Markov Decision Process for Video Generation

Vladyslav Yushchenko[1*] Nikita Araslanov[2] Stefan Roth[2]
[1]iNTENCE automotive electronics GmbH, [2]TU Darmstadt

Abstract

We identify two pathological cases of temporal inconsistencies in video generation: video freezing and video looping. To better quantify the temporal diversity, we propose a class of complementary metrics that are effective, easy to implement, data agnostic, and interpretable. Further, we observe that current state-of-the-art models are trained on video samples of fixed length thereby inhibiting long-term modeling. To address this, we reformulate the problem of video generation as a Markov Decision Process (MDP). The underlying idea is to represent motion as a stochastic process with an infinite forecast horizon to overcome the fixed length limitation and to mitigate the presence of temporal artifacts. We show that our formulation is easy to integrate into the state-of-the-art MoCoGAN framework. Our experiments on the Human Actions and UCF-101 datasets demonstrate that our MDP-based model is more memory efficient and improves the video quality both in terms of the new and established metrics.

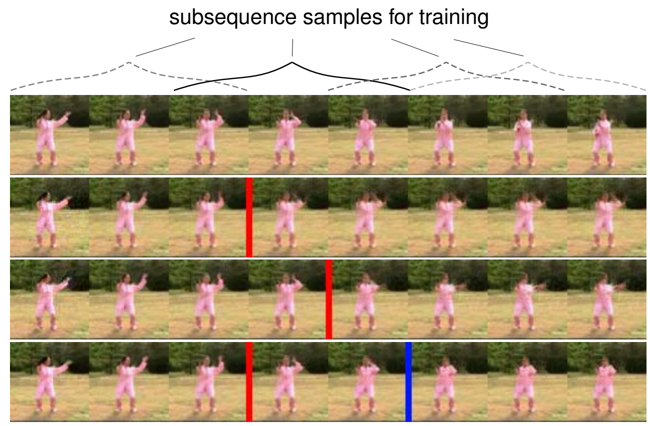


Figure 1. **Problem illustration on a Tai Chi sequence.** Every 6th frame is shown. **Top row**: The ground truth video is a non-repetitive action sequence. **Second row**: Even when trained only on one video, MoCoGAN [29] can only reproduce the sequence until the training length, marked by the red boundary, and the motion freezes thereafter. **Third row**: Increasing the training length comes at increased memory costs and only delays the freezing. **Last row**: Our MDP approach uses shorter training sequences yet extends the movement duration, indicated by the blue boundary.

1. Introduction

Video synthesis is a very challenging problem [12, 17, 19, 29, 34], arguably even more challenging than the already difficult image generation task [5, 11, 20]. The temporal dimension of the data introduces an additional mode of variation, since feasible motions are dependent on the object category and the scene appearance. Consequently, the evaluation of video synthesis methods should account not only for the quality of individual frames but also for their temporal coherence, motion realism, and diversity.

In this work, we take a closer look at the temporal quality of *unconditional* video generators, represented by the state-of-the-art MoCoGAN approach [29]. Note that this subcategory of video generation is different from future frame prediction [13, 16], which takes a number of initial frames as input. We only rely on the training data as input instead.[1]

We find that the common training strategy of sampling a fixed-length video subsequence at training time often leads to degenerate solutions. As illustrated in Fig. 1, the MoCoGAN model exhibits temporal artifacts as soon as the video sequence length at inference time exceeds the length of the temporal window at training time. We establish two common types of such artifacts. If the model continues to predict the last frame without change, we refer to that as *freezing*. On the other hand, *looping* occurs when the exact subsequence of frames is continually repeated.

To address these limitations, we make two main contributions. First, to tackle the detrimental effect of fixed-length video training, we reformulate video generation as a *Markov Decision Process* (MDP). This reformulation allows approximating an infinite forecast horizon in order to optimize every generated frame w.r.t. to its long-term effect on future frames. One benefit of our MDP formulation is that it is model-agnostic. We evaluate it by applying it to the state-of-the-art MoCoGAN [29], which requires only a minor modification of the original design and does not signif-

*This work was done while VY was at TU Darmstadt.

[1]Note that the model is still conditioned on the particular training data distribution, hence not truly "unconditional". Still, we adhere to the common terminology used in the literature.

icantly increase the model capacity. Second, we propose a *family of evaluation metrics* to detect and measure the temporal artifacts. Our new metrics are model-free, simple to implement, and offer an easy interpretation. In contrast to the Inception Score (IS) [20] or the recent Fréchet Video Distance (FVD) [30], the proposed metrics do not require model pre-training and, hence, do not build upon a data-sensitive prior. Our experiments show that our MDP-based formulation leads to a consistent improvement of the video quality, both in terms of the artifact mitigation as well as on the more common metrics, the IS and FVD scores.

2. Related Work

Video generation models can be divided into two main categories: *conditional* and *unconditional*. Exemplified by the task of future frame prediction, conditional models historically preceded the latter and some of their features lend themselves to unconditional prediction. Therefore, we first give a brief overview of conditional approaches.

Conditional video generation. One of the first network-based models for motion dynamics used a temporal extension of Restricted Boltzmann Machines (RBMs) [24, 27] with a focus on resolving the intractable inference [25]. The increasing volume of video data for deep learning shifted the attention to learning suitable representations and enabling some control over the generated frames [6]. Srivastava *et al*. [23] show that unsupervised sequence-to-sequence pre-training with LSTMs [8] enhances the performance on the supervised frame prediction task. Patch-based quantization of the output space [18] or predicting pixel motion [4, 15] can improve the frame appearance at larger resolutions. In contrast, Kalchbrenner *et al*. [9] predict pixel-wise intensities and extend the context model of PixelCNNs [31] to the temporal domain. A coarse-to-fine strategy allows to decouple the structure from the appearance [32, 33], or dedicate individual stages of a pipeline to multiple scales [16].

