HARP: Autoregressive Latent Video Prediction with High-Fidelity Image Generator

Younggyo Seo[1,*] **Kimin Lee**[2,†] **Fangchen Liu**[2] **Stephen James**[2,‡] **Pieter Abbeel**[2]
[1]KAIST [2]UC Berkeley

Abstract

Video prediction is an important yet challenging problem; burdened with the tasks of generating future frames and learning environment dynamics. Recently, autoregressive latent video models have proved to be a powerful video prediction tool, by separating the video prediction into two sub-problems: pre-training an image generator model, followed by learning an autoregressive prediction model in the latent space of the image generator. However, successfully generating high-fidelity and high-resolution videos has yet to be seen. In this work, we investigate how to train an autoregressive latent video prediction model capable of predicting high-fidelity future frames with minimal modification to existing models, and produce high-resolution (256x256) videos. Specifically, we scale up prior models by employing a high-fidelity image generator (VQ-GAN) with a causal transformer model, and introduce additional techniques of top-k sampling and data augmentation to further improve video prediction quality. Despite the simplicity, the proposed method achieves competitive performance to state-of-the-art approaches on standard video prediction benchmarks with fewer parameters, and enables high-resolution video prediction on complex and large-scale datasets. Videos are available at https://sites.google.com/view/harp-videos/home*.

1 Introduction

Figure 1: Selcted 256×256 video sample generated by HARP on RoboNet (Dasari et al., 2019).

Video prediction can enable agents to learn useful representations for predicting the future consequences of the decisions they make, which is crucial for solving the tasks that require long-term planning, including robotic manipulation (Finn & Levine, 2017; Kalashnikov et al., 2018) and autonomous driving (Xu et al., 2017). Despite the recent advances in improving the quality of video

---

prediction (Finn et al., 2016; Babaeizadeh et al., 2018; Denton & Fergus, 2018; Lee et al., 2018; Weissenborn et al., 2020; Babaeizadeh et al., 2021), learning an accurate video prediction model remains notoriously difficult problem and requires a lot of computing resources, especially when the inputs are video sequences with high-resolution (Villegas et al., 2019; Clark et al., 2019; Luc et al., 2020). This is because the video prediction model should excel at both tasks of generating high-fidelity images and learning the dynamics of environments, though each task itself is already a very challenging problem.

Recently, autoregressive latent video prediction methods (Rakhimov et al., 2021; Yan et al., 2021; 2022) have been proposed to improve the efficiency of video prediction, by separating video prediction into two sub-problems: first pre-training an image generator (*e.g.,* VQ-VAE; Oord et al. 2017), and then learning the autoregressive prediction model (Weissenborn et al., 2020; Chen et al., 2020) in the latent space of the pre-trained image generator. However, the prior works are limited in that they only consider relatively low-resolution videos (up to 128×128 pixels) for demonstrating the efficiency of the approach; it is questionable that such experiments can fully demonstrate the benefit of operating in the latent space of image generator instead of pixel-channel space.

In this paper, we present **H**igh-fidelity **A**uto**R**egressive latent video **P**rediction (HARP), which scales up the previous autoregressive latent video prediction methods for high-fidelity video prediction. The main principle for the design of HARP is simplicity: we improve the video prediction quality with minimal modification to existing methods. First, for image generation, we employ a high-fidelity image generator, *i.e.,* vector-quantized generative adversarial network (VQ-GAN; Esser et al. 2021). This improves video prediction by enabling high-fidelity image generation (up to 256×256 pixels) on various video datasets. Then a causal transformer model (Chen et al., 2020), which operates on top of discrete latent codes, is trained to predict the discrete codes from VQ-GAN, and autoregressive predictions made by the transformer model are decoded into future frames at inference time.

We highlight the main contributions of this paper below:

- We show that our autoregressive latent video prediction model, HARP, can predict high-resolution (256×256 pixels) frames on robotics dataset (*i.e.,* Meta-World (Yu et al., 2020)) and large-scale real-world robotics dataset (*i.e.,* RoboNet (Dasari et al., 2019)).
- We show that HARP can leverage the image generator pre-trained on ImageNet for training a high-resolution video prediction model on complex, large-scale Kinetics-600 dataset (Carreira et al., 2018).
- HARP achieves competitive or superior performance to prior state-of-the-art video prediction models on widely-used BAIR Robot Pushing (Ebert et al., 2017) and KITTI driving (Geiger et al., 2013) video prediction benchmarks.

