2020 IEEE International Conference on Robotics and Automation (ICRA)
31 May - 31 August, 2020. Paris, France

Action-conditioned Benchmarking of Robotic Video Prediction Models: a Comparative Study

Manuel Serra Nunes[1], Atabak Dehban[1,2], Plinio Moreno[1] and José Santos-Victor[1]

***Abstract*— A defining characteristic of intelligent systems is the ability to make action decisions based on the anticipated outcomes. Video prediction systems have been demonstrated as a solution for predicting how the future will unfold visually, and thus, many models have been proposed that are capable of predicting future frames based on a history of observed frames (and sometimes robot actions). However, a comprehensive method for determining the fitness of different video prediction models at guiding the selection of actions is yet to be developed.**

Current metrics assess video prediction models based on human perception of frame quality. In contrast, we argue that if these systems are to be used to guide action, necessarily, the actions the robot performs should be encoded in the predicted frames. In this paper, we are proposing a new metric to compare different video prediction models based on this argument. More specifically, we propose an action inference system and quantitatively rank different models based on how well we can infer the robot actions from the predicted frames. Our extensive experiments show that models with high perceptual scores can perform poorly in the proposed action inference tests and thus, may not be suitable options to be used in robot planning systems.

I. Introduction

An important stepping stone on the path towards intelligent robotic agents is providing them with the ability to explore their environment and to learn from interaction. Visual data, in the form of video, plays a central role in this problem and has led to great success in problems such as unsupervised learning of object keypoints [1] and action recognition [2]. In this direction, the next step should be for the robot to be able to learn the inherent workings of a real world environment and to understand how different bodies move, deform and influence each other.

As suggested by Srivastava *et al*. [3], if a robot has the ability to predict the imminent future based on an observed sequence of visual queues, then it must have acquired a representation of the spatial and temporal dynamics of the world. Predicting future video frames is perhaps the most straightforward materialization of this idea as the better a system can make predictions about future observations, the better the acquired feature representation must be [4]. For example, a robot that is able to predict that a falling stick will become occluded by a box, must understand (1) where the trajectory of the stick will take it, (2) be capable of perceiving depth, and (3) recognize the box in the foreground is opaque.

[1]Institute for Systems and Robotics, Instituto Superior Técnico, Universidade de Lisboa, Lisbon, Portugal
{mserranunes,adehban,plinio,jasv}@isr.tecnico.ulisboa.pt
[2]Champalimaud Centre for the Unknown, Lisbon, Portugal

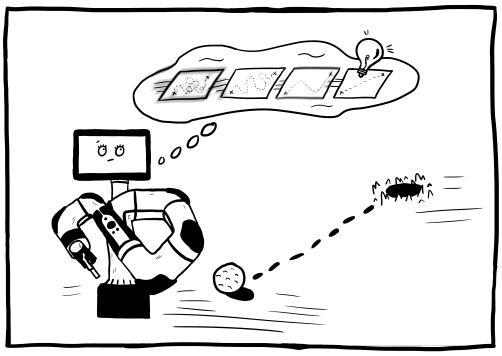

Fig. 1: A sawyer robot imagining different possible outcomes of executing a sequence of actions in the hypothetical task of pushing a ball to a whole. Image credit: Teresa Serra Nunes.

From a robotic perspective, if the predictions consider the actions of the agent itself, *i.e*. are conditioned on the action, then the representation should also help understand how performing an action in a given situation will affect the future appearance of the scene and thus, guide action decisions, an idea illustrated in fig. 1.

Similarly, the idea of anticipating sensory inputs to optimize action response in the human brain is studied under the theory of prospective coding, which explains the phenomenon by which representations of future states influence event perception and generation [5], [6]. Forward generative models constitute a fundamental part of predictive coding theory, especially in the domain of human action planning [7] where the latency between the stimulus of the retina and the corresponding acknowledgement by the responsible region of the brain makes it difficult to select appropriate actions in response to rapidly evolving events. The capabilities of Video Prediction (VP) systems to serve as forward models have been explored *e.g*. with the introduction of an architecture that learns a policy for solving OpenAI Gym [8] Reinforcement Learning (RL) problems using an encoder of observed video frames and a MDN-RNN to predict future visual codes, given current and past observations and executed actions [9]. Finn and Levine [10] use a VP model to continuously sample the expected future given different sequences of actions. The sequence that maximizes the likelihood of the robot achieving the goal of pushing an object to a specified location is selected at each time step to be executed. This type of model-based control is an active area of research in RL and large-scale datasets of robotic experiment such as RobotNet

978-1-7281-7395-5/20/$31.00 ©2020 IEEE

[11], developed concurrently with this work, should allow future breakthroughs.

