Generative Ai :Text to Video Generation

Mallarapu Yagnesh Naidu, Modugula Gopinath, Midde Venkat Sai, Y.V.Niranjan Reddy and Hiren Raithatha

March 7, 2024

Generative AI :Text to video Generation

By YAGNESH MALLARAPU

WORD COUNT	3971	TIME SUBMITTED	28-FEB-2024 11:14AM
		PAPER ID	107200140

Generative AI: Text to video generation

1.Mallarapu Yagnesh Naidu
ERP - 210304124030
CSE-AI
P.I.E.T
Parul University, Vadodara
210304124030@paruluniversity.ac.in

2.Modugula Gopinath
ERP - 210304124079
CSE-AI
P.I.E.T
Parul University, Vadodara
210304124079@paruluniversity.ac.in

3.Midde Venkat Sai
ERP - 210304124074
CSE-AI
P.I.E.T
Parul University, Vadodara
210304124074@paruluniversity.ac.in

4.Y.V. Niranjan Reddy
ERP - 210304124449
CSE-AI
P.I.E.T
Parul University, Vadodara
210304124449@paruluniversity.ac.in

5.HirenRaiththa
Faculty of PU
CSE-AI
P.I.E.T
Parul University, Vadodara
hiren.raithatha19387@paruluniversity.
ac.in

Abstract

We present Stable Video Diffusion — a latent video diffusion model for high-resolution, state-of-the-art text-to-video and image-to-video generation. Recently, latent diffusion models trained for 2D image synthesis have been turned into generative video models by inserting temporal layers and finetuning them on small, high-quality video datasets. However, training methods in the literature vary widely, and the field has yet to agree on a unified strategy for curating video data. In this paper, we identify and evaluate three different stages for successful training of video LDMs: text-to-image pretraining, video pretraining, and high-quality video finetuning. Furthermore, we demonstrate the necessity of a well-curated pretraining dataset for generating high-quality videos and present a systematic curation process to train a strong base model, including captioning and filtering strategies. We then explore the impact of finetuning our base model on high-quality data and train a text-to-video model that is competitive with closed-source video generation. We also show that our base empirical study on the effect of data curation during video.

1.Introduction

Driven by advances in generative image modelling with diffusion models, there has been significant recent progress on generative video models both in research and real-world applications. Broadly, these models are either trained from scratch or finetuned (partially or fully) from pretrained image models with additional temporal layers inserted. Training is often carried out on a mix of image and video datasets.

While research around improvements in video modelling has primarily focused on the exact arrangement of the spatial and temporal layers, none of the afore mentioned works investigate the influence of data selection. This is surprising, especially since the significant impact of the training data distribution on generative models is undisputed. Moreover, for generative image modelling, it is known that pretraining on a large and diverse dataset and finetuning on a smaller but higher quality dataset significantly improves the performance. Since many previous approaches to video modelling have successfully drawn on techniques from the image domain, it is noteworthy that the effect of data and training strategies, i.e., the separation of video pretraining at lower resolutions and high-quality finetuning, has yet to be studied. This work directly addresses these previously uncharted territories. We believe that the significant contribution of data selection is heavily underrepresented in today's video research landscape despite being well-recognized among practitioners when training video models at scale. Thus, in contrast to previous works, we draw on simple latent video diffusion baselines for which we fix architecture and training scheme and assess the effect of data curation. To this end, we first identify three different video training stages that we find crucial for good performance: text-to-image pre training, video pretraining on a large dataset at low resolution, and high-resolution video finetuning on a much smaller dataset with higher-quality videos. Borrowing from large scale image model training, we introduce a systematic approach to curate video data at scale and present an empirical study on the effect of data curation during video pretraining. Our main findings imply that pretraining on well-curated datasets leads to significant performance improvements that persist after high-quality finetuning.

