

Efficient Video Generation on Complex Datasets

Aidan Clark

DeepMind
London, UK

aidanclark@google.com

Jeff Donahue

DeepMind
London, UK

jeffdonahue@google.com

Karen Simonyan

DeepMind
London, UK

simonyan@google.com

Abstract

Generative models of natural images have progressed towards high fidelity samples by the strong leveraging of scale. We attempt to carry this success to the field of video modeling by showing that large Generative Adversarial Networks trained on the complex Kinetics-600 dataset are able to produce video samples of substantially higher complexity than previous work. Our proposed model, Dual Video Discriminator GAN (DVD-GAN), scales to longer and higher resolution videos by leveraging a computationally efficient decomposition of its discriminator. We evaluate on the related tasks of video synthesis and video prediction, and achieve new state of the art Fréchet Inception Distance on prediction for Kinetics-600, as well as state of the art Inception Score for synthesis on the UCF-101 dataset, alongside establishing a strong baseline for synthesis on Kinetics-600.

1 Introduction



Figure 1: Selected frames from videos generated by a DVD-GAN trained on Kinetics-600 at 256×256 , 128×128 , and 64×64 resolutions (top to bottom).

Modern deep generative models can produce realistic natural images [10, 25, 28, 34, 39] when trained on high-resolution and diverse datasets. Generation of natural *video* is an obvious further challenge for generative modeling, but one that is plagued by increased data complexity and computational requirements. For this reason, much prior work on video generation has revolved around relatively simple datasets, or tasks where strong temporal conditioning information is available.



Figure 2: Generated video samples with interesting behavior. In raster-scan order: **a)** A generated *camera change*. **b)** Zooming in on an object. **c)** A well structured moving object. **d)** Fine detail from a pen being left on paper. The first sample is 128×128 , all others are 64×64 .

We focus on the tasks of video synthesis and video prediction (defined in Section 2.1), and aim to extend the strong results of generative image models to the video domain. We build upon the state-of-the-art BigGAN architecture [10], introducing a number of video-specific modifications including efficient separable attention and a spatio-temporal decomposition of the discriminator. This allows us to train on Kinetics-600 – a complex dataset of natural videos an order of magnitude larger than the commonly used datasets. The resulting model, Dual Video Discriminator GAN (DVD-GAN), is able to generate temporally coherent, high-resolution videos of notable fidelity (Figure 1).

Our contributions are as follows:

- We propose DVD-GAN – a scalable generative model of natural video which produces high-quality samples at resolutions up to 256×256 and lengths up to 48 frames.
- We achieve state of the art for video synthesis on UCF-101 and prediction on Kinetics-600.
- We establish class-conditional video synthesis on Kinetics-600 as a new benchmark for generative video modeling, and report DVD-GAN results as a strong baseline.

2 Background

2.1 Video Synthesis and Prediction

Generative video modeling is a widely explored problem which has included work on VAEs [5, 16, 30, 22], auto-regressive models [38, 48, 24, 62], normalizing flows [29], and GANs [33, 59, 42, 41]. Exact formulations of the task differ in the type of the conditioning signal provided. At one extreme lies unconditional video synthesis where the task is to generate *any* video following the training distribution. Another extreme is occupied by strongly-conditioned models, including generation conditioned on another video for content transfer [7, 68], per-frame segmentation masks [61], or pose information [60, 58, 66]. In the middle ground there are tasks which are more structured than unconditional generation, and yet are more challenging from a modeling perspective than strongly-conditioned generation (which gets a lot of information about the generated video through its input). The objective of *class-conditional video synthesis* is to generate a video of a given category (e.g., “riding a bike”) while *future video prediction* is concerned with generation of continuing video given initial frames. These problems differ in several aspects, but share a common requirement of needing to generate realistic temporal dynamics, and in this work we restrict ourselves to these two problems.

2.2 Generative Adversarial Networks

Generative Adversarial Networks (GANs) [19] are a class of generative models defined by a minimax game between a *Discriminator* \mathcal{D} and a *Generator* \mathcal{G} . The original objective was proposed by [19], and many improvements have since been suggested, mostly targeting improved training stability [4, 67, 10, 20, 36]. We use the hinge formulation of the objective [32, 10] which is optimized by gradient descent (ρ is the elementwise ReLU function):

$$\mathcal{D}: \min_{\mathcal{D}} \mathbb{E}_{x \sim \text{data}(x)} [\rho(1 - \mathcal{D}(x))] + \mathbb{E}_{z \sim p(z)} [\rho(1 + \mathcal{D}(\mathcal{G}(z)))] , \quad \mathcal{G}: \max_{\mathcal{G}} \mathbb{E}_{z \sim p(z)} [\mathcal{D}(\mathcal{G}(z))].$$

GANs have well-known limitations including a tendency towards limited diversity in generated samples (a phenomenon known as mode collapse) and the difficulty of quantitative evaluation due to the lack of an explicit likelihood measure over the data. Despite these downsides, GANs have produced some of the highest fidelity samples across many visual domains [25, 10].

2.3 Multiple GAN Discriminators for Video

High-quality videos typically contain coherent objects that progress consistently in time. An active area of prior work considers decompositions which model the texture and spatial consistency of objects separately from their temporal dynamics. One approach is to split \mathcal{G} into foreground and background models [59, 47], while another considers explicit or implicit optical flow in either \mathcal{G} or \mathcal{D} [42, 37]. A third approach is to decompose \mathcal{D} into sub-networks that judge image quality separately from motion. For instance, MoCoGAN [54] contains a separate discriminator for individual frames in addition to the discriminator operating on slices of the whole video; other models discriminate groups of frames [65, 49] or sub-batches at different resolutions [41]. A benefit of some of these approaches is that \mathcal{D} no longer processes the entire batch of videos at full resolution.

2.4 Kinetics-600

Kinetics is a large dataset of 10-second high-resolution YouTube clips [26, 2] originally created for the task of human action recognition. We use the second iteration of the dataset, Kinetics-600 [13], which consists of 600 classes with at least 600 videos per class for a total of around 500,000 videos.¹ Kinetics videos are diverse and unconstrained, which allows us to train large models without being concerned with the overfitting that occurs on small datasets with fixed objects interacting in specified ways [18, 9]. Among prior work, the closest dataset (in terms of subject and complexity) which is consistently used is UCF-101 [46]. We focus on Kinetics-600 because of its larger size (almost 50x more videos than UCF-101) and its increased diversity (600 instead of 101 classes – not to mention increased intra-class diversity). Nevertheless for comparison with prior art we train on UCF-101 and set a new state-of-the-art Inception Score there. Kinetics contains many artifacts expected from YouTube, including cuts (as in Figure 2a), title screens and visual effects. Except when specifically described, we choose frames with stride 2 (meaning we skip every other frame). This allows us to generate videos with higher complexity without incurring extra computational cost.

2.5 Evaluation Metrics

Designing metrics for measuring the quality of generative models (GANs in particular) is an active area of research [43, 8]. In this work we report the two most commonly used metrics, Inception Score (IS) [44] and Fréchet Inception Distance (FID) [21]. The standard instantiation of these metrics is intended for generative image models, and uses an Inception model [51] for image classification or feature extraction. For videos, we use the publicly available Inflated 3D Convnet (I3D) network trained on Kinetics-600 [12]. Our Fréchet Inception Distance is therefore very similar to the Fréchet Video Distance (FVD) [56], although our implementation is different and more aligned with the original FID metric.²

3 Dual Video Discriminator GAN

Our primary contribution is Dual Video Discriminator GAN (DVD-GAN) which is able to generate high-resolution and temporally consistent videos. It extends a large image generation model (BigGAN [10]) to video while introducing several techniques to accelerate training. An overview of the DVD-GAN architecture is given in Figure 3 and a detailed description is in Appendix A.2. Unlike some of the prior work, our generator contains no explicit priors for foreground, background or motion (optical flow); instead, we rely on a high-capacity neural network to learn this in a data-driven manner. DVD-GAN contains both self-attention and an RNN but is not autoregressive in time or in space. While the RNN produces features for each frame sequentially, afterwards all frames are produced by the ResNet in parallel, generating all pixels in each frame jointly. In other words, the

¹Kinetics is occasionally pruned and so we cannot give an exact size. The dataset is available at [2].