The frames of a distant future cannot be extrapolated deterministically due to the stochastic nature of the problem [1, 13, 36] (*i.e*. there are multiple feasible futures for a given initial frame). In practice, this manifests itself through frame blurring – a gradual loss of details in the frame. To alleviate this effect, Mathieu *et al*. [16] used an adversarial loss [5]. Liang *et al*. [14] further show that adversarial learning of the pixel flows leads to better generalisation.

Unconditional video generation. These more recent methods are based on the GAN framework [5] and incorporate some of the insights from their conditional counterparts. For example, Vondrick *et al*. [34] decouple the active foreground from a static background by using an architecture with two parallel generator streams. Saito *et al*. [19] use two generators to disentangle the video representation into distinct temporal and spatial domains. Following [32], the state-of-the-art MoCoGAN of Tulyakov *et al*. [29] decomposes the latent representation into content and motion parts for finer control over the generated scene. In addition, the discriminator in the MoCoGAN model is separated into image and video modules. While the image module targets the visual quality of individual frames, the focus of the video discriminator is the temporal coherence.

Evaluating unconditional video generators. Borrowed from the image generation literature [20], the Inception Score (IS) has become one of the established metrics for quality assessment in videos [19, 29, 34]. IS incorporates the entropy of the class distributions obtained from a separately trained classifier. Therefore, it is only meaningful if the training data distribution of the classifier matches the one on which it will be evaluated later. Following [7], Unterthiner *et al*. [30] recently proposed the Fréchet Video Distance (FVD) that compares the distributions of *feature embeddings* of real and generated data.

However, these metrics provide only a holistic measure of the video quality and do not allow for a detailed assessment of its individual properties. One of the desirable qualitative traits of video generators is their *ability to produce realistic videos of arbitrary length*. Yet, the established experimental protocol evaluates only on video sequences of a fixed length. Indeed, some previous work [19, 34] is even tailored to a pre-defined video length, both at training and at inference time.

3. MDP for Video Generation

To motivate MDP for video generation, we first review MoCoGAN [29] and discuss its limitations. After a short presentation of the MDP formalism (*c.f*. [26] for a comprehensive introduction), we then integrate MDP into MoCoGAN to incorporate knowledge of the infinite-time horizon into the generative process.

3.1. Preliminaries

MoCoGAN. Figure 2a illustrates the main components of MoCoGAN: the generator, the image discriminator, and the video discriminator. At every timestep, the stochastic generator G emits one frame x_t and maintains a recurrent state h_t perturbed by random noise. The image discriminator D_I provides feedback for a single image; the video discriminator D_V evaluates a contiguous subsequence of frames $\mathbf{x}_t$ of a pre-defined length $|\mathbf{x}_t| = K$. The training objective is specified by the familiar max-min game

$$\max_{G} \min_{D_I, D_V} \mathop{\mathbb{E}}_{x_t, \mathbf{x}_t} \Big[\mathcal{L}_I(x_t^{\text{real}}, x_t^{\text{fake}}) + \mathcal{L}_V(\mathbf{x}_t^{\text{real}}, \mathbf{x}_t^{\text{fake}}) \Big], \quad (1)$$

where x_t^{real} and $\mathbf{x}_t^{\text{real}}$ are samples from the training data, the generator provides x_t^{fake} and $\mathbf{x}_t^{\text{fake}}$, and $\mathcal{L}_I$ and $\mathcal{L}_V$ are defined by the scalar scores of D_I and D_V [5, 29].

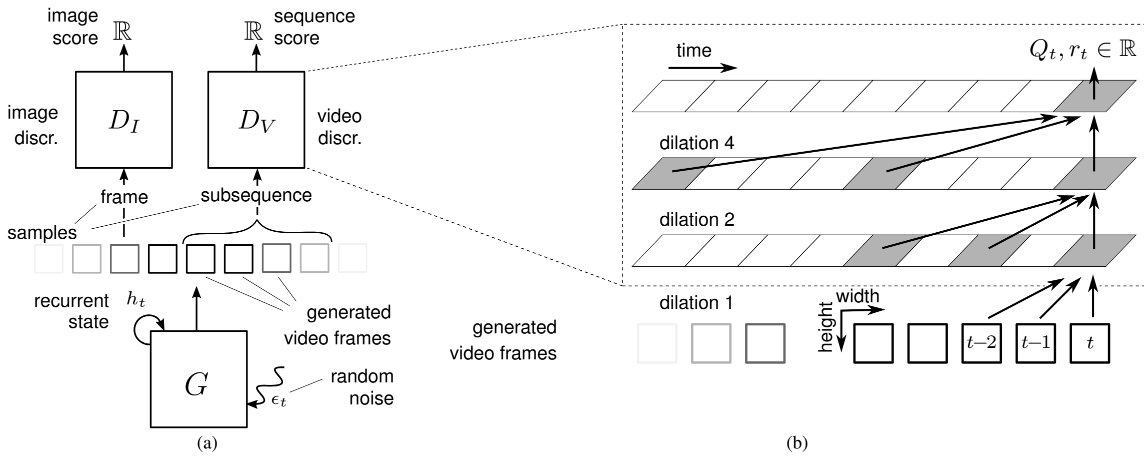


Figure 2. **The original MoCoGAN architecture** (a) and our **proposed modification** of D_V for modeling the MDP (b). Our MDP reformulation follows the TCN design [2]: a sequence of 3D-convolutional layers with layer-specific dilations and strides. The input to the next convolutional layer is the output of the previous one. The last layer produces the $\{r_t\}_{1\leq t\leq K}$ and the $\{Q_t\}_{1\leq t\leq K}$, *i.e.* the Q-values are produced by the same network, D_V.