In these applications, the ability to select the best possible action is very dependent on how well the VP model can anticipate future observations based on the robot's actions and the current status of the scene. Having a metric that can rank video prediction models based on how well they perform as a forward model is therefore of fundamental importance.

As will be described in section II, most state-of-the-art work in video prediction measures the performance of the models using metrics designed to reflect human perception of quality. While these metrics might be useful for applications such as precipitation nowcasting [12] or semantic segmentation prediction [13], we argue that they are not necessarily adequate in action oriented applications such as robotic planning where the quality of the video prediction model should be measured by how well it can guide action decisions from the predicted frames.

Inspired by this notion, we propose a new, simple metric for ranking video prediction models from a robotic standpoint. Given a sequence of predicted frames, we train a model to infer the action performed at each time step. The ability of our inference machine to recognize the correct sequence of actions only from the predicted frames should indicate that the representation of the world held by the video prediction model is correctly encoding action features and it is able to understand the consequences of executing a given action at the current state of the environment.

This paper has the following three contributions:

- we propose a novel action-based quality assessment metric for robotic video prediction models;
- we apply the metric on several different models and quantitatively rank them from a robotic action-inference perspective;
- we qualitatively compare our method with other metrics and show that in most cases our quality measure can independently assess models, providing new insights about their performance as action-conditioned video prediction models.

In addition, we provide the implementation of our experiments to facilitate assessment of VP models that we did not consider or will be proposed in the future[1].

The rest of the paper is organized as follows: in section II we examine the related work on different VP models with an emphasis on the research directions that we consider in this paper. In section III we explain the details of of our method, metrics we used, and several design choices we considered that made this work possible. Section IV begins with a description of the dataset used in our experiments. It then continues by discussing how well different methods could compete in our metric and how do they compare in terms of other metrics already developed in the literature. Finally, we draw our conclusions and discuss promising future research directions in section V.

[1]https://github.com/m-serra/action-inference-for-video-prediction-benchmarking

II. Related Lines of Research

A. Video Prediction

The importance of anticipation and predictive sensori processing has long been regarded as crucial in controlling neural and cognitive processes such as perception, decision making and motion in both humans and animals, with studies on the subject dating at least as far back as the 19^{th} century [14] and extending into the 21^{st} century [15], [16]. In the field of robotics, these concepts inspired the development of sensori-motor networks [17] which emulate the interaction between the visual and motor systems in organisms to predict future visual stimulus. Santos *et al*. [18] applied sensori-motor networks to small image patches to predict the next time step's stimulus.

However, when the problem is extended to more generic settings involving observations of a complete scene and longer temporal sequences, the high dimensionality of the data and the rich, complex, and uncertain dynamics of the real world become a bigger hurdle. In recent years, research in neural networks has mitigated these problems, with the development of Convolutional Neural Networks (CNNs), that reduce the dimensionality burden in image processing, Recurrent Neural Networkss (RNNs) which capture the information contained in sequential data, and a combination of the two in the Convolutional Long Short Term Memorys (ConvLSTMs) [12]. All these systems have been widely used in the field of video prediction.

Influential work on video prediction by Mathieu *et al*. [4] focused on improving the quality of the generated video by experimenting with different loss functions as an alternative to ℓ_2, which is hypothesized to cause blurry predictions. One of the most meaningful contributions to video prediction was perhaps the introduction of the concept of pixel motion by Finn *et al*. [10], Xue *et al*. [19] and De Brabandere *et al*. [20] which liberates the system from having to predict every pixel from scratch by instead modelling pixel motion from previous images and applying it to the most recent observation. Since then several authors have continued the work in this direction: Babaeizadeh *et al*. [21] account for the stochasticity of the world by conditioning the predictions on stochastic variables while Lee *et al*. [22] explore how the introduction of a Generative Adversarial Network (GAN) improves the visual quality of predictions.

Other lines of research have included motion and content decomposition [23], predicting transformations on feature maps [24], [19], and biologically inspired approaches [25] which propose a hierarchical architecture that emulates the top-down and bottom-up transmission of local predictions and errors in predictive coding theories of human perception.

In this work we focus on action-conditioned video prediction models as it is presumable that those are the most suited models for use in robotic planning. We select 1) Convolutional Dynamic Neural Advection (CDNA): a deterministic model based on pixel-motion modelling [26], 2) Stochastic

Adversarial Video Prediction (SAVP): which also models pixel motion but introduces variational and adversarial terms to the loss, to try to improve prediction quality and account for the variability in the environment [22], 3) a variant of SAVP in which the adversarial term is suppressed, 4) Stochastic Variational Video Prediction (SV2P): an extension of CDNA conditioned on stochastic variables, 5) and finally we test Stochastic Video Generation with Learned Prior (SVG-LP): the stochastic, action-free model of [27].

