World Model Learning using Graphical Models

Praveen Venkatesh
pvenkat2

Vibhakar Mohta
vmohta

Abstract

Model-based RL, such as DreamerV3, have shown great promise in recent times for solving real-world tasks. Part of learning a good control policy in MBRL requires learning a strong world model. A common approach to world model learning is learning a latent transition model that predicts the next latent state given the current latent state and actions. But oftentimes, we would like our world models to be interpretable, that is, have the capability to generate open-loop world image reconstructions, which is not guaranteed without optimizing for the reconstruction objective. Our work explores the use of diffusion models (DDPMs) as latent transition models that operate on latent representations of a frozen variational autoencoder to guarantee reconstructability. We show that our diffusion world model is capable of learning accurate latent transition dynamics over latent spaces optimized for reconstructability. We compare our model with graph neural networks and fully connected network baselines and show that the expressivity of diffusion models allows for learning complex transition functions. Code for our project can be found at: https://github.com/vib2810/diffusion_world_models

1 Introduction

Model-based reinforcement learning requires learning a strong world model in order to learn an optimal control policy. Learning accurate world models has been explored extensively in the literature and is an active field of research. Our work aims to explore learning world models capable of reconstructing the world image. This is an extremely hard task, as it involves learning a transition function over latent representations that are optimized for reconstructability instead of ease of transition learning. Moreover, such transition models should be capable of modeling multimodal distributions conditioned on action labels for extensibility to complex stochastic environments, which often have multimodal transition dynamics.

2 Background and Literature Review

Literature on world model learning is divided into two major approaches: one that focuses on learning accurate world models and then planning for an optimal policy, or other approaches that focus on jointly learning world and reward models, in essence, learning a world model that gives the maximum reward. The second kind of approach often combines other techniques for collecting data in a "smart" way so that we are able to learn better world models. Examples of this include using curiosity-based exploration [1] [7], policy model learning [6] [3], and planning for disagreement [14].

Model learning is a complex problem as, at any timestep, you might not have sufficient information in the observations to predict the next state given the action. Early work on world model learning proposed compressing the observations into a latent state that encodes information about the global world state [2]. Various methods have been explored in the literature to learn the latent state; some of them include random CNN features, variational autoencoders, and inverse dynamics model features [1]. The encoder is either held fixed or trained in a self-supervised way, and the latent space transition model is then learned using the (s, a, s') labeled transition data.

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

Instead of modeling the transition model with an MLP, [8] proposed using a graph neural network as a transition model. Such an architecture adds inductive biases like node interaction and permutation invariance to the architecture, which are good to have in low-data regimes. The authors also propose training the model using a contrastive loss when the negative samples are chosen from the previous state. Contrastive loss has been extensively explored in literature for tasks like self-supervised learning like MoCo [4], and CLIP [12]. They have also been used to enforce learning a good latent space in model learning [13] [9].

This project aims to build world models that are capable of reconstructing visual representations. We aim to do this by first training a variational autoencoder to learn latent representations suitable for image reconstruction. Next, to model the latent transition dynamics, we propose the use of the highly expressive denoising diffusion probabilistic models (DDPM), which are capable of learning multimodal conditional distributions. These models have been shown to be extremely stable in training and robust to hyperparameters. The contributions of our work are two-fold: 1) Showcase the use of diffusion models as latent dynamics models, and 2) Demonstrate the use of VAEs coupled with latent diffusion models to generate interpretable open loop world model predictions.

