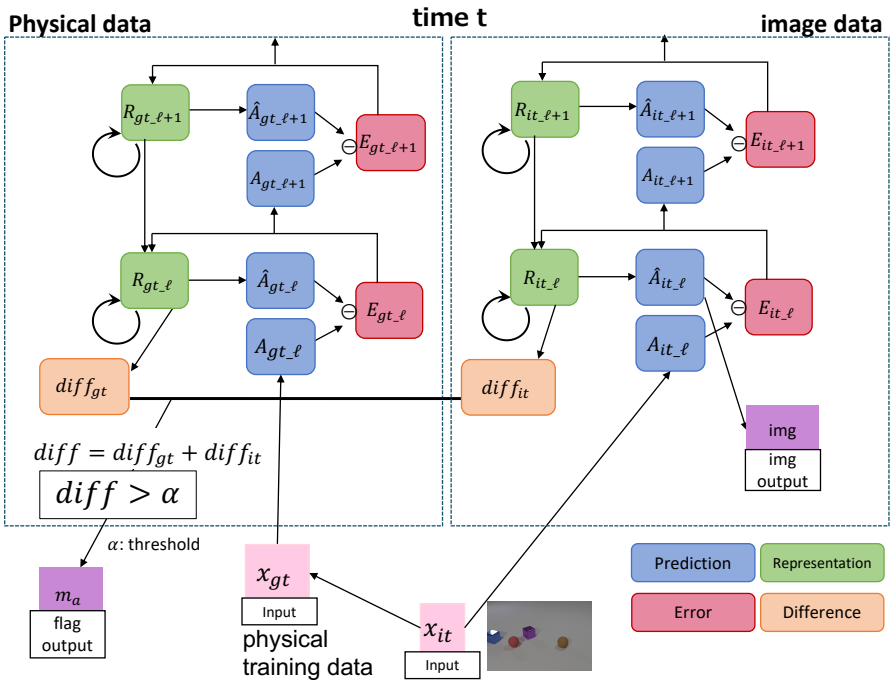

Fig. 3. Schematic diagram of a change point prediction model.

The proposed model is a parallel hierarchical structure of two PredNet models, one of which predicts physical phenomena of the environment by representing them as graphs, and the other of which predicts them by visual information of the environment. The proposed model also incorporates the change point discrimination flag m, which is a mechanism of VTA. The input information consists of two datasets: the CLEVRER dataset with image information x_{it} and the physical training dataset x_{gt} with physical properties generated from CLEVRER. The output information consists of two pieces of information: the predicted image ("img output" in Figure 3), which was sequentially predicted for the image, and the change point m_a ("flag output" in Figure 3), which was computed by the inference of the embedding vector representing the physical property. The change point m_a serves as an indicator flag, signifying when the cumulative value of physical and image data has significantly changed and takes the value 0 or 1. Both mechanisms learn by error propagation to higher levels, minimizing the differences between prediction $\hat{A}$ derived from the representation tier R and actual observation A. To determine the change point m, the difference $diff$ between the representation layer R at time $t-1$ and time t is calculated for the physical and image data, respectively, such that the change point m_a becomes 1 if the difference $diff$ exceeds a threshold value α.

The algorithmic updates are expounded upon in Algorithm 1, along with Equations (1) through (11). In this instance, R represents the layer of representation, A represents the layer of prediction, $\hat{A}$ signifies the generated prediction content derived from the representation layer R, and E represents the layer of error. Furthermore, it denotes the variable used in the processing of the image, while gt denotes the variable utilized in the processing of physical information. Equation (12) illustrates the training loss. λ_t and λ_ℓ are weighting factors for time and layer, respectively, and n is the number of units in the ℓ-th layer.

