Proceedings of CONF-APMM 2026 Symposium: Applied Mathematics and Quantitative Modeling
DOI: 10.54254/2753-8818/2026.PU36720

Latent Space Representation Learning Based on Variational Autoencoders

Xiaozhou Gao

New Channel Guangzhou School, Guangzhou, China
gxiaozhou@outlook.com

Abstract. World models constitute an important research direction in artificial intelligence, aiming to enable machines to perceive, understand, and predict the external environment in a manner similar to humans. Representation learning in latent space, serving as the foundation of the perception layer in world models, determines the model's capacity for compressing and understanding environmental states. The Variational Autoencoder (VAE) is a deep generative model that achieves effective mapping from high-dimensional observational data to low-dimensional latent states through variational inference. This paper first introduces the fundamental concepts of latent space and the working principles of autoencoders, and then elaborates on the evolution from traditional autoencoders to variational autoencoders. Subsequently, it analyzes the core mathematical framework of the VAE, including variational inference, the evidence lower bound, and the reparameterization trick. Finally, it discusses the application value and existing limitations of this model in the perception layer of world models. The study demonstrates that the VAE possesses significant advantages in state compression and uncertainty modeling, but suffers from issues such as blurry image generation and oversimplified posterior distribution assumptions. Future research may focus on the integration of diffusion models with variational autoencoders and causal representation learning.

Keywords: Variational Autoencoder, Latent Space, Representation Learning, World Model

1. Introduction

The concept of world models was first proposed by Ha and Schmidhuber in 2018, whose core idea is to construct an internal model that simulates the operating principles of the external environment, thereby enabling agents to learn and make decisions through internal "imagination" [1]. The three-layer mathematical structure of world models comprises the perception representation layer, the prediction memory layer, and the decision control layer, among which the perception representation layer is responsible for compressing high-dimensional sensory inputs into low-dimensional latent states, serving as the foundation of the entire architecture [2].

In the perception representation layer, how to effectively compress high-dimensional data such as images and videos into compact and meaningful low-dimensional representations is a key problem. As a probabilistic generative model, the variational autoencoder introduces the variational inference framework to transform the encoding process from deterministic mapping to probabilistic

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2753-8818/2026.PU36720

distribution modeling, thereby enabling the quantification of representation uncertainty [3]. Compared with traditional autoencoders, the Variational Autoencoder (VAE) can not only reconstruct input data but also generate new plausible samples, which gives it unique advantages in constructing the latent space of world models.

This paper aims to systematically expound the mathematical principles and applications of variational autoencoders in latent space representation learning. By analyzing the core mathematical framework, it reveals the value and limitations of VAEs in world models, providing a reference for subsequent research.

2. Latent space and autoencoders

2.1. Fundamental concepts of latent space

Latent space refers to the low-dimensional abstract representation space obtained by compressing high-dimensional observational data through some mapping. Taking a color image with a resolution of 224×224 as an example, its original data dimensionality is 224×224×3 = 150,528, whereas the latent space may only require tens to hundreds of dimensions to effectively express the core semantic information of the image [4]. This dimensionality reduction not only decreases computational cost but, more importantly, extracts the essential structural features of the data.

Mathematically, the latent space can be defined as a continuous space $S \subseteq R^d$, where d is much smaller than the original observation dimensionality. The encoder function φ maps the observation o to the latent state $s = s_\varphi(o)$, and the decoder function θ reconstructs the latent state back to the observation $\hat{o} = o_\theta(s)$. The entire encoding-decoding process can be expressed as:

$$\hat{o} = o_\theta(s_\varphi(o)) \tag{1}$$

where o is the input observation and ô is the reconstructed output. Ideally, ô should be as close to o as possible.

2.2. Structure and limitations of autoencoders

An autoencoder consists of two components: an encoder and a decoder. The encoder compresses the input data into a latent representation, while the decoder reconstructs the input from the latent representation. The training objective is to minimize the reconstruction error:

$$L_{AE} = \|o - o_\theta(s_\varphi(o))\|^2 \tag{2}$$

Although traditional autoencoders can effectively compress data, they have notable limitations. First, their latent space is an irregular, unstructured space, making it impossible to guarantee that randomly sampled points in the latent space will generate meaningful outputs. Second, autoencoders only learn deterministic mappings and cannot express data uncertainty. These limitations motivated the proposal of variational autoencoders.

