Navigation Dreams

Lan Feng
MAVT
ETH Zurich
Switzerland
lafeng@student.ethz.ch

Xiang Liu
MAVT
ETH Zurich
Switzerland
xianliu@student.ethz.ch

Abstract: Autonomously navigating a robot with images as input presents a significant challenge in perception and planning. Images capture potentially important details, such as complex geometry, body movement, and other visual cues. In order to successfully solve the navigation task from only images, algorithms must be able to model the scene and its dynamics using only this channel of information. The world model has shown success in the camera-based navigation problem. We investigate the connection between the quality of a world model and the navigation task's performance and improve the world model's performance based on our discovery. To this end, we propose a systematic world model evaluation method and a feature-based world model. We find that the deep features captured by a deep neural network can be used to evaluate a world model's quality better, and the feature-based world model performs better than the image-based world model.

Keywords: World Model, Navigation, Transformer

1 Introduction

Autonomous robot navigation involves complex perception and planning methods. When using only monocular image sensor data as input, how to extract meaningful information from the images becomes challenging. Recent works [1] have shown success in robot navigation by using the world model for visual feature extraction. However, the relationship between the world model and performance improvement remains unclear due to insufficient evaluation metrics such as pixel loss do not translate well to consistency, and the commonly used straight-forward sequence-to-sequence error advantages "blurred" predictions, failing to account for the stochasticity of real phenomena.

The contributions of our work can be summarized as follows:

- We propose a systematic evaluation method, which combines best of many and latent feature comparison, to evaluate dreams' quality more comprehensively.
- We propose a feature-based world model that encodes richer features from the environment, and outperforms the existing image-based world model in terms of dreams' quality.

2 Related Work

GAN In [2], Ian *et al* propose a framework for estimating generative models via an adversarial process. The generator learns to generate plausible data, while the discriminator learns to distinguish the generator's fake data from real data.

Pixel Loss A per-pixel loss function is used as a metric for understanding differences between images on a pixel level. The loss function measures the differences between output pixel values in an image. While the function is valuable for understanding interpolation on a pixel level, it is argued that it does not as accurately address qualities of the image that are important or meaningful.

Perceptual Loss [3] Perceptual loss functions are used when comparing two different images that look similar, like the same photo but shifted by one pixel. The function is used to compare high-level differences, like content and style discrepancies, between images. The perceptual loss function is a more commonly used component as it often provides more accurate results regarding style transfer.

Best of Many In [4], Apratim *et al* propose an objective with a "Best-of-Many-Samples" reconstruction cost, which means only the best sample from the generator is considered while computing the reconstruction loss.

3 Better Evaluation Method

The current evaluation method takes one sequence of images generated by the world model, computes its pixel loss with respect to the ground truth sequence, then judges its performance solely based on the mean squared error of this pixel loss. However, randomness is involved when generating sequences as these are outputs from a variational auto-encoder, and this uncertainty may destabilize the evaluation process. Besides, the pixel loss adopted may not be well aligned with our natural understanding of images, as it is not shift-invariant and encodes trivial details that may not be informative for the navigation task.

