Exploration of the Intuitive Physics through Latent Space Disentanglement

Turgay Yıldız
Department of Cognitive Science, Graduate School of Informatics
Middle East Technical University (METU)
yildiz.turgay@metu.edu.tr

***Abstract*—The ability to make physical inferences is a fundamental aspect of human cognition. However, the computational mechanisms underlying these inferences remain poorly understood and under-explored. In recent years, many efforts have been made to address this challenge. Within the field of AI research, a primary focus has been to understand and integrate this capability into machine intelligence using Bayesian models and deep neural networks. Notably, there has been a significant rise in the use of deep generative models for this purpose. This project aims to investigate intuitive physics through the lens of latent space disentanglement by employing a Beta-Variational Autoencoder (Beta-VAE).**

***Index Terms*—Intuitive Physics, Beta Variational Autoencoder, Beta-VAE, Deep Generative Models, Latent Space, Disentanglement, Generative Factors**

I. Introduction

Every day, we encounter numerous physical events in our daily lives that require us to make rapid physical inferences. For instance, when stacking plates in the kitchen, we intuitively arrange them in a way that prevents them from toppling over. Even human infants demonstrate an understanding of whether their toy tower-blocks will fall if placed in an unstable configuration. Despite these abilities, the mechanisms behind how we make such rapid inferences remain unclear. While there are many theoretical approaches to these phenomena, as Battaglia et al. (2013) [3] highlight, the computational foundations of these rapid physical inferences are still poorly understood. Consequently, the literature explores these phenomena through various lenses, such as physical scene understanding (Battaglia et al., 2013) [3] , physically-grounded abstract social events for machine social perception (Netanyahu et al., 2021) [5], and intuitive physics as a core component of human intelligence (Lake et al., 2017) [4].

The aforementioned studies primarily focus on the concept of intuitive physics and are grounded in domain knowledge from fields such as cognitive science and developmental psychology. In addition, there are significant engineering efforts and research initiatives by leading AI companies aimed at advancing intuitive physics studies, as seen in the works of Rezaei-Shoshtari et al. (2021) [6] and Lerer et al. (2016) [7].

The current project builds on these studies but diverges in its objectives. Unlike previous research, which often aims to predict or understand physical inferences, this study seeks to gain control over generative factors by leveraging latent space disentanglement. While some prior studies, such as Piloto et al. (2022) [8], have explored intuitive physics using latent space representations, to the best of my knowledge, none have explicitly employed the concept of disentanglement. Therefore, this project investigates intuitive physics through latent space disentanglement to explore whether it is possible to control the generative factors underlying intuitive physics.

II. Literature Review

1) ***Intuitive Physics****:* Intuitive physics refers to our innate ability to interact with and understand the physical world. It encompasses cognitive concepts such as *object permanence*, *solidity*, and *continuity*. For example, intuitive physics can be conceptualized as the understanding of physical principles like object permanence (Lake et al., 2017) [4], which allows infants to recognize that objects continue to exist even when they are out of sight. This cognitive ability involves the awareness that objects persist over time, despite temporary occlusion.

Moreover, infants demonstrate an ability to disregard physically impossible trajectories (Lake et al., 2017) [4]. They show an understanding of how objects behave in accordance with fundamental physical laws, such as gravity and friction. Additionally, infants exhibit comprehension of properties like solidity (Lake et al., 2017) [4] and other analogous characteristics. Intuitive physics can also serve as a foundation for building models that provide a more causally accurate understanding of the world. For instance, Battaglia et al. (2013) [3] developed the Intuitive Physics Engine (IPE), a probabilistic simulation tool designed to model object behavior and trajectories under various conditions.

The term *intuitive physics* is predominantly used in cognitive science and developmental psychology research. A common assumption in these research is that a significant gap exists between human and machine intelligence, partly due to the inability of machines or AI models to understand physics in the same way humans do (Piloto et al., 2022 [8]; Lake et al., 2017 [4]; Battaglia et al., 2013 [3]). As a result, many studies aim to bridge this gap by developing AI models that can comprehend physics as humans do. This is achieved

through the creation of new datasets and the application of domain-specific knowledge. The ultimate goal of such research is to replicate human perception, learning, and cognitive processes in machine systems, effectively creating machines that can see, learn, and think like humans (Lake et al., 2011 [2]; Lake et al., 2015 [1]; Lake et al., 2017 [4]).