$$A_l^{it} = \begin{cases} x_{it} & \text{if} \quad l = 0 \\ MaxPool(ReLU(Conv(E_{l-1}^{it}))) & l > 0 \end{cases} \tag{1}$$

$$A_l^{gt} = \begin{cases} x_{gt} & \text{if} \quad l = 0 \\ MaxPool(ReLU(Conv(E_{l-1}^{gt}))) & l > 0 \end{cases} \tag{2}$$

$$\hat{A}_l^{it} = ReLU(Conv(R_l^{it})) \tag{3}$$

$$\hat{A}_l^{gt} = ReLU(Conv(R_l^{gt})) \tag{4}$$

$$E_l^{it} = [ReLU(A_l^{it} - \hat{A}_l^{it}); ReLU(\hat{A}_l^{it} - A_l^{it})] \tag{5}$$

$$E_l^{gt} = [ReLU(A_l^{gt} - \hat{A}_l^{gt}); ReLU(\hat{A}_l^{gt} - A_l^{gt})] \tag{6}$$

$$R_l^{it} = ConvLSTM(E_l^{it-1}, R_l^{it-1}, Upsample(R_{l+1}^{it})) \tag{7}$$

$$R_l^{gt} = ConvLSTM(E_l^{gt-1}, R_l^{gt-1}, Upsample(R_{l+1}^{gt})) \tag{8}$$

$$diff_{it} = R_l^{it} - R_l^{it-1} \tag{9}$$

$$diff_{gt} = R_l^{gt} - R_l^{gt-1} \tag{10}$$

$$diff = diff_{it} + diff_{gt} \tag{11}$$

$$L_{train} = \sum_t \lambda_t \sum_l \frac{\lambda_l}{n_l} \sum_{n_l} E_l^t \tag{12}$$

Algorithm 1 Calculation of change point prediction model

Require: x_{it}, x_{gt}
  $A_0^{it} \leftarrow x_{it}, A_0^{gt} \leftarrow x_{gt}$
  $E_l^0, R_l^0 \leftarrow 0$
  for $t = 1$ **to** T **do**
    for $l = L$ **to** 0 **do**
      if $l = L$ **then**
        $R_L^{it} = \text{ConvLSTM}(E_L^{it-1}, R_L^{it-1})$
        $R_L^{gt} = \text{ConvLSTM}(E_L^{gt-1}, R_L^{gt-1})$
      else
        $R_l^{it} = \text{ConvLSTM}(E_l^{it-1}, R_l^{it-1}, \text{Upsample}(R_{l+1}^{it}))$
        $R_l^{gt} = \text{ConvLSTM}(E_l^{gt-1}, R_l^{gt-1}, \text{Upsample}(R_{l+1}^{gt}))$
      end if
    end for
    for $l = 0$ **to** L **do**
      if $l = 0$ **then**
        $\hat{A}_0^{it} = \text{SatLU}(\text{ReLU}(\text{Conv}R_0^{it}))$
        $\hat{A}_0^{gt} = \text{SatLU}(\text{ReLU}(\text{Conv}R_0^{gt}))$
      else
        $\hat{A}_l^{it} = \text{ReLU}(\text{Conv}R_l^{it})$
        $\hat{A}_l^{gt} = \text{ReLU}(\text{Conv}R_l^{gt})$
      end if
      $E_l^{it} = [\text{ReLU}(A_l^{it} - \hat{A}_l^{it}); \text{ReLU}(\hat{A}_l^{it} - Ait_l)]$
      $E_l^{gt} = [\text{ReLU}(A_l^{gt} - \hat{A}_l^{gt}); \text{ReLU}(\hat{A}_l^{gt} - A_l^{gt})]$
      if $l < L$ **then**
        $A_{l+1}^{it} = \text{MaxPool}(\text{Conv}(E_{it}^l))$
        $A_{l+1}^{gt} = \text{MaxPool}(\text{Conv}(E_{gt}^l))$
      end if
      $diff_{it} = R_l^{it} - R_l^{it-1}$
      $diff_{gt} = R_l^{gt} - R_l^{gt-1}$
      $diff = diff_{it} + diff_{gt}$
      if $diff > \alpha$ **then**
        $m_a = 1$
      else
        $m_a = 0$
      end if
    end for
  end for

6 Experiment

6.1 Change point extraction in predictive inference

To verify the effectiveness of our proposed model, an experiment was conducted to see if the model can correctly extract the change point of the next step state. The dataset we used was the CLEVRER dataset and physical training data generated from CLEVRER.