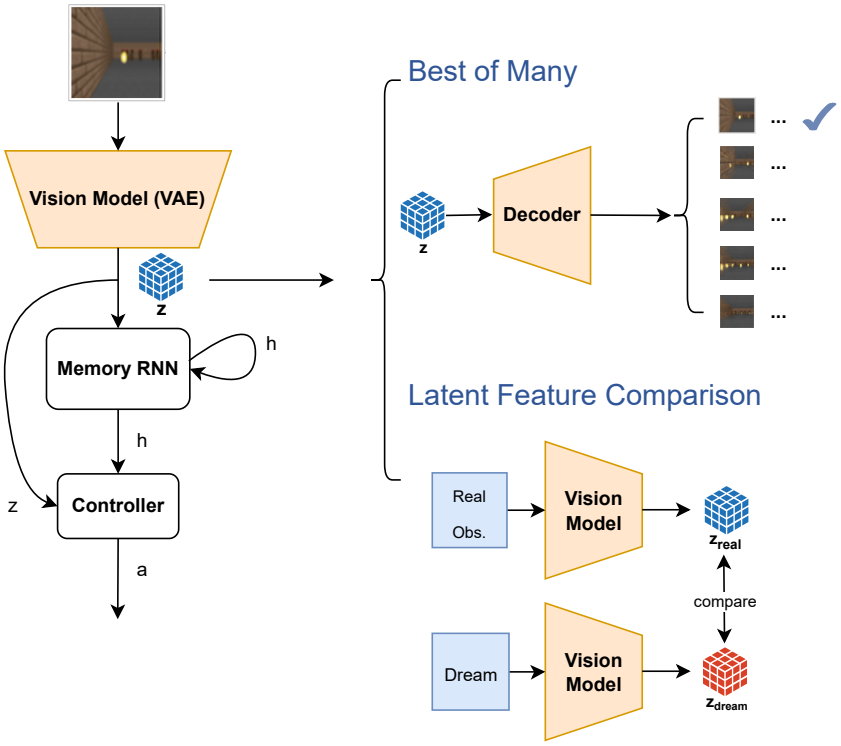


Figure 1: Framework of evaluation method

3.1 Best of Many

To tackle the randomness of dream generation, instead of just generating one sequence, we generate multiple sequences and pick the best one to mitigate this problem.

Due to limited computational resources, we use best of 10 in practice. As shown in Fig. 2, the best and worst sequence is picked out of 10 sequences generated by the world model. It can be observed that the best sequence emulates the movement and appearance of the object better than

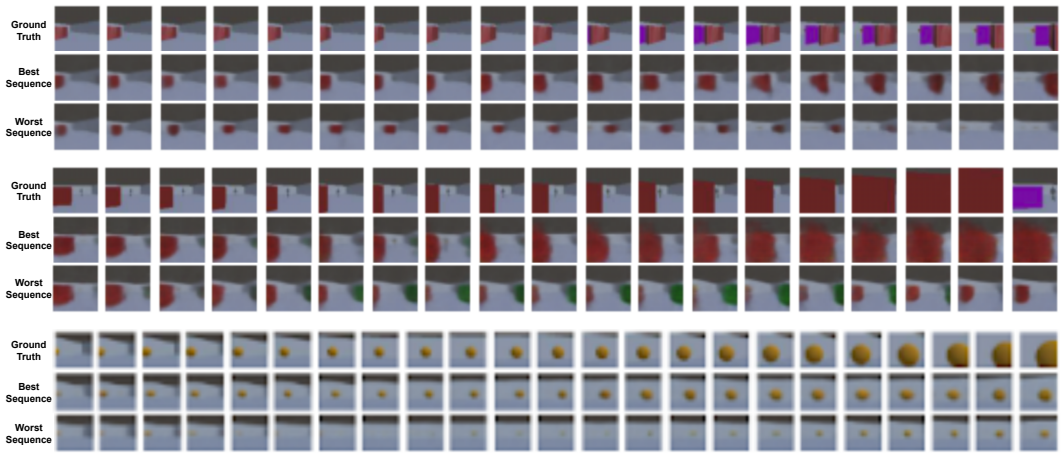


Figure 2: Image sequences of best of 10, where the best sequence is visually closer to the worst sequence

the worst sequence. Thus, the randomness involved in the generation process significantly impacts dream quality.

However, we also observed that the worst sequence performs better in some cases.

Figure 3: Image sequences of best of 10 where the worst sequence outperforms the best sequence

As shown in Fig. 3, the worst sequence somehow predicts the appearance of the human, but the best sequence almost renders an empty scene. However, if we compute the pixel loss of ground truth with respect to the best sequence, their difference is equivalent to subtracting a human from the background. However, suppose we compute the pixel loss of the ground truth with respect to the worst sequence. In that case, their difference is equivalent to subtracting two humans from the background, as the human predicted by the worst sequence is not perfectly aligned with the ground truth, which enlarges the pixel difference, neglecting the fact that they look visually closer.

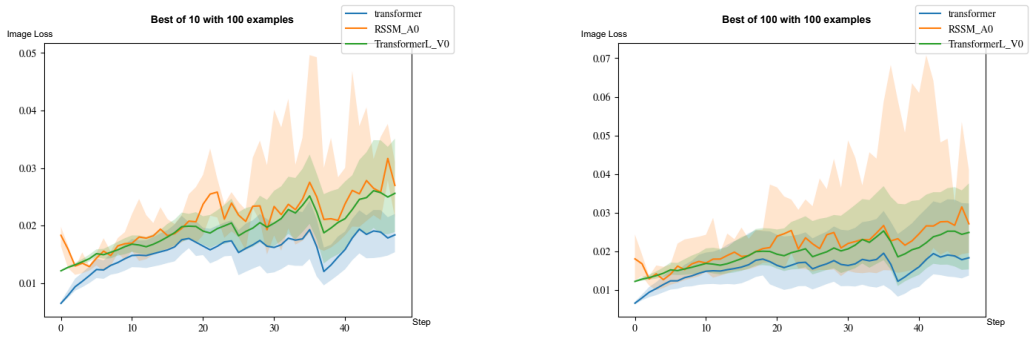
We also conducted an ablation study of the number of generated sequences with respect to the best image error.