2 RELATED WORK

Video prediction. Video prediction aims to predict the future frames conditioned on images (Michalski et al., 2014; Ranzato et al., 2014; Srivastava et al., 2015; Vondrick et al., 2016; Lotter et al., 2017), texts (Wu et al., 2021b), and actions (Oh et al., 2015; Finn et al., 2016), which would be useful for several applications, *e.g.,* model-based RL (Hafner et al., 2019; Kaiser et al., 2020; Hafner et al., 2021; Rybkin et al., 2021; Seo et al., 2022a;b), and simulator development (Kim et al., 2020; 2021). Various video prediction models have been proposed with different approaches, including generative adversarial networks (GANs; Goodfellow et al. 2014) known to generate high-fidelity images by introducing adversarial discriminators that also considers temporal or motion information (Aigner & Körner, 2018; Jang et al., 2018; Kwon & Park, 2019; Clark et al., 2019; Luc et al., 2020; Skorokhodov et al., 2022; Yu et al., 2022), latent video prediction models that operates on the latent space (Babaeizadeh et al., 2018; Denton & Fergus, 2018; Lee et al., 2018; Villegas et al., 2019; Wu et al., 2021a; Babaeizadeh et al., 2021), and autoregressive video prediction models that operates on pixel space by predicting the next pixels in an autoregressive way (Kalchbrenner et al., 2017; Reed et al., 2017; Weissenborn et al., 2020).

Figure 2: Illustration of our approach. We first train a VQ-GAN model that encodes frames into discrete latent codes. Then the discrete codes are flattened following the raster scan order, and a causal transformer model is trained to predict the next discrete codes in an autoregressive manner.

Autoregressive latent video prediction. Most closely related to our work are autoregressive latent video prediction models that separate the video prediction problem into image generation and dynamics learning. Walker et al. (2021) proposed to learn a hierarchical VQ-VAE (Razavi et al., 2019) that extracts multi-scale hierarchical latents then train SNAIL blocks (Chen et al., 2018) that predict hierarchical latent codes, enabling high-fidelity video prediction. However, this involves a complicated training pipeline and a video-specific architecture, which limits its applicability. As simple alternatives, Rakhimov et al. (2021); Yan et al. (2021; 2022) proposed to first learn a VQ-VAE (Oord et al., 2017) and train a causal transformer with 3D self-attention (Weissenborn et al., 2020) and factorized 2D self-attention (Child et al., 2019), respectively. These approaches, however, are limited in that they only consider low-resolution videos. We instead present a simple high-resolution video prediction method that incorporates the strengths of both prior approaches.

3 PRELIMINARIES

We aim to learn a video prediction model that predicts the future frames $\mathbf{x}_{c:T} = (\mathbf{x}_c, ..., \mathbf{x}_{T-1})$ conditioned on the first c frames of a video $\mathbf{x}_{<c} = (\mathbf{x}_0, \mathbf{x}_1, ..., \mathbf{x}_{c-1})$, where $\mathbf{x}_t \in \mathbb{R}^{H \times W \times N_{ch}}$ is the frame at timestep t. Optionally, one can also consider conditioning the prediction model on actions $\mathbf{a} = (\mathbf{a}_0, ..., \mathbf{a}_{T-1})$ that the agents would take.

3.1 AUTOREGRESSIVE VIDEO PREDICTION MODEL

Autoregressive video prediction model (Weissenborn et al., 2020) approximates the distribution of a video in a pixel-channel space. Given a video $\mathbf{x} \in \mathbb{R}^{T \times H \times W \times N_{ch}}$, the joint distribution over pixels conditioned on the first c frames is modelled as the product of channel intensities N_{ch} and all $N_p = T \cdot H \cdot W$ pixels except $N_c = c \cdot H \cdot W$ pixels of conditioning frames:

$$p(\mathbf{x}_{c:T} \,|\, \mathbf{x}_{<c}) = \prod_{i=N_c-1}^{N_p-1} \prod_{k=0}^{N_{ch}-1} p(\mathbf{x}_{\pi(i)}^k | \mathbf{x}_{\pi(<i)}, \mathbf{x}_{\pi(i)}^{<k}), \tag{1}$$

where π is a raster-scan ordering over all pixels from the video (we refer to Weissenborn et al. (2020) for more details), $\mathbf{x}_{\pi(<i)}$ is all pixels before $\mathbf{x}_{\pi(i)}$, $\mathbf{x}_{\pi(i)}^k$ is k-th channel intensity of the pixel $\mathbf{x}_{\pi(i)}$, and $\mathbf{x}_{\pi(i)}^{<k}$ is all channel intensities before $\mathbf{x}_{\pi(i)}^k$.