B. Assessment of video prediction models

A common trend in video prediction models is the evaluation of model performance based on metrics designed to mirror human perception of quality in image and video, *i.e.*, Quality of Experience (QoE). This is a subjective concept, which depends not only on the data fidelity of the reconstructed image or video but also on the personal experience and expectations of the viewer [28]. The standard measure for QoE is the Mean Opinion Score (MOS) which is the average quality rating, given by a sample of viewers. QoE prediction is an active area of research in which proposed methods are usually compared to the Peak Signal to Noise Ratio (PSNR) benchmark. PSNR is a logarithmic measure of the mean squared error between a distorted image. Its mathematical simplicity and convenient optimization properties make it one of the most popular metrics for image quality [29]. However, PSNR compares images pixel by pixel, not taking into account the content, leading to pathological cases [28] in which it fails at approximating human judgement.

An alternative metric that addresses this problem is the Structural Similarity (SSIM) Index [30], which is founded on the principle that signals that are close in space have strong dependencies between each other and that the human visual system is highly adapted for extracting this structural information. SSIM indices are calculated using a sliding window which produces an index map. This index is 1 if the structure of corresponding patches of the two images is the same and the final SSIM score corresponds to the average of the index map. More recently, Learned Perceptual Image Patch Similarity (LPIPS) metrics, based on learned features of neural networks such as VGG have shown remarkable capabilities as a perceptual distance metric [31].

Inspired by the developments in image generation, methods that are specifically designed for assessing realism in generated video have also been proposed [32]. *E.g.* the Fréchet Video Distance (FVD) [33] accounts for visual quality, temporal coherence, and diversity by measuring the distance between the distribution that originated the observed data and the distribution from which the predicted video is generated, instead of comparing pixels or image patches.

In this work we compare the performance of VP models on our proposed metric with performance on commonly used PSNR and SSIM, and on FVD.

III. Methods

In this section we present a simple method for ranking VP models based on their capacity to guide a robotic agent's action decisions, reflected by the performance of an action inference system. We start by assuming that the better the dynamics representation of the agent is at encoding action features, the better it will be for planning actions based on the expected outcome. Under this assumption, the problem turns into evaluating how well a VP model is encoding action features and assigning it a score based on such evaluation.

With this in mind, we hypothesise that the capacity to observe a sequence of predicted frames and infer the executed actions should be an indicator that the VP model is correctly encoding action features. To better illustrate the idea, first consider a failure case: if the VP model generates a sequence of predicted frames that do not correspond to the executed actions by the robot, then no action inference model can recognize the correct set of actions from the predicted images, resulting in a low action inference score. On the other hand, if the VP model understands the consequences of the input actions, then the frames it predicts should correctly reflect the action and its consequences, allowing an inference model to recognize the actual executed actions and attain a high score.

A. Video prediction

In order to compare how the proposed metric correlates with PSNR, SSIM and FVD, we start by selecting a group of VP models from prior work to be tested using our metric. By comparing model performance under metrics designed to predict human quality perception with our metric, designed to assess the capabilities of the model to guide action decisions, we intend to answer the question *"Does a good video prediction from a human perspective correspond to a good video prediction from the standpoint of a robot making decisions?"*. This is an interesting question considering that a change in model ranking, when compared to PSNR or SSIM, may not only influence the choice of the VP in an action planning experiment but also indicate that the best representation for a robot to make a decision may not resemble anything a human may recognize [34], [35] and inspire new lines of research such as optimizing for losses other than ground truth similarity.

The selection of the tested VP models described in section II was made with the goal of covering the main approaches to robotic video prediction, which opens up the possibility of identifying the most significant features of a VP model used in a robotic planning context.

B. Action inference model

To assess the quality of models, we first train a simple convolutional neural network to infer the actions executed between every two frames using predicted videos. The actions are assumed to be continuous and multidimensional, to be representative of most robotic control action-spaces. Each pair of frames is concatenated along the channels dimension and given to the network as input, as illustrated in fig. 2.

Because action dynamics should not change over time, model parameters are shared across all time steps of a sequence. While a RNN would typically have been useful for learning the sequence of executed actions, we choose to input a window of two frames at a time, cutting off any temporal correlation between actions. This forces the inference model to identify actions from the frames instead of focusing on learning the temporal action distribution. The option for a window size of two frames is due to the fact that in the selected dataset the robot's actions are randomly updated every two frames. For datasets with different conditions, however, the window size parameter can control the temporal information received by the network without shifting the attention of the model from the frames, and it is expected that bigger windows should result in better action inference.

