Short review of state-of-the-art methods on Video Generation.

Anonymous CVPR submission

Paper ID *****

Abstract

This short review provides a summary of state-of-the-art video generation methods and an algorithm for developing a video generation network, which is presented in the final sections.

1. Literature Review

Conventionally, auto-regressive models are preferred over GANs for generating visual data [19, 20, 26] due to their explicit density modeling and stable training. The outputs of an auto-regressive model are usually generated in a pixel-by-pixel manner. This method requires high computational power. Hence, they are mostly used for low-resolution images. CNN [21], pixel RNN [19], image transformer [23], iGPT [5], and Video Transformer [33] have employed instances of such methods.

Video generation from text differs from video prediction, where only the previous frames are used to generate the forthcoming ones. FitVid [2] from GoogleBrain is an example of a video prediction model. FitVid solved the under/over-fitting problem with the use of data augmentation and achieved state-of-the-art performance in video prediction. It has no complex structure, no attention mechanism, no curriculum learning, and no training scheduling. It comprises two LSTM layers for modeling the dynamics and a few convolutional layers.
[1] uses stochastic methods to model the inherent uncertainty in the task of video prediction by considering motion history. It uses generative models with latent variables for appearance and motion. there are other models *e.g*. SVG [7] and SRVP [10] that study video prediction. [4] implemented a network for generating videos out of an input image through sampling out of latent variables. This framework is adversarially trained in an unsupervised manner. Fig. 1 shows a network structure that generates videos from an input image. In the following sections, we review some of the state-of-the-art models on text-to-video generation, which stem from the task of video prediction.

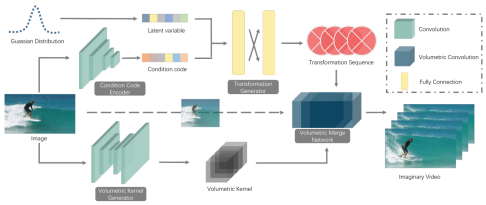


Figure 1. Video Imagination from a Single Image with Transformation Generation [4].

1.1. Models

Much research has recently focused on implementing networks capable of generating videos. [14, 22] combined GANS with 3D convolution layers to generate low-resolution, fixed-length videos. [3] used conditional filters to generate videos of varying lengths. [6] utilized LSTM cells with 2D convolutional networks to retain both temporal coherence and quality of the generated frames. However, the main problem with these works is that they only perform well on small datasets and suffer from low generalization ability. In [12], generating videos from text using transformer-based architectures has been investigated. This architecture receives an input text and tokenizes it. Ultimately, these embedded tokens provided to the transformer-based network create a set of images that form a low-framerate video. The low-quality video gets interpolated recursively by entering into a bidirectional attention network shown in Fig. 2.

1.1.1 Google Imagen [11]

Submitted on 5 October 2020, Imagen is the most recent T2V model. The following procedure describes the input pipeline, First, the input text enters a fully convolutional network, resulting in a low-resolution, low-framerate video. In the subsequent layers, using temporal super-resolution doubles the number of frames resulting in smoother animation. The next stage up-samples the resulting tensor and increases the spatial resolution. Eventually, the final two successive stages double the number of frames and

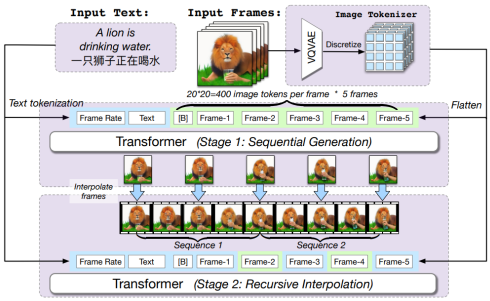


Figure 2. Multi-frame-rate hierarchical generation framework in CogVideo [12].

increase the image's resolution. The structure mentioned above makes use of fully-convolutional temporal and spatial super-resolution and v-parameterization of diffusion models.

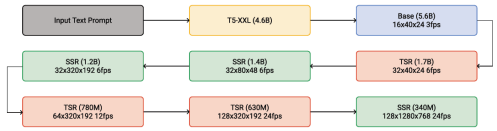


Figure 3. The cascaded sampling pipeline starting from a text prompt input to generating a 5.3- second, $1280 \cdot 768$ video at 24fps. "SSR" and "TSR" denote spatial and temporal super-resolution The videos are labeled as $frames \cdot width \cdot height$. In practice, the text embedding is injected into all models, not just the base model. [11].