A general motion and multi-view prior

Drawing on these findings, we apply our proposed curation scheme to a large video dataset comprising roughly 600 million samples and train a strong pretrained text-to-video base model, which provides a general motion representation. We exploit this and finetune the base model on a smaller, high-quality dataset for high-resolution downstream tasks such as text to-video (see Figure 1, top row) and image-to-video, where we predict a sequence of frames from a single conditioning image (see Figure 1, mid rows). Human preference studies reveal that the resulting model outperforms state-of-the-art image-to-video models. Furthermore, we also demonstrate that our model pro videos a strong multi-view prior and can serve as a base to finetune a multi-view diffusion model that generates multiple consistent views of an object in a feedforward manner and outperforms specialized novel view synthesis methods such as Zero123XL and SyncDreamer. Finally, we demonstrate that our model allows for explicit motion control by specifically prompting the temporal layers with motion cues and also via training LoRA modules on datasets resembling specific motions only, which can be efficiently plugged into the model. To summarize, our core contributions are threefold: (i) We present a systematic data curation workflow to turn a large uncurated video collection into a quality dataset for generative video modelling. Using this workflow, we (ii) train state-of-the-art text-to-video and image-to-video models, outperforming all prior models. Finally, we (iii) probe the strong prior of motion and 3D understanding in our models by conducting domain-specific experiments. Specifically, we provide evidence that pretrained video diffusion models can be turned into strong multi-view generators, which may help overcome the data scarcity typically observed in the 3D domain.

2. Background

Latent Video Diffusion Models Video-LDMs train the main generative model in a latent space of reduced computational complexity. Most related works make use of a pretrained text-to-image model and insert temporal mixing layers of various forms into the pretrained architecture. Ge et al. addition ally relies on temporally correlated noise to increase temporal consistency and ease the learning task. In this work, we follow the architecture proposed in Blattmann et al. and insert temporal convolution and attention layers after every spatial convolution and attention layer. In contrast to works that only train temporal layers or are completely training-free, we finetune the full model. For text to-video synthesis in particular, most works directly condition the model on a text prompt or make use of an additional text-to-image prior. In our work, we follow the former approach and show that the resulting model is a strong general motion prior, which can easily be finetuned into an image-to-video or multi-view synthesis model. Additionally, we introduce micro-conditioning on frame rate. We also employ the EDM-framework and significantly shift the noise schedule towards higher noise values, which we find to be essential for high-resolution finetuning.

3. Curating Data for HQ Video Synthesis

In this section, we introduce a general strategy to train a state-of-the-art video diffusion model on large datasets of videos. To this end, we (i) introduce data processing and cu ration methods, for which we systematically analyze the impact on the quality of the final model, and (ii) identify three different training regimes for generative video modelling. In particular, these regimes consist of • Stage I: image pretraining, i.e. a 2D text-to-image diffusion model. • Stage II: video pretraining, which trains on large amounts of videos. • Stage III: video finetuning, which refines the model on a small subset of high-quality videos at higher resolution.

3.1. Data Processing and Annotation

We collect an initial dataset of long videos which forms the base data for our video pretraining stage. To avoid cuts and fades leaking into synthesized videos, we apply a cut detection pipeline1 in a cascaded manner at three different FPS levels. After applying our cut-detection pipeline, we obtain a significantly higher number (∼4×) of clips, indicating that many video clips in the unprocessed dataset contain cuts beyond those obtained from metadata. Next, we annotate each clip with three different synthetic captioning methods: First, we use the image captioner CoCa to annotate the mid-frame of each clip and use V-BLIP to obtain a video-based caption. Finally, we generate a third description of the clip via an LLM-based summarization of the first two captions. The resulting initial dataset, which we dub Large Video Dataset (LVD), consists of 580M annotated video clip pairs, forming 212 years of content. However, further investigation reveals that the resulting dataset contains examples that can be expected to degrade the performance of our final video model, such as clips with less motion, excessive text presence, or generally low aesthetic value. We therefore additionally annotate our dataset with dense optical flow, which we calculate at 2 FPS and with which we filter out static scenes by removing any videos whose average optical flow magnitude is below a certain threshold. Indeed, when considering the motion distribution of LVD via optical flow scores, we identify a subset of close-to-static clips therein. Moreover, we apply optical character recognition to weed out clips containing large amounts of written text. Lastly, we annotate the first, middle, and last frames of each clip with CLIP embeddings from which we calculate aesthetics scores as well as text-image similarities.

3.2. Stage I: Image Pretraining

We consider image pretraining as the first stage in our training pipeline. Thus, in line with concurrent work on video models, we ground our initial model on a pre trained image diffusion model- namely Stable Diffusion to equip it with a strong visual representation. To analyze the effects of image pretraining, we train and compare two identical video models as detailed in App. D on a 10M subset of LVD; one with and one without pre trained spatial weights. We compare these models using a human preference study which clearly shows that the image-pretrained model is preferred in both quality and prompt-following.