Physical training dataset The two datasets we used were CLEVRER dataset [33] and a dataset representing physical properties of real-world objects – the procedure for creating the physical properties dataset is shown in Figure 4.

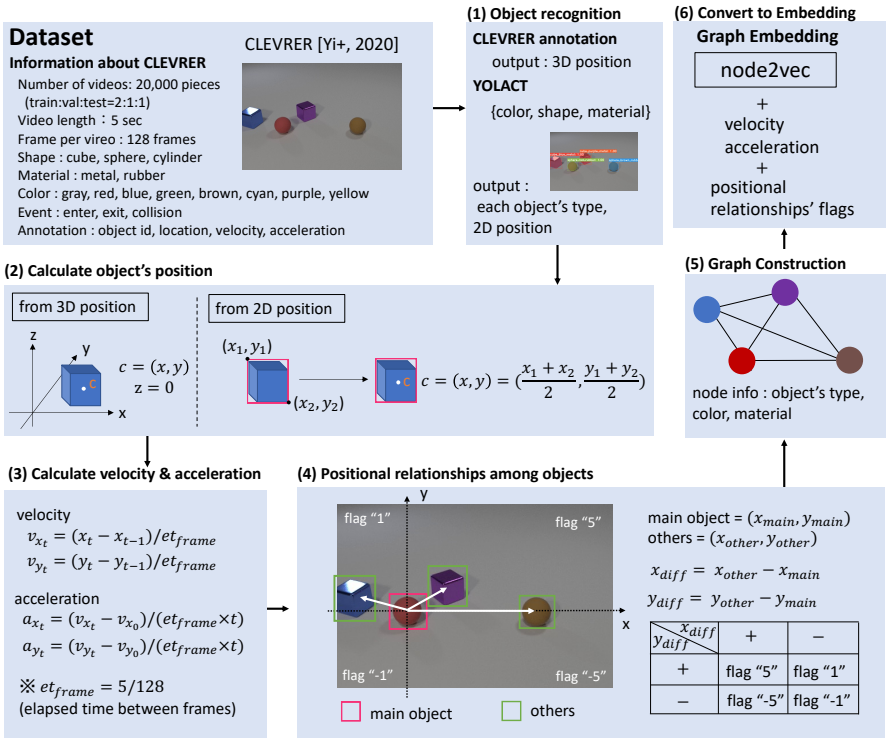

Fig. 4. Steps to create physical training dataset.

Table 1 shows the experimental settings of the model, which is the same as used in the previous study [20].

Results & Discussion The results of the change point prediction accuracy in the proposed model are shown in Table 2. The physical data in Table 2 shows the results

Table 1. Experimental Settings.

Number of training data	600,000
Number of validation data	60,000
Number of times studied	500,000
#Layers	4
Size of convolutional filter	3×3 (for all conventions)
#Channels	From lower module, 3, 48, 96, 192
Optimization	Adam [15]
Learning rate decay	0.0001
α	5

obtained from the data set created in Figure 4, and the annotation data in the table shows the results obtained from the CLEVRER annotation dataset.

Table 2. Accuracy of the proposed prediction model.

Validation range	i	ii	iii	iv	v	vi
physical data	33.3	**50**	50	33.3	**66.7**	**50**
annotation data	**66.7**	**50**	**66.7**	**40**	50	**50**

The results show that the accuracy of the physical data is equivalent to that of the annotation data, which is the supervised data, in predicting the change points. As a result example, the predicted images and flags for region i are shown in Figure 5. As the predicted image is also accurately generated, it can be said that this model is proficient in generating both the predicted image and flag of the next time's change point.

6.2 Text generation of prediction results

The proposed model made two predictions, one for physical data and one for visual data. Humans apprehend and acquire knowledge of the real-world by perceiving it and engaging in predictions and inferences. Furthermore, linking language to the physical world enables us to gain a more profound comprehension of reality and our prior experiences. Put differently, human intelligence can be conveyed through symbol manipulation using language that pertains to the real-world. Therefore, research on comprehending the physical world through machine learning technology should express reasoning as a language, with the aim of linking the recognition of real-world objects, understanding of physical properties, and prediction using language. This study generated embedded vectors, extracted as change points in physical data, as a form of language information. Additionally, only collisions were used as change points for the generation.