DOI: 10.54254/2753-8818/2026.PU36720

3. Mathematical framework of variational autoencoders

3.1. Variational inference and the evidence lower bound

The core innovation of the variational autoencoder lies in transforming the encoding process from deterministic mapping to probabilistic distribution modeling. Specifically, the encoder no longer outputs a single deterministic latent state but instead outputs a probability distribution $q_\varphi(s|o)$, typically assumed to be a Gaussian distribution:

$$q_\varphi(s|o) = N(\mu_\varphi(o), \sigma_\varphi(o)^2) \tag{3}$$

where $\mu_\varphi(o)$ and $\sigma_\varphi(o)$ denote the mean and standard deviation of the distribution, respectively, parameterized by a neural network.

The training objective of the VAE is to maximize the log-likelihood of the observed data, $\log p(o)$. Since directly computing this quantity requires integrating over all possible latent states, which is intractable in practice, variational inference is introduced to approximate it through the Evidence Lower Bound (ELBO):

$$\log p(o) \geq E_{q(s|o)}[\log p(o|s)] - D_{KL}(q(s|o) \| p(s)) \tag{4}$$

The right-hand side of Equation (4) consists of two terms. The first term, $E_{q(s|o)}[\log p(o|s)]$, is the reconstruction term, measuring the decoder's ability to reconstruct observations from latent states. The second term, D_{KL}, is the Kullback-Leibler divergence, measuring the difference between the posterior distribution $q(s|o)$ output by the encoder and the prior distribution $p(s)$, which is typically assumed to be a standard normal distribution $N(0, I)$. The KL divergence is computed as:

$$D_{KL} = -\tfrac{1}{2} \Sigma (1 + \log(\sigma^2) - \mu^2 - \sigma^2) \tag{5}$$

By maximizing the ELBO, the model simultaneously optimizes the data reconstruction capability and the structural regularity of the latent space.

3.2. Reparameterization trick

Sampling directly from the distribution $q_\varphi(s|o)$ would prevent gradients from backpropagating through the stochastic node, thereby hindering end-to-end training. To address this issue, Kingma and Welling proposed the reparameterization trick [3]. This trick decomposes the stochastic sampling process into a deterministic transformation and an independent noise component:

$$s = \mu_\varphi(o) + \sigma_\varphi(o) \odot \varepsilon, \quad \varepsilon \sim N(0, I) \tag{6}$$

where $\odot$ denotes the element-wise (Hadamard) product and ε is a noise vector sampled from the standard normal distribution. Through this transformation, gradients can propagate normally via $\mu_\varphi(o)$ and $\sigma_\varphi(o)$, enabling the entire network to be trained end-to-end using standard stochastic gradient descent.

DOI: 10.54254/2753-8818/2026.PU36720

4. Applications and limitations of variational autoencoders in world models

4.1. Applications of VAEs in the perception layer of world models

In the V-M-C three-layer mathematical structure of world models, the variational autoencoder serves the core function of the perception representation layer [1]. In the original world model framework, Ha and Schmidhuber used a VAE to compress high-dimensional image observations into low-dimensional latent vectors, which were then fed into a recurrent neural network for dynamic prediction [1]. This design significantly reduced the computational complexity of the subsequent prediction and decision layers.

Building upon the VAE-based perception framework, Hafner et al. introduced the Recurrent State-Space Model (RSSM) in PlaNet, which separates deterministic recurrent hidden states from stochastic latent variables, enabling world models to simultaneously model the deterministic and stochastic aspects of the environment [5]. The Dreamer series inherited and refined this architecture for more effective latent imagination and policy learning [6, 7]. The capabilities of the VAE in state compression, uncertainty quantification, and latent space structuring have made it a standard component of the perception layer in world models.