3.3. Stage II: Curating a Video Pretraining Dataset

A systematic approach to video data curation. For multimodal image modelling, data curation is a key element of many powerful discriminative and generative models. However, since there are no equally powerful off-the-shelf representations available in the video domain to filter out unwanted examples, we rely on human preferences as a signal to create a suitable pre training dataset. Specifically, we curate subsets of LVD using different methods described below and then consider the human-preference-based ranking of latent video diffusion models trained on these datasets.

3.4. Stage III: High-Quality Finetuning

In the previous section, we demonstrated the beneficial effects of systematic data curation for video pretraining. However, since we are primarily interested in optimizing the performance after video finetuning, we now investigate how these differences after Stage II translate to the final performance after Stage III. Here, we draw on training techniques from latent image diffusion modelling and increase the resolution of the training examples. Moreover, we use a small finetuning dataset comprising 250K pre-captioned video clips of high visual fidelity.

4. Training Video Models at Scale

In this section, we borrow takeaways and present results of training state-of-the-art video models at scale. We first use the optimal data strategy inferred from ablations to train a powerful base model at 320 × 576 in App. D.2. We then perform finetuning to yield several strong state-of-the-art models for different tasks such as text-to-video in Section 4.2, image-to-video in Section 4.3 and frame interpolation in Section 4.4. Finally, we demonstrate that our video-pretraining can serve as a strong implicit 3D prior, by tuning our image-to-video models on multi-view generation in Section 4.5 and outperform con current work, in particular Zero123XL and Sync Dreamer in terms of multi-view consistency.

4.1. Pretrained Base Model

Ours 0.7 0.6 0.5 0.4 0.3 0.2 User Preference 0.1 0.0 baseline Ours vs Pika Ours vs Gen2 Figure 6. Our 25 frame Image to-Video model is preferred by human voters over GEN-2 and PikaLabs. As discussed in Section 3.2, our video model is based on Stable Diffusion. Recent works show that it is crucial to adopt the noise schedule when training image diffusion models, shifting towards more noise for higher-resolution images. As a first step, we finetune the fixed discrete noise schedule from our image model towards continuous noise using the network preconditioning proposed in Karras et al. for images of size 256 × 384. After inserting temporal layers, we then train the model on LVD-F on 14 frames at resolution 256 × 384. We use the standard EDM noise schedule for 150k iterations and batch size 1536. Next, we finetune the model to generate 14320 ×576 frames for 100k iterations using batch size 768. We find that it is important to shift the noise schedule towards more noise for this training stage, confirming results by Hoogeboom et al. for image models. For further training details, see App. D. We refer to this model as our base model which can be easily finetuned for a variety of tasks as we show in the following sections. The base model has learned a powerful motion representation, for example, it significantly outperforms all baselines for zero-shot text to-video generation on UCF-101. Evaluation details can be found in App. E.

4.2. High-Resolution Text-to-Video Model

We finetune the base text-to-video model on a high-quality video dataset of ∼ 1M samples. Samples in the dataset generally contain lots of object motion, steady camera motion, and well-aligned captions, and are of high visual quality al together. We finetune our base model for 50k iterations at resolution 576 × 1024 (again shifting the noise schedule towards more noise) using batch size 768. Samples are, more can be found in App. E.

4.3. High Resolution Image-to-Video Model

Besides text-to-video, we finetune our base model for image-to-video generation, where the video model receives a still input image as a conditioning. Accordingly, we re place text embeddings that are fed into the base model with the CLIP image embedding of the conditioning. Additionally, we concatenate a noise-augmented version of the conditioning frame channel-wise to the input of the UNet. We do not use any masking techniques and simply copy the frame across the time axis. We finetune two models, one predicting 14 frames and another one predicting 25 frames; implementation and training details can be found in App. D. We occasionally found that standard vanilla classifier-free guidance can lead to artifacts: too little guidance may result in inconsistency with the conditioning frame while too much guidance can result in oversaturation. Instead of using a constant guidance scale, we found it helpful to linearly increase the guidance scale across the frame axis (from small to high). Details can be found in App. D. Samples are more can be found in App. E. In Section 4.5 we compare our model with state-of-the art, closed-source video generative models, in particular GEN-2 and PikaLabs, and show that our model is preferred in terms of visual quality by human voters. Details on the experiment, as well as many more image-to video samples, can be found in App. E.