²We use ‘avgpool’ features (rather than logits) by default, our I3D model is trained on Kinetics-600 (rather than Kinetics-400), and we pre-calculate ground-truth statistics on the entire training set.

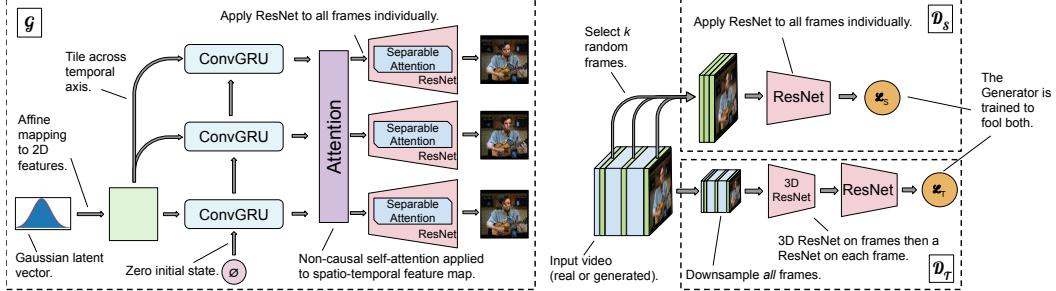


Figure 3: Simplified architecture diagram of \mathcal{G} (left) and $\mathcal{D}_S/\mathcal{D}_T$ (right). More details in A.2.

pixels of each frame do not directly depend on other pixels in the video, as would be the case for auto-regressive models.

3.1 Dual Discriminators

Given a video with length T , height H , and width W , DVD-GAN employs two discriminators for its assessment: a *Spatial Discriminator* \mathcal{D}_S and a *Temporal Discriminator* \mathcal{D}_T . \mathcal{D}_S critiques single frame content and structure by randomly sampling k full-resolution frames and processing them individually. We use $k = 8$ and show in Section 4.3 this choice improves performance. \mathcal{D}_S 's final score is the sum of the per-frame scores, similar to TGANv2 [41].

The temporal discriminator \mathcal{D}_T must provide \mathcal{G} with the learning signal to generate movement (something not critiqued by \mathcal{D}_S). To make the model scalable, we would like to achieve this without processing the entire video at full resolution. Rather than subsampling the batch [41] or selecting only a subset of frames [41, 54], we apply a spatial downsampling function $\phi(\cdot)$ to the whole video and feed its output to \mathcal{D}_T . We choose ϕ to be a 2×2 average pooling function, but discuss alternatives in Section 4.3. This results in an architecture where the discriminators do not process the entire high-resolution video (since \mathcal{D}_S processes only $k \times H \times W$ pixels and \mathcal{D}_T only $T \times \frac{H}{2} \times \frac{W}{2}$), and yet together they ensure that \mathcal{G} learns to generate high-resolution and temporally consistent videos.

\mathcal{D}_S is similar to the per-frame discriminator \mathcal{D}_I in MoCoGAN [54]. However MoCoGAN's analog of \mathcal{D}_T looks at full resolution videos, whereas \mathcal{D}_S is the only source of learning signal for high-resolution details in DVD-GAN. For this reason, \mathcal{D}_S is essential when ϕ is not the identity, unlike in MoCoGAN where the per-frame discriminator is less crucial.

3.2 Separable Self-Attention

The Transformer module of self-attention [57] is a popular architectural building block whose global receptive field allows for propagation of information across the entire feature map. However, directly applying it to large video features is prohibitive since self-attention requires the computation and storage of the attention weight matrix of size $(HWT)^2$. To circumvent this limitation, we introduce an efficient self-attention variant that we term *Separable Attention*. Instead of attending to all locations of our features simultaneously, we apply three attention layers in a row, each subsequently attending over the height, width and time axis. This reduces the size of the largest tensor we need to store in memory from being proportional to $(HWT)^2$ to max $\{H^2WT, HW^2T, HWT^2\}$. This can be seen as a special case of Factorized Attention concurrently introduced in [14].

4 Experiments and Analysis

Our training setup is directly modeled on that of BigGAN [10]: a detailed description can be found in Appendix A.3. Each DVD-GAN was trained on slices of TPUs v3 pods [1] using between 32 and 512 replicas with an Adam [27] optimizer for up to 300,000 update steps (though commonly we evaluate models at the final checkpoint before collapse – often between 100k and 250k steps). We rely on the TF-Replicator [11] framework for data parallel training. Models took between 12 and 96 hours to train.



Figure 4: Each quadrant has initial frames from 12 samples for a fixed class. From the top-left clockwise the classes are: **jetskiing**, **playing basketball**, **curling** (**sport**), **snatch weight lifting**.



Figure 5: All 48 frames (in raster-scan order) from a 64×64 sample from a soccer class.

4.1 Class-Conditional Video Synthesis

Our primary results concern the problem of *Class-Conditional Video Synthesis*. We provide our results for the UCF-101 and Kinetics-600 datasets. With Kinetics-600 emerging as a new benchmark for generative video modelling, our results establish a strong baseline for future work.

4.1.1 Kinetics-600 Results

Table 1: FID/IS for DVD-GAN on Kinetics-600 Video Synthesis.

(# Frames / Resolution)	FID (\downarrow)	IS (\uparrow)	
		No Truncation	With Truncation
12/ 64×64	0.91	61.10	129.9
48/ 64×64	12.92	97.62	219.05
12/ 128×128	2.16	55.09	80.32
48/ 128×128	31.50	111.19	222.07
12/ 256×256	3.35	59.74	64.05

In Table 1 we show the main result of this paper: benchmarks for Video Synthesis on Kinetics-600. We consider a range of resolutions and video lengths, and measure Inception Score and Fréchet Inception Distance (FID) for each (as described in Section 2.5). There is no prior work with which to quantitatively compare these results (for comparative experiments see Section 4.1.2 and Section 4.2.1), but we believe these samples to show a level of fidelity not yet achieved in datasets as complex as Kinetics-600 (see samples from each row in Appendix D.1). Because all videos are resized for the I3D network (to 224×224), it is meaningful to compare metrics across equal length videos at different resolutions. Neither IS nor FID are comparable across videos of different lengths, and should be treated as separate metrics.

Generating longer and larger videos is a more challenging modeling problem, which is conveyed by the metrics (in particular, comparing 12-frame videos across 64×64 , 128×128 and 256×256 resolutions). Nevertheless, DVD-GAN is able to generate plausible videos at all resolutions and with actions spanning up to 4 seconds (48 frames). As can be seen in Appendix D.1, smaller videos display



Figure 6: Single generated frames from 128x128 Kinetics across different classes.