3.2 VECTOR QUANTIZED VARIATIONAL AUTOENCODER

VQ-VAE (Oord et al., 2017) consists of an encoder that compresses images into discrete representations, and a decoder that reconstructs images from these discrete representations. Formally, given an image $x \in \mathbb{R}^{H \times W \times N_{ch}}$, the encoder E encodes x into a feature map $z_e(x) \in \mathbb{R}^{H' \times W' \times N_z}$ consisting of a series of latent vectors $z_{\pi'(i)}(x) \in \mathbb{R}^{N_z}$, where π' is a raster-scan ordering of the feature map

$z_e(x)$ of size $|\pi'| = H' \cdot W'$. Then $z_e(x)$ is quantized to discrete representations $z_q(x) \in \mathbb{R}^{|\pi'| \times N_z}$ based on the distance of latent vectors $z_{\pi'(i)}(x)$ to the prototype vectors in a codebook $C = \{e_k\}_{k=1}^K$ as follows:

$$z_q(x) = (e_{q(x,1)}, e_{q(x,2)}, \cdots, e_{q(x,|\pi'|)}),$$
$$\text{where } q(x,i) = \text{argmin}_{k \in [K]} \|z_{\pi'(i)}(x) - e_k\|_2, \tag{2}$$

where $[K]$ is the set $\{1, \cdots, K\}$. Then the decoder G learns to reconstruct x from discrete representations $z_q(x)$. The VQ-VAE is trained by minimizing the following objective:

$$\mathcal{L}_{\texttt{VQVAE}}(x) = \underbrace{\|x - G(z_q(x))\|_2^2}_{\mathcal{L}_{\texttt{recon}}} + \underbrace{\|sg\,[z_e(x)] - z_q(x)\|_2^2}_{\mathcal{L}_{\texttt{codebook}}} + \underbrace{\beta \cdot \|sg\,[z_q(x)] - z_e(x)\|_2^2}_{\mathcal{L}_{\texttt{commit}}}, \tag{3}$$

where the operator sg refers to a stop-gradient operator, $\mathcal{L}_{\texttt{recon}}$ is a reconstruction loss for learning representations useful for reconstructing images, $\mathcal{L}_{\texttt{codebook}}$ is a codebook loss to bring codebook representations closer to corresponding encoder outputs h, and $\mathcal{L}_{\texttt{commit}}$ is a commitment loss weighted by β to prevent encoder outputs from fluctuating frequently between different representations.

3.3 VECTOR QUANTIZED GENERATIVE ADVERSARIAL NETWORK

VQ-GAN (Esser et al., 2021) is a variant of VQ-VAE that (a) replaces the $\mathcal{L}_{\texttt{recon}}$ in (3) by a perceptual loss $\mathcal{L}_{\texttt{LPIPS}}$ (Zhang et al., 2018), and (b) introduces an adversarial training scheme where a patch-level discriminator D (Isola et al., 2017) is trained to discriminate real and generated images by maximizing following loss:

$$\mathcal{L}_{\texttt{GAN}}(x) = [\log D(x) + \log(1 - D(G(z_q(x))))]. \tag{4}$$

Then, the objective is given as below:

$$\min_{E,G,C} \max_D \mathbb{E}_{x \sim p(x)} \big[\big(\mathcal{L}_{\texttt{LPIPS}} + \mathcal{L}_{\texttt{codebook}} + \mathcal{L}_{\texttt{commit}} \big) + \lambda \cdot \mathcal{L}_{\texttt{GAN}} \big], \tag{5}$$

where $\lambda = \frac{\nabla_{G_L}[\mathcal{L}_{\texttt{LPIPS}}]}{\nabla_{G_L}[\mathcal{L}_{\texttt{GAN}}] + \delta}$ is an adaptive weight, ∇_{G_L} is the gradient of the inputs to the last layer of the decoder G_L, and $\delta = 10^{-6}$ is a scalar introduced for numerical stability.

4 METHOD

We present HARP, a video prediction model capable of predicting high-fidelity future frames. Our method is designed to fully exploit the benefit of autoregressive latent video prediction model that separates the video prediction into image generation and dynamics learning. The full architecture of HARP is illustrated in Figure 2.