We find that MoCoGAN's samples exhibit looping and freezing patterns (see Sec. 5.4 for results and analyses). The intuitive reason comes from the specifics of training: to save memory, the training samples contain only *subsequences* of the complete video. As a result, the gradient signal from the video discriminator is unaware of the frames following the subsequence. The predefined length of the subsequence ultimately determines the maximum length of a sample with a non-repeating pattern.[2]

MDP. In an MDP defined by the tuple (S, A, T, π, r), the *agent* interacts with the environment by performing *actions*, $a_t \in A$, based on the current state, $s_t \in S$. The environment specifies the outcome of the action by returning a *reward*, $r(s_t, a_t)$, and the next state, $s_{t+1} = T(s_t, a_t)$. The goal of the agent is to find the optimal policy $\pi^* : S \rightarrow A$, maximizing the discounted cumulative reward

$$\pi^* = \arg\max_{\pi} \sum_{t=0}^{\infty} \gamma^t r(s_t, a_t), \quad a_t \sim \pi(s_t), \qquad (2)$$

where $\gamma \in (0, 1)$ is the *discount factor* to ensure the convergence of the sum.

In the context of an MDP the generator G plays the role of the agent's policy. The frames predicted by G are the actions. The hidden recurrent state h_t becomes the agent's state s_t. The additive noise at every timestep determines the transition function T. A frame incurs a reward r_t as the score provided by the discriminators. Due to the deterministic mapping $s_t \rightarrow a_t$, the MoCoGAN's G corresponds to a deterministic policy [21] (*i.e.* the sampling in Eq. (2) becomes an equality). The optimization task for the agent is a search for the optimal policy π^*:

$$\max_{\pi} \; r(s_t, a_t) + \mathop{\mathbb{E}}_{a=\pi(s_t)} \Bigg[\sum_{i=t+1}^{\infty} \gamma^{i-t} r(s_i, a) \Bigg]. \qquad (3)$$

Observe that the MoCoGAN objective for D_V is equivalent to only the first term of Eq. (3), the immediate reward, since the D_V computes only a single score for a given video sample. In contrast, we also consider the future rewards, *i.e.* the second term of Eq. (3). To this end, we decompose the score of the video generator into immediate rewards associated with individual frames. We then learn a utility Q-function approximating the expected cumulative reward, $\mathbb{E}\big[\sum_t \gamma^t r_t\big]$. Its definition is also known as Bellman's optimality principle:

$$Q(s_t, a_t) = r(s_t, a_t) + \max_{a=\pi(s_t)} Q(s_{t+1}, a). \qquad (4)$$

By training the generator to maximize the Q-function instead of just the immediate reward, we arrive at an approximate solution of Eq. (3). In the next section, we detail how MoCoGAN can be extended to this setup.

3.2. Integrating MDP into MoCoGAN

We need the model implementing the MDP to comply with two requirements:

(a) The *Markov property* needs to be fulfilled, *i.e.* the next state s_{t+1} given the previous state s_t is conditionally independent from the past history $s_{i<t}$.
(b) By *causality*, the immediate reward r_t is a function of the current state s_t and the action a_t and incorporates no knowledge about future actions.

The MoCoGAN generator already satisfies the Markov property using a parametrized RNN mapping from the cur-

[2]To verify this, we also trained the MoCoGAN model on longer subsequences and found the breaking point to occur at a correspondingly later timestep.

rent state to the next. However, the video discriminator has to be modified to satisfy the second requirement. This modification is straightforward to implement and leads to a variant of the Temporal Convolutional Network (TCN) [2].

Figure 2b gives an overview of the proposed MDP-extension for the video discriminator. The key property of this design is that the t^{th} output – a scalar – corresponds to a temporal receptive field of the frames up to the t^{th} timestep. In this way the immediate reward will capture only the relevant motion history. Fortunately, adapting the MoCoGAN video discriminator to this architecture is straightforward (*c.f.* supplemental material for more details).

To implement Eq. (4), alongside r_t we also predict another time-dependent scalar, the Q-value. As discussed in Sec. 3.1, the purpose of the Q-value is to approximate the expected cumulative reward, $\mathbb{E}\big[\sum_{t=0}^{\infty}\gamma^t r(s_t, a_t)\big]$. We use the squared difference loss, defined for each timestep by

$$\mathcal{L}_{Q,t} = \left\| \frac{1}{K-t+1}\sum_{i=t}^{K}\gamma^{i-t}r_i - Q_t \right\|_2^2, \ 1 \leq t \leq K, \quad (5)$$

where $\gamma \in (0, 1)$ is the discounting factor specifying the lookahead span: larger values encourage the Q-value to account for the future outcome far ahead; low values focus the Q-value on the immediate effect of the current frames.

Our TCN-based D_V ensures that the parameters for predicting Q_t are now *shared* for all t. As a result, Eq. (5) forces even the last Q_K to incorporate knowledge of rewards *beyond* the temporal window of size K. Hence, by maximizing Q_K, the generator will implicitly maximize the rewards for $t > K$. Contrast this to the original D_V producing a single score for the complete K-frame sequence: due to lack of causality, the generator is "unaware" that at inference time the requested video length may exceed K.

Note that the definition in Eq. (5) is confined to a limited time window of length K to ensure that the memory consumption remains manageable. Now, our task is to train the generator by maximizing the Q-value incorporating the long-term effects of individual predictions. However, since we keep K fixed, each consecutive Q_t in Eq. (5) will be optimized w.r.t. to the sum containing one term fewer. That is, Q_1 will approximate a sum of K immediate rewards, Q_2 a sum of $K-1$ terms, *etc.* As a result, Q_1 incorporates the effect of the 1st frame on $K-1$ future frames, whereas Q_{K-1} will only observe the influence of the $(K-1)^{\text{th}}$ frame on the last prediction. It is therefore evident that the Q-values are not equally informative for modeling the long-term dependencies as supervision to the generator.