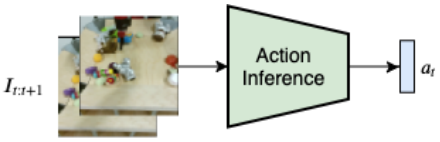

Fig. 2: Action inference network. At each time step the network receives a pair of frames and outputs a multidimensional recognized action.

IV. Experiments and Results

A. Experimental setup

We conduct our experiments using the BAIR robot push dataset [36] which consists of a robotic arm pushing a collection of objects on a table. This dataset was collected in the context of visual planning and video prediction and has since become a benchmark in the field [27], [37], [38], [39]. The dataset contains 43520 examples of random movements, as exemplified by a birds eye view of the gripper trajectory during a sample in fig. 3. Videos are 30 frames long, with 64×64 RGB images, collected at 10 Hz. The dataset also provides the commanded action sequences, a 4-dimensional array representing the joint velocities and whether the gripper is open or closed, and a 3-dimensional array representing the Cartesian coordinates of the gripper. All tested VP models were pre-trained by the respective authors with exception of CDNA, which was trained on over 200000 iterations, using scheduled sampling [40]. At training time, models receive 2 context frames and actions (with the exception of SVG-LP which only receives the frames) and predict video up to time step 12, with each prediction being fed back as input for the next time step.

In our work, after training a forward pass is made over the entire training set and the generated predictions, this time generalizing until step 30, are saved as a new dataset for subsequent training of the action inference model. Having a dataset of predictions for each VP model, the action inference network is trained on the 28 frame long predictions. In our experiments, we define the actions being inferred as the displacements Δx and Δy of the robot's gripper along the x and y axis, between every two time steps. The ground truth

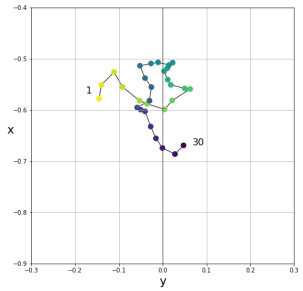

Fig. 3: A sample trajectory of the gripper illustrating the random nature of the actions.

targets for the actions are directly extracted from the BAIR dataset gripper state sequences by subtracting consecutive temporal positions for both axis. This results in an action sequence of length 27 for each 28-frame predicted video.

A characteristic of the BAIR dataset which has particular effect on the results is the fact that joint velocities are only updated every two frames. Even though the gripper position still changes at every time step, the variance of the change, *i.e.* the variance of Δx and Δy is higher on time steps in which joint velocities are updated. This aspect of the data is depicted in fig. 4 where the Δy targets of the test set are scattered, revealing an alternating standard deviation. In practice, this alternating nature results in the action inference network not experiencing all types of actions the same way, therefore becoming better fit to some situations than others. For this reason, results are presented separately for odd and even frames corresponding to time steps in fig. 4.

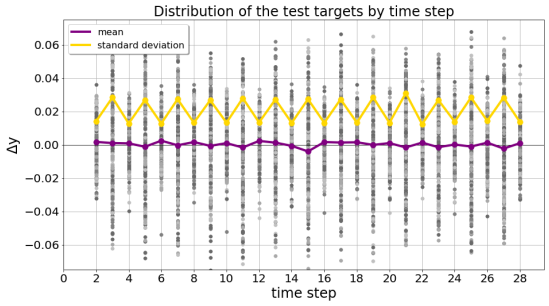

Fig. 4: Distribution of the test targets Δy, revealing a characteristic alternating standard deviation.

B. Quantitative Comparison

We start by evaluating the selected group of VP models on some of the traditionally used metrics described in section II and on the recently proposed FVD. As opposed to the methodology adopted by some of the previous work [22], [27] in which 100 possible futures are sampled and the best score of the group is reported, we choose to sample a single time, in order to better approximate the conditions of a robot planning actions. This approach has especial impact

on action-free models like SVG-LP, that are exposed to greater uncertainty. Regarding the action-conditioned models, the results displayed in fig. 5 are in line with previous reports, indicating that models have better performance when conditioned on both actions and stochastic variables, as is the case with SAVP-VAE and SV2P. On the other hand, the addition of an adversarial loss term seems to affect performance negatively, which reflects on SAVP having a lower PSNR/SSIM score than a deterministic model like CDNA despite the high visual appeal of the predicted frames.