4.3.1 Camera Motion LoRA

To facilitate controlled camera motion in image-to-video generation, we train a variety of camera motion LoRAs within the temporal attention blocks of our model; see App. D for exact implementation details. We train these additional parameters on a small dataset with rich camera motion metadata. In particular, we use three subsets of the data for which the camera motion is categorized as “horizontally moving”, “zooming”, and “static”. We show samples of the three models for identical conditioning frames; more samples can be found in App. E.

4.4. Frame Interpolation

To obtain smooth videos at high frame rates, we finetune our high-resolution text-to-video model into a frame interpolation model. We follow Blattmann et al. and concatenate the left and right frames to the input of the UNet via masking. The model learns to predict three frames within the two conditioning frames, effectively increasing the frame rate by four. Surprisingly, we found that a very small number of iterations (≈ 10k) suffices to get a good model. Details and samples can be found in App. D and App. E, respectively.

4.5. Multi-View Generation

To obtain multiple novel views of an object simultaneously, we finetune our image-to-video SVD model on multi-view datasets. Datasets. We finetuned our SVD model on two datasets, where the SVD model takes a single image and outputs a sequence of multi-view images: (i) A subset of Obja verse [14] consisting of 150K curated and CC-licensed synthetic 3D objects from the original dataset. For each object, we rendered 360◦ orbital videos of 21 frames with 7 randomly sampled HDRI environment map and elevation angles between [−5◦,30◦]. We evaluate the resulting models on an unseen test dataset consisting of 50 sampled objects from Google Scanned Objects (GSO) dataset. and (ii) MVImgNet consisting of casually captured multi view videos of general household objects. We split the videos into ∼200K train and 900 test videos. We rotate the frames captured in portrait mode to landscape orientation. The Objaverse-trained model is additionally conditioned on the elevation angle of the input image, and outputs orbital videos at that elevation angle. The MVImgNet-trained models are not conditioned on pose and can choose an arbitrary camera path in their generations. For details on the pose conditioning mechanism, see App. E.

Models

We refer to our finetuned Multi-View model as SVD-MV. We perform an ablation study on the importance of the video prior of SVD for multi-view generation. To this effect, we compare the results from SVD MV i.e. from a video prior to those finetuned from an image prior i.e. the text-to-image model SD2.1 (SD2.1 MV), and that trained without a prior i.e. from random initialization (Scratch-MV). In addition, we compare with the current state-of-the-art multi-view generation models of Zero123, Zero123XL, and SyncDreamer.

Metrics

We use the standard metrics of Peak Signal-to Noise Ratio (PSNR), LPIPS, and CLIP Similarity scores (CLIP-S) between the corresponding pairs of ground truth and generated frames on 50 GSO test objects.

Training

We train all our models for 12k steps (~16 hours) with 8 80GB A 100 GPUs using a total batch size of 16, with a learning rate of 1e-5.

Results

The average metrics on the GSO test dataset. The higher performance of SVD-MV compared to SD2.1-MV and Scratch-MV clearly demonstrates the advantage of the learned video prior in the SVD model for multi-view generation. In addition, as in the case of other models finetuned from SVD, we found that a very small number of iterations (≈ 12k) suffices to get a good model. Moreover, SVD-MV is competitive w.r.t state-of the-art techniques with lesser training time (12k iterations in 16 hours), whereas existing models are typically trained for much longer (for example, SyncDreamer was trained for four days specifically on Objaverse). The convergence of different finetuned models. After only 1k iterations, SVD-MV has much better CLIP-S and PSNR scores than its image-prior and no-prior counterparts. A qualitative comparison of multi-view generation results on a GSO test object and Figure 10 on an MVImgNet test object. As can be seen, our generated frames are multi-view consistent and realistic. More details on the experiments, as well as more multi-view generation samples, can be found in App. E

5. Conclusion

We present Stable Video Diffusion (SVD), a latent video diffusion model for high-resolution, state-of-the-art text-to video and image-to-video synthesis. To construct its pre training dataset, we conduct a systematic data selection and scaling study, and propose a method to curate vast amounts of video data and turn large and noisy video collection into suitable datasets for generative video models. Furthermore, we introduce three distinct stages of video model training which we separately analyze to assess their impact on the final model performance. Stable Video Diffusion provides a powerful video representation from which we finetune video models for state-of-the-art image-to-video synthesis and other highly relevant applications such as LoRAs for camera control. Finally we provide a pioneering study on multi-view finetuning of video diffusion models and show that SVD constitutes a strong 3D prior, which obtains state of-the-art results in multi-view synthesis while using only a 8 fraction of the compute of previous methods. We hope these findings will be broadly useful in the generative video modelling literature. A discussion on our work's broader impact and limitations can be found in App. A.