In Fig. 4, the shaded area indicates the variation range of image error, which is outlined by the error of the best and the worst sequence. Moreover, the solid line in the middle indicates the mean image error averaged among 100 examples. It can be observed that the variation range grows as the step progresses, which is reasonable since the farther from the context, the harder the prediction gets. Also, the transformer-based model performs the best in terms of stability and precision.

As we can observe from Fig. 4, there is only a slight change in the variation range for best of 10 and best of 100, and the mean and the best sequence do not change significantly. Thus, we stick with best of 10 out of computational consideration in the following experiment.

3.2 Latent Feature Comparison

In order to tackle the problem that occurred in Fig. 3, we propose to compare model performance at the feature level. Inherited from the idea of perceptual loss, features encoded by the vision model



(a) Best of 10

(b) Best of 100

Figure 4: Comparison of Best of 10 and Best of 100

can represent high-level observation of the image. Thus, in the perfect setting, comparing features may be comparable to the scenario where a human inspector manually picks the best sequence out.

However, if images are fed into different encoders, then their latent features should not be used to compare, as different encoders will project the same image into different spaces. Therefore, feeding the image into the same encoder is essential for the validity of this method, which is the case in perceptual loss [3], where images are fed into the same VGG16 to obtain latent features.

Then we compare the performance of the image-based transformer model (`TransformerL_V0`) with the feature-based transformer model (`TransformerZ_V0`) introduced in section 4, where they have the same ConvVAE image encoder.

We leverage these two models to generate dream sequences and plot their image loss and feature loss, respectively, to see which metric better represents the dream quality, as shown in Fig. 5.

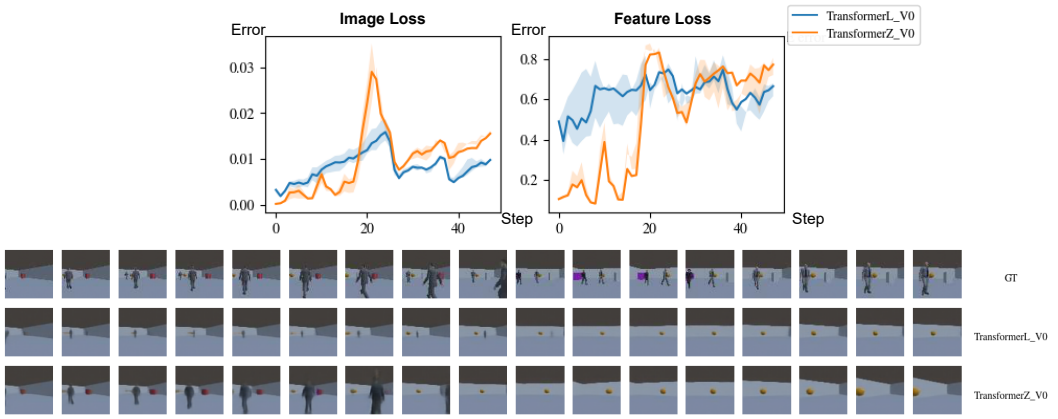
In Fig. 5a, we can observe from the generated image sequence below that `TransformerZ_V0` emulates the appearance and movement of the human in the ground truth sequence much better in the first couple of steps but performs almost the same as `TransformerL_V0` in the subsequent steps, so intuitively the dream quality metric of these two models should show a significant gap at first and then show a much smaller gap at last. As we can observe from the loss plot above, the image loss almost shows the same gap across all steps, whereas feature loss matches our intuition much better.

In the case of Fig. 5b, `TransformerZ_V0` also performs much better in matching the ground truth with respect to keeping colorful boxes and the human. However, suppose we compute the pixel loss of generated sequences with the ground truth. In that case, their numerical gap will be very small as the predicted objects in these images only occupy a small portion of pixels, which will not cause a significant improvement in image loss even if the model successfully predicts them as shown in the Image Loss plot. However, in the Feature Loss plot, the numerical gap is more significant and stable in terms of variation range, which can be instantiated by the change of vertical coordinate range in Image Loss and Feature Loss across these two examples.