4.1 HIGH-FIDELITY IMAGE GENERATOR

We utilize the VQ-GAN model (Esser et al., 2021) that has proven to be effective for high-resolution image generation as our image generator (see Section 3 for the formulation of VQ-GAN). Specifically, we first pre-train the image generator then freeze the model throughout training to improve the efficiency of learning video prediction models. The notable difference to a prior work that utilize 3D convolutions to temporally downsample the video for efficiency (Yan et al., 2021) is that our image generator operates on single images; hence our image generator solely focus on improving the quality of generated images. Importantly, this enables us to utilize the VQ-GAN model pre-trained on a wide range of natural images, *e.g.,* ImageNet, without training the image generator on the target datasets, which can significantly reduce the training cost of high-resolution video prediction model.

4.2 AUTOREGRESSIVE LATENT VIDEO PREDICTION MODEL

To leverage the VQ-GAN model for video prediction, we utilize the autoregressive latent video prediction architecture that operates on top of the discrete codes. Specifically, we extract the discrete codes $\mathbf{z}(\mathbf{x}) = (\mathbf{z}(\mathbf{x}_1), ..., \mathbf{z}(\mathbf{x}_T))$ using the pre-trained VQ-GAN, where $\mathbf{z}(\mathbf{x}_t) =$

(a) RoboNet

(b) Kinetics-600

Figure 3: 256×256 future frames predicted by HARP trained on (a) RoboNet (Dasari et al., 2019) and (b) Kinetics-600 (Carreira et al., 2018) datasets.

$\left(q_{(\mathbf{x}_t,1)}, q_{(\mathbf{x}_t,2)}, ..., q_{(\mathbf{x}_t,|\pi'|)}\right)$ is the discrete code extracted from the frame $\mathbf{x}_t$ as in (2). Then, instead of modelling the distribution of video $p(\mathbf{x})$ in the pixel-channel space as in (1), we learn the distribution of the video in the discrete latent representation space:

$$p(\mathbf{z}(\mathbf{x}_{c:T}|\mathbf{x}_{<c})) = \prod_{i=0}^{N_d-1} p(\mathbf{z}_{\pi'(i)}(\mathbf{x})|\mathbf{z}_{\pi'(<i)}(\mathbf{x})), \tag{6}$$

where $N_d = (T - C) \cdot H' \cdot W'$ is the total number of codes from $\mathbf{x}_{c:T}$. Due to its simplicity, we utilize the causal transformer architecture (Yan et al., 2021) where the output logits from input codes are trained to predict the next discrete codes.

4.3 ADDITIONAL TECHNIQUES

Top-k sampling. To improve the video prediction quality of latent autoregressive models whose outputs are sampled from the probability distribution over a large number of discrete codes, we utilize the top-k sampling (Fan et al., 2018) that randomly samples the output from the top-k probable discrete codes. By preventing the model from sampling rare discrete codes from the long-tail of a probability distribution and predicting future frames conditioned on such discrete codes, we find that top-k sampling improves video prediction quality, especially given that the number of discrete encodings required for future prediction is very large, *e.g.,* 2,560 on RoboNet (Dasari et al., 2019) up to 6,400 on KITTI dataset (Geiger et al., 2013) in our experimental setup.

Data augmentation. We also investigate how data augmentation can be useful for improving the performance of autoregressive latent video prediction models. Since the image generator model is not trained with augmentation, we utilize a weak augmentation to avoid the instability coming from aggressive transformation of input frames, *i.e.,* translation augmentation that moves the input images by m pixels along the X or Y direction.

5 EXPERIMENTS

We design our experiments to investigate the following:

- Can HARP predict high-resolution future frames (up to 256×256 pixels) on various video datasets with different characteristics?
- How does HARP compare to state-of-the-art methods with large end-to-end networks on standard video prediction benchmarks in terms of quantitative evaluation?
- How does the proposed techniques affect the performance of HARP?

Table 1: Quantitative evaluation on (a) BAIR Robot Pushing (Ebert et al., 2017) and (b) KITTI driving dataset (Geiger et al., 2013). We observe that HARP can achieve competitive performance to state-of-the-art methods with large end-to-end networks on these benchmarks.