Table 2: IS on UCF-101 (higher is better)

Method	IS (\uparrow)
VGAN [59]	$8.31 \pm .09$
TGAN [42]	$11.85 \pm .07$
MoCoGAN [54]	$12.42 \pm .03$
ProgressiveVGAN [3]	$14.56 \pm .05$
TGANv2 [41]	$24.34 \pm .35$
DVD-GAN (ours)	32.97 ± 1.7

Table 3: DVD-GAN-FP’s FVD scores on Video Prediction for 16 frames of Kinetics-600 without frame skipping. The final row represents a *Video Synthesis* model generating 16 frames without frame skipping.

Method	Training Set FVD (\downarrow)	Test Set FVD (\downarrow)
Video Transformer [62]	-	170 ± 5
DVD-GAN-FP	99.32 ± 0.55	103.78 ± 1.17
DVD-GAN	32.3 ± 0.82	31.1 ± 0.56

high quality textures, object composition and movement. At higher resolutions, generating coherent objects becomes more difficult (movement consists of a much larger number of pixels), but high-level details of the generated scenes are still extremely coherent, and textures (even complicated ones like the side of an ice rink in Figure 1a) are generated well. It is further worth noting that the 48-frame models do not see more high resolution frames than the 12-frame model (due to the fixed choice of $k = 8$ described in Section 3.1), yet nevertheless learn to generate high resolution images across all 48 frames.

4.1.2 Video Synthesis on UCF-101

We further verify our results by testing the same model on UCF-101 [46], a smaller dataset of 13,320 videos of human actions across 101 classes that has previously been used for video synthesis and prediction [42, 41, 54]. Our model produces samples with an IS of 32.97, significantly outperforming the state of the art (see Table 2 for quantitative comparison and Appendix B.2 for more details).

4.2 Future Video Prediction

Future Video Prediction is the problem of generating a sequence of frames which directly follow from one (or a number) of initial conditioning frames. Both this and video synthesis require \mathcal{G} to learn to produce realistic scenes and temporal dynamics, however video prediction further requires \mathcal{G} to analyze a video and discover elements in the scene which will evolve over time. In this section, we use the Fréchet Video Distance exactly as [56] using the logits of an I3D network trained on Kinetics-400 as features. This allows for direct comparison to prior work. Our model, DVD-GAN-FP (Frame Prediction), is slightly modified to facilitate the changed problem, and details of these changes are given in Appendix A.4.

4.2.1 Frame-Conditional Kinetics

For direct comparison with concurrent work on autoregressive video models [62] we consider the generation of 11 frames of Kinetics-600 at 64×64 resolution conditioned on 5 frames, where the videos for training are **not** taken with any frame skipping. We show results for all these cases in Table 3. Our frame-conditional model **DVD-GAN-FP** outperforms the (limited) prior work on frame-conditional prediction for Kinetics. The final row labeled DVD-GAN is the FVD of 16-frame

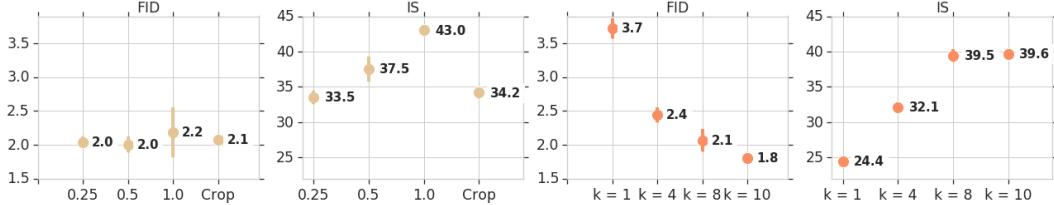


Figure 7: The effect of ϕ in \mathcal{D}_T (left two) and k in \mathcal{D}_S (right two). FID is similar for any choice of ϕ , while IS declines as downsampling increases. Increasing k improves both with diminishing returns.

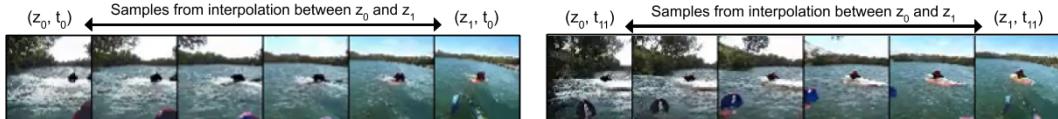


Figure 8: A latent interpolation under a shared class at the first (left 6) and last (right 6) frames.

Video Synthesis samples (no conditioning frames) without any frame skipping, which is notably better. This means that the synthesis model is able to generate higher quality videos from scratch than a model which must infer plausible video continuations.

4.2.2 BAIR Robot Pushing

We further evaluate DVD-GAN-FP on the BAIR Robot Pushing [18] dataset, a dataset of stationary videos of a robot arm manipulating a range of small objects. Here our model gets competitive FVD with prior work, but does not achieve state of the art performance. More details are given in Appendix B.1.

4.3 Dual Discriminator Input

We analyze several choices for k (the number of frames per sample in the input to \mathcal{D}_S) and ϕ (the downsampling function for \mathcal{D}_T). We expect setting ϕ to the identity or $k = T$ to result in the best model, but we are interested in the maximally compressive k and ϕ that reduce discriminator input size (and the amount of computation), while still producing a high quality generator. For ϕ , we consider: 2×2 and 4×4 average pooling, the identity (no downsampling), as well as a ϕ which takes a random half-sized crop of the input video (as in [41]). Results can be seen in Figure 7. For each ablation, we train three identical DVD-GANs with different random initializations on 12-frame clips of Kinetics-600 at 64×64 resolution for 100,000 steps. We report mean and standard deviation (via the error bars) across each group for the whole training period. For k , we consider 1, 2, 8 and 10 frames.

4.4 Truncation Curves and Interpolations

We expect \mathcal{G} to produce samples of higher quality from latents near the center or mean of the distribution (zero). This is the idea behind the Truncation Trick [10]. Like BigGAN, we find that DVD-GAN is amenable to truncation. We also experiment with interpolations in the latent space (Figure 8) and in the class embedding (Figure 9) (see also Appendix D.2). In both cases, interpolations are evidence that \mathcal{G} has learned a relatively smooth mapping from the latent space to real videos: this would be impossible for a network that has only memorized the training data, or which is only capable of generating a few exemplars per class. Note that while all latent vectors along an interpolation are valid (and therefore \mathcal{G} should produce a reasonable sample), at no point during training is \mathcal{G} asked to generate a sample halfway between two classes. Nevertheless \mathcal{G} is able to interpolate between even very distinct classes.

5 Conclusion

We approached the challenging problem of modeling natural video by introducing a GAN capable of capturing the complexity of a large video dataset. We showed that on UCF-101 and frame-conditional Kinetics-600 it quantitatively achieves the new state of the art, alongside qualitatively producing sample videos with high complexity and diversity. We further wish to emphasize the benefit of training generative models on large and complex video datasets, such as Kinetics-600. We envisage the strong baselines we established on this dataset with DVD-GAN will be used as a reference point by the generative modeling community moving forward. While much remains to be done before realistic videos can be consistently generated in an unconstrained setting, we believe DVD-GAN is a step in that direction.

Acknowledgments

We would like to thank Eric Noland and João Carreira for help with the Kinetics dataset and Marcin Michalski and Karol Kurach for helping us acquire data and models for Fréchet Video Distance comparison. We would further like to thank Sander Dieleman, Jacob Walker and Tim Harley for useful discussions and feedback on the paper.