Dataset To generate language from the embedded vectors predicted by the change-point prediction model, it is necessary to learn new linguistic information. For this purpose, we developed a language dataset consisting of a pair of data: an embedded

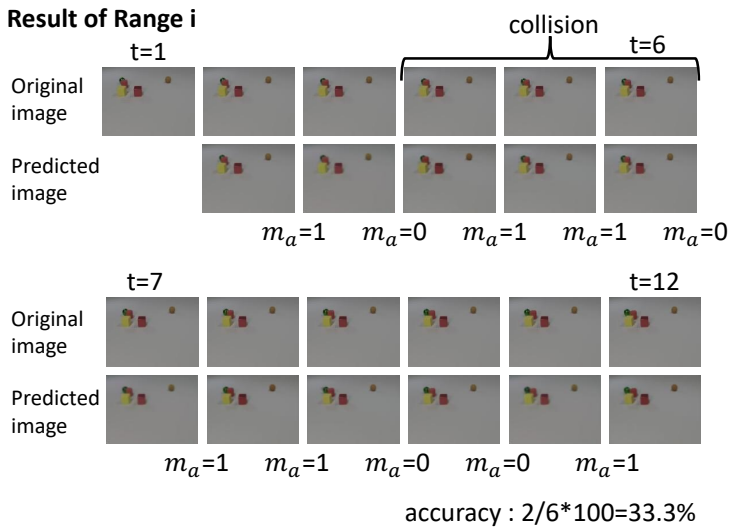

Fig. 5. Predicted change point extraction results in range i.

vector of graph representations representing physical properties and a sentence describing the state of the graph. Although the experiment was conducted in Japanese, this paper covers both English and Japanese. The graph's embedding vector representation was created from the CLEVRER annotation data using the procedure illustrated in Figure 4. The paired sentences were devised to fit into nine templates of three (before collision, collision, and after collision) × three (type of sentences). The correct answer for each image was three sentences. The details of the templates are as follows: Two objects A and B collide with each other, and A and B are "{gray, red, blue, green, brown, water, purple, yellow} {sphere, cylinder, cube }." For example, "Red sphere" and "Blue cylinder" are now included. In addition to the collision data, we also created a dataset for when the objects were approaching before the collision and when they were leaving after the collision. The approaching time was five frames before the collision, and the leaving time was five frames after the collision. An example of the generated pair dataset is shown in Figure 6.

Text generation model The text generation model utilized only the decoder component of the Transformer [27]. The decoder architecture is depicted in Figure 7. Although conventional transformers are based on an encoder-decoder model, this study adopts the embedding vector prediction result of the graph in the change-point prediction model of the proposed model as the encoder output. This prediction result is employed as input from the encoder to the decoder. The paired data generated in Figure 4 was utilized to train the decoder, with the number of paired sentence data set at 219,303 (nine sentences × 24,367 collisions) and the predictive graph embeddings for test data set at 10,965. The training settings are detailed in Table 3.

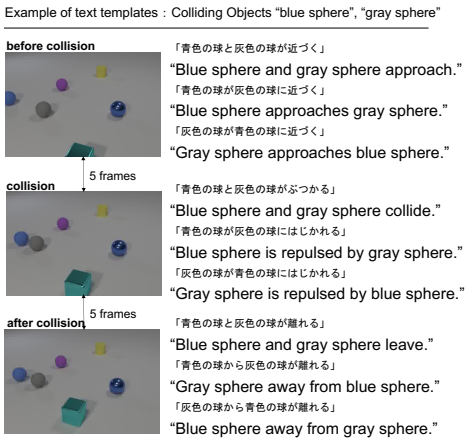

Fig. 6. Example of text templates.

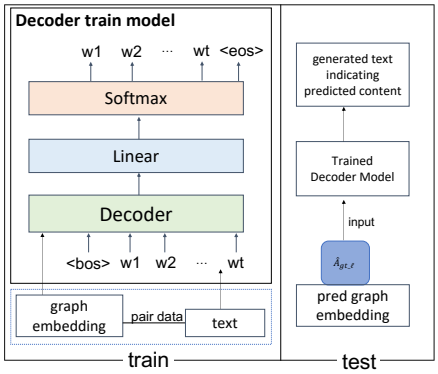

Fig. 7. Schematic diagram of the text generation model.