*2) **Variational Autoencoders (VAEs):*** Variational Autoencoders (VAEs) are deep generative models designed to generate new samples by leveraging a compressed representation of data known as the *latent space*. One of the main advantages of VAEs is their relative ease of training and greater stability compared to other generative models, such as Generative Adversarial Networks (GANs). However, a notable limitation of VAEs is their tendency to produce outputs that are often less sharp or detailed, resulting in lower-quality reconstructions compared to GANs.

A specialized version of VAEs, called Beta-VAE, is employed in this project. Beta-VAE introduces a hyperparameter, β, which controls the trade-off between reconstruction accuracy and the disentanglement of latent variables. One key advantage of Beta-VAE is its ability to suppress the Kullback-Leibler (KL) divergence term in the loss function during the initial stages of training. This allows the model to focus on learning effective reconstructions early on, without being dominated by the KL loss. As training progresses, the KL term is gradually reinforced, encouraging the latent space to adopt a more structured and disentangled representation. This approach can lead to a modest improvement in reconstruction quality while promoting better interpretability of the latent space.

Additionally, Beta-VAE's emphasis on *disentanglement* makes it particularly suitable for tasks requiring control over specific *generative factors*, such as those in intuitive physics. By isolating and manipulating individual latent variables, Beta-VAE enables more precise generation and analysis of data, which can be critical for understanding complex physical interactions.

*3) **Latent Space and Disentanglement:*** The architecture of Variational Autoencoders (VAEs) includes a *bottleneck* layer, which is intentionally designed to reduce the dimensionality of the input data. This bottleneck compresses high-dimensional inputs into a lower-dimensional latent space, making the data more manageable and controllable. By reducing dimensionality, VAEs can capture the most essential features of the data while discarding redundant or less relevant information.

The Kullback-Leibler (KL) divergence term in the VAE loss function plays a critical role in shaping the latent space. It encourages the latent space vector, denoted as $\mathbf{z}$, to approximate a Gaussian distribution. This ensures that the latent space is continuous and structured, enabling more meaningful representations of the data. Additionally, the introduction of noise at the bottleneck introduces stochasticity into the model. This stochasticity, combined with the ability to manipulate the generative factors of the latent vector $\mathbf{z}$, allows for creative and controlled generation of outputs.

Disentanglement refers to the property of the latent space where changes in the values of the latent vector $\mathbf{z}$ correspond to unique and independent changes in the generative factors of the output. In other words, each dimension of the latent vector should control a distinct and interpretable feature of the generated data. For example, in an image generation task, one dimension of $\mathbf{z}$ might control the color of an object, while another controls its size. If changes in $\mathbf{z}$ lead to entangled or overlapping changes in the output (e.g., altering both color and size simultaneously), the latent space is said to be entangled. Disentanglement is a desirable property because it enables fine-grained control over the generative process and improves the interpretability of the model.

Achieving disentanglement is particularly important in tasks like intuitive physics, where understanding and controlling specific physical properties (e.g., object trajectories or stability) is crucial. By disentangling the latent space, models can better simulate and manipulate physical interactions in a way that aligns with human-like reasoning.

*4) **Intuitive Physics and Latent Space Disentanglement:*** While the existing literature extensively explores intuitive physics, most studies focus on predicting future outcomes of physical interactions. For instance, Rezaei-Shoshtari et al. (2021) [6] utilize a Multi-modal Variational Autoencoder (MVAE) to predict the results of physical interactions, while Piloto et al. (2022) [8] employ a VAE to generate latent codes representing objects and then use Long Short-Term Memory (LSTM) networks to predict the evolution of these object codes over time. These architectures are primarily designed to model concepts such as continuity and to integrate these principles into AI frameworks.

In contrast to these approaches, this project shifts the focus from prediction to generation. Specifically, it aims to explore the generative factors within the latent space to gain control over intuitive physics. Rather than predicting future states, the goal is to generate a range of possible future scenarios across multiple dimensions. This approach allows for a more flexible and creative exploration of physical interactions, enabling the model to simulate diverse outcomes based on controlled manipulations of the latent space.