References

- [1] Cloud TPU. <https://cloud.google.com/tpu/>. Accessed: 2019.
- [2] Kinetics. <https://deepmind.com/research/open-source/open-source-datasets/kinetics/>. Accessed: 2019.
- [3] D. Acharya, Z. Huang, D. P. Paudel, and L. Van Gool. Towards high resolution video generation with progressive growing of sliced Wasserstein GANs. *arXiv:1810.02419*, 2018.
- [4] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein GAN. *arXiv:1701.07875*, 2017.
- [5] M. Babaeizadeh, C. Finn, D. Erhan, R. H. Campbell, and S. Levine. Stochastic variational video prediction. *ICLR*, 2018.
- [6] N. Ballas, L. Yao, C. Pal, and A. Courville. Delving deeper into convolutional networks for learning video representations. *arXiv:1511.06432*, 2015.
- [7] A. Bansal, S. Ma, D. Ramanan, and Y. Sheikh. Recycle-GAN: Unsupervised video retargeting. In *ECCV*, September 2018.
- [8] S. Barratt and R. Sharma. A note on the Inception Score. *arXiv:1801.01973*, 2018.
- [9] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. In *ICCV*, 2005.
- [10] A. Brock, J. Donahue, and K. Simonyan. Large scale GAN training for high fidelity natural image synthesis. *ICLR*, 2019.
- [11] P. Buchlovsky, D. Budden, D. Grewe, C. Jones, J. Aslanides, F. Besse, A. Brock, A. Clark, S. G. Colmenarejo, A. Pope, et al. TF-Replicator: Distributed machine learning for researchers. *arXiv:1902.00465*, 2019.



Figure 9: Frames from two videos generated from the same z under two distinct classes (top two rows) and the interpolation between them at a single selected frame (bottom row).

- [12] J. Carreira and A. Zisserman. Quo vadis, action recognition? A new model and the Kinetics dataset. In *ICCV*, 2017.
- [13] J. Carreira, E. Noland, A. Banki-Horvath, C. Hillier, and A. Zisserman. A short note about Kinetics-600. *arXiv:1808.01340*, 2018.
- [14] R. Child, S. Gray, A. Radford, and I. Sutskever. Generating long sequences with sparse transformers. *arXiv:1904.10509*, 2019.
- [15] H. De Vries, F. Strub, J. Mary, H. Larochelle, O. Pietquin, and A. C. Courville. Modulating early visual processing by language. In *NeurIPS*, 2017.
- [16] E. Denton and R. Fergus. Stochastic video generation with a learned prior. *ICML*, 2018.
- [17] V. Dumoulin, J. Shlens, and M. Kudlur. A learned representation for artistic style. In *ICLR*, 2017.
- [18] F. Ebert, C. Finn, A. X. Lee, and S. Levine. Self-supervised visual planning with temporal skip connections. *arXiv:1710.05268*, 2017.
- [19] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *NeurIPS*, 2014.
- [20] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville. Improved training of Wasserstein GANs. In *NeurIPS*, 2017.
- [21] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*, 2017.
- [22] J.-T. Hsieh, B. Liu, D.-A. Huang, L. Fei-Fei, and J. C. Niebles. Learning to decompose and disentangle representations for video prediction. In *NeurIPS*. 2018.
- [23] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv:1502.03167*, 2015.
- [24] N. Kalchbrenner, A. van den Oord, K. Simonyan, I. Danihelka, O. Vinyals, A. Graves, and K. Kavukcuoglu. Video pixel networks. In *ICML*, 2017.
- [25] T. Karras, S. Laine, and T. Aila. A style-based generator architecture for generative adversarial networks. *arXiv:1812.04948*, 2018.
- [26] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, and A. Zisserman. The Kinetics human action video dataset. *arXiv:1705.06950*, 2017.
- [27] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv:1412.6980*, 2014.
- [28] D. P. Kingma and P. Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. In *NeurIPS*, 2018.
- [29] M. Kumar, M. Babaeizadeh, D. Erhan, C. Finn, S. Levine, L. Dinh, and D. Kingma. VideoFlow: A flow-based generative model for video. *arXiv:1903.01434*, 2019.
- [30] A. X. Lee, R. Zhang, F. Ebert, P. Abbeel, C. Finn, and S. Levine. Stochastic adversarial video prediction. *arXiv:1804.01523*, 2018.
- [31] J. Lei Ba, J. R. Kiros, and G. E. Hinton. Layer normalization. *arXiv:1607.06450*, 2016.
- [32] J. H. Lim and J. C. Ye. Geometric GAN. *arXiv:1705.02894*, 2017.
- [33] M. Mathieu, C. Couprie, and Y. LeCun. Deep multi-scale video prediction beyond mean square error. *arXiv:1511.05440*, 2015.
- [34] J. Menick and N. Kalchbrenner. Generating high fidelity images with subscale pixel networks and multidimensional upscaling. In *ICLR*, 2019.
- [35] T. Miyato and M. Koyama. cGANs with projection discriminator. *arXiv:1802.05637*, 2018.
- [36] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida. Spectral normalization for generative adversarial networks. *arXiv:1802.05957*, 2018.
- [37] K. Ohnishi, S. Yamamoto, Y. Ushiku, and T. Harada. Hierarchical video generation from orthogonal information: Optical flow and texture. In *AAAI*, 2018.

- [38] M. Ranzato, A. Szlam, J. Bruna, M. Mathieu, R. Collobert, and S. Chopra. Video (language) modeling: a baseline for generative models of natural videos. *arXiv:1412.6604*, 2014.
- [39] A. Razavi, A. van den Oord, and O. Vinyals. Generating diverse high-fidelity images with VQ-VAE-2. *arXiv preprint arXiv:1906.00446*, 2019.
- [40] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015.
- [41] M. Saito and S. Saito. TGANv2: Efficient training of large models for video generation with multiple subsampling layers. *arXiv:1811.09245*, 2018.
- [42] M. Saito, E. Matsumoto, and S. Saito. Temporal generative adversarial nets with singular value clipping. In *ICCV*, 2017.
- [43] M. S. Sajjadi, O. Bachem, M. Lucic, O. Bousquet, and S. Gelly. Assessing generative models via precision and recall. In *NeurIPS*, 2018.
- [44] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved techniques for training GANs. In *NeurIPS*, 2016.
- [45] A. M. Saxe, J. L. McClelland, and S. Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. *arXiv:1312.6120*, 2013.
- [46] K. Soomro, A. R. Zamir, and M. Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv:1212.0402*, 2012.
- [47] C. Spampinato, S. Palazzo, P. D’Oro, F. Murabito, D. Giordano, and M. Shah. VOS-GAN: Adversarial learning of visual-temporal dynamics for unsupervised dense prediction in videos. *arXiv:1803.09092*, 2018.
- [48] N. Srivastava, E. Mansimov, and R. Salakhutdinov. Unsupervised learning of video representations using LSTMs. In *ICML*, 2015.
- [49] X. Sun, H. Xu, and K. Saenko. A two-stream variational adversarial network for video generation. *arXiv:1812.01037*, 2018.
- [50] I. Sutskever, J. Martens, and G. E. Hinton. Generating text with recurrent neural networks. In *ICML*, 2011.
- [51] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the Inception architecture for computer vision. In *CVPR*, 2016.
- [52] S. Tokui, K. Oono, S. Hido, and J. Clayton. Chainer: a next-generation open source framework for deep learning. In *Workshop on Systems for ML and Open Source Software at NeurIPS*, 2015.
- [53] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3D convolutional networks. In *ICCV*, 2015.
- [54] S. Tulyakov, M.-Y. Liu, X. Yang, and J. Kautz. MoCoGAN: Decomposing motion and content for video generation. In *CVPR*, 2018.
- [55] D. Ulyanov, A. Vedaldi, and V. Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv:1607.08022*, 2016.
- [56] 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:1812.01717*, 2018.
- [57] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *NeurIPS*, 2017.
- [58] R. Villegas, J. Yang, Y. Zou, S. Sohn, X. Lin, and H. Lee. Learning to generate long-term future via hierarchical prediction. In *ICML*, 2017.
- [59] C. Vondrick, H. Pirsiavash, and A. Torralba. Generating videos with scene dynamics. In *NeurIPS*, 2016.
- [60] J. Walker, K. Marino, A. Gupta, and M. Hebert. The pose knows: Video forecasting by generating pose futures. In *ICCV*, 2017.
- [61] T. Wang, M. Liu, J. Zhu, G. Liu, A. Tao, J. Kautz, and B. Catanzaro. Video-to-video synthesis. *arXiv:1808.06601*, 2018.