Table 3. Experiment Setting.

batch size	8
Embedding	128
hidden layer	512
Optimization	Adam[15]

Results & Discussion We confirmed that the embedded representations of the predicted graphs made correct predictions about the real-world by generating a language sentence describing the observed real-world situation. The four ranges that were examined for description were those shown in Figure 2, i, ii, iv, and vi, which indicate the time of the collision.

Range of i. In range i, a green sphere collides with a red cylinder and the assumed correct statement is shown in Figure 8. The generated sentence was "A green cylinder is repelled by a red cylinder." The sentence was correct about the color of the object, but incorrect about its shape.

Range of ii. In range ii, a green cylinder collides with a brown cube and the assumed correct statement is shown in Figure 8. The sentence generated was "A green cylinder collides with a brown cube." The sentence was correct for both color and shape of the objects.

Range of iv. In range iv, a gray sphere collides with a blue cube and the assumed correct statement is shown in Figure 8. The generated sentence was "A grey sphere is repelled by blue cube." The sentence was correct for the color of the object, but incorrect for the shape.

Range of vi. In range vi, a cyan cube collides with a blue sphere and the assumed correct statement is shown in Figure 8. The generated sentence was "A cyan cube collides with a blue sphere," which was incorrect for the object's color and shape. Unlike the other results, range vi produced incorrect judgments for both color and shape of the object. Figure 9 depicts the objects' transition up to the collision in

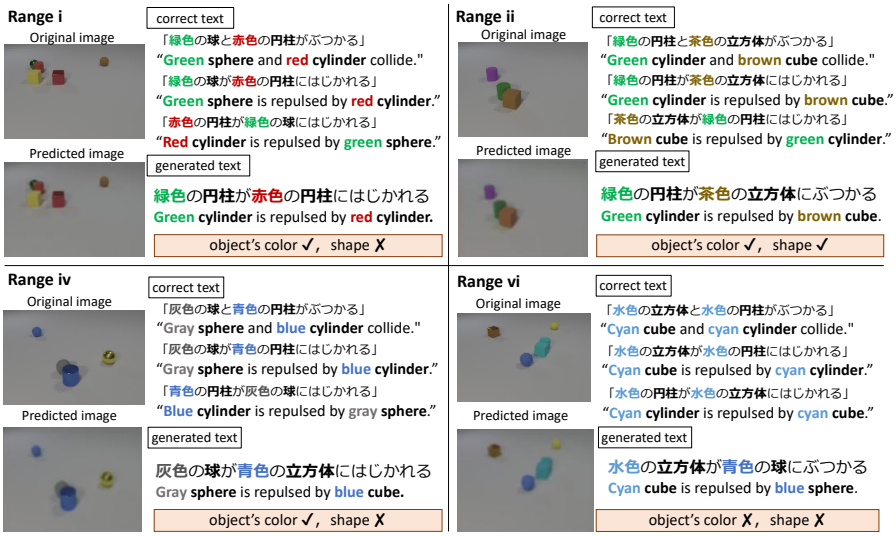

Fig. 8. Example of text generation for prediction results.

range vi, which includes the "cyan cube" and "cyan cylinder" colliding objects. It is noticeable that the "cyan cube" passed through the "blue sphere" without collision. The infinitesimally small distance between the cyan cube and the blue sphere led to the incorrect prediction of their collision. It is likely that considering the cyan cylinder hidden behind the cyan cube, and both objects being of the same color, contributed to the failure to generate a description accurately. To improve the text generation accuracy, it is necessary to improve the points where objects of the same color are regarded as the same object and where incorrect collision predictions are made.

Accuracy verification with BLEU.

The accuracy of the generated text is evaluated by BLEU [24]. BLEU@n is a measure

Table 4. BLEU evaluation.