To achieve this, I designed a Beta-Variational Autoencoder (Beta-VAE) and trained it on a tower-blocks dataset. The Beta-VAE architecture was chosen for its ability to promote disentanglement in the latent space, which is crucial for

isolating and controlling specific generative factors related to intuitive physics. The results of this experiment, including insights into the model's ability to generate physically plausible scenarios, will be elaborated in the subsequent sections.

This work represents a novel contribution to the field by emphasizing control and generation over prediction, opening new avenues for exploring intuitive physics in AI systems. By leveraging disentangled latent spaces, this project aims to bridge the gap between human-like physical reasoning and machine intelligence.

III. Dataset

The dataset used in this study (Fig. 1) is derived from Lerer et al. (2016) [7]. The dataset consists of 493 video recordings of wooden cubes stacked in various configurations, captured at 60 frames per second. The cubes were spray-painted in four distinct colors: red, green, blue, and yellow. Manufacturing imperfections introduced a degree of randomness to the stability of the stacked blocks, ensuring variability in the physical properties of the configurations.

The dataset includes towers of 2, 3, and 4 blocks, with 115, 139, and 239 examples respectively. Each configuration was manually stacked against a white bedsheet, and a tripod-mounted DSLR camera was used to film the blocks falling after a white pole, initially held against the top block, was rapidly lifted upward. This procedure was applied uniformly to both stable and unstable configurations to avoid bias. The motion of the blocks became noticeable only after the pole was several inches away from the top block, ensuring consistent recording conditions.

The dataset is balanced between stable and unstable configurations, making it suitable for studying physical stability and intuitive physics. It provides a rich resource for analyzing the dynamics of stacked objects and their collapse, as well as for training and evaluating models that simulate or predict physical interactions.

The original images were concatenated, with each sequence containing between 50 to 150 images. However, for the purposes of this study, the images were discretized to make them compatible with the Beta-VAE architecture. This preprocessing step ensures that the dataset can be effectively used for latent space disentanglement and generative modeling tasks.

IV. Model

The proposed model is a Beta-Variational Autoencoder (Beta-VAE), a deep generative model designed to learn disentangled representations in the latent space. The architecture consists of an encoder, a decoder, and a bottleneck with a reparameterization mechanism, which together enable the model to map high-dimensional input

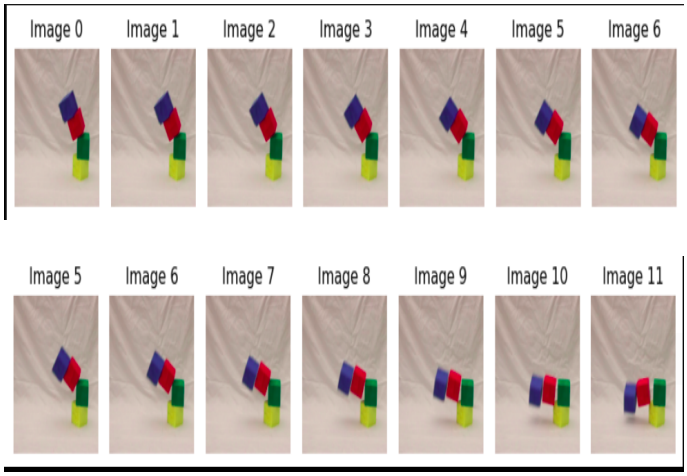

Fig. 1: Block Towers Dataset (Lerer et al., 2016)

data into a lower-dimensional latent space and reconstruct it effectively.

Model's architecture is below (Fig. 2).