1.1.2 META AI Make-a-Video [28]

This research paper submitted on 29 September 2022 belongs to Meta AI. The input text is first converted into an image. Subsequently, using the created image, a low-framerate movie is generated. At the following stages, there is an improvement in either the framerate by frame interpolation modules or in the quality with the help of spatiotemporal super-resolution convolution and attention layers. Fig. 4 shows the overall structure of the network.

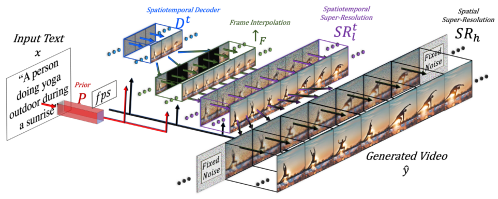


Figure 4. Make-A-Video structure from Meta AI. [28].

1.1.3 Temporal Video Generation [18]

The main component of this architecture is a latent code sequence generator that uses 3D multivariate Gaussian distributions to generate output samples. The generator's structure uses temporal residual connections that extract temporal features both from the past and the future.

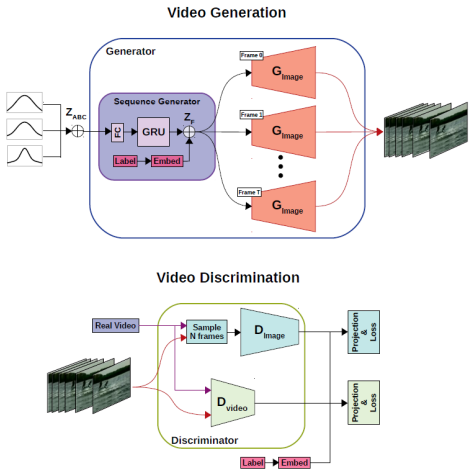


Figure 5. TS-GAN framework.

1.1.4 GODIVA [34]

GODIVA is a pre-trained model for T2V[1] generation, which is a result of a collaboration between Microsoft Research Asia and Duke university. It uses auto-regressive modeling principles and also a 3D sparse attention mechanism. This mechanism decreases the computational complexity. The same collaboration led to the publishing of another model called NUWA, introduced in the next section. GODIVA utilizes VQ-VAE as the encoder. The idea of employing VQ-VAEs for videos has already been used in the task of video prediction in [24, 27, 31, 32, 37, 38], but GODIVA is the first to use this structure for T2V tasks. According to Fig. 8, which demonstrates the results of training several networks on the Kinetics and Youtube videos datasets, the outputs look blurry in TFGAN, [3] whose network architecture is shown in Fig. 6. The metrics even indicate less promising results in T2V [15]. However, GODIVA's results show higher quality both in temporal and spatial domains. GODIVA is pre-trained on Howto100M, a large-scale text-video dataset with more than 136 million text-video pairs using 64 V100 GPUs. It is also fine-tuned on the MSR-VTT dataset with 8 V100 GPUs comprising 10000 human-annotated text-video pairs.

[1] text-to-video

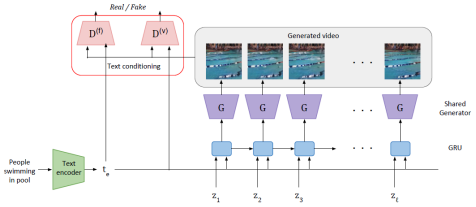


Figure 6. TF-GAN architecture.

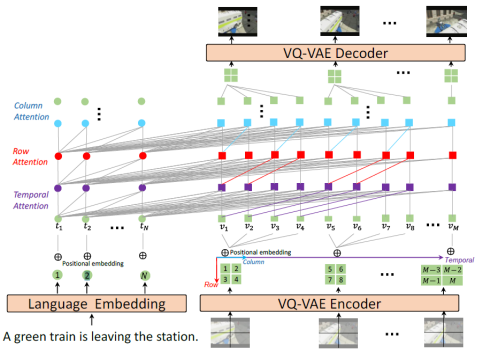


Figure 7. A simple illustration of our GODIVA model with a three-dimensional sparse attention mechanism for text-to-video generation task. The video is auto-regressively predicted with the consideration of four aspects: the input text, the same position of the previously generated frames, the same rows on the same frame, same columns on the same column.