- [62] D. Weissenborn, O. Täckström, and J. Uszkoreit. Scaling autoregressive video models. *arXiv:1906.02634*, 2019.
- [63] Y. Wu and K. He. Group normalization. In *ECCV*, 2018.
- [64] Y. Wu, S. Zhang, Y. Zhang, Y. Bengio, and R. R. Salakhutdinov. On multiplicative integration with recurrent neural networks. In *NeurIPS*, 2016.
- [65] Y. Xie, E. Franz, M. Chu, and N. Thuerey. tempoGAN: A temporally coherent, volumetric GAN for super-resolution fluid flow. *TOG*, 2018.
- [66] C. Yang, Z. Wang, X. Zhu, C. Huang, J. Shi, and D. Lin. Pose guided human video generation. In *ECCV*, 2018.
- [67] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena. Self-attention generative adversarial networks. *arXiv:1805.08318*, 2018.
- [68] Y. Zhou, Z. Wang, C. Fang, T. Bui, and T. L. Berg. Dance dance generation: Motion transfer for internet videos. *arXiv:1904.00129*, 2019.

A Experiment Methodology

A.1 Dataset Processing

For all datasets we randomly shuffle the training set for each model replica independently. Experiments on the BAIR Robot Pushing dataset are conducted in the native resolution of 64×64 , where for UCF-101 we operate at a (downsampled) 128×128 resolution. This is done by a bilinear resize such that the video’s smallest dimension is mapped to 128 pixels (maintaining aspect ratio). From this we take a random 128-pixel crop along the other dimension. We use the same procedure to construct datasets of different resolutions for Kinetics-600. All three datasets contain videos with more frames than we generate, so we take a random sequence of consecutive frames from the resized output.

A.2 Architecture Description

Our model adopts many architectural choices from BigGAN [10] including our nomenclature for describing network width, which is determined by the product of a channel multiplier ch with a constant for each layer in the network. The layer-wise constants for \mathcal{G} are $[8, 8, 8, 4, 2]$ for 64×64 videos and $[16, 16, 8, 4, 2, 1]$ for 128×128 (identical to BigGAN). The width of the i -th layer is given by the product of ch and the i -th constant and all layers prior to the residual network in \mathcal{G} use the initial layer’s multiplier and we refer to the product of that and ch as ch_0 . ch for \mathcal{D} is always 128, where for \mathcal{G} it is 128 in smaller models and 64 for models at greater than 64×64 resolution, or which generate 48 frames. The corresponding ch lists for both \mathcal{D}_T and \mathcal{D}_S are $[2, 4, 8, 16, 16]$ for 64×64 resolution and $[1, 2, 4, 8, 16, 16]$ for 128×128 .

The input to \mathcal{G} consists of a Gaussian latent noise $z \sim \mathcal{N}(0, I)$ and a learnt embedding $e(y)$ of the desired class y . Both inputs are 120-dimensional vectors. \mathcal{G} first computes an affine transformation of $[z; e(y)]$ to a $[4, 4, ch_0]$ -shaped tensor, which is treated as the input to a recurrent neural network for each of T timesteps. In the majority of our experiments we use a Multiplicative Convolutional Gated Recurrent Unit [6, 64, 50] in \mathcal{G} , whose update rule for input x_t and previous output h_{t-1} is given by the following:

$$\begin{aligned} g &= (W_{gh} \star_1 h_{t-1} + b_{gh}) \odot (W_{gx} \star_1 x_t + b_{gx}) \\ r &= \sigma(W_r \star_3 [g; x_t] + b_r) \\ u &= \sigma(W_u \star_3 [g; x_t] + b_u) \\ c &= \rho(W_c \star_3 [x_t; r \odot g] + b_c) \\ h_t &= u \odot g + (1 - u) \odot c \end{aligned}$$

In these equations σ and ρ are the elementwise sigmoid and ReLU functions respectively, the \star_n operator represents a convolution with a kernel of size $n \times n$, and the \odot operator is an elementwise

multiplication. Brackets are used to represent a feature concatenation. We found that the choice of the recurrent network architecture has a non-negligible effect on the sample quality. For the majority of our experiments we generate less than 48 frames, and in that setting the Multiplicative Convolutional Gated Recurrent Unit performs best. However, for videos with 48 frames the standard ConvGRU (where $g = h_{t-1}$) performs better. We believe that further architectural improvements in \mathcal{G} 's RNN should enable high-quality video modeling across all lengths with the same architecture, and we leave this to future work.

The result of the RNN are features of shape $[T, 4, 4, ch_0]$ which are fed as the input to a non-causal self-attention block [57]. It has a single attention head and is not separable (as described in Section 3.2) since it operates on a small number of features. Finally, each frame is generated from the corresponding temporal slice of the attention's output by a residual network nearly identical to that of BigGAN; though following BigGAN-deep we double the number of overall blocks and pass the entire conditioning vector $[z; e(y)]$ to each block (without a hierarchical split). The residual network is applied to each frame individually (i.e., the time axis is folded into the batch axis before the forward pass), but we do not reduce over the time dimension when calculating Batch Normalization statistics. This prevents the network from utilizing the Batch Normalization layers to pass information between timesteps. The residual network upscales each frame largely independently, with the exception of its self-attention layer which, following BigGAN, is placed before the final residual block. Due to the large size of the tensors, we use the separable self-attention block described in Section 3.2.

The spatial discriminator \mathcal{D}_S functions almost identically to BigGAN's discriminator. A score is calculated for each of the uniformly sampled k frames (we default to $k = 8$) and the \mathcal{D}_S output is the sum over per-frame scores. The temporal discriminator \mathcal{D}_T has a similar architecture, but pre-processes the real or generated video with a 2×2 average-pooling downsampling function ϕ . Furthermore, the first two residual blocks of \mathcal{D}_T are 3-D, where every convolution is replaced with a 3-D convolution with a kernel size of $3 \times 3 \times 3$. The rest of the architecture follows BigGAN [10].

A.3 Training Details

Sampling from DVD-GAN is very efficient, as the core of the generator architecture is a feed-forward convolutional network: two 64×64 48-frame videos can be sampled in less than 150ms on a single TPU core. The dual discriminator \mathcal{D} is updated twice for every update of \mathcal{G} [21] and we use Spectral Normalization [67] for all weight layers (approximated by the first singular value) and orthogonal initialization of weights [45]. Sampling is carried out using the exponential moving average of \mathcal{G} 's weights, which is accumulated with decay $\gamma = 0.9999$ starting after 20,000 training steps. The model is optimized using Adam [27] with batch size 512 and a learning rate of $1 \cdot 10^{-4}$ and $2 \cdot 10^{-4}$ for \mathcal{G} and \mathcal{D} respectively. Class conditioning in \mathcal{D} [35] is projection-based whereas \mathcal{G} relies on class-conditional Batch Normalization [23, 15, 17]: equivalent to standard Batch Normalization without a learned scale and offset, followed by an elementwise affine transformation where each parameter is a function of the noise vector and class conditioning.