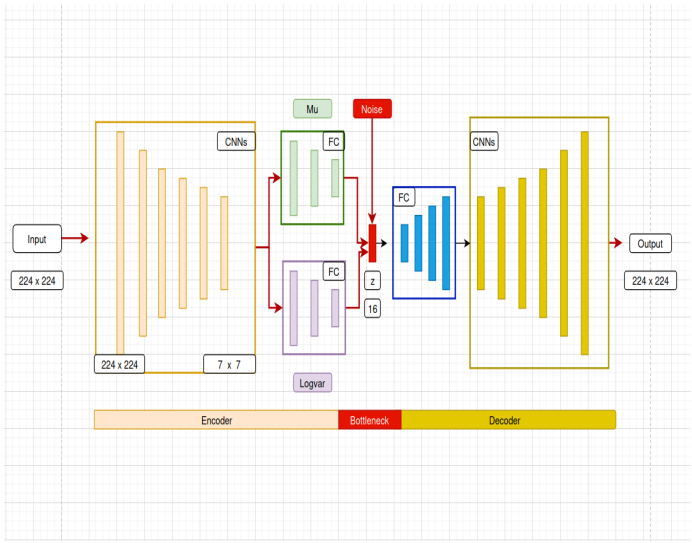

Fig. 2: Model's architecture

Below, I describe the key components of the model.

A. *Encoder*

The encoder is responsible for compressing the input data into a latent space representation. It consists of a convolutional neural network (CNN) followed by fully connected (FC) layers:

- **CNN Encoder:** The CNN encoder comprises six convolutional blocks, each consisting of convolutional layers, batch normalization, and LeakyReLU activation functions. Max-pooling layers are used to progressively reduce the spatial dimensions of the input. The final block flattens the output into a 1D vector.

- **Fully Connected Layers (Mu and LogVar):** The flattened output is passed through two separate FC networks to compute the mean (μ) and log-variance of the latent space distribution. Each FC network consists of three hidden layers with batch normalization, LeakyReLU activation, and dropout for regularization. The final layer outputs a vector of size 16.

B. *Latent Space and Reparameterization*

The latent space is modeled as a Gaussian distribution with mean *mu* and log-variance *logvar*. The reparameterization trick is used to sample from this distribution:

$$z = \mu + \epsilon \times \sigma, \quad where \quad \sigma = \exp(0.5 \times logvar)$$
$$and \quad \epsilon \sim \mathcal{N}(0, 1).$$

This allows for differentiable sampling during training.

C. *Decoder:*

The decoder reconstructs the input data from the latent space representation. It consists of an FC network followed by a transposed convolutional network, transposed CNN decoder.

- **FC Decoder** The FC decoder maps the latent vector **z** back to a higher-dimensional space. It consists of three hidden layers with batch normalization, LeakyReLU activation, and dropout. The fourth layer outputs a tensor of size : first_dim x 128 x 7 x 7.
- **Transposed CNN Decoder** The transposed CNN decoder upsamples the tensor back to the original input dimensions (from 7x7 to 224x224). It consists of five transposed convolutional blocks, each using a 4x4 kernel and stride 2 for upsampling. Batch normalization and LeakyReLU activation are applied after each block. The final (sixth) layer uses a 5x5 kernel to produce the reconstructed output with 3 channels.

D. *Loss Function*

The model is trained using a combination of reconstruction loss and KL divergence loss:

$$L = Reconstruction\ Loss + \beta \times KL\ Divergence$$

where the reconstruction loss is the Mean Squared Error (MSE) and the KL divergence term encourages the latent space to approximate a standard Gaussian distribution:

$$KL = -0.5 \cdot \sum (1 + logvar - \mu^2 - exp(logvar))$$

The hyperparameter β controls the trade-off between reconstruction accuracy and latent space disentanglement.

E. *Weight Initialization*

All weights in the model are initialized using Kaiming initialization with a LeakyReLU nonlinearity. This ensures stable and efficient training.

F. *Optimization*

The Adam optimizer is used with an initial learning rate of 0.001. A learning rate scheduler (with $\gamma = 0.1$) is employed to progressively reduce the learning rate every 20 epochs.

G. *Key Features*

- **Disentanglement:** The Beta-VAE promotes disentangled representations in the latent space, enabling control over individual generative factors.
- **Stochasticity:** The reparameterization trick introduces stochasticity, allowing the model to generate diverse outputs.
- **Regularization:** Dropout and batch normalization are used throughout the network to prevent overfitting and improve generalization.

H. *Implementation Details*

The model is implemented using PyTorch. The encoder and decoder are modular, allowing for easy customization of the architecture. The model is trained on a dataset of stacked block configurations, with the goal of learning disentangled representations of physical stability and object interactions.