Figure 8. GODIVA visual results

1.1.5 NUWA [35]

NUWA is a unified, multi-modal pre-trained model capable of manipulating existing visual data and generating new ones. NUWA supports texts, images, and videos using an adaptive transformer framework that receives 1D,2D, or 3D data as the input. The decoder can conduct eight visual synthesis tasks, including T2I, S2I, I2I, TI2I, T2V, S2V, V2V, and TV2V. The transformer uses a 3D Nearby Attention or (3DNA) mechanism capable of considering both spatial and temporal content information. This mechanism not only reduces the computational complexity but also improves the visual quality of the generated results. NUWA is trained using the Kinetics [13] dataset. The training results are compared to two other models in Tab. 1, verifying that NUWA outperforms other structures. Also, an illustration of the results can be seen in Fig. 9, which indicates that NUWA outperforms other structures in generating unseen videos, such as playing gulf at the swimming pool. This feature is called zero-shot capability.

Model	Acc	FID-img	FID-vid	CLIPSIM
T2V(64X64)	42.6	82.13	14.65	0.2853
SC(128X128)	74.7	33.51	7.34	0.2915
TFGAN(128X128)	76.2	31.76	7.19	0.2961
NUWA(128X128)	77.9	28.46	7.05	0.3012
Make-A-Video	-	-	13.17	0.3049
CogVideo (English)	-	-	23.59	0.2631

Table 1. The quantitative results of T2V using four state-of-the-art models trained on the Kinetics dataset. (The last two rows are trained on the MSR-VTT dataset.)

Figure 9. Results of eight tasks over an unseen text from NUWA, GODIVA, TFGAN, and T2V.

1.1.6 VideoGPT [36]

VideoGPT is a conceptually simple architecture that uses VQ-VAEs to create down-sampled latent representations of video frames with the help of convolutional layers and axial self-attention. It utilizes positional embedding both in the time and spatial domains. The results are comparable to

that of GANs and state-of-the-art models. The network has been trained on the BAIR Robot dataset resulting in high-fidelity outputs.

2. Proposed Algorithm

The idea behind the proposed algorithm is based on work in [9]. VQ-GAN is a kind of VQ-VAE that exploits a discriminator acting as an extra supervisor; This causes improvements in the quality of the generated data. As represented in Fig. 10, at the first stage of training, images are encoded using the encoder of a VQ-VAE, creating a representation of the frames sampled from the videos of the training set. The input sentence is also embedded into a latent space to assist output generation by defining a loss function measuring the distance between the representations created by the T2I unit.

A codebook is a mapper that quantizes the latent representations created by the encoder. It gets trained during the training phase of the VAE by considering a codebook mapping loss in the total loss term along the GAN loss. We may also consider some extra points in the latent space close to the created representations to retain the ability to create new and unseen examples. These representations can be concatenated to enter the decoder, where a reconstruction of the input is created. The reconstructed images' qualities are assessed by a reconstruction loss which can be a pixel-wise MSE. 3DNA-based attention can be used both for the input text and the input video frames a sequence generator to generate codebook vectors that make sense and go along with the sequences already seen in the codebook. In order to train the structure, first, the GAN and the codebook sections must be trained. The generator gets trained to map the input images into a latent representation with the help of a GAN loss and a reconstruction loss. Then the generator weights are kept constant so that the discriminator can be trained to learn the generator's flaws. This is executed recursively.

After freezing weights of the encoder, the decoder, and the codebook, it is time for the transformer to get trained. The proposed transformer uses a three-dimensional sparse attention model to be used both for images and texts, which means that it exploits temporal, column, and row information. 3DNA-based attention enables the network to be used for other types of sequences so that the network can learn to focus on the important regions of the frames. A transformer has an auto-regressive structure and works best with sequences. That is why it is suitable for generating sequences that best match the codebook examples.

The transformer in this structure acts as a sequence generator. It gets trained after the VQ-GAN to generate sequences of latent representations comparable to the codebook. Its training phase is as follows: some codebook vectors get randomly modified after the codebook training. These adjusted latent representations enter the transformer to generate images. The discriminator measures how real the created images are. As a result, the transformer weights get trained to generate sequences that match the ones stored in the codebook.

In other words, the transformer is supposed to correct the randomly adjusted quantized vectors.