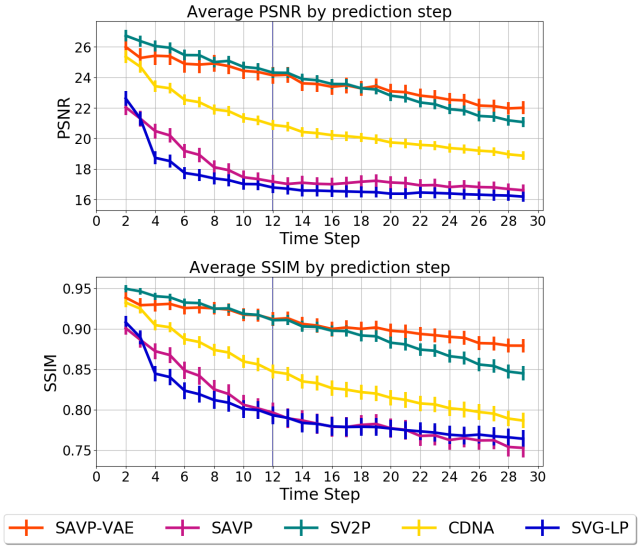

Fig. 5: Average PSNR and SSIM over the test set with 95% confidence interval. Results were reproduced with modification from [26], [27], [22].

We compute the FVD values for the test set predictions in table I using batches of size 32 and discard the two context frames to only consider the predictions of length 28. This approach is different from the one proposed in [33], therefore resulting in higher FVD values but preserving model rankings.

For each VP model's predictions dataset, the action inference model that produces the best validation score during training is selected. To measure how well it can identify the executed actions, we compute the R^2 goodness-of-fit metric which in our particular case represents the percentage of change in ground truth action variance that is explained by the inferred actions. A model that perfectly identifies the executed actions will have a score of 1.0 whereas a model that simply outputs the mean Δx and Δy will have a score of 0.0. It is worth noting that while R^2 may not be a strong enough statistic for comparing different regression models, the focus of this work is to assess the predictions made by different VP models, using the same training regime for the inference model. In our experiments R^2 is computed along the 256 test examples for each time step and the evolution of the metric over time is reported in fig. 6. The Mean Absolute Error (MAE) is also presented in fig. 7, computed in the same way as R^2, for each time step. The most immediate

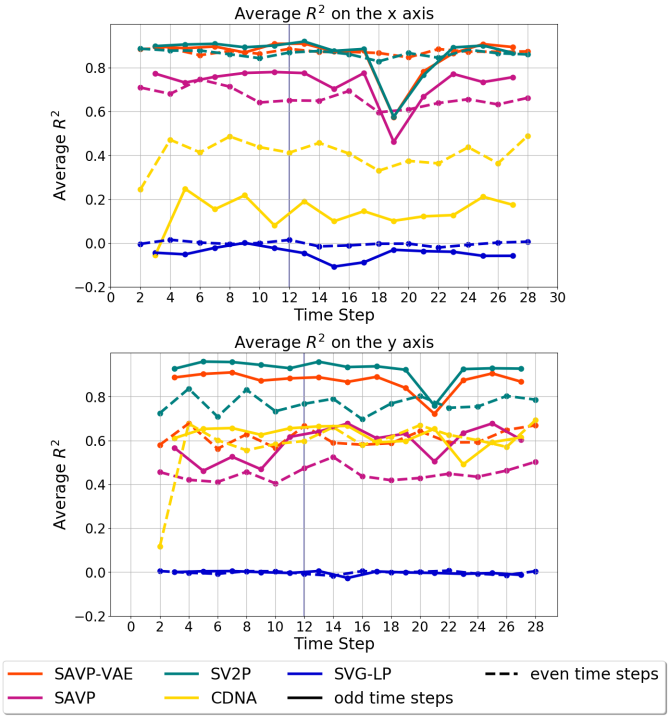

Fig. 6: R^2 results over time for predictions made by different VP models. Odd and even time steps are shown separately.

characteristic in the temporal evolution of action inference that arises from an initial analysis of figures 6 and 7 is that the temporal downgrade artefact in performance observed in PSNR and SSIM is not manifested in the capacity of the model to recognize the actions, with exception of results for CDNA. This quality of the metric stems from the fact that the parameters of the inference model are shared across all time steps, a choice based on the fact that action dynamics do not change over time and therefore VP models should have a consistent action encoding for all time steps. For this reason, a VP model that encodes actions in a consistent manner should allow the inference network to better learn how to recognize actions and will therefore display stable R^2 and MAE values across time, as is verified for SAVP and SAVP-VAE. On the other hand, because video predictions made by CDNA have changing dynamics, starting with good resolution and transitioning to blurry images as time advances, it is difficult for the action inference model to learn to identify actions.