References

[1] Jie An, Songyang Zhang, Harry Yang, Sonal Gupta, Jia Bin Huang, Jiebo Luo, and Xi Yin. Latent-shift: Latent diffusion with temporal shift for efficient text-to-video generation. arXiv preprint arXiv:2304.08477, 2023.

[2] Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Jackson Kernion, Ka mal Ndousse, Catherine Olsson, Dario Amodei, Tom Brown, JackClark, SamMcCandlish, ChrisOlah, andJared Kaplan. A general language assistant as a laboratory for alignment, 2021.

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

[4] Youngmin Baek, Bado Lee, Dongyoon Han, Sangdoo Yun, and Hwalsuk Lee. Character region awareness for text detection. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 9365–9374, 2019.

[5] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. Training a helpful and harm less assistant with reinforcement learning from human feedback, 2022.

[6] Max Bain, Arsha Nagrani, Gul Varol, and Andrew Zisser man. Frozen in time: A joint video and image encoder for end-to-end retrieval, 2022.

[7] Andreas Blattmann, Timo Milbich, Michael Dorkenwald, and Bjorn Ommer. ipoke: Poking a still image for controlled stochastic video synthesis. In 2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021, 2021.

[8] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja

Fidler, and Karsten Kreis. Align your Latents: High-Resolution Video Synthesis with Latent Diffusion Models. arXiv:2304.08818, 2023.

[9] Tim Brooks, Janne Hellsten, Miika Aittala, Ting-Chun Wang, Timo Aila, Jaakko Lehtinen, Ming-Yu Liu, Alexei A Efros, and Tero Karras. Generating long videos of dynamic scenes. 2022.

[10] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 6299–6308, 2017.

[11] Lluis Castrejon, Nicolas Ballas, and Aaron Courville. Improved conditional vrnns for video prediction. In The IEEE International Conference on Computer Vision (ICCV), 2019.

[12] Xiaoliang Dai, Ji Hou, Chih-Yao Ma, Sam Tsai, Jialiang Wang, Rui Wang, Peizhao Zhang, Simon Vandenhende, Xi aofang Wang, Abhimanyu Dubey, Matthew Yu, Abhishek Kadian, Filip Radenovic, Dhruv Mahajan, Kunpeng Li, Yue Zhao, Vladan Petrovic, Mitesh Kumar Singh, Simran Mot wani, Yi Wen, Yiwen Song, Roshan Sumbaly, Vignesh Ra manathan, Zijian He, Peter Vajda, and Devi Parikh. Emu: Enhancing image generation models using photogenic needles in a haystack, 2023.

[13] Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Chris tian Laforte, Vikram Voleti, Samir Yitzhak Gadre, et al. Objaverse-XL: A universe of 10m+ 3d objects. arXiv preprint arXiv:2307.05663, 2023.

[14] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 13142–13153, 2023.

[15] Emily Denton and Rob Fergus. Stochastic video generation with a learned prior. In Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsm¨assan, Stockholm, Sweden, July 10-15, 2018, 2018.

[16] Prafulla Dhariwal and Alex Nichol. Diffusion Models Beat GANsonImage Synthesis. arXiv:2105.05233, 2021.

[17] Michael Dorkenwald, Timo Milbich, Andreas Blattmann, Robin Rombach, Konstantinos G. Derpanis, and Bjorn Om mer. Stochastic image-to-video synthesis using cinns. In IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021, 2021.

[18] Laura Downs, Anthony Francis, Nate Koenig, Brandon Kinman, Ryan Hickman, Krista Reymann, Thomas B McHugh, and Vincent Vanhoucke. Google scanned objects: A high-quality dataset of 3d scanned household items. In 2022 International Conference on Robotics and Automation (ICRA), pages 2553–2560. IEEE, 2022.

[19] Arpad E. Elo. The Rating of Chessplayers, Past and Present. Arco Pub., New York, 1978.

[20] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. arXiv preprint arXiv:2012.09841, 2020.