	BLEU@2	BLEU@3	BLEU@4
score	79.7	74.5	68.8

of how well each correct and generated sentence matches in the n-gram. The evaluation results of the generated sentences using the BLEU evaluation metric are presented in Table 4. Since there were three correct answers for each generated sentence, the average of each score was used as the BLEU score for the generated sentence. The BLEU scores were computed for Japanese sentences, and the generated sentences achieved scores of 80 for the 2-gram, 75 for the 3-gram, and 69 for the 4-gram, indicating that they were able to generate informative and accurate sentences about the observed environment to a certain extent.

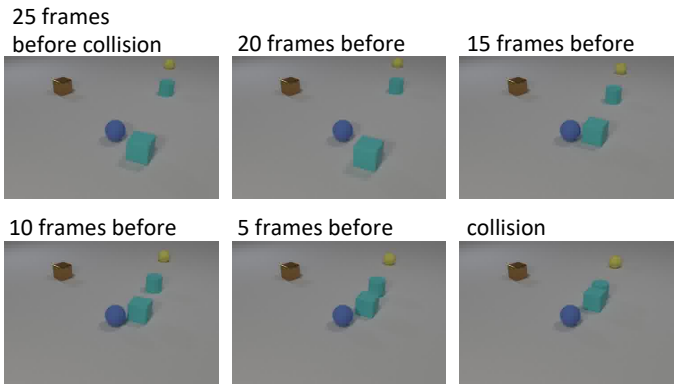

Fig. 9. Object transition status of range vi.

7 Conclusions

In this study, we constructed a model that emulates the structures in the human brain, which can predict the observed environment visually and physically. The predictive model was able to appropriately retrieve change points occurring in the next step, such as object collisions in the environment. Moreover, we generated descriptions from the predicted physical attributes of the environment and calculated the BLEU score, resulting in a language generation capability with a certain degree of accuracy. Based on this outcome, we assert that this model is capable of not only visual prediction but also physical prediction. The outputs of this model and language generation have allowed us to establish a link between the recognition of real-world objects and the understanding and prediction of their physical properties, mediated through the use of language. On the other hand, we believe that there is still room for improvement in both prediction model and language generator since the target dataset is less complex than the actual environment perceived by humans. As future work, we aim to enhance the model and expand the number of language datasets to allow language generation for various physical properties other than collisions.

Acknowledgements This work was supported by the Japan Society for the Promotion of Science KAKENHI Grant Numbers JP22J21786, JP22KJ1355, 23H03453 and JSPS Bilateral Program Number JPJSBP120213504.

References

1. Bear, D.M., Wang, E., Mrowca, D., Binder, F.J., Tung, H.Y.F., Pramod, R.T., Holdaway, C., Tao, S., Smith, K., Sun, F.Y., Fei-Fei, L., Kanwisher, N., Tenenbaum, J.B., Yamins, D.L.K., Fan, J.E.: Physion: Evaluating physical prediction from vision in humans and machines (Jun 2021)
2. Chang, Z., Zhang, X., Wang, S., Ma, S., Gao, W.: STIP: A SpatioTemporal Information-Preserving and Perception-Augmented model for High-Resolution video prediction (Jun 2022)