The performance of the action inference on predictions made by different models indicates, based on figures 6 and 7 and on table I, that the model that is better encoding action features and would therefore be the most suited in robotic planning problems is SV2P, closely followed by SAVP-VAE, implying that conditioning on stochastic variables is beneficial but the introduction of the adversarial loss for better image quality removes attention from optimal encoding of action features. These models even outperform the ground truth oracle, supporting the argument that the stochastic variables should be accounting for non observable

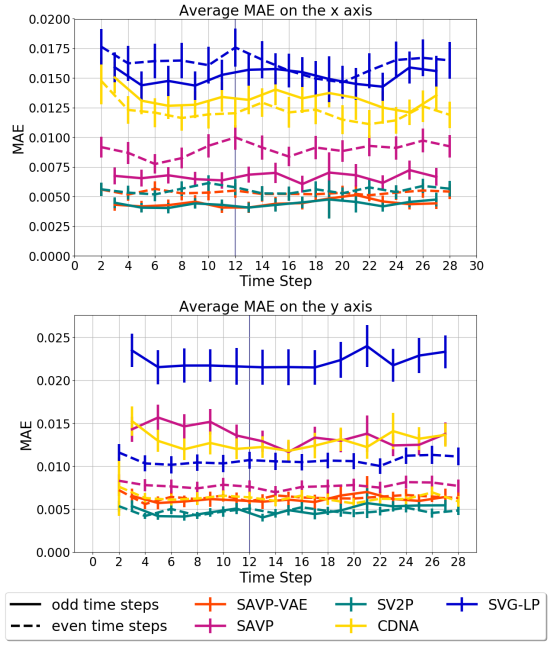

Fig. 7: Average MAE results with 95% confidence interval for predictions made by different VP models. Odd and even time steps are shown separately.

aspects of the scene and that some blurring of the background may actually help the inference network focus on the action features. On the other hand, the action-free SVG-LP model has an R^2 value of approximately 0 and an MAE value of 0.163 which corresponds to the variance of the data. This indicates, as observed in section IV-C, that the inference model is unable to identify the actions and limits itself to predicting a constant average. The origin of this result is that an action-free stochastic model from which a single prediction is sampled, may produce a future that is different from the ground truth, causing recognized actions to not match the targets and preventing the model from learning a meaningful mapping during training.

In general, and as reported by [33], PSNR and SSIM present a very high correlation as both of them are based on frame by frame comparisons with the original data. Furthermore, because most VP models use an ℓ_2 term in the loss function, these are biased metrics. We also verify that multiple ranking changes occur between our proposed score and FVD, including SV2P scoring the best in action recognition while having an FVD value close to that of SVG-LP, which for being action-free has the lowest score under our metric. These results show that the ability to recognize actions from predicted images doesn't necessarily correlate with previously proposed metrics and that action inference may offer a valuable perspective for choosing the best model in a planning scenario.

C. Qualitative Comparison

To better understand how the R^2 and MAE results reflect on action inference in practice, examples of inferred action sequences are showed against the ground truth in figure 8. We find the best and worst R^2 score on action inference

TABLE I: FVD, R^2 and MAE values for each VP model.

Model	FVD Value	R2 Score	MAE Value
CDNA	943.5	0.4339	0.0111
SAVP	738.3	0.5953	0.0092
SAVP-VAE	**409.8**	0.8427	0.0056
SV2P	691.1	**0.8719**	**0.0049**
SVG-LP	728.2	-0.0582	0.0160
Oracle	0.0	0.8203	0.0058

from video predictions from any VP model and then plot the inferred actions from that video sequence. The best performing model in the selected examples was SAVP-VAE, while the worst was SVG-LP. A visual interpretation of the extent to which the inference network is successful at recognizing the actions is in conformance with the results of section IV-B, with the SAVP-VAE and SV2P models performing the best, followed by SAVP and CDNA.

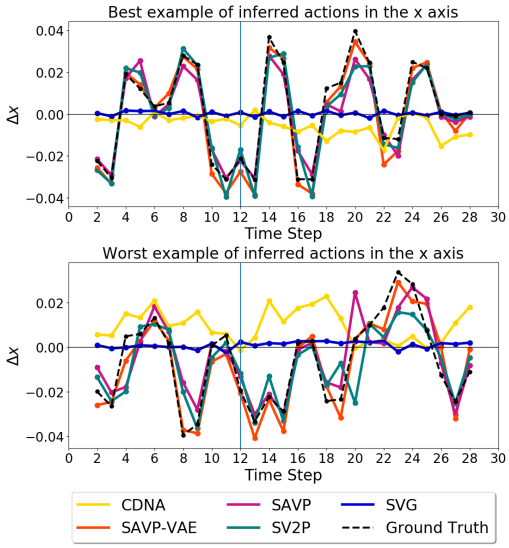

Fig. 8: Best and worst examples of inferred Δx.