A video's frame samples can be generated simultaneously, but if they are created frame-wisely, a temporal coherency loss term must be added to the structure. Briefly, the proposed structure uses two discriminators; one determines how fake the generated video frame is, and the other measures the temporal coherency of the generated frames. Also, a perceptual loss is defined to determine the visual realisticity and relative matching of the output according to [34].

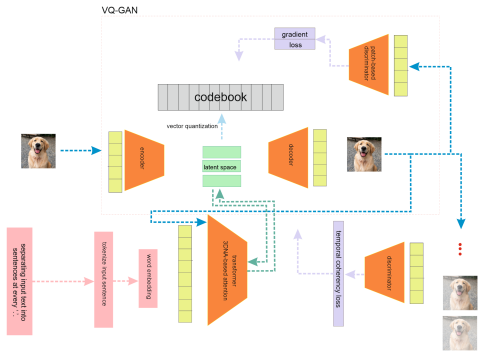


Figure 10. Proposed Structure

3. Advantages of the Proposed Algorithm

Considering previous discussions in the introduction, pixel-by-pixel models [5, 19, 21, 33] have explicit density modeling and stable training, which pushes us to prefer them over GANs [16, 17, 29, 30]. However, pixel-by-pixel data generation is problematic for high-dimensional data synthesis. It is evident that using 3D convolution models to generate videos requires considerable memory. Also, scaling the dimensions up for images of higher resolution is challenging.

After introducing VQ-VAEs and auto-regressive models for visual synthesis tasks, the idea of converting images into discrete visual tokens became widespread, as it enabled the researchers to conduct large-scale pre-training *e.g.* in DALL-E [27] and CogView [8], and for the tasks of video prediction in [25, 36] and video generation in GODIVA [34].

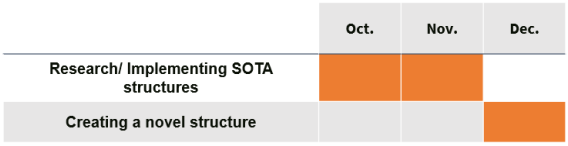
Also, the study in [35] showed that employing VQ-GANs instead of VQ-VAEs leads to creation of higher-quality videos. According to this paper, the loss function of

the encoder in VQ-VAE forces the reconstructed image to be exactly like the original one, which causes limitations in the generalization capabilities of VQ-VAEs. Hence by adding a perceptual loss term to the general loss function and an extra discriminator to the model, the results show more satisfactory generalization capabilities and semantic matching by eliminating exact similarity constraints.

4. Project Schedule

The Gantt chart considering the time needed to research and trying various implementations is shown in Tab. 2.

Table 2. Estimation of the RD timeline.



References

[1] Adil Kaan Akan, Erkut Erdem, Aykut Erdem, and Fatma Güney. Slamp: Stochastic latent appearance and motion prediction, October 2021. 1

[2] Mohammad Babaeizadeh, Mohammad Saffar, Suraj Nair, Sergey Levine, Chelsea Finn, and Dumitru Erhan. Fitvid: Overfitting in pixel-level video prediction, 06 2021. 1

[3] Yogesh Balaji, Martin Renqiang Min, Bing Bai, Rama Chellappa, and Hans Peter Graf. Conditional gan with discriminative filter generation for text-to-video synthesis. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 1995–2001. International Joint Conferences on Artificial Intelligence Organization, 7 2019. 1, 2

[4] Baoyang Chen, Wenmin Wang, and Jinzhuo Wang. Video imagination from a single image with transformation generation, 2017. 1

[5] Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pretraining from pixels, 13–18 Jul 2020. 1, 4

[6] Kangle Deng, Tianyi Fei, Xin Huang, and Yuxin Peng. Irc-gan: Introspective recurrent convolutional gan for text-to-video generation, 7 2019. 1

[7] Emily Denton and Rob Fergus. Stochastic video generation with a learned prior, 10–15 Jul 2018. 1

[8] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, and Jie Tang. Cogview: Mastering text-to-image generation via transformers, 2021. 4

[9] Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis, 2020. 4

[10] Jean-Yves Franceschi, Edouard Delasalles, Mickael Chen, Sylvain Lamprier, and Patrick Gallinari. Stochastic latent residual video prediction, 13–18 Jul 2020. 1