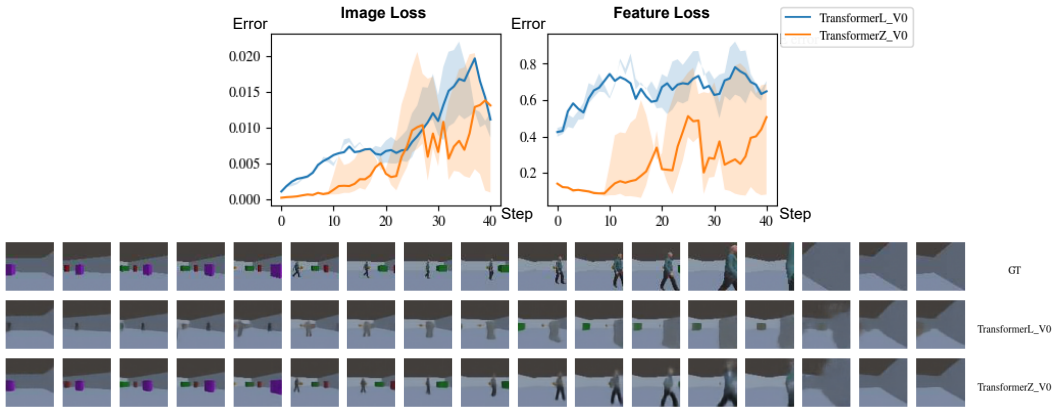
In conclusion, the feature loss behaves more consistently with our natural interpretation of the image sequence, so using feature loss rather than image loss will be a better practice in evaluating dream quality.

4 Better World Model

In this part, we will introduce the `TransformerZ_V0` model mentioned above. [1] proposes to use a transformer-based world model to generate dream sequences. They use an image-based transformer (`TransformerL_V0`), which takes image sequences as input and output. They first encode raw images to latent features by a ConvVAE and then feed the features to a transformer. A ConvVAE



(a)



(b)

Figure 5: Examples of Image Loss and Feature Loss

decoder decodes the output latent features to get dream sequences. Since both the input and output are images, the model is trained by minimizing pixel MSE loss.

However, we notice that encoding and decoding may cause massive information loss. For example, the reconstructed images generated by an autoencoder are usually blurrier than the original images. Besides, minimizing image loss will cause another problem. On the one hand, small objects in an image will weigh less than large objects in pixel MSE loss, which means the model will focus on memorizing the background of the environment instead of dynamic objects. On the other hand, pixel MSE loss requires the model to memorize insignificant details in the images, such as color, appearance, and texture, which may not be so crucial for navigation. Finally, the ultimate goal of a world model is to provide good features for downstream control tasks. Minimizing pixel MSE loss does not directly lead to information-rich features.

Considering the above, we propose a feature-based transformer world model, a new architecture that solves the above problems.

4.1 Image-based Transformer

An image-based transformer is trained to predict 48 future images based on history 16 images. The generation process of the image-based transformer is done in an auto-regressive way. An array with length of 32 that contains all the history images. The transformer predicts the next single image based on the history images array and then appends the new image to the array. Finally, the model is trained to minimize MSE loss between predicted images and authentic images

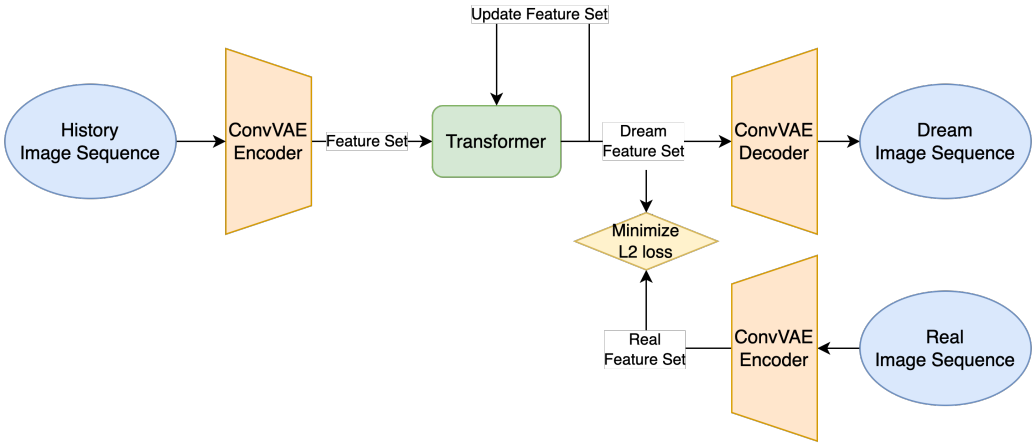


Figure 6: Feature-based Transformer Model

4.2 Feature-based Transformer

As shown in Fig. 6, a feature-based transformer is trained to predict 48 future image features based on 16 history images. We use a pre-trained ConvVAE to encode all the images to get image features. During training, the weights of ConvVAE are frozen to make the image features invariant. When the generation process is done, all the predicted features are decoded together to get image sequences. Since we maintained a feature set, the model can be trained to minimize feature differences between generated and real ones.