To reflect this observation in our training, we introduce an additional discounting factor $\beta \in [0, 1]$ that shifts the weight of the long-term supervision to the first frames, but offsets the reliance on the Q-value for the last predictions. Concretely, the new term in the generator loss is

$$\mathcal{L}_T = \frac{1}{K}\sum_{t=1}^{K}\beta^t Q_t. \quad (6)$$

To summarize, extending the original MoCoGAN training objective (Eq. 1) into our MDP-based GAN yields

$$\begin{aligned}\min_{D_I, D_V} \mathop{\mathbb{E}}_{x_t, \mathbf{x}_t} \Big[&\mathcal{L}_I(x_t^{\text{real}}, x_t^{\text{fake}}) + \mathcal{L}_V(\mathbf{x}_t^{\text{real}}, \mathbf{x}_t^{\text{fake}}) \\ &+ \sum_{t=1}^{K}\big(\mathcal{L}_{Q,t}(\mathbf{x}_t^{\text{real}}) + \mathcal{L}_{Q,t}(\mathbf{x}_t^{\text{fake}})\big)\Big], \end{aligned} \quad (7a)$$

$$\max_{G} \mathop{\mathbb{E}}_{x_t, \mathbf{x}_t} \big[\mathcal{L}_I(x_t^{\text{fake}}) + \mathcal{L}_V(\mathbf{x}_t^{\text{fake}}) + \mathcal{L}_T\big]. \quad (7b)$$

Here, we split the original objective in Eq. (1) into the discriminator- and generator-specific losses for illustrative purposes although the joint nature of the max-min optimization problem remains. Following standard practice [5], we optimize the new objective by alternately updating the discriminators using Eq. (7a) and the generator using Eq. (7b).

4. Quantifying Temporal Diversity

Motivated by our observation of the looping and freezing artifacts (see Fig. 1), we propose an interpretable way to quantify the temporal diversity of the video. Here, our assumption is that realistic videos comprise a predominantly unique sequence of frames. The idea then is to compare the predicted frame to the preceding ones: if there is a match, this indicates a re-occurring pattern in the sequence.

Let $X = (x_t)_{t=1..N}$ be a sequence of frames predicted by the model. Our diversity measure relies on a distance function of choice between arbitrary frames $d(x_i, x_j)$ as

$$t\text{-}d = \frac{1}{N}\sum_{i=2}^{N}\min_{j<i} d(x_i, x_j), \quad (8)$$

where we use prefix "*t*-" for disambiguation. Eq. (8) essentially finds the most similar preceding frame and averages the distance over all such pairs in the sequence. The obvious dual of this metric is to replace the distance function $d(\cdot,\cdot)$ in Eq. (8) with a similarity measure $s(\cdot,\cdot)$ and substitute the $\min$ for the $\max$ operation. In this work, we use two instantiations of Eq. (8): the *t-DSSIM* employs the structural similarity (SSIM) [35] in the distance function $\text{DSSIM} = \frac{1}{2}(1-\text{SSIM})$; *t-PSNR* utilizes the peak signal-to-noise ratio (PSNR) as a similarity measure. Hence, higher *t*-DSSIM and lower *t*-PSNR indicate higher diversity of frames within a sequence. We show next that despite its apparent simplicity, our proposed metric effectively captures deficiencies in frame diversity.

	Configuration	IS ↑	FVD ↓	t-DSSIM ↑	t-PSNR ↓
Tai Chi	Original	1.63 ± 0.05	**115.3 ± 6.9**	**0.013**	**36.50**
	Looping-FWD	**2.03 ± 0.03**	336.7 ± 13.5	0.0062	∞
	Looping-BWD	1.69 ± 0.03	541.7 ± 19.4	0.0062	∞
	Freezing	1.55 ± 0.05	254.4 ± 15.5	0.0062	∞
UCF-101	Original	**40.74 ± 0.20**	472.8 ± 18.5	0.073	27.10
	Original + ϵ	36.69 ± 0.23	**444.8 ± 17.2**	**0.107**	**25.44**
	Looping-FWD	38.59 ± 0.22	597.2 ± 13.5	0.034	∞
	Looping-BWD	35.16 ± 0.78	737.7 ± 40.0	0.034	∞
	Freezing	32.45 ± 0.22	667.3 ± 17.8	0.034	∞

Table 1. **Comparison of IS, FVD, *t*-DSSIM, and *t*-PSNR metrics** for ground-truth videos and videos with purposely crafted artifacts. The Gaussian noise ϵ is drawn from $\mathcal{N}(\mu = 0, \sigma^2 = 0.03)$.

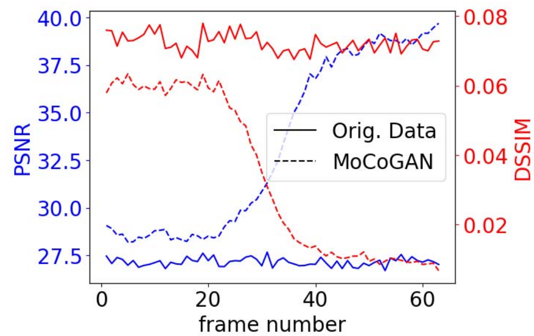


Figure 3. ***t*-PSNR and *t*-DSSIM decomposed as functions of time.** In contrast to the ground truth, the diversity of the MoCoGAN samples vanishes with time.

5. Experiments

5.1. Datasets

Following the established evaluation protocol from previous studies [19, 29], we use the following benchmarks:

(1) **Human actions** [3]: The dataset contains 81 videos of 9 people performing 9 actions, *e.g.* walking, jumping, etc. All videos are extracted with 25 fps and downscaled to 64×64 pixels. We also add a flipped copy of each video sequence to the training set. Following Tulyakov *et al.* [29] we used only 4 action classes, which amounts to 72 videos for training in total.

(2) **UCF-101** [22]: This dataset consists of 13 220 videos with 101 classes of human actions grouped into 5 categories: human-object and human-human interaction, body motion, playing musical instruments, and sports. This dataset is challenging due to a high diversity of scenes, motion dynamics, and viewpoint changes.

(3) **Tai-Chi**: The dataset contains 72 Tai Chi videos taken from the UCF-101 dataset.[3] All videos are centered on the performer and downscaled to 64×64 pixels. We use this dataset for our ablation studies as it has moderate complexity, yet represents real-world motion.