3 Methods

3.1 Contrastive World Model Learning Formulation

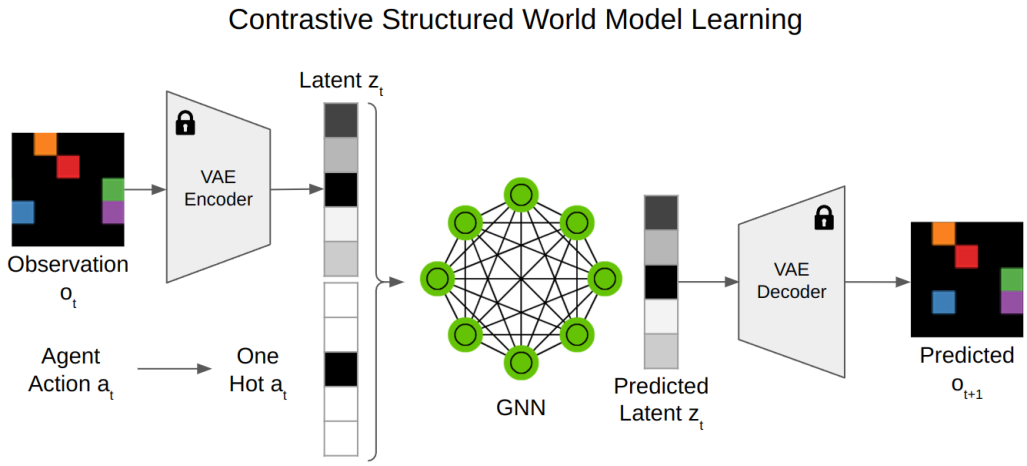

Figure 1: Diffusion World Models

In this project, we are specifically looking at building world models and exploring the use of PGM-based approaches to world modeling. Here, we build on an approach presented in [8] to utilize a GNN for modeling transitions atop latent representations learned by deep networks. We have implemented two baselines:

- Contrastive Structured World Model: Here, we use a CNN-based observation encoder, a GNN-based transition model, and a contrastive loss function to optimize for predicting the residual embedding required to transition to the next states from the current state.
- Auto Encoder: Here, instead of using a contrastive loss function, we utilize pixel-based reconstruction loss to learn both a latent space embedding as well as the transition model.

In both cases, a GNN-based transition model is used to predict the next state from the given current state and action vector.

For the cSWM model, a contrastive loss is applied in an object-factorized manner (explicitly defining nodes for each object). As shown in the paper, this explicit structure allows the model to perform significantly better than regular auto-encoders.

3.2 Diffusion World Models

3.2.1 Formulation

We formulate world model learning as a Denoising Diffusion Probabilistic Model (DDPM) [5]. We utilize a diffusion model as they have shown the ability to express complex multimodal action distributions and possess stable training behavior while requiring significantly low task-specific hyperparameter tuning.

We use a DDPM to approximate the conditional distribution $p(o_{t+1}|o_t, a_t)$. Given the previous observation and action, the model predicts the distribution of the next observation. Instead of operating on the image space directly, our diffusion model operates on a latent representation of the world state, learned by a variational autoencoder. A well-trained VAE guarantees these representations to be expressive enough for the visual world state to be reconstructable. Therefore, our models are defined as

$$\text{Encoder (VAE): } p(z_t|o_t) \tag{1}$$
$$\text{Transition Model (DDPM): } p(z_{t+1}|z_t, a_t) \tag{2}$$
$$\text{Decoder (VAE): } p(o_t|z_t) \tag{3}$$

Starting from z^k sampled from random Gaussian noise, DDPM performs T steps of denoising to produce a sequence of intermediate next latent predictions with decreasing levels of noise, $z^k, z^{k-1} \dots z^0$, where z^0 represents the denoised next latent prediction. The process can be represented as:

$$z_{t+1}^{k-1} = \alpha(z_{t+1}^k - \gamma\epsilon_\theta(z_{t+1}^k, z_t, a_t, k) + N(0, \sigma^2 I)) \tag{4}$$

Here α, γ, σ are computed based on a noise schedule, which is a hyperparameter of the model. We use a neural network $\epsilon_\theta(z_{t+1}^k, z_t, a_t, k)$ that predicts the noise added at timestep k given the noisy predicted latent z_{t+1}^k, previous observations z_t and action a_t. The model is trained using the simplified latent diffusion objective

$$\mathcal{L} = \mathcal{MSE}(\epsilon^k, \epsilon_\theta(z_{t+1}^k, z_t, a_t, k)) \tag{5}$$