V. RESULTS

Following plots are reconstruction and KL losses (Fig. 3):

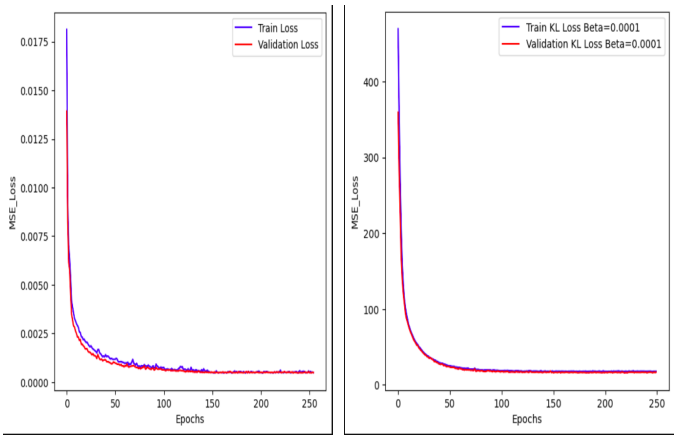

Fig. 3: Reconstruction and KL losses: The KL loss weight, β, was initialized to 0.0001 and gradually increased to 50 during training

This plot (Fig. 4) illustrates the original images (top) and reconstructed images (bottom):

When an image is passed through the encoder part of the model, it is mapped to a latent vector **z** in the latent space. To analyze the disentanglement of the latent space, we perform *traversals* by systematically varying one dimension of **z** while keeping the other dimensions fixed. These perturbed latent vectors are then passed through the decoder, which generates new images. The resulting images reflect the influence of the

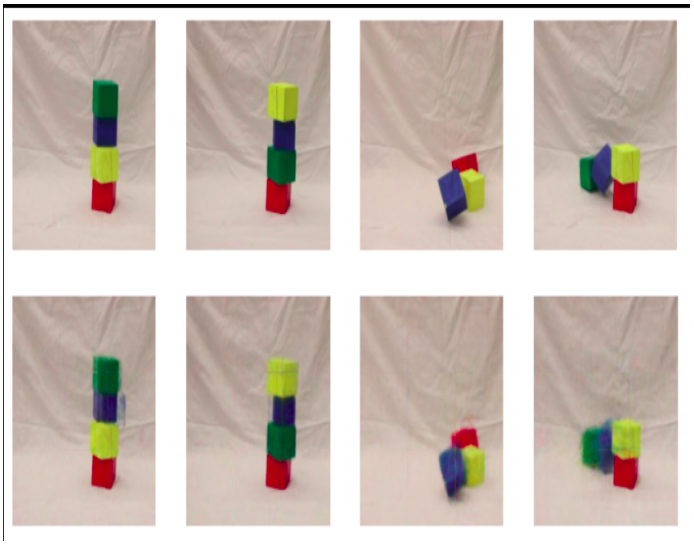

Fig. 4: Original (top) versus Reconstructed (bottom) Images

specific latent dimension being varied, demonstrating how changes in that dimension correspond to meaningful and interpretable transformations in the output.

The following plots (Fig. 5) illustrate the latent space disentanglement by showing the effects of traversals on the generated images when perturbing the latent vector $\mathbf{z}$ obtained from the encoder.