3. Chen, Z., Yi, K., Li, Y., Ding, M., Torralba, A., Tenenbaum, J.B., Gan, C.: ComPhy: Compositional physical reasoning of objects and events from videos (May 2022)
4. Ding, M., Chen, Z., Du, T., Luo, P., Tenenbaum, J.B., Gan, C.: Dynamic visual reasoning by learning differentiable physics models from video and language (Oct 2021)
5. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale (Oct 2020)
6. Duan, J., Dasgupta, A., Fischer, J., Tan, C.: A survey on machine learning approaches for modelling intuitive physics (Feb 2022)
7. Gao, Z., Tan, C., Wu, L., Li, S.Z.: SimVP: Simpler yet better video prediction (Jun 2022)
8. Ge, J., Liu, Y., Gui, J., Fang, L., Lin, M., Kwok, J.T.Y., Huang, L., Luo, B.: Learning the relation between similarity loss and clustering loss in Self-Supervised learning (Jan 2023)
9. Ha, D., Schmidhuber, J.: World models (Mar 2018)
10. Hafner, D., Lillicrap, T., Ba, J., Norouzi, M.: Dream to control: Learning behaviors by latent imagination (Dec 2019)
11. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (Nov 1997)
12. Johnson, J., Hariharan, B., van der Maaten, L., Fei-Fei, L., Zitnick, C.L., Girshick, R.B.: CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. CoRR **abs/1612.06890** (2016), http://arxiv.org/abs/1612.06890
13. Kandukuri, R.K., Achterhold, J., Moeller, M., Stueckler, J.: Physical representation learning and parameter identification from video using differentiable physics. Int. J. Comput. Vis. **130**(1), 3–16 (Jan 2022)
14. Kim, T., Ahn, S., Bengio, Y.: Variational temporal abstraction. CoRR **abs/1910.00775** (2019), http://arxiv.org/abs/1910.00775
15. Kingma, Ba: Adam: A method for stochastic optimization. arXiv:1412. 6980 [cs] (Jan 2017)
16. LeCun, Y.: A path towards autonomous machine intelligence
17. Lee, S., Kim, H.G., Choi, D.H., Kim, H.I., Ro, Y.M.: Video prediction recalling long-term motion context via memory alignment learning (Apr 2021)
18. Li, Z., Zhu, X., Lei, Z., Zhang, Z.: Deconfounding physical dynamics with global causal relation and confounder transmission for counterfactual prediction. AAAI **36**(2), 1536–1545 (Jun 2022)
19. Lin, Z., Li, M., Zheng, Z., Cheng, Y., Yuan, C.: Self-Attention ConvLSTM for spatiotemporal prediction. AAAI **34**(07), 11531–11538 (Apr 2020)
20. Lotter, Kreiman, Cox: Deep predictive coding networks for video prediction and unsupervised learning. arXiv:1605. 08104 [cs, q-bio] (Feb 2017)
21. Lotter, W., Kreiman, G., Cox, D.: A neural network trained to predict future video frames mimics critical properties of biological neuronal responses and perception (May 2018)
22. Mao, J., Yang, X., Zhang, X., Goodman, N., Wu, J.: CLEVRER-Humans: Describing physical and causal events the human way (Oct 2022)
23. Pan, M., Zhu, X., Wang, Y., Yang, X.: Iso-Dream: Isolating and leveraging noncontrollable visual dynamics in world models (May 2022)

24. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. pp. 311–318. ACL '02, Association for Computational Linguistics, USA (Jul 2002)
25. Piloto, L.S., Weinstein, A., Battaglia, P., Botvinick, M.: Intuitive physics learning in a deep-learning model inspired by developmental psychology. Nature Human Behaviour **6**(9), 1257–1267 (2022). https://doi.org/10.1038/s41562-022-01394-8
26. Tang, Q., Zhu, X., Lei, Z., Zhang, Z.: Intrinsic physical concepts discovery with Object-Centric predictive models (Mar 2023)
27. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. CoRR **abs/1706.03762** (2017), http://arxiv.org/abs/1706.03762
28. Wang, Y., Gao, Z., Long, M., Wang, J., Yu, P.S.: PredRNN++: Towards a resolution of the Deep-in-Time dilemma in spatiotemporal predictive learning (Apr 2018)
29. Wang, Y., Wu, H., Zhang, J., Gao, Z., Wang, J., Yu, P.S., Long, M.: PredRNN: A recurrent neural network for spatiotemporal predictive learning (Mar 2021)
30. Wu, B., Yu, S., Chen, Z., Tenenbaum, J.B., Gan, C.: STAR: A benchmark for situated reasoning in Real-World videos (Jan 2022)
31. Ye, T., Wang, X., Davidson, J., Gupta, A.: Interpretable intuitive physics model. In: Proceedings of (ECCV) European Conference on Computer Vision. pp. 89 – 105 (September 2018)
32. Yi, Gan, Li, Kohli, Wu, Torralba, Tenenbaum: CLEVRER: CoLlision events for video REpresentation and reasoning. arXiv:1910. 01442 [cs] (Mar 2020)
33. Yi, K., Gan, C., Li, Y., Kohli, P., Wu, J., Torralba, A., Tenenbaum, B.J.: Clevrer: Collision events for video representation and reasoning. ICLR (2020)