[11] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P. Kingma, Ben Poole, Mohammad Norouzi, David J. Fleet, and Tim Salimans. Imagen video: High definition video generation with diffusion models, 2022. 1, 2

[12] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*, 2022. 1, 2

[13] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset, 2017. 3

[14] Yitong Li, Martin Min, Dinghan Shen, David Carlson, and Lawrence Carin. Video generation from text, Apr. 2018. 1

[15] Yitong Li, Martin Min, Dinghan Shen, David Carlson, and Lawrence Carin. Video generation from text, Apr. 2018. 2

[16] Gaurav Mittal, Tanya Marwah, and Vineeth N. Balasubramanian. Sync-DRAW. In *Proceedings of the 25th ACM international conference on Multimedia*. ACM, oct 2017.

[17] Gaurav Mittal, Tanya Marwah, and Vineeth N. Balasubramanian. Sync-draw: Automatic video generation using deep recurrent attentive architectures. In *Proceedings of the 25th ACM International Conference on Multimedia*, MM '17, page 1096–1104, New York, NY, USA, 2017. Association for Computing Machinery. 4

[18] Andres Munoz, Mohammadreza Zolfaghari, Max Argus, and Thomas Brox. Temporal shift gan for large scale video generation, 2020. 2

[19] Aaron van den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. Pixel recurrent neural networks, 2016. 1, 4

[20] Aaron van den Oord, Nal Kalchbrenner, Oriol Vinyals, Lasse Espeholt, Alex Graves, and Koray Kavukcuoglu. Conditional image generation with pixelcnn decoders, 2016. 1

[21] Aaron van den Oord, Nal Kalchbrenner, Oriol Vinyals, Lasse Espeholt, Alex Graves, and Koray Kavukcuoglu. Conditional image generation with pixelcnn decoders, 2016. 1, 4

[22] Yingwei Pan, Zhaofan Qiu, Ting Yao, Houqiang Li, and Tao Mei. To create what you tell: Generating videos from captions, 2018. 1

[23] Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Łukasz Kaiser, Noam Shazeer, Alexander Ku, and Dustin Tran. Image transformer, 2018. 1

[24] Ruslan Rakhimov, Denis Volkhonskiy, Alexey Artemov, Denis Zorin, and Evgeny Burnaev. Latent video transformer, 2020. 2

[25] Ruslan Rakhimov, Denis Volkhonskiy, Alexey Artemov, Denis Zorin, and Evgeny Burnaev. Latent video transformer, 2020. 4

[26] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation, 2021. 1

[27] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation, 18–24 Jul 2021. 2, 4

[28] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, Devi Parikh, Sonal Gupta, and Yaniv Taigman. Make-a-video: Text-to-video generation without text-video data, 2022. 2

[29] Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, and Jan Kautz. Mocogan: Decomposing motion and content for video generation, 2017. 4

[30] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Generating videos with scene dynamics, 2016. 4

[31] Jacob Walker, Ali Razavi, and Aäron van den Oord. Predicting video with vqvae, 2021. 2

[32] Jacob Walker, Ali Razavi, and Aäron van den Oord. Predicting video with vqvae, 2021. 2

[33] Dirk Weissenborn, Oscar Täckström, and Jakob Uszkoreit. Scaling autoregressive video models, 2019. 1, 4

[34] Chenfei Wu, Lun Huang, Qianxi Zhang, Binyang Li, Lei Ji, Fan Yang, Guillermo Sapiro, and Nan Duan. GODIVA: generating open-domain videos from natural descriptions, 2021. 2, 4

[35] Chenfei Wu, Jian Liang, Lei Ji, Fan Yang, Yuejian Fang, Daxin Jiang, and Nan Duan. NÜwa: Visual synthesis pre-training for neural visual world creation, 2021. 3, 4

[36] Wilson Yan, Yunzhi Zhang, Pieter Abbeel, and Aravind Srinivas. Videogpt: Video generation using vq-vae and transformers, 2021. 3, 4

[37] Yunzhi Zhang, Wilson Yan, Pieter Abbeel, and Aravind Srinivas. Videogen: Generative modeling of videos using {vq}-{vae} and transformers, 2021. 2

[38] Yunzhi Zhang, Wilson Yan, Pieter Abbeel, and Aravind Srinivas. Videogen: Generative modeling of videos using {vq}-{vae} and transformers, 2021. 2