(a) BAIR Robot Pushing

Method[b]	Params	FVD ($\downarrow$)
LVT	50M	125.8
SAVP	53M	116.4
DVD-GAN-FP	—[†]	109.8
VideoGPT	82M	103.3
TrIVD-GAN-FP	—[†]	103.3
Video Transformer	373M	94.0
FitVid	302M	**93.6**
HARP (ours)	89M	99.3

(b) KITTI

Method[4]	Params	FVD ($\downarrow$)	LPIPS ($\downarrow$)
SVG	298M	1217.3	0.327
GHVAE	599M	552.9	0.286
FitVid	302M	884.5	0.217
HARP (ours)	89M	**482.9**	**0.191**

[†] Not available

5.1 HIGH-RESOLUTION VIDEO PREDICTION

Implementation. We utilize up to 8 Nvidia 2080Ti GPU and 20 CPU cores for training each model. For training VQ-GAN (Esser et al., 2021), we first train the model without a discriminator loss $\mathcal{L}_{\texttt{GAN}}$, and then continue the training with the loss following the suggestion of the authors. For all experiments, VQ-GAN downsamples each frame into 16×16 latent codes, *i.e.,* by a factor of 4 for frames of size 64×64 frames, and 16 for frames of size 256×256. For training a transformer model, the VQ-GAN model is frozen so that its parameters are not updated. We use Sparse Transformers (Child et al., 2019) as our transformer architecture to accelerate the training. For hyperparameterse, we use $k = 10$ for sampling at inference time.

Setup. For all experiments, VQ-GAN downsamples each frame into 16×16 latent codes, *i.e.,* by a factor of 4 for frames of size 64×64 frames, and 16 for frames of size 256×256. For training a transformer model, the VQ-GAN model is frozen so that its parameters are not updated. As for hyperparameter, we use $k = 10$ for sampling at inference time, but no data augmentation for high-resolution video prediction experiments. We investigate how our model works on large-scale real-world RoboNet dataset (Dasari et al., 2019) consisting of more than 15 million frames, and Kinetics-600 dataset consisting of more than 400,000 videos, which require a large amount of computing resources for training even on 64×64 resolution (Babaeizadeh et al., 2021; Clark et al., 2019). For RoboNet experiments, we first train a VQ-GAN model, and then train a 12-layer causal transformer model that predicts future 10 frames conditioned on first two frames and future ten actions. For Kinetics-600 dataset, to avoid the prohibitively expensive training cost of high-resolution video prediction models on this dataset and fully exploit the benefit of employing a high-fidelity image generator, we utilize the ImageNet pre-trained VQ-GAN model. As we train the transformer model only for autoregressive prediction, this enables us to train a video prediction model in a very efficient manner.

Results. First, we provide the predicted frames on the held-out test video of RoboNet dataset in Figure 3a, where the model predicts the high-resolution future frames where a robot arm is moving around various objects of different colors and shapes. Furthermore, Figure 3b shows that Kinetics-600 pre-trained model can also predict future frames on the test natural videos[c], which demonstrates that leveraging the large image generator pre-trained on a wide range of natural images can be a promising recipe for efficient video prediction on large-scale video datasets.

---

[b]Baselines are SVG (Villegas et al., 2019), GHVAE (Wu et al., 2021a), FitVid (Babaeizadeh et al., 2021), LVT (Rakhimov et al., 2021), SAVP (Lee et al., 2018), DVD-GAN-FP (Clark et al., 2019), VideoGPT (Yan et al., 2021), TrIVD-GAN-FP (Luc et al., 2020), and Video Transformer (Weissenborn et al., 2020).

[c]Videos with CC-BY license: Figure 3b top and bottom

Table 2: FVD scores of HARP with varying (a) the number of codes to use for top-k sampling, (b) number of layers, and (c) magnitude m of data augmentation.

(a) Effects of k

Dataset	k	FVD ($\downarrow$)
BAIR	No top-k	104.4
	100	103.6
	10	**99.3**
KITTI	No top-k	578.1
	100	557.7
	10	**482.9**

(b) Effects of layers

Dataset	Layers	FVD ($\downarrow$)
BAIR	6	111.8
	12	**99.3**
KITTI	6	520.1
	12	**482.9**