A.4 Architecture Extension to Video Prediction

In order to provide results on future video prediction problems we describe a simple modification to DVD-GAN to facilitate the added conditioning. A diagram of the extended model is in Figure 10.

Given C conditioning frames $\{f_1, \dots, f_C\}$, our modified DVD-GAN-FP passes each frame separately through a deep residual network identical to \mathcal{D}_S . The resulting features for each conditioning frame are concatenated in the channel dimension and a 1×1 convolution reduces the channel number to 512. The spatial dimensions of the output of this network exactly match the shape of the recurrent state in \mathcal{G} 's RNN, so we pass the output of this ResNet as the initial state. Both \mathcal{D}_T and \mathcal{D}_S operate on the concatenation of the conditioning frames and the output of \mathcal{G} , meaning that the discriminators do not receive any extra information detailing that the first C frames are special. To further facilitate the preservation of small details from the conditioning frames we add U-Net [40] style skip connections between the intermediate outputs of the residual blocks processing the conditioning frames and the intermediate features of \mathcal{G} . To be precise, the features from the i -th block of the conditioning frame's network are concatenated with the i -th block of \mathcal{G} 's residual network, and the resulting features are passed through a single 3×3 convolution to compress the number of channels back to the normal

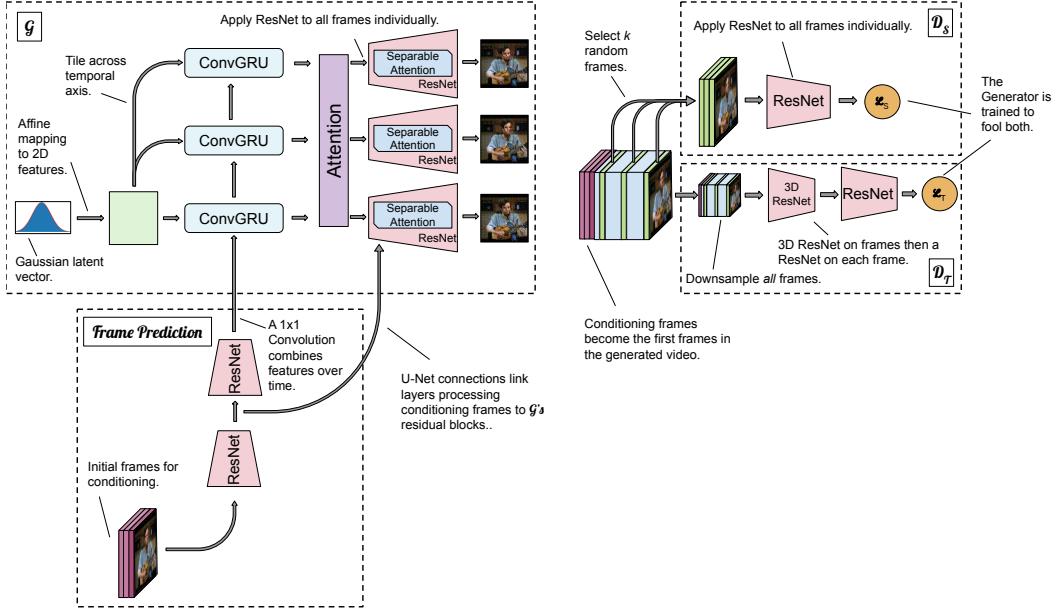


Figure 10: An architecture diagram describing the changes for the frame conditional model.

size before continuing with the next block. Finally, our video prediction variant does not condition on any class information, allowing us to directly compare with prior art.

B Further Experiments

B.1 BAIR Robot Pushing

We test future video prediction on the single-class BAIR Robot Pushing Dataset [18], a dataset of stationary videos of a robot arm moving around a set of changing objects. In order for direct comparison with previous results reported in [56] (for the SAVP model [30]) and [62], we consider generating 15 frames conditioned on a single starting frame. Results are reported in Table 4. We find that DVD-GAN-FP is competitive with previous adversarial approaches, but worse than concurrent results from autoregressive models [62].

B.2 UCF-101

UCF-101 [46] is a dataset of 13,320 videos of human actions across 101 classes that has previously been used for video synthesis and prediction [42, 41, 54]. We report Inception Score (IS) calculated with a C3D network [53] for quantitative comparison with prior work.³ Our model produces samples with an IS of 32.97, significantly outperforming the state of the art (see Table 2). The DVD-GAN architecture on UCF-101 is identical to the model used for Kinetics, and is trained on 16-frame 128×128 clips from UCF-101.

However, it is worth mentioning that our improved score is, at least partially, due to memorization of the training data. In Figure 11 we show interpolation samples from our best UCF-101 model. Like Figure 8, we sample 2 latents (left and rightmost columns) and show samples from the linear

Table 4: FVD on BAIR Robot Pushing (lower is better). Results for Video Transformer are from [62], all other results are as reported in [56].

Method	FVD
SV2P [5]	262.5
SAVP [30]	116.4
Video Transformer [62]	94 ± 2
DVD-GAN-FP (ours)	127.1

³We use the Chainer [52] implementation of Inception Score for C3D available at <https://github.com/pfnet-research/tgan>.



Figure 11: The first frames of interpolations between UCF-101 samples. Each row is a separate interpolation. Contrast with Figure 8 and Appendix D.2.

interpolation in latent space along each row. Here we show 4 such interpolations (the first frame from each video). Unlike Figure 8 which smoothly transitions from one sample to the other, we see abrupt jumps in the latent space between highly distinct samples, and little intra-video diversity between samples in each group. It can be further seen that some generated samples highly correlate with samples from the training set.

We show this both as a failure of the Inception Score metric, the commonly reported value for class-conditional video synthesis on UCF-101, but also as strong signal that UCF-101 is not a complex or diverse enough dataset to facilitate interesting video generation. Each class is relatively small, and reuse of clips from shared underlying videos means that the intra-class diversity can be restricted to just a handful of videos per class. This suggests the need for larger, more diverse and challenging datasets for generative video modelling, and we believe that Kinetics-600 provides a better benchmark for this task.

C Miscellaneous Experiments

Here we detail a number of modifications or miscellaneous results we experimented with which did not produce a conclusive result.

- We experimented with several variations of normalization which do not require calculating statistics over a batch of data. Group Normalization [63] performed best, almost on a par with (but worse than) Batch Normalization. We further tried Layer Normalization [31], Instance Normalization [55], and no normalization, but found that these significantly underperformed Batch Normalization.
- We found that removing the final Batch Normalization in \mathcal{G} , which occurs after the ResNet and before the final convolution, caused a catastrophic failure in learning. Interestingly, just removing the Batch Normalization layers within \mathcal{G} 's residual blocks still led to good (though slightly worse) generative models. In particular, variants without Batch Normalization in the residual blocks often achieve significantly higher IS (up to 110.05 for 64×64 12 frame samples – twice normal). But these models had substantially worse FID scores (1.22 for the aforementioned model) – and produced qualitatively worse video samples.
- Early variants of DVD-GAN contained Batch Normalization which normalized over all frames of all batch elements. This gave \mathcal{G} an extra channel to convey information across time. It took advantage of this, with the result being a model which required batch statistics in order to produce good samples. We found that the version which normalizes over timesteps independently worked just as well and without the dependence on statistics.
- Models based on the residual blocks of BigGAN-deep trained faster (in wall clock time) but slower with regards to metrics, and struggled to reach the accuracy of models based on BigGAN's residual blocks.