The benefits of the feature-based transformer are apparent. Minimizing feature loss helps to ignore trivial details in the images. Besides, the feature-based transformer has a better optimization objective since the downstream task of our world model is to navigate a robot, whose input is the features encoded by the world model.

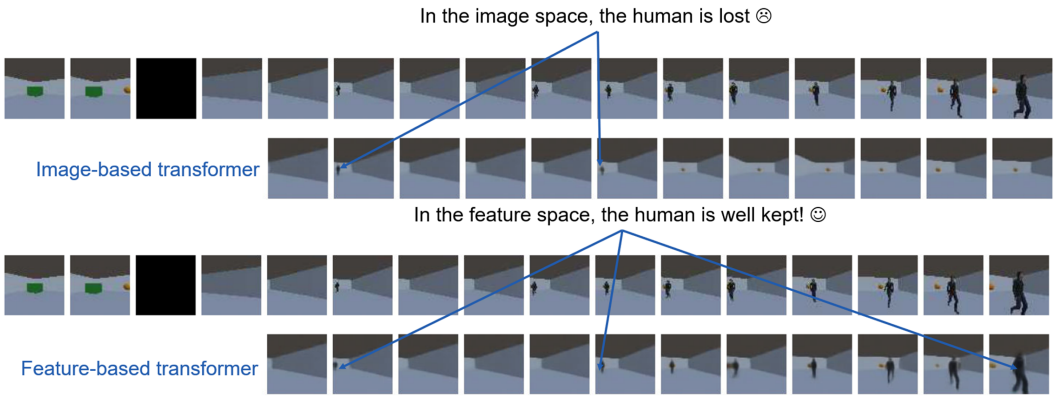


Figure 7: Dream sequences comparison of Image-based and Feature-based model

5 Experiments

5.1 Qualitative Results

As shown in Fig. 7, the dream sequence generated by the image-based transformer loses the human in the ground truth sequence, which is caused by the information loss during encoding and decoding, whereas the feature-based transformer almost perfectly predicts the appearance and movement of the human, thus more information is preserved in the feature space.

More examples showing feature-based model performs better than the image-based model are shown in Fig. 9.

5.2 Quantitative Results

Figure 8: Error plot of `TransformerL_V0` and `TransformerZ_V0` averged among 100 examples with best of 10

In Fig. 8, we test `TransformerL_V0` and `TransformerZ_V0` on 100 examples with best of 10, and average their error among examples. We can observe that for the first 20 steps `TransformerZ_V0` performs significantly better than `TransformerL_V0`, which are instantiated by both Image Loss and Feature Loss. Predicting the subsequent 20 steps will be sufficient for the navigation task.

As for the last 20 steps, `TransformerZ_V0` performs slightly worse than `TransformerL_V0` in terms of Image Loss, which is because that `TransformerZ_V0` is optimizing with respect to feature loss, whereas `TransformerL_V0` is optimizing with respect to image loss and will blur the image to achieve lower image loss.

6 Conclusion

In a nutshell, we combine best of many with latent feature comparison to evaluate the world model more comprehensively than the typical one sequence image loss evaluation.

Besides, we propose a feature-based world model and compare its performance with an image-based world model. Then combining our systematic evaluation method, we can conclude that our feature-based world model generally performs better than the image-based world model on both qualitative and quantitative results. Though evaluating a model trained to optimize features in feature space may advantage its performance, we also see a noticeable improvement in the qualitative evaluation as shown in Fig. 9, which should reduce this concern. To further reduce, we could use a different encoder for evaluation than the one with which `TransformerZ_V0` is trained.

In future work, we could confirm the advantage of `TransformerZ_V0` by deploying it to downstream navigation tasks and see if it improves the policy performance.

Figure 9: Additional examples of Image-based and Feature-based model. `TransformerL_V0` is image-based, `TransformerZ_V0` is feature-based, `TransformerZ_V0` performs generally better than `TransformerL_V0`.

Acknowledgments

We are grateful to Daniel Dugas[1] for the timely instruction and technical support throughout the semester.

References

[1] D. Dugas, O. Andersson, R. Siegwart, and J. J. Chung. Navdreams: Towards camera-only rl navigation among humans. *arXiv preprint arXiv:2203.12299*, 2022.

[2] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[3] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. *CoRR*, abs/1603.08155, 2016. URL http://arxiv.org/abs/1603.08155.

[4] A. Bhattacharyya, M. Fritz, and B. Schiele. "best-of-many-samples" distribution matching. *CoRR*, abs/1909.12598, 2019. URL http://arxiv.org/abs/1909.12598.

[1] daniel@dugas.ch