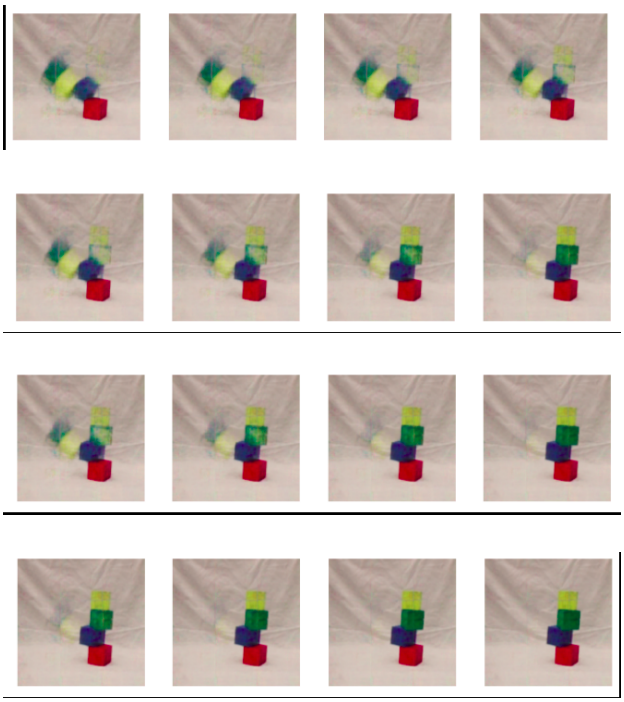

Fig. 5: Traversals

VI. Conclusions

From the experimental results and plots, it is evident that the Beta-VAE model struggles to learn latent representations that effectively capture the generative factors underlying intuitive physics. Despite extensive efforts to improve the model's performance—such as generating over one million synthetic images of falling block-towers and replacing the CNN decoder with a Vision Transformer network—the results did not show significant improvement. Additionally, experiments with Vector-Quantized VAEs (VQ-VAEs) yielded even poorer results, likely due to the discrete nature of their latent space, which is ill-suited for modeling the continuous dynamics inherent in physical systems.

Further attempts to integrate physics-based constraints into the VAE framework also did not produce satisfactory outcomes. These challenges suggest that the fundamental trade-off between reconstruction accuracy and KL divergence in VAEs limits their ability to explore and model intuitive physics effectively. While such models may excel in tasks like image generation, they fall short in capturing the continuous and causal nature of physical phenomena.

In light of these findings, it appears that physics-integrated or physics-guided models hold greater promise for studying intuitive physics. By explicitly incorporating physical principles into the learning process, these models may overcome the limitations of traditional VAEs and provide a more robust framework for understanding and simulating physical interactions. Future work should focus on developing hybrid approaches that combine the strengths of deep generative models with domain-specific knowledge of physics, enabling more accurate and interpretable representations of intuitive physics.

REFERENCES

[1] Brenden M. Lake, Ruslan Salakhutdinov, and Joshua B. Tenenbaum (2015). "Human-level concept learning through probabilistic program induction." *Science*, 350(6266), 1332-1338. DOI: 10.1126/science.aab3050. Available at: https://www.science.org/doi/abs/10.1126/science.aab3050.

[2] Brenden M. Lake, Ruslan Salakhutdinov, Jason Gross, and Joshua B. Tenenbaum (2011). "One shot learning of simple visual concepts." *Proceedings of the Annual Meeting of the Cognitive Science Society*, 33. Available at: https://escholarship.org/uc/item/4ht821jx.

[3] Peter W. Battaglia, Jessica B. Hamrick, and Joshua B. Tenenbaum (2013). "Simulation as an engine of physical scene understanding." *Proceedings of the National Academy of Sciences*, 110(45), 18327-18332. DOI: 10.1073/pnas.1306572110. Available at: https://www.pnas.org/doi/abs/10.1073/pnas.1306572110.

[4] Brenden M. Lake, Tomer D. Ullman, Joshua B. Tenenbaum, and Samuel J. Gershman (2017). "Building Machines That Learn and Think Like People." *Behavioral and Brain Sciences*, 40. DOI: 10.1017/s0140525x16001837.

[5] Netanyahu, A., Shu, T., Katz, B., Barbu, A., & Tenenbaum, J.B. (2021). PHASE: PHysically-grounded Abstract Social Events for Machine Social Perception. *ArXiv*, https://doi.org/10.1609/aaai.v35i1.16167.

[6] Rezaei-Shoshtari, S., Hogan, F. R., Jenkin, M., Meger, D., & Dudek, G. (2021). Learning intuitive physics with multimodal generative models. *arXiv* preprint arXiv:2101.04454. https://arxiv.org/abs/2101.04454

[7] Lerer, A., Gross, S., & Fergus, R. (2016). Learning physical intuition of block towers by example. *In Proceedings of the 33rd International Conference on Machine Learning* (Vol. 48, pp. 430–438). https://dl.acm.org/doi/10.5555/3045390.3045437

[8] Piloto, L.S., Weinstein, A., Battaglia, P. et al. Intuitive physics learning in a deep-learning model inspired by developmental psychology. Nat Hum Behav 6, 1257–1267 (2022). https://doi.org/10.1038/s41562-022-01394-8