5.2. Overview

We first verify that t-PSNR and t-DSSIM effectively quantify the temporal artifacts. We then employ these metrics to analyze the MoCoGAN model [29] w.r.t. these artifacts. Next, we study the effect of the time-horizon hyper-parameters, γ and β, of our MDP approach. Finally, we validate our approach on the Human Actions dataset and on the more challenging UCF-101 dataset. We compare our model to TGAN [19] and MoCoGAN, where we find a consistent improvement of the temporal diversity over the baseline.

[3]Note that the Tai Chi subset used in the evaluation of MoCoGAN [29] is not publicly available and could not be obtained due to licensing restrictions.

We compute the IS following Saito *et al.* [19], who trained the C3D network [28] on the Sports-1M dataset [10] and then further finetuned on UCF-101 [22]. For FVD we use the original implementation by Unterthiner *et al.* [30]. To manage computational time, we calculate the FVD for the first 16 frames, sampled from 256 videos, and derive the FVD mean and variance from 4 trials, similar to IS.

5.3. Metric evaluation

We design a set of proof-of-concept experiments to study the properties of the newly introduced t-PSNR and t-DSSIM. Concretely, we synthesize the looping and freezing patterns in the ground-truth videos from UCF-101 and Tai Chi. We construct 16 frames by sampling 8 frames directly from the dataset and completing the sequence with an artifact counterpart. `Looping-FWD` contains a repeating subsequence from the original video (`Original`), whereas `Looping-BWD` reverses the frame order. The size of the re-occurring subsequence in `Freezing` is one. To put the results in context, we also compare to the mainstream IS as well as the recent FVD scores and study the robustness of all metrics to additive Gaussian noise $\epsilon \sim \mathcal{N}(\mu, \sigma^2)$. The results are summarized in Table 1.

We observe that t-PSNR and t-DSSIM correlate well with the more sophisticated IS and FVD. Recall that both IS and FVD require training a network on videos of fixed length, hence *(i)* can be computed only for short-length videos, due to GPU constraints; *(ii)* may be misleading (*e.g.* Tai Chi results in Table 1) when the training data for the inception network is different from the evaluated data. By contrast, t-PSNR and t-DSSIM prove to be faithful in quantifying the artifacts we study, as they are data-agnostic and accommodate videos of arbitrary length. However, our metrics are permutation invariant, do not assess the quality of the frames themselves, and are not robust to random noise. Hence we stress their *complementary role* to IS and FVD as a measure of the *overall* video quality.

Metric	Tai Chi	MoCoGAN $K = 16$	MDP model				
			$\gamma = 0.0$ $\beta = 0.0$	$\gamma = 0.7$ $\beta = 0.7$	$\gamma = 0.7$ $\beta = 0.9$	$\gamma = 0.9$ $\beta = 0.7$	$\gamma = 0.9$ $\beta = 0.9$
IS $\uparrow$	1.63 ± 0.05	4.49 ± 0.04	$\mathbf{4.52 \pm 0.05}$	4.15 ± 0.04	4.24 ± 0.06	3.92 ± 0.07	3.99 ± 0.04
FVD $\downarrow$	118 ± 5	828 ± 38	1108 ± 50	787 ± 10	782 ± 40	$\mathbf{744 \pm 40}$	809 ± 22
t-DSSIM $\uparrow$	0.0135	0.0031	0.0031	0.0024	**0.0037**	0.0035	0.0035
t-PSNR $\downarrow$	36.48	45.37	57.34	50.16	44.87	**44.39**	45.06

Table 2. **Results of the ablation study of the MDP approach on the Tai Chi dataset.** Our MDP configurations assume a selection of hyperparameters β and γ. For comparison, we include the results from the MoCoGAN baseline. By leveraging the long-term rewards, our MDP model improves the temporal diversity (*t*-PSNR and *t*-DSSIM) and FVD scores at the cost of a slight drop in IS.

Figure 4. **Tai Chi comparison** between MoCoGAN (**top row**) exhibiting the freezing artifact, and our MDP model (**bottom row**) generating perceivable motion (*e.g*. torso).

5.4. MoCoGAN: a case study

Here, we study the temporal diversity of the MoCoGAN model [29] using our *t*-PSNR and *t*-DSSIM scores.

We train MoCoGAN[4] on the UCF-101 dataset with temporal windows of size $K = 16$, and apply our temporal metrics to the samples from the generator. To enable a more detailed view of the temporal dynamics, we inspect the video samples as a function of time in Fig. 3 by plotting the values of the summands in Eq. (8) for each timestep. To rule out the possibility of any degenerate phenomena in the original data, we also plot the corresponding curves of the ground-truth sequences alongside. This clearly shows that MoCoGAN exhibits a vanishing diversity of video frames – a pattern that is not found in the training data.

5.5. MDP approach: an ablation study

Here, we perform an ablation study of our MDP approach by varying the time-horizon hyperparameters, γ and β, introduced in Sec. 3.2. Recall that γ controls the timespan of the future predictions modeled by the Q-value: lower values imply a shorter time horizon, whereas higher values encourage the model to learn long-term dependencies. Parameter β, on the other hand, specifies how accounting for the long-term effect is distributed over the timesteps. High values specify equal distribution; lower values force the model to encode the long-term effects more in the earlier than in the later timesteps. As a boundary case, we also consider $\beta = 0$ and $\gamma = 0$ to gage the effect of the architecture change in the video discriminator (TCN), which is needed to implement reward causality (*c.f*. Sec. 3.2). As quantitative measures, we use the Inception Score (IS) [20], the Fréchet Video Distance (FVD) [30], as well our temporal metrics, *t*-DSSIM and *t*-PSNR, introduced in Sec. 4.