4.2. Limitations and improvement directions

Despite the widespread application of variational autoencoders in world models, several limitations remain. First, due to the regularization effect of the KL divergence term, VAEs tend to produce blurry reconstructions, which perform poorly in scenarios requiring fine visual details [3, 8]. Second, the assumption of a Gaussian posterior distribution is overly simplistic, making it difficult to capture the multimodal characteristics of real data distributions.

To address these limitations, subsequent studies have proposed various improvement directions. Diffusion models generate high-quality samples through an iterative denoising process and can serve as a replacement for the VAE decoder [9, 10]. The Joint Embedding Predictive Architecture (JEPA) abandons pixel-level reconstruction and directly models predictive relationships in an abstract semantic space, providing a new paradigm for representation learning [11]. Causal representation learning attempts to move from statistical correlation toward causal understanding, enhancing model interpretability [12].

5. Conclusion

This paper systematically expounded on the application of latent space representation learning based on variational autoencoders in world models. Starting from the fundamental concepts of latent space, it analyzed the structure and limitations of traditional autoencoders, revealing their fundamental deficiencies in probabilistic modeling and structured latent space. It then elaborated on the core mathematical framework of the VAE, demonstrating that the combination of variational inference, the evidence lower bound, and the reparameterization trick can achieve effective mapping from high-dimensional observations to low-dimensional latent states while quantifying representation uncertainty.

The study shows that the VAE plays a critical role in the perception layer of world models by providing compressed, structured, and uncertainty-aware state representations. However, the trade-off between reconstruction quality and latent space regularization remains a persistent challenge. Regarding room for improvement, this paper did not delve into the fusion mechanisms between

DOI: 10.54254/2753-8818/2026.PU36720

diffusion models and VAEs, nor did it address the concrete implementation of causal representation learning in multimodal environments. Future research may focus on the following directions: designing hybrid architectures of diffusion models and VAEs to improve generation quality, modeling non-Gaussian posterior distributions to capture multimodal data characteristics, and developing interpretable representation learning frameworks that move beyond statistical correlation toward genuine causal understanding.

References

[1] Ha, D., & Schmidhuber, J. (2018). Recurrent world models facilitate policy evolution. *Advances in Neural Information Processing Systems*. https://papers.nips.cc/paper/7512-recurrent-world-models-facilitate-policy-evolution

[2] Wang, J., Cui, Y., & Zhang, Y. (2024). Overview of world models. *Journal of Zhengzhou University (Natural Science Edition), 56*(5), 1-12. DOI: 10.13705/j.issn.1671-6841.2024078

[3] Kingma, D. P., & Welling, M. (2013). Auto-encoding variational Bayes. *arXiv preprint arXiv:1312.6114*.

[4] Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press. Cambridge.

[5] Hafner, D., Lillicrap, T., Fischer, I., et al. (2019). Learning latent dynamics for planning from pixels. *International Conference on Machine Learning*. http://proceedings.mlr.press/v97/hafner19a.html

[6] Hafner, D., Lillicrap, J., Ba, J., & Norouzi, M. (2019). Dream to control: Learning behaviors by latent imagination. *arXiv preprint arXiv:1912.01603*.

[7] Hafner, D., Pasukonis, J., Ba, J., & Lillicrap, J. (2023). Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*.

[8] Higgins, I., Matthey, L., Pal, A., et al. (2017). beta-VAE: Learning basic visual concepts with a constrained variational framework. *International Conference on Learning Representations*.

[9] Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*. https://proceedings.neurips.cc/paper_files/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf

[10] Rombach, R., Blattmann, A., Lorenz, D., et al. (2022). High-resolution image synthesis with latent diffusion models. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10684-10695.

[11] LeCun, Y. (2022). A path towards autonomous machine intelligence. *OpenReview*. https://openreview.net/pdf?id=BZ5a1r-kVsf

[12] Schölkopf, B., Locatello, F., Bauer, S., et al. (2021). Toward causal representation learning. *Proceedings of the IEEE, 109*(5), 612-634.