(c) Effects of m

Dataset	m	FVD ($\downarrow$)
KITTI	0	980.1
	2	497.0
	4	**482.9**
	8	523.4

5.2 COMPARATIVE EVALUATION ON STANDARD BENCHMARKS

Setup. For quantitative evaluation, we first consider the BAIR robot pushing dataset (Ebert et al., 2017) consisting of roughly 40k training and 256 test videos. Following the setup in prior work (Yan et al., 2021), we predict 15 future frames conditioned on one frame. We also evaluate our method on KITTI driving dataset (Geiger et al., 2013), where the training and test datasets are split by following the setup in Villegas et al. (2019). Specifically, the test dataset consists of 148 video clips constructed by extracting 30-frame clips and skipping every 5 frames, and the model is trained to predict future ten frames conditioned on five frames and evaluated to predict future 25 frames conditioned on five frames. For hyperparameters, We use $k = 10$ for both datasets and data augmentation with $m = 4$ is only applied to KITTI as there was no sign of overfitting on BAIR dataset. For evaluation metrics, we use LPIPS (Zhang et al., 2018) and FVD (Unterthiner et al., 2018), computed using 100 future videos for each ground-truth test video, then reports the best score over 100 videos for LPIPS, and all videos for FVD, following Babaeizadeh et al. (2021); Villegas et al. (2019).

Results. Table 1 shows the performances of our method and baselines on test sets of BAIR Robot Pushing and KITTI driving dataset. We observe that our model achieves competitive or superior performance to state-of-the-art methods with large end-to-end networks, *e.g.,* HARP outperforms FitVid with 302M parameters on KITTI driving dataset. Our model successfully extrapolates to unseen number of future frames (*i.e.,* 25) instead of 10 future frames used in training on KITTI dataset. This implies that transformer-based video prediction models can also predict arbitrary number of frames at inference time. In the case of BAIR dataset, HARP achieves the similar performance of FitVid with 302M parameters, even though our method only requires 89M parameters.

Analysis. We investigate how the top-k sampling, number of layers, and magnitude m of data augmentation affect the performance. Table 2a shows that smaller k leads to better performance, implying that the proposed top-k sampling is effective for improving the performance by discarding rare discrete codes that might degrade the prediction quality at inference time. As shown in Table 2b, we observe that more layers leads to better performance on BAIR dataset, which implies our model can be further improved by scaling up the networks. Finally, we find that (i) data augmentation on KITTI dataset is important for achieving strong performance, similar to the observation of Babaeizadeh et al. (2021), and (ii) too aggressive augmentation leads to worse performance.

6 DISCUSSION

In this work, we present HARP that employs a high-fidelity image generator for predicting high-resolution future frames, and achieves competitive performance to state-of-the-art video prediction methods with large end-to-end networks. We also demonstrate that HARP can leverage the image generator pre-trained on a wide range of natural images for video prediction, similar to the approach in the context of video synthesis (Tian et al., 2021). We hope this work inspires more investigation into leveraging recently developed pre-trained image generators (Oord et al., 2017; Chen et al., 2020; Esser et al., 2021) for high-fidelity video prediction.

| | (a) RoboNet | (b) Kinetics-600 |

Figure 4: Failure cases in our experiments. (a) Interaction with the objects is ignored. (b) The model repeats the first frame while a person is moving right in the ground-truth frames.

Finally, we report the failure cases of video prediction with HARP and discuss the possible extensions to resolve the issue. A common failure case for video prediction on RoboNet dataset is ignoring the interaction between a robot arm and objects. For example, in Figure 4a, our model ignores the objects and only predicts the movement of a robot arm. On the other hand, common failure case for Kinetics-600 is a degenerate video prediction, where a model just repeats the conditioning frame without predicting the future, as shown in Figure 4b. These failure cases might be resolved by training more larger networks similar to the observation in the field of natural language processing, *e.g.,* GPT-3 (Brown et al., 2020), or might necessitate a new architecture for addressing the complexity of training autoregressive latent prediction models on video datasets.

7 ACKNOWLEDGEMENTS

REFERENCES

Sandra Aigner and Marco Körner. Futuregan: Anticipating the future frames of video sequences using spatio-temporal 3d convolutions in progressively growing gans. *arXiv preprint arXiv:1810.01325*, 2018.

Mohammad Babaeizadeh, Chelsea Finn, Dumitru Erhan, Roy H Campbell, and Sergey Levine. Stochastic variational video prediction. In *International Conference on Learning Representations*, 2018.

Mohammad Babaeizadeh, Mohammad Taghi Saffar, Suraj Nair, Sergey Levine, Chelsea Finn, and Dumitru Erhan. Fitvid: Overfitting in pixel-level video prediction. *arXiv preprint arXiv:2106.13195*, 2021.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

Joao Carreira, Eric Noland, Andras Banki-Horvath, Chloe Hillier, and Andrew Zisserman. A short note about kinetics-600. *arXiv preprint arXiv:1808.01340*, 2018.

Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In *International Conference on Machine Learning*, 2020.

---

[d]https://cirrascale.com

Xi Chen, Nikhil Mishra, Mostafa Rohaninejad, and Pieter Abbeel. Pixelsnail: An improved autoregressive generative model. In *International Conference on Machine Learning*, 2018.

Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. Generating long sequences with sparse transformers. *arXiv preprint arXiv:1904.10509*, 2019.

Aidan Clark, Jeff Donahue, and Karen Simonyan. Adversarial video generation on complex datasets. *arXiv preprint arXiv:1907.06571*, 2019.

Sudeep Dasari, Frederik Ebert, Stephen Tian, Suraj Nair, Bernadette Bucher, Karl Schmeckpeper, Siddharth Singh, Sergey Levine, and Chelsea Finn. Robonet: Large-scale multi-robot learning. In *Conference on Robot Learning*, 2019.

Emily Denton and Rob Fergus. Stochastic video generation with a learned prior. In *International Conference on Machine Learning*, 2018.

Frederik Ebert, Chelsea Finn, Alex X Lee, and Sergey Levine. Self-supervised visual planning with temporal skip connections. In *Conference on Robot Learning*, 2017.

Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

Angela Fan, Mike Lewis, and Yann Dauphin. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, 2018.

Chelsea Finn and Sergey Levine. Deep visual foresight for planning robot motion. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2017.

Chelsea Finn, Ian Goodfellow, and Sergey Levine. Unsupervised learning for physical interaction through video prediction. In *Advances in neural information processing systems*, 2016.

Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 2013.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, 2014.

Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In *International Conference on Machine Learning*, 2019.

Danijar Hafner, Timothy Lillicrap, Mohammad Norouzi, and Jimmy Ba. Mastering atari with discrete world models. In *International Conference on Learning Representations*, 2021.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017.

Yunseok Jang, Gunhee Kim, and Yale Song. Video prediction with appearance and motion conditions. In *International Conference on Machine Learning*, 2018.

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. In *International Conference on Learning Representations*, 2020.

Dmitry Kalashnikov, Alex Irpan, Peter Pastor, Julian Ibarz, Alexander Herzog, Eric Jang, Deirdre Quillen, Ethan Holly, Mrinal Kalakrishnan, Vincent Vanhoucke, et al. Qt-opt: Scalable deep reinforcement learning for vision-based robotic manipulation. In *Conference on Robot Learning*, 2018.

Nal Kalchbrenner, Aäron Oord, Karen Simonyan, Ivo Danihelka, Oriol Vinyals, Alex Graves, and Koray Kavukcuoglu. Video pixel networks. In *International Conference on Machine Learning*, 2017.

Seung Wook Kim, Yuhao Zhou, Jonah Philion, Antonio Torralba, and Sanja Fidler. Learning to simulate dynamic environments with gamegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

Seung Wook Kim, Jonah Philion, Antonio Torralba, and Sanja Fidler. Drivegan: Towards a controllable high-quality neural simulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

Yong-Hoon Kwon and Min-Gyu Park. Predicting future frames using retrospective cycle gan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

Alex X Lee, Richard Zhang, Frederik Ebert, Pieter Abbeel, Chelsea Finn, and Sergey Levine. Stochastic adversarial video prediction. *arXiv preprint arXiv:1804.01523*, 2018.

William Lotter, Gabriel Kreiman, and David Cox. Deep predictive coding networks for video prediction and unsupervised learning. In *International Conference on Learning Representations*, 2017.

Pauline Luc, Aidan Clark, Sander Dieleman, Diego de Las Casas, Yotam Doron, Albin Cassirer, and Karen Simonyan. Transformation-based adversarial video prediction on large-scale data. *arXiv preprint arXiv:2003.04035*, 2020.

Vincent Michalski, Roland Memisevic, and Kishore Konda. Modeling deep temporal dependencies with recurrent grammar cells. In *Advances in Neural Information Processing Systems*, 2014.

Junhyuk Oh, Xiaoxiao Guo, Honglak Lee, Richard Lewis, and Satinder Singh. Action-conditional video prediction using deep networks in atari games. In *Advances in Neural Information Processing Systems*, 2015.

Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *Advances in Neural Information Processing Systems*, 2017.

Ruslan Rakhimov, Denis Volkhonskiy, Alexey Artemov, Denis Zorin, and Evgeny Burnaev. Latent video transformer. In *International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, 2021.

MarcAurelio Ranzato, Arthur Szlam, Joan Bruna, Michael Mathieu, Ronan Collobert, and Sumit Chopra. Video (language) modeling: a baseline for generative models of natural videos. *arXiv preprint arXiv:1412.6604*, 2014.

Ali Razavi, Aaron van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. In *Advances in Neural Information Processing Systems*, 2019.

Scott Reed, Aäron Oord, Nal Kalchbrenner, Sergio Gómez Colmenarejo, Ziyu Wang, Yutian Chen, Dan Belov, and Nando Freitas. Parallel multiscale autoregressive density estimation. In *International Conference on Machine Learning*, 2017.

Oleh Rybkin, Chuning Zhu, Anusha Nagabandi, Kostas Daniilidis, Igor Mordatch, and Sergey Levine. Model-based reinforcement learning via latent-space collocation. In *International Conference on Machine Learning*, 2021.

Younggyo Seo, Danijar Hafner, Hao Liu, Fangchen Liu, Stephen James, Kimin Lee, and Pieter Abbeel. Masked world models for visual control. *arXiv preprint arXiv:2206.14244*, 2022a.

Younggyo Seo, Kimin Lee, Stephen L James, and Pieter Abbeel. Reinforcement learning with action-free pre-training from videos. In *International Conference on Machine Learning*, 2022b.

Ivan Skorokhodov, Sergey Tulyakov, and Mohamed Elhoseiny. Stylegan-v: A continuous video generator with the price, image quality and perks of stylegan2. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Nitish Srivastava, Elman Mansimov, and Ruslan Salakhudinov. Unsupervised learning of video representations using lstms. In *International Conference on Machine Learning*, 2015.

Yu Tian, Jian Ren, Menglei Chai, Kyle Olszewski, Xi Peng, Dimitris N Metaxas, and Sergey Tulyakov. A good image generator is what you need for high-resolution video synthesis. In *International Conference on Learning Representations*, 2021.

Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018.

Ruben Villegas, Arkanath Pathak, Harini Kannan, Dumitru Erhan, Quoc V Le, and Honglak Lee. High fidelity video prediction with large stochastic recurrent neural networks. *Advances in Neural Information Processing Systems*, 2019.

Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Generating videos with scene dynamics. In *Advances in Neural Information Processing Systems*, 2016.

Jacob Walker, Ali Razavi, and Aäron van den Oord. Predicting video with vqvae. *arXiv preprint arXiv:2103.01950*, 2021.

Dirk Weissenborn, Oscar Täckström, and Jakob Uszkoreit. Scaling autoregressive video models. In *International Conference on Learning Representations*, 2020.

Bohan Wu, Suraj Nair, Roberto Martin-Martin, Li Fei-Fei, and Chelsea Finn. Greedy hierarchical variational autoencoders for large-scale video prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021a.

Chenfei Wu, Lun Huang, Qianxi Zhang, Binyang Li, Lei Ji, Fan Yang, Guillermo Sapiro, and Nan Duan. Godiva: Generating open-domain videos from natural descriptions. *arXiv preprint arXiv:2104.14806*, 2021b.

Huazhe Xu, Yang Gao, Fisher Yu, and Trevor Darrell. End-to-end learning of driving models from large-scale video datasets. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017.

Wilson Yan, Yunzhi Zhang, Pieter Abbeel, and Aravind Srinivas. Videogpt: Video generation using vq-vae and transformers. *arXiv preprint arXiv:2104.10157*, 2021.

Wilson Yan, Ryo Okumura, Stephen James, and Pieter Abbeel. Patch-based object-centric transformers for efficient video generation. *arXiv preprint arXiv:2206.04003*, 2022.

Sihyun Yu, Jihoon Tack, Sangwoo Mo, Hyunsu Kim, Junho Kim, Jung-Woo Ha, and Jinwoo Shin. Generating videos with dynamics-aware implicit generative adversarial networks. In *International Conference on Learning Representations*, 2022.

Tianhe Yu, Deirdre Quillen, Zhanpeng He, Ryan Julian, Karol Hausman, Chelsea Finn, and Sergey Levine. Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning. In *Conference on Robot Learning*, 2020.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018.