The results in Table 2 show that by leveraging the increasing values of the time-horizon hyperparameters, our model clearly improves the temporal diversity in terms of *t*-PSNR and *t*-DSSIM. Moreover, we also observe that the TCN baseline ($\gamma = 0$, $\beta = 0$) performs worse than the original MoCoGAN in terms of temporal diversity. This is easily understood when considering that the TCN alone does not have any lookahead into the future (*c.f*. Fig. 2b). However, once we enable taking the future rewards into account by virtue of our MDP formulation, we not only reach but actually surpass the temporal diversity of the baseline MoCoGAN, as expected.

The somewhat inferior IS and FVD scores might be due to their sensitivity to the data prior, as discussed in Sec. 5.3. This hypothesis is also supported by a qualitative comparison between MoCoGAN and our MDP model. Figure 4 gives one such example; more results can be found in the supplemental material. While we observe no notable difference in per-frame quality, the motion between consecutive frames from our MDP model is more apparent than the samples from MoCoGAN (*e.g*., the torso of the performer).

5.6. Human Actions and UCF-101

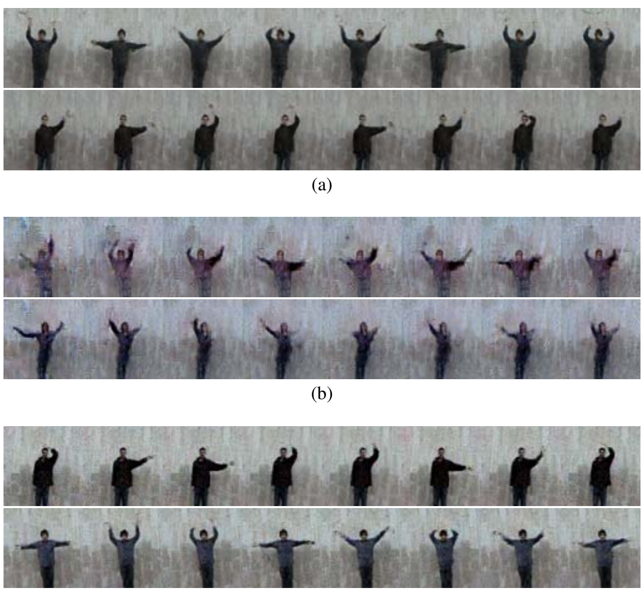
We perform further experiments on the Human Actions and the more challenging UCF-101 datasets.[5] We select $\gamma = 0.9, \beta = 0.7$ for our MDP model, which provide a good trade-off between the improved *t*-PSNR, *t*-DSSIM, FVD and only a slight drop of IS on Tai Chi (*c.f*. Sec. 5.2). For reference, we train the TCN baseline, `MDP-0`, by setting $\gamma = 0$ and $\beta = 0$ to decouple the influence of modeling

[4]We use the publicly available code provided by the MoCoGAN authors at `https://github.com/sergeytulyakov/mocogan`.

[5]To ensure a fair comparison, we use the same inception network for IS and FVD and train other methods [19, 29] using the authors' implementation (*c.f*. supplemental material for details).

Model	K	Human Actions				UCF-101			
		IS ↑	FVD ↓	*t*-DSSIM ↑	*t*-PSNR ↓	IS ↑	FVD ↓	*t*-DSSIM ↑	*t*-PSNR ↓
Raw dataset	–	3.39 ± 0.08	49 ± 2	0.0815	23.35	40.80 ± 0.26	452 ± 49	0.0723	28.34
TGAN (Normal)	16	2.90 ± 0.04	977 ± 31	–	–	8.11 ± 0.07	1686 ± 24	–	–
TGAN (SVC)	16	**3.65 ± 0.10**	**227 ± 10**	–	–	11.91 ± 0.21	1324 ± 23	–	–
MoCoGAN	16	3.53 ± 0.02	300 ± 8	0.0259	33.76	11.15 ± 0.10	1351 ± 49	0.0337	33.29
MoCoGAN-D_V^+	16	3.51 ± 0.02	245 ± 6	0.0243	34.79	11.48 ± 0.15	1314 ± 45	0.0358	33.61
MoCoGAN	24	3.47 ± 0.02	318 ± 9	0.0254	35.72	10.49 ± 0.09	1352 ± 49	0.0387	32.63
MDP-0 (ours)	16	3.55 ± 0.03	1413 ± 15	0.0559	33.31	6.16 ± 0.08	2147 ± 87	0.0160	47.36
MDP (ours)	16	3.55 ± 0.02	641 ± 8	0.0604	30.12	11.86 ± 0.11	**1277 ± 56**	0.0370	32.77
MDP (ours)	24	3.49 ± 0.03	686 ± 12	**0.0661**	**29.39**	**12.14 ± 0.18**	1293 ± 58	**0.0454**	**31.05**

Table 3. **Comparison of our two MDP models to the state of the art.** Temporal metrics are calculated for 64 frames. Our MDP model consistently improves the temporal video quality in terms of *t*-PSNR, *t*-DSSIM, and IS. Moreover, it is more memory efficient as it is comparable to MoCoGAN $K = 24$ and can produce videos of arbitrary length in contrast to TGAN. Note that since TGAN [19] can only generate videos of 16 frames, we do not compute *t*-PSNR and *t*-DSSIM for this model here.



(a)

(b)

(c)

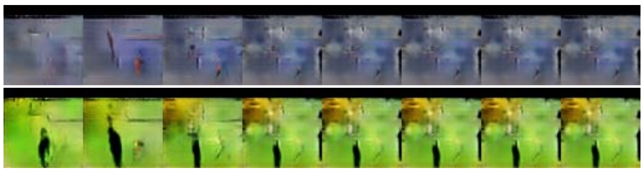
Figure 5. **Random samples on Human Actions.** (a) MoCoGAN, (b) MDP-0, (c) MDP. Disabling MDP leads to poorer video quality in (b), while modelling long-term rewards leads to comparable per-frame quality of the samples from our MDP model (c) w.r.t. MoCoGAN baseline (a), also reflected by IS, yet tangibly higher temporal diversity measured by *t*-PSNR and *t*-DSSIM. From the video sequence of 64 frames, every 8th frame is shown.

the long-term effects from the changes in the MoCoGAN architecture to comply with reward causality. We also train our MDP model and MoCoGAN on an extended temporal window $K = 24$. Recall that higher K require more GPU memory, but give the model an advantage, since it observes longer sequences at training time. Therefore, we aim to mitigate the artifacts while keeping K constant.