3.2.2 Architecture

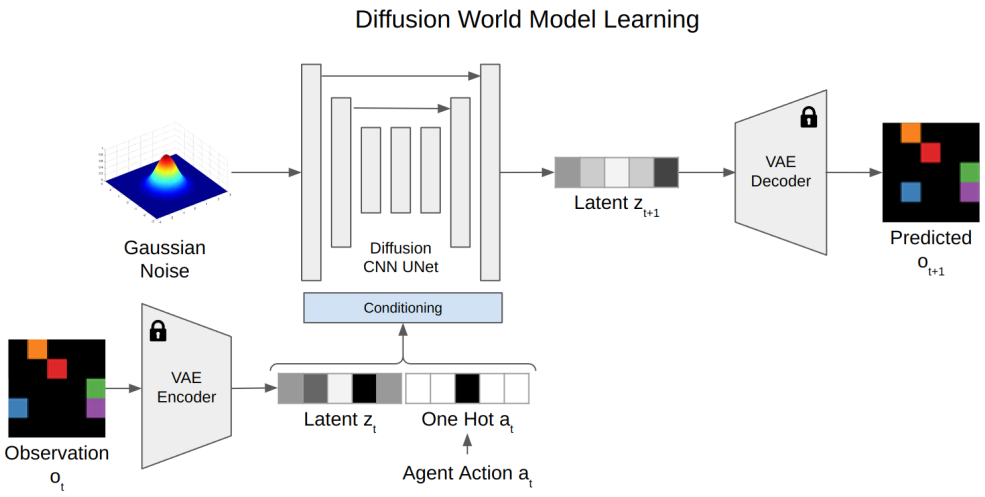

Figure 2: Diffusion World Models

An overview of the architecture is shown in Fig 2. As a first step, we train the variational autoencoder, which learns to compress the observation image into a latent space. The encoder and decoder of the model are frozen while training the diffusion model.

Algorithm 1 Train Diffusion World Model

Require: Training dataset D, number of timesteps T, variance schedule β_t, noise model $\epsilon_\theta(x, t)$

Initialize model parameters θ
for epoch = 1 to num_epochs **do**
  for batch (o_t, a, o_{t+1}) in D **do**
    Get latents: $z_t = Encoder(o_t)$, $z_{t+1} = Encoder(o_{t+1})$
    Sample timestep $t \sim \text{Uniform}(1, T)$ and noise $\epsilon \sim \mathcal{N}(0, I)$
    Get noisy latent: $z^k_{t+1} = \sqrt{1-\beta_t} z_{t+1} + \sqrt{\beta_t}\epsilon$
    Predict noise: $\epsilon_\theta(z^k_{t+1}, z_t, a_t, k)$
    Compute loss: $\mathcal{L} = \|\epsilon - \epsilon_\theta\|_2^2$
    Backpropagate and update θ using optimizer
  end for
end for

The noise prediction network is a fully convolutional UNet with Mish [10] activations. It encodes conditioning into the convolutional block using FiLM modulation that predicts a per-channel scale and bias, as formulated in [11]. We use the *squared cosine* noise schedule to model the DDPM process. The model is optimized using an adaptive momentum estimation with weight decay (AdamW) optimizer, utilizing a cosine learning rate schedule with warm restarts to escape local minima convergence. Finally, to speed up training and make the model more stable, an exponential moving average (EMA) copy of the model weights is maintained and used for inference.

4 Results

We evaluate a vanilla CNN-based Autoencoder, and a Contrastive Structured World Model on Space Invaders, 3 Body Problem, and 2D shapes environments. We tabulate quantitative metrics in table 4. They are described as follows:

- Space Invaders: A popular Atari-based game (SpaceInvadersDeterministic-v4 in OpenAI Gym).
- 3 Body Problem: This is a problem from physics where the task is to predict the next state for 3 planetary bodies in space. This is a challenging problem since it is a chaotic system.
- 2D Shapes: This is a custom environment where pushing a block allows collision between blocks. If an action causes a collision between blocks, the state doesn't change.