V. Conclusions and Future Work

In this work we proposed a novel method for evaluating the quality of VP models from a robotic standpoint. We compared different existing video prediction models using our metric, showing that good performance on metrics that mirror human perception of quality does not necessarily imply that the model holds a good representation of action-effect. In the future we plan to introduce better datasets that include states of the environment other than gripper position, such as objects positions and speeds, allowing the assessment of models based on more comprehensive states. Developing action or object-state aware cost functions for training VP models is another possible future line of research.

Acknowledgements

Work partially supported by the Portuguese Foundation for Science and Technology (FCT) [UID/EEA/50009/2019], LARSyS-FCT Plurianual funding 2020-2023 and the Robotics, Brain and Cognition Lab, part of the Portuguese Roadmap of Research Infrastructures.

REFERENCES

[1] T. Kulkarni, A. Gupta, C. Ionescu, S. Borgeaud, M. Reynolds, A. Zisserman, and V. Mnih, "Unsupervised learning of object keypoints for perception and control," in *Advances in Neural Information Processing Systems (NIPS)*, 2019.

[2] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Advances in Neural Information Processing Systems (NIPS)*, 2014, pp. 568–576.

[3] N. Srivastava, E. Mansimov, and R. Salakhudinov, "Unsupervised learning of video representations using lstms," in *International Conference on Machine Learning (ICML)*, 2015, pp. 843–852.

[4] M. Mathieu, C. Couprie, and Y. LeCun, "Deep multi-scale video prediction beyond mean square error," in *International Conference on Learning Representations (ICLR)*, 2015.

[5] K. Friston, "A theory of cortical responses," *Philosophical transactions of the Royal Society B: Biological sciences*, vol. 360, no. 1456, pp. 815–836, 2005.

[6] S. Schütz-Bosbach and W. Prinz, "Prospective coding in event representation," *Cognitive processing*, vol. 8, no. 2, pp. 93–102, 2007.

[7] T. H. FitzGerald, P. Schwartenbeck, M. Moutoussis, R. J. Dolan, and K. Friston, "Active inference, evidence accumulation, and the urn task," *Neural Computation*, vol. 27, no. 2, pp. 306–328, 2015.

[8] G. Brockman, V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang, and W. Zaremba, "Openai gym," 2016.

[9] D. Ha and J. Schmidhuber, "Recurrent world models facilitate policy evolution," in *Advances in Neural Information Processing Systems (NIPS)*, 2018.

[10] C. Finn and S. Levine, "Deep visual foresight for planning robot motion," in *IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2017, pp. 2786–2793.

[11] S. Dasari, F. Ebert, S. Tian, S. Nair, B. Bucher, K. Schmeckpeper, S. Singh, S. Levine, and C. Finn, "Robonet: Large-scale multi-robot learning," *arXiv preprint arXiv:1910.11215*, 2019.

[12] S. Xingjian, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-c. Woo, "Convolutional lstm network: A machine learning approach for precipitation nowcasting," in *Advances in Neural Information Processing Systems (NIPS)*, 2015, pp. 802–810.

[13] P. Luc, N. Neverova, C. Couprie, J. Verbeek, and Y. LeCun, "Predicting deeper into the future of semantic segmentation," in *IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 648–657.

[14] W. James, *The Principles of Psychology*. New York: Dover Publications, 1890, vol. Vol. 1.

[15] G. Pezzulo, M. A. van der Meer, C. S. Lansink, and C. M. Pennartz, "Internally generated sequences in learning and executing goal-directed behavior," *Trends in cognitive sciences*, vol. 18, no. 12, pp. 647–657, 2014.

[16] R. P. Rao and D. H. Ballard, "Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects," *Nature neuroscience*, vol. 2, no. 1, p. 79, 1999.

[17] D. M. Wolpert, J. Diedrichsen, and J. R. Flanagan, "Principles of sensorimotor learning," *Nature Reviews Neuroscience*, vol. 12, no. 12, p. 739, 2011.

[18] R. Santos, R. Ferreira, Â. Cardoso, and A. Bernardino, "Sensorimotor networks vs neural networks for visual stimulus prediction," in *4th International Conference on Development and Learning and on Epigenetic Robotics*. IEEE, 2014, pp. 287–292.