The quantitative results are summarized in Table 3. For both the Human Actions and UCF-101 datasets, we observe a consistent improvement of our MDP model in terms of temporal diversity measured by *t*-PSNR and *t*-DSSIM. Moreover, our model also outperforms MoCoGAN in terms of IS on both datasets, as well as FVD on the UCF-101 dataset. This can be explained by the more varied nature of motion on these datasets compared to the Tai Chi dataset, which makes taking into account future frames more important. On the Human Actions dataset, the FVD score for our model is inferior to MoCoGAN. Recall from Sec. 5.2, that for IS and FVD metrics we did not fine-tune the inception classifiers on the Human Actions dataset, which impedes the interpretability of the scores on this dataset. A visual inspection of the per-frame quality (*c.f.* Fig. 5 for examples) reveals no perceptual loss compared to the baseline model. In contrast, disabling MDP modeling (`MDP-0`) leads to a clear deterioration in video quality.

On both datasets, our model with $K = 16$ is also superior to MoCoGAN with $K = 24$ in terms of IS and FVD, and reaches on par performance in terms of *t*-PSNR and *t*-DSSIM. Yet, our MDP-based formulation is significantly more memory efficient, since extending the temporal window at training incurs addition memory costs. Concretely, at training time the MDP model with $K = 16$ consumes roughly 20% more memory than MoCoGAN, whereas setting $K = 24$ for the original MoCoGAN incurs a 50% higher memory footprint. Note that simply increasing the number of parameters of D_V in MoCoGAN is less effective than our proposed MDP approach (see MoCoGAN-D_V^+ in Tab. 5.6). Also, our MDP model with $K = 24$ improves further over $K = 16$ on UCF-101 and regarding the temporal metrics on Human Actions. A visual inspection of the samples from Human Actions did not reveal any perceptible difference to MoCoGAN or our MDP with $K = 16$, despite

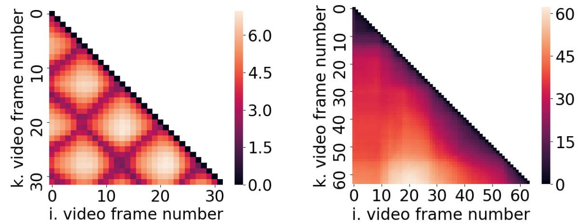
(a)



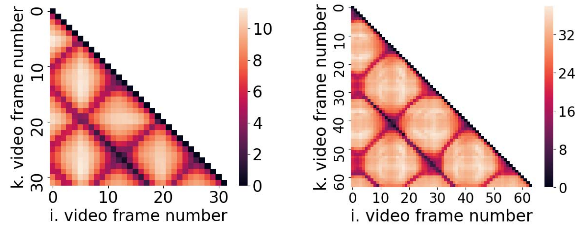
(b)

(c)

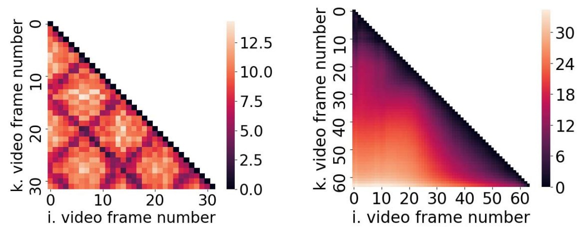
Figure 6. **Random samples of the MoCoGAN baseline and MDP models on UCF-101.** (a) MoCoGAN with looping artifact. (b) Our MDP-0 without modeling future rewards exhibits a freezing pattern. (c) Our MDP model. In (c), while the first sample has some looping, the second does not have temporal artifacts. From the video sequence of 64 frames, every 8$^{\text{th}}$ frame is shown.



(a)



(b)



(c)

Figure 7. **Heatmap comparison between ground truth, MoCoGAN, and our MDP models** trained on the Human Actions dataset **(left)** and UCF-101 **(right)** (different scales). (a) ground truth, (b) MoCoGAN, (c) MDP ($\gamma = 0.9, \beta = 0.7$). Our MDP model alleviates the looping artifact on Human Actions, where it can still appear natural. On the more complex UCF-101, our MDP is able to approximate the temporal quality of the ground truth.

the inferior IS and FVD scores; we believe this to be an artifact of the evaluation specifics. The IS score of our MDP model is slightly inferior only to TGAN [19]. However, TGAN can produce video sequences of only fixed length, whereas our MDP model can generate videos of arbitrary length, owing to the recurrent generator.

The qualitative results in Fig. 6 show that our model can generate complex scenes from UCF-101 that are visually comparable to the MoCoGAN samples. Similar to our observation on Human Actions, `MDP-0` produces poorer samples, which asserts the efficacy of the underlying MDP. Since the interpretation of the UCF-101 results is difficult, we examine a visualization of a pairwise L_1-distance between two frames in the video, shown in Fig. 7. The distance matrix can be represented as a lower triangular two-dimensional heatmap, owing to the symmetry of L_1. We observe that while MoCoGAN exhibits a looping pattern, our MDP approach tends to preserve the temporal qualities of the ground-truth datasets. Note that some samples in Human Actions can be naturally periodic (*e.g.* hand-waving), hence, we do not expect our model to dispense with the looping pattern completely. The overall results suggest that modeling long-term dependencies with an MDP consistently leads to more diverse motion dynamics, which becomes more apparent in increasingly complex scenes.