D Generated Samples

It is difficult to accurately convey complicated generated video through still frames. Where provided, we recommend readers view the generated videos themselves via the provided links. We refer to videos within these batches by row/column number where the video in the 0th row and column is in the top left corner.

D.1 Synthesis Samples

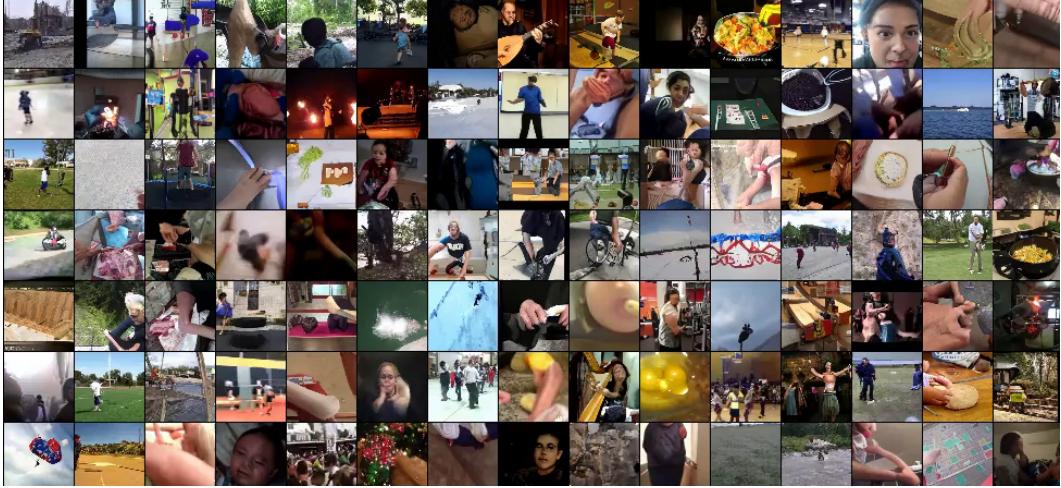


Figure 12: The first frames from a random batch of samples from DVD-GAN trained on 12 frames of 64×64 Kinetics-600. Full samples at https://drive.google.com/open?id=1YJtaQgVDnt_r35xKghelgd4V8Po-Ueaz.

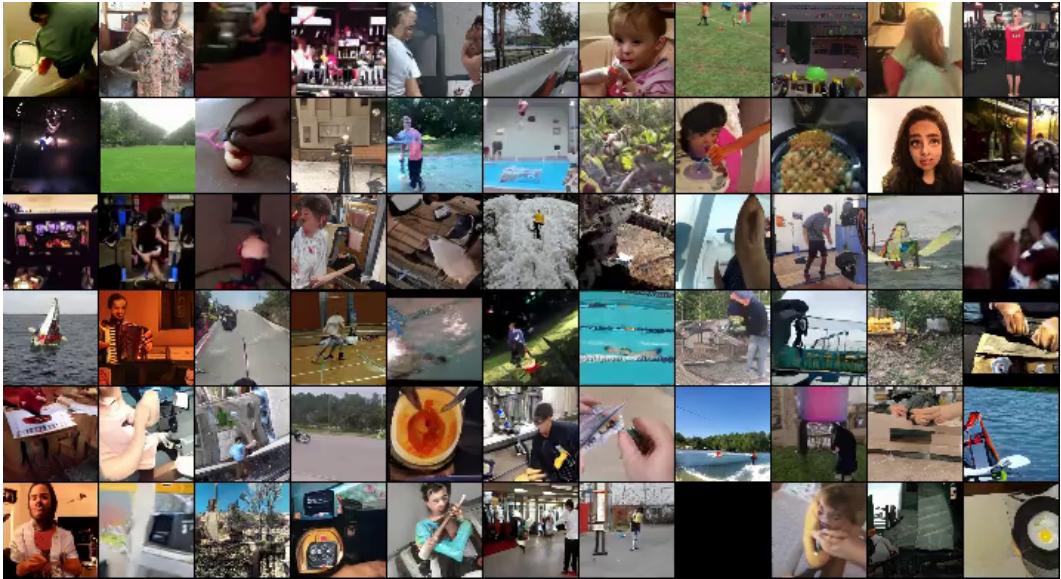


Figure 13: The first frames from a random batch of samples from DVD-GAN trained on 48 frames of 64×64 Kinetics-600. Full samples at <https://drive.google.com/open?id=18pcN8W1AH-1VbGMCrR5VzllXsqbEchL0>. The video in row 1, column 1 showcases DVD-GANs capability to remember details of the generation which are "off-screen", while the video in row 1, column 5 displays a complex cause-and-effect relationship between the diver and the water splash.

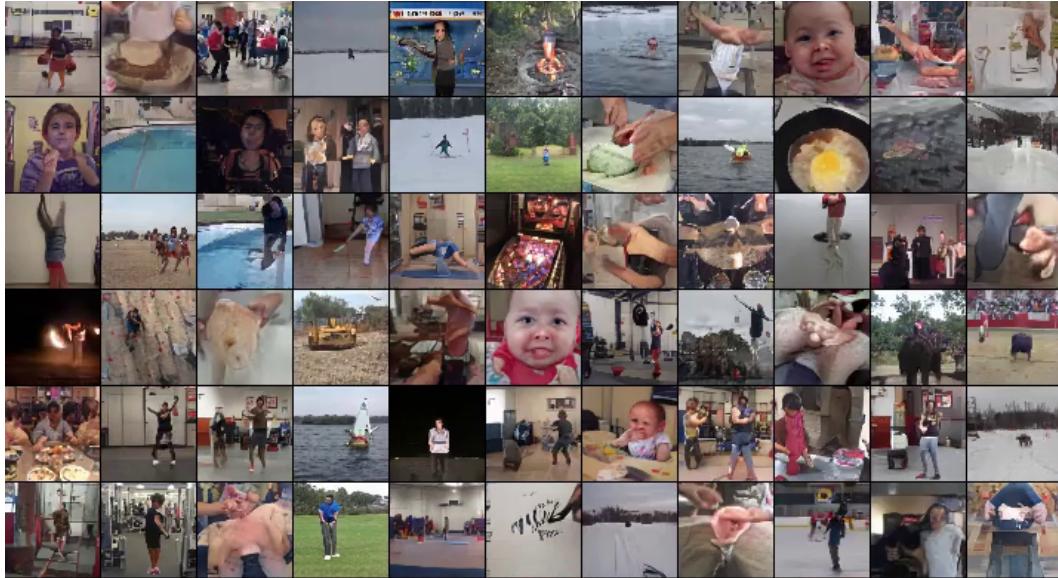


Figure 14: The first frames from a random batch of **truncated** samples from DVD-GAN trained on 48 frames of 64×64 Kinetics-600. Full samples at <https://drive.google.com/open?id=1XbqdFD70JrWSkOsW4v31QCQexpjxlALB>. These samples were taken from a Gaussian distribution with $\sigma = 0.44$.