[19] T. Xue, J. Wu, K. Bouman, and B. Freeman, "Visual dynamics: Probabilistic future frame synthesis via cross convolutional networks," in *Advances in Neural Information Processing Systems (NIPS)*, 2016, pp. 91–99.

[20] X. Jia, B. De Brabandere, T. Tuytelaars, and L. V. Gool, "Dynamic filter networks," in *Advances in Neural Information Processing Systems (NIPS)*, 2016, pp. 667–675.

[21] M. Babaeizadeh, C. Finn, D. Erhan, R. H. Campbell, and S. Levine, "Stochastic variational video prediction," *International Conference on Learning Representations (ICLR)*, 2017.

[22] A. X. Lee, R. Zhang, F. Ebert, P. Abbeel, C. Finn, and S. Levine, "Stochastic adversarial video prediction," *arXiv preprint arXiv:1804.01523*, 2018.

[23] R. Villegas, J. Yang, S. Hong, X. Lin, and H. Lee, "Decomposing motion and content for natural video sequence prediction," in *International Conference on Learning Representations (ICLR)*, 2017.

[24] J. Oh, X. Guo, H. Lee, R. L. Lewis, and S. Singh, "Action-conditional video prediction using deep networks in atari games," in *Advances in Neural Information Processing Systems (NIPS)*, 2015, pp. 2863–2871.

[25] W. Lotter, G. Kreiman, and D. Cox, "Deep predictive coding networks for video prediction and unsupervised learning," in *International Conference on Learning Representations (ICLR)*, 2016.

[26] C. Finn, I. Goodfellow, and S. Levine, "Unsupervised learning for physical interaction through video prediction," in *Advances in Neural Information Processing Systems (NIPS)*, 2016, pp. 64–72.

[27] E. Denton and R. Fergus, "Stochastic video generation with a learned prior," in *International Conference on Machine Learning (ICML)*, 2018.

[28] S. Winkler and P. Mohandas, "The evolution of video quality measurement: From psnr to hybrid metrics," *IEEE Transactions on Broadcasting*, vol. 54, no. 3, pp. 660–668, 2008.

[29] Q. Huynh-Thu and M. Ghanbari, "The accuracy of psnr in predicting video quality for different video scenes and frame rates," *Telecommunication Systems*, vol. 49, no. 1, pp. 35–48, 2012.

[30] Z. Wang, L. Lu, and A. C. Bovik, "Video quality assessment based on structural distortion measurement," *Signal processing: Image communication*, vol. 19, no. 2, pp. 121–132, 2004.

[31] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 586–595.

[32] P. Bhattacharjee and S. Das, "Temporal coherency based criteria for predicting video frames using deep multi-stage generative adversarial networks," in *Advances in Neural Information Processing Systems (NIPS)*, 2017, pp. 4268–4277.

[33] T. Unterthiner, S. van Steenkiste, K. Kurach, R. Marinier, M. Michalski, and S. Gelly, "Towards accurate generative models of video: A new metric & challenges," *arXiv preprint arXiv:1812.01717*, 2018.

[34] A. Mordvintsev, C. Olah, and M. Tyka, "Inceptionism: Going deeper into neural networks," 2015. [Online]. Available: https://research.googleblog.com/2015/06/inceptionism-going-deeper-into-neural.html

[35] C. Olah, A. Satyanarayan, I. Johnson, S. Carter, L. Schubert, K. Ye, and A. Mordvintsev, "The building blocks of interpretability," *Distill*, 2018, https://distill.pub/2018/building-blocks.

[36] F. Ebert, C. Finn, A. X. Lee, and S. Levine, "Self-supervised visual planning with temporal skip connections," 2017.

[37] L. Castrejon, N. Ballas, and A. Courville, "Improved conditional vrnns for video prediction," in *IEEE International Conference on Computer Vision (ICCV)*, 2019.

[38] F. Ebert, C. Finn, S. Dasari, A. Xie, A. Lee, and S. Levine, "Visual foresight: Model-based deep reinforcement learning for vision-based robotic control," *arXiv preprint arXiv:1812.00568*, 2018.

[39] M. Kumar, M. Babaeizadeh, D. Erhan, C. Finn, S. Levine, L. Dinh, and D. Kingma, "Videoflow: A conditional flow-based model for stochastic video generation," in *International Conference on Learning Representations (ICLR)*, 2020.

[40] S. Bengio, O. Vinyals, N. Jaitly, and N. Shazeer, "Scheduled sampling for sequence prediction with recurrent neural networks," in *Advances in Neural Information Processing Systems (NIPS)*, 2015, pp. 1171–1179.