6. Conclusions and Future Work

We revealed two pathological cases in the videos synthesized by the state-of-the-art MoCoGAN model, namely *freezing* and *looping*. To quantify the temporal diversity, we proposed an interpretable class of metrics. We showed that the SSIM- and PSNR-based metrics, *t*-PSNR and *t*-DSSIM, effectively complement IS and FVD to quantify temporal artifacts. Next, we traced the artifacts to the limited training length, which inhibits long-term modeling of the video sequences. As a remedy, we reformulated video generation as an MDP and incorporated it into MoCoGAN. We showed the efficacy of our MDP model on the challenging UCF-101 dataset both in terms of our temporal metrics, as well as in IS and FVD scores. Maintaining the recurrent state between the training iterations or imposing a tractable prior on the state suggest promising extensions of this work toward generating long-sequence videos.

Acknowledgements. The authors thank Sergey Tulyakov and Masaki Saito for helpful clarifications.

References

[1] Mohammad Babaeizadeh, Chelsea Finn, Dumitru Erhan, Roy H. Campbell, and Sergey Levine. Stochastic variational video prediction. In *ICLR*, 2018.

[2] Shaojie Bai, J. Zico Kolter, and Vladlen Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv:1803.01271 [cs.CV]*, 2018.

[3] Moshe Blank, Lena Gorelick, Eli Shechtman, Michal Irani, and Ronen Basri. Actions as space-time shapes. In *ICCV*, 2005.

[4] Chelsea Finn, Ian J. Goodfellow, and Sergey Levine. Unsupervised learning for physical interaction through video prediction. In *NIPS*2016*.

[5] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*2014*.

[6] Jiawei He, Andreas M. Lehrmann, Joseph Marino, Greg Mori, and Leonid Sigal. Probabilistic video generation using holistic attribute control. In *ECCV*, 2018.

[7] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *NIPS*2017*.

[8] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[9] Nal Kalchbrenner, Aäron van den Oord, Karen Simonyan, Ivo Danihelka, Oriol Vinyals, Alex Graves, and Koray Kavukcuoglu. Video pixel networks. In *ICML*, 2017.

[10] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *CVPR*, 2014.

[11] Diederik P. Kingma and Max Welling. Auto-encoding variational Bayes. In *ICLR*, 2014.

[12] Bernhard Kratzwald, Zhiwu Huang, Danda Pani Paudel, Acharya Dinesh, and Luc Van Gool. Improving video generation for multi-functional applications. *arXiv:1711.11453 [cs.CV]*, 2017.

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

[14] Xiaodan Liang, Lisa Lee, Wei Dai, and Eric P. Xing. Dual motion GAN for future-flow embedded video prediction. In *ICCV*, 2017.

[15] Ziwei Liu, Raymond A. Yeh, Xiaoou Tang, Yiming Liu, and Aseem Agarwala. Video frame synthesis using deep voxel flow. In *ICCV*, 2017.

[16] Michaël Mathieu, Camille Couprie, and Yann LeCun. Deep multi-scale video prediction beyond mean square error. In *ICLR*, 2016.

[17] Katsunori Ohnishi, Shohei Yamamoto, Yoshitaka Ushiku, and Tatsuya Harada. Hierarchical video generation from orthogonal information: Optical flow and texture. In *AAAI*, 2018.

[18] Marc'Aurelio Ranzato, Arthur Szlam, Joan Bruna, Michaël Mathieu, Ronan Collobert, and Sumit Chopra. Video (language) modeling: a baseline for generative models of natural videos. *arXiv:1412.6604 [cs.LG]*, 2014.

[19] Masaki Saito, Eiichi Matsumoto, and Shunta Saito. Temporal generative adversarial nets with singular value clipping. In *ICCV*, 2017.

[20] Tim Salimans, Ian J. Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training GANs. In *NIPS*2016*.

[21] David Silver, Guy Lever, Nicolas Heess, Thomas Degris, Daan Wierstra, and Martin A. Riedmiller. Deterministic policy gradient algorithms. In *ICML*, 2014.

[22] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv:1212.0402 [cs.CV]*, 2012.

[23] Nitish Srivastava, Elman Mansimov, and Ruslan Salakhutdinov. Unsupervised learning of video representations using LSTMs. In *ICML*, 2015.

[24] Ilya Sutskever and Geoffrey E. Hinton. Learning multilevel distributed representations for high-dimensional sequences. In *AISTATS*, 2007.

[25] Ilya Sutskever, Geoffrey E Hinton, and Graham W. Taylor. The recurrent temporal restricted Boltzmann machine. In *NIPS*2009*.

[26] Richard S. Sutton and Andrew G. Barto. *Introduction to Reinforcement Learning*. MIT Press, 1998.

[27] Graham W. Taylor, Geoffrey E Hinton, and Sam T. Roweis. Modeling human motion using binary latent variables. In *NIPS*2017*, 2007.

[28] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3D convolutional networks. In *ICCV*, 2015.

[29] Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, and Jan Kautz. MoCoGAN: Decomposing motion and content for video generation. In *CVPR*, 2018.

[30] Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric and challenges. *arXiv:1812.01717 [cs.CV]*, 2018.

[31] Aäron van den Oord, Nal Kalchbrenner, Lasse Espeholt, Koray Kavukcuoglu, Oriol Vinyals, and Alex Graves. Conditional image generation with PixelCNN decoders. In *NIPS*2016*.

[32] Ruben Villegas, Jimei Yang, Seunghoon Hong, Xunyu Lin, and Honglak Lee. Decomposing motion and content for natural video sequence prediction. In *ICLR*, 2017.

[33] Ruben Villegas, Jimei Yang, Yuliang Zou, Sungryull Sohn, Xunyu Lin, and Honglak Lee. Learning to generate long-term future via hierarchical prediction. In *ICML*, 2017.

[34] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Generating videos with scene dynamics. In *NIPS*2016*.

[35] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Processing*, 13(4):600–612, 2004.

[36] Tianfan Xue, Jiajun Wu, Katherine L. Bouman, and Bill Freeman. Visual dynamics: Probabilistic future frame synthesis via cross convolutional networks. In *NIPS*2016*.