We evaluate the Mean Square Errors for Reconstruction and Latent space observation, Hits at Rank 1 (H@1) and the Mean Reciprocal Rank (MRR) for each model in the given environments. MRR and H@1 metrics are evaluated in the latent space as we wish to measure how well the world model is performing. They are defined as follows:

- Hits at Rank 1 - Essentially a substitute for accuracy, where we measure if the next state predicted is correct (within a threshold).
- Mean Reciprocal Rank - Defined as the average inverse rank. $MRR = \frac{1}{N}\sum_1^N \frac{1}{rank_n}$, where $rank_n$ is the rank of the nth sample.
- MSE Reconstruction - Evaluates how well the end-to-end model is able to reconstruct the visual observations.
- MSE Latent - Evaluates the MSE between the encoded observations in the latent space.

While we have performed baseline analysis on 3 environments as described above, we restrict ourselves to a simple 2D block-pushing environment for our proposed model for consistency across our baselines. We tabulate our results in table 4. As seen in the result section, our diffusion model performs almost 20x better than the autoencoder with the GNN transition model. This is also reflected in the almost pixel-perfect reconstructions observed in fig 3. Thus, we can see that diffusion models can indeed capture the transition dynamics of an environment, and more effectively than an autoencoder even with strong underlying structural biases. We also notice that the cSWM is able

Table 1: Preliminary Results

Environment	Models	H@1	MRR
Space Invaders	Contrastive Structured World Model	0.54	0.72
	Auto Encoder	0.01	0.07
3 Body Problem	Contrastive Structured World Model	1	1
	Auto Encoder	0.671	0.8
2D Shapes	Contrastive Structured World Model	0.9995	0.99975
	Auto Encoder	0.5319	0.57431

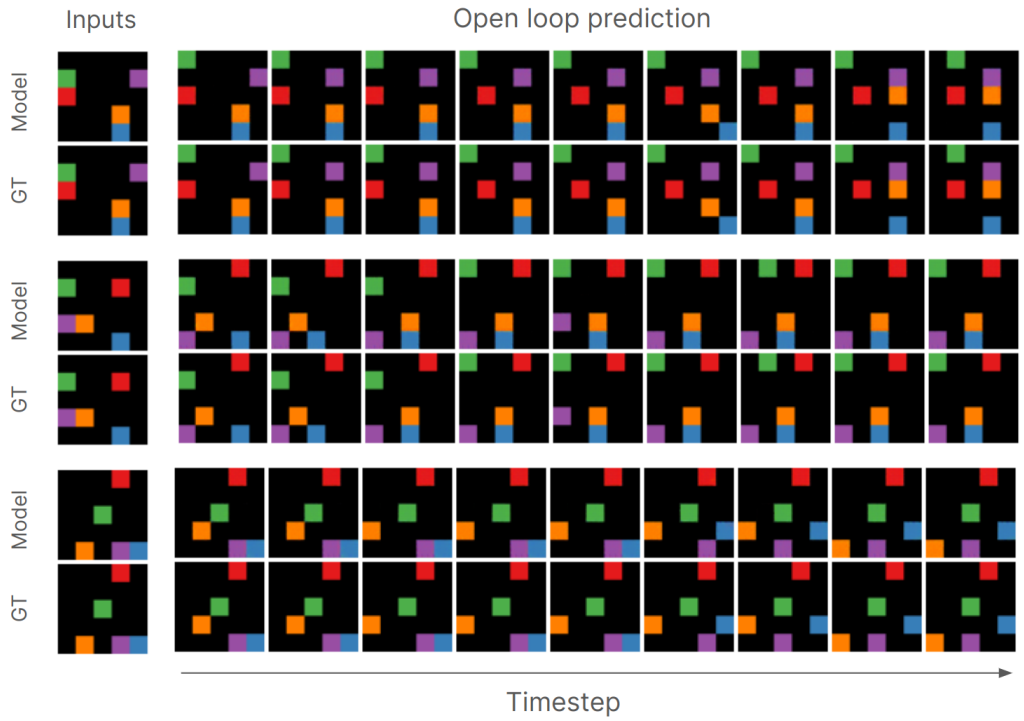

Figure 3: Successful open loop predictions using our diffusion model. We sample actions from the dataset and perform multiple successive open-loop transitions using latent diffusion. The latent vector is reconstructed using the frozen decoder from the VAE.

to reconstruct the latent spaces very effectively (in fact, more effectively than our LDM). However, we believe that since there is no explicit objective to allow the latent space of the cSWM to be reconstructible, the learned latent representation does not fully capture the model of the environment. We believe that further experiments are required to validate this hypothesis.