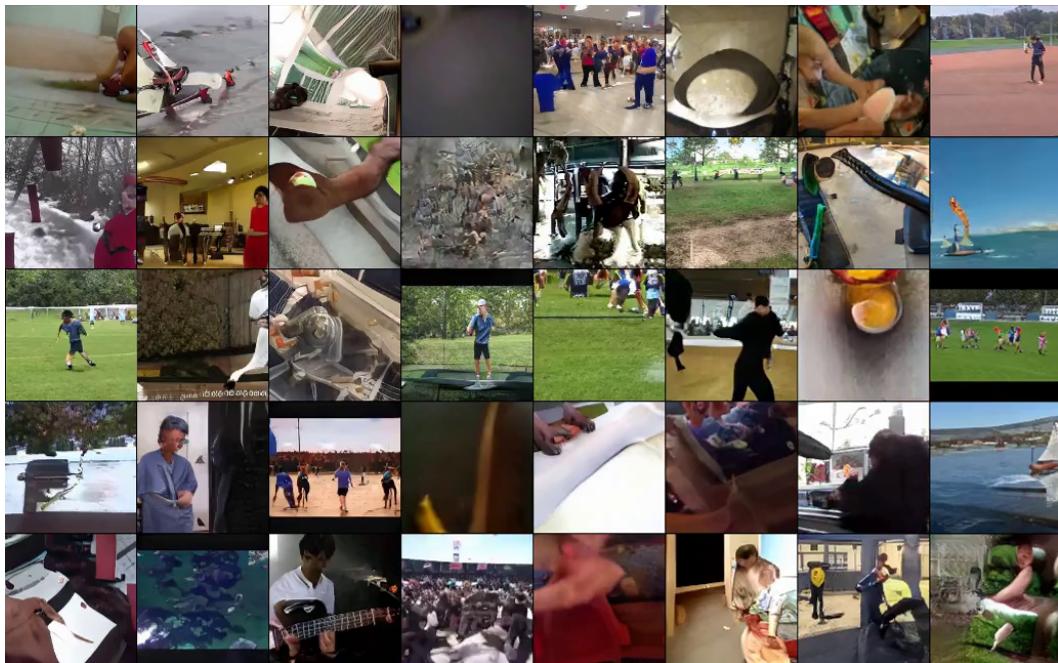


Figure 15: The first frames from a random batch of samples from DVD-GAN trained on 12 frames of 128×128 Kinetics-600. Full samples at <https://drive.google.com/open?id=15MkrAkP3B9U4n12CgWUdlzr-YROEDRSk>

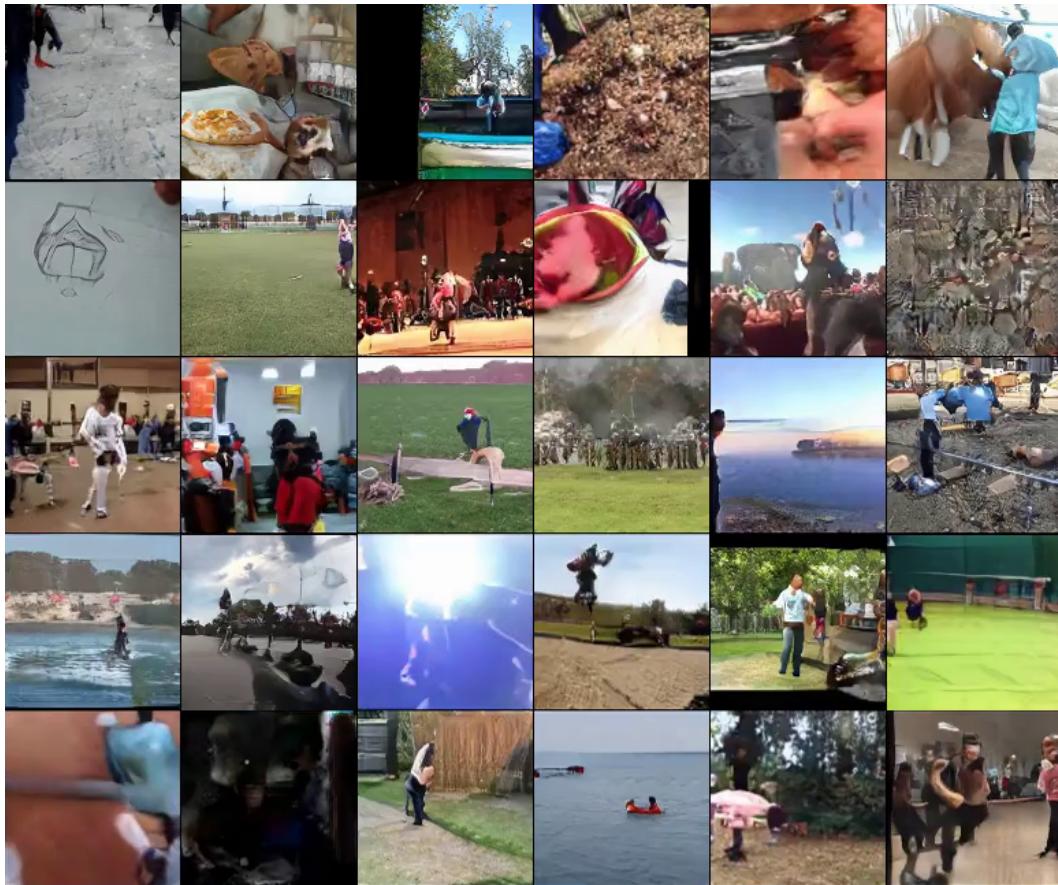


Figure 16: The first frames from a random batch of samples from DVD-GAN trained on 48 frames of 128×128 Kinetics-600. Full samples at <https://drive.google.com/open?id=19kXShENC-7KC-VjkIR3GixcdLVgSGSW5>. The sample in row 1, column 5 is a stereotypical example of a degenerate sample occasionally produced by DVD-GAN.



Figure 17: The first frames from a random batch of samples from DVD-GAN trained on 12 frames of 256×256 Kinetics-600. Full samples at <https://drive.google.com/open?id=1wagcMpBAnIfYSEgnOoAbEJoqmHTnprcr>.

D.2 Interpolation Samples

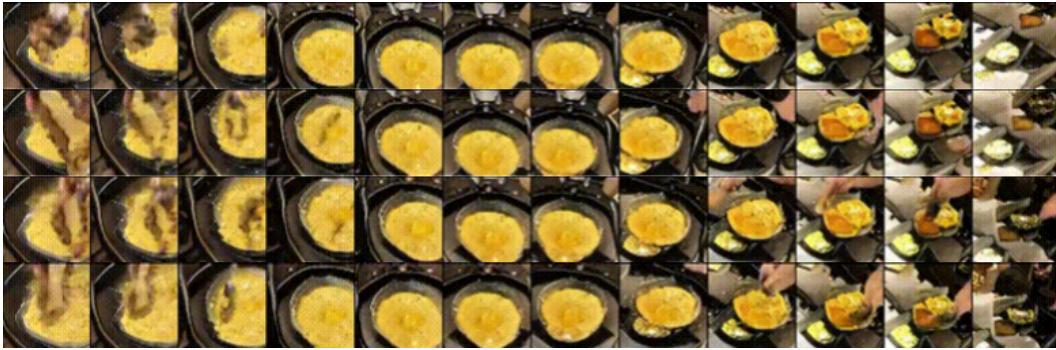


Figure 18: An example *intra-class* interpolation. Each column is a separate video (the vertical axis is the time dimension). The left and rightmost columns are randomly sampled latent vectors and are generated under a shared class. Columns in between represent videos generated under the same class across the linear interpolation between the two random samples. Note the smooth transition between videos at all four timesteps displayed here.



Figure 19: Another example of *intra-class* interpolation.



Figure 20: An example of *class* interpolation. As before, each column is a sequence of timesteps of a single video. Here, we sample a **single** latent vector, and the left and rightmost columns represent generating a video of that latent under two different classes (swimming and fire dancing). Columns in between represent videos of that same latent generated across an interpolation of the class embedding. Even though at no point has DVD-GAN been trained on data under an interpolated class, it nevertheless produces reasonable samples.