Table 2: Final Results on 2D shapes

Model	H@1	MRR	MSE (Reconstruction)	MSE (Latent)
Contrastive Structured World Model	0.9995	0.99975	NA	1.57E-07
Auto Encoder + GNN Transition	0.4848	0.5237	0.0001135	0.0433
Latent Diffusion Model (Ours)	0.346	0.5941	0.000136	0.00205

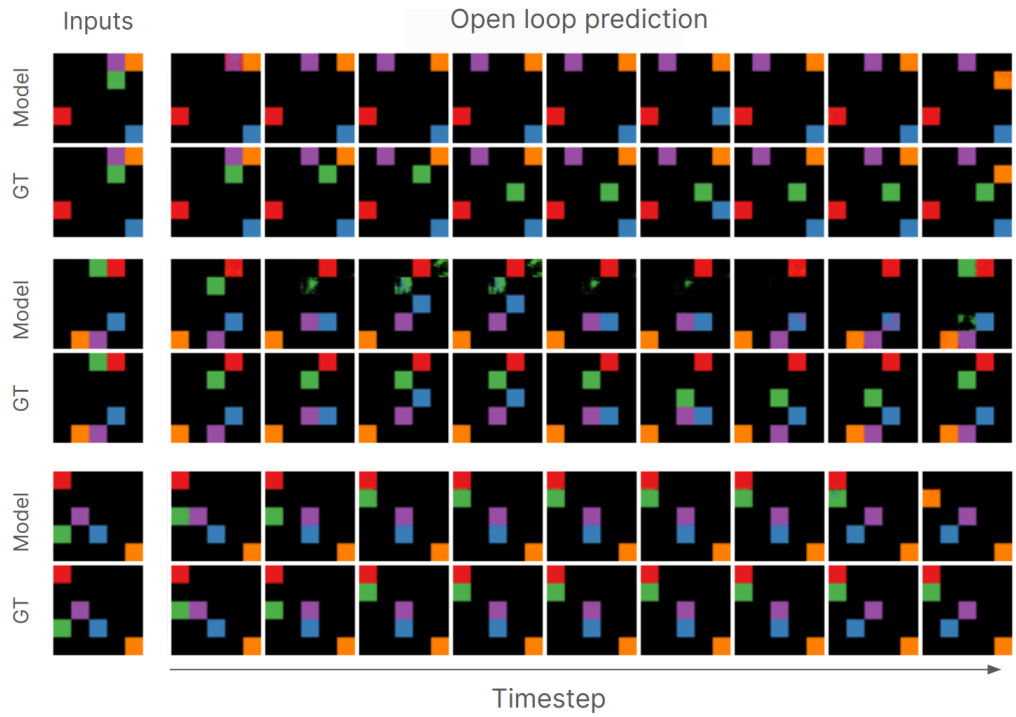

Figure 4: Failure modes of our diffusion model. Row 1 - we see that the green square disappears in step 2, and the successive predictions do nothing. Row 2 - The green square is not reconstructed accurately. Row 3 - At the last time step, the green square changes color to orange, and the red square disappears.

5 Discussion and Analysis

As seen in the table (tab 4), we can see that the autoencoder with GNN optimized on pixel-based reconstruction performs worse compared to the cSWM trained with a contrastive loss. We can thus infer that the underlying latent representation learned is insufficient to capture the actual nature of the task at hand.

Some qualitative examples of open-loop predictions upon sampling a random observation and performing actions are shown in 3. Here, we observe that even over long horizons, the transition model is able to allow for accurate reconstruction of the world from a low-dimensional latent vector. While unable to fit in the extent of this report, we performed experiments that showed consistency even over 100 timesteps.

We also note some failure cases in fig 4. Here, we can see that the LDM is occasionally unable to reconstruct the latent space accurately. Common failures include the sudden disappearance of certain blocks, or blocks changing colour. While non-sharp reconstructions are rare, one instance of the same in also documented in 4. We believe that further training could help ameliorate this issue, as our diffusion models are very expressive and require large amounts of data and time to train well.

6 Learnings and Failed Experiments

While experimenting with diffusion models as latent transition models, we had to integrate a bunch of tricks, without which the model failed to produce good results. The first trick was normalizing the VAE latent, as the diffusion sampling process is generally clipped between -1 and 1 for stability, and the latent representations are not guaranteed to be within this range. Another trick that we used to improve model convergence was FiLM conditioning, which infuses the conditioning through per-channel scaling and biasing, improving the overall learning process. The use of one hot action

labels was extremely crucial as well, as we found that using class labels directly generally leads to worse performance. We also explored using the diffusion model to predict observation deltas, i.e., $p(z_{t+1} - z_t | z_t, a_t)$ and found that both formulations lead to similar convergence and performance.

Furthermore, we experimented with freezing a trained VAE, and training just the GNN-based transition model using an MSE reconstruction objective. This did not yield successful results, and we noticed diverging latent space losses while noticing improved reconstruction loss. We believe that since the objective is not directly optimizing over latent space predictions, this indirect objective leads to ineffective optimization.

7 Limitations and Future Work

While learning a transition model atop VAE latents is a feasible solution towards learning reconstruction-capable world models, it might at many times be extremely complicated, as the latent representations are not guaranteed to be transition learning friendly. This remains a major limitation of our work and remains a future direction of exploration. Unlocking the VAE encoder and decoder and jointly optimizing for both objectives seems to be a plausible direction for future work. Another direction would be to enforce reconstructabilty on latents learned by contrastive methods. This could be done by adding a reconstruction loss with the contrastive objective.

Further work can be carried out to understand how these models perform in varying data regimes. In our models, we have used a lot of data sampled randomly from the environment available to us. In real-world scenarios, such an abundance of data may not be available.

8 Teammates and Work Division

Vibhakar Mohta focussed on integrating latent diffusion models with VAEs. Praveen Venkatesh focused on setting up the infrastructure to develop and test our baselines and running and managing experiments.

References

[1] Yuri Burda, Harri Edwards, Deepak Pathak, Amos Storkey, Trevor Darrell, and Alexei A Efros. Large-scale study of curiosity-driven learning. *arXiv preprint arXiv:1808.04355*, 2018.

[2] David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018.

[3] Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. *arXiv preprint arXiv:1912.01603*, 2019.

[4] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

[5] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[6] Lukasz Kaiser, Mohammad Babaeizadeh, Piotr Milos, Blazej Osinski, Roy H Campbell, Konrad Czechowski, Dumitru Erhan, Chelsea Finn, Piotr Kozakowski, Sergey Levine, et al. Model-based reinforcement learning for atari. *arXiv preprint arXiv:1903.00374*, 2019.

[7] Kuno Kim, Megumi Sano, Julian De Freitas, Nick Haber, and Daniel Yamins. Active world model learning with progress curiosity. In *International conference on machine learning*, pages 5306–5315. PMLR, 2020.

[8] Thomas Kipf, Elise Van der Pol, and Max Welling. Contrastive learning of structured world models. *arXiv preprint arXiv:1911.12247*, 2019.

[9] Michael Laskin, Aravind Srinivas, and Pieter Abbeel. Curl: Contrastive unsupervised representations for reinforcement learning. In *International conference on machine learning*, pages 5639–5650. PMLR, 2020.

[10] Diganta Misra. Mish: A self regularized non-monotonic activation function. *arXiv preprint arXiv:1908.08681*, 2019.

[11] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[12] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[13] Kandan Ramakrishnan, R James Cotton, Xaq Pitkow, and Andreas S Tolias. Generalization properties of contrastive world models. *arXiv preprint arXiv:2401.00057*, 2023.

[14] Ramanan Sekar, Oleh Rybkin, Kostas Daniilidis, Pieter Abbeel, Danijar Hafner, and Deepak Pathak. Planning to explore via self-supervised world models. In *International conference on machine learning*, pages 8583–8592. PMLR, 2020.