Efficient text-to-video generation with latent diffusion models

Hongxun Ding, Mingjian Zhu

DBGroup, SUSTech

Abstract. Text-to-video generation aims to generate videos from input text descriptions. It remains challenging due to the scarcity and weak relevance of text-video data as well as the high variation in videos, which can cause misalignment between the text and its temporal counterparts. Existing methods such as VQVAE and autoregressive transformers are widely used for text-to-video generation. In contrast, our project aims to address this task from a fresh perspective. Specifically, we propose to use the latent diffusion model to generate video from the input text as it has shown promising results in enhancing the realism of generated videos. Our goal is to generate videos from input text that have low FVD (Fréchet Video Distance) and high IS (Inception Score).

Keywords: text-to-video generation, latent diffusion model

1 Introduction

Text-to-image production has lately undergone a revolution thanks to autoregressive transformers like Cog View. It makes sense to look at the autoregressive transformers' potential for text-to-video production. This fundamental framework was used in earlier publications, such as VideoGPT, which confirmed its superiority to GAN-based approaches, but which are still far from perfect.

One issue is that the generated video frames frequently gradually veer away from the text prompt, making it difficult for the generated characters to carry out the intended actions. Typical autoregressive models may be effective in synthesising movies with regular or random patterns, such as speaking by randomly moving lips, but they fall short when given a text stimulus such as "a lion is drinking water." The main distinction between the two cases is that in the former, the first frame already contains enough data to support subsequent changes, whereas in the latter, the model must precisely comprehend the action "drink" in order to generate the desired action—the lion lifts the glass to its lip, downs it after drinking, before repeating the process.

Why do autoregressive transformers have no trouble comprehending text-image relationships but have trouble comprehending text-action relationships in videos? We postulate that the primary causes are the datasets and how they are used.

The main distinction between the two cases is that in the former, the first frame already contains enough data to support subsequent changes, whereas in the latter, the model must precisely comprehend the action "drink" in order to generate the desired action—the lion lifts the glass to its lip, downs it after drinking, before repeating the process. Why do autoregressive transformers have no trouble comprehending text-image relationships but have trouble comprehending text-action relationships in videos? We postulate that the primary causes are the datasets and how they are used.

Second, there is a wide range in the length of videos. The alignment between the text and its temporal counterparts in the video is destroyed by previous models' splitting the movie into several training clips with a set number of frames and a length of two. If a video of someone "drinking" is broken up into four separate movies showing people "holding a glass," "raising," "drinking," and "putting down," all with the same text, the model will be unable to understand the true definition of what it means to "drink."

2 Related work

2.1 Text-to-Image Generation

Text-to-image (T2I) generation has been a topic of research for some time. Initially, unconditional generative adversarial networks (GANs) were used for T2I generation. Later, GAN variants focused on progressive generation or better text-image alignment. DALLE, a pioneering work, considered T2I generation as a sequence-to-sequence translation problem using a discrete variational auto-encoder (VQVAE) and Transformer. Since then, additional variants have been proposed. Make-A-Scene explores controllable T2I generation using semantic maps, while Parti aims for more diverse content generation through an encoder-decoder architecture and an improved image tokenizer.

Denoising Diffusion Probabilistic Models (DDPMs) have been successfully leveraged for T2I generation. GLIDE trained a T2I and an upsampling diffusion model for cascade generation. GLIDE's proposed classifier-free guidance has been widely adopted in T2I generation to improve image quality and text faithfulness. DALLE-2 leverages the CLIP latent space and a prior model, while VQ-diffusion and stable diffusion perform T2I generation in the latent space instead of the pixel space to improve efficiency.

2.2 Text-to-Video Generation

Early works on video generation were mainly focused on simple domains, such as moving digits or specific human actions. Sync-DRAW is the first T2V generation approach that leverages a VAE with recurrent attention. Later, some works extended GAN from image generation to T2V generation. GODIVA is the first

to use 2D VQVAE and sparse attention for T2V generation, supporting more realistic scenes. NUWA extends GODIVA and presents a unified representation for various generation tasks in a multitask learning scheme.

To further improve the performance of T2V generation, CogVideo is built on top of a frozen CogView-2 T2I model by adding additional temporal attention modules. More recently, there has been an increase in the use of the diffusion model for video generation. Video Diffusion Models (VDM) uses a space-time factorized U-Net with joint image and video data training.

2.3 Diffusion model

Diffusion Models are probabilistic models that are designed to learn the probability distribution of data, represented as $p(x)$, by denoising a normally distributed variable gradually. The model is trained to perform the reverse process of a fixed Markov Chain of length T.

Lately, Diffusion Probabilistic Models (DM) have shown cutting-edge findings in sample quality and density estimation. When these models' neural underpinnings are implemented as UNets, they naturally adapt to the inductive biases of image-like data, which gives these models their generative capacity. When a reweighted target is utilized for training, the best synthesis quality is often attained. In this instance, the DM functions as a lossy compressor and allows for the trade-off of compression efficiency for picture quality. Nevertheless, the disadvantage of evaluating and improving these models in pixel space is low inference speed and very large training costs. Advanced sampling techniques and hierarchical methods can only partially address the first issue, but training on high-resolution picture data always necessitates the calculation of pricey gradients.

3 Method

We note that although diffusion models allow to ignore perceptually irrelevant details by undersampling the corresponding loss terms, they still require expensive function evaluations in pixel space, which causes huge demands in computation time and energy resources. This is to lower the computational demands of training diffusion models towards high-resolution image synthesis. By explicitly separating the compressive from the generative learning phases, we suggest avoiding this problem. We use an autoencoding approach to do this, which learns a space that is perceptually comparable to the picture space but has a far lower computational complexity.

Such a strategy has the following benefits: I By leaving the high-dimensional picture space, we are able to get DMs that employ sampling on a low-dimensional space, which is significantly more computationally efficient. We take use of the inductive bias that DMs have acquired through their UNet design, which makes

them especially useful for data with spatial structure and eliminates the requirement for harsh, quality-decreasing compression levels as required by prior techniques. In the end, we are left with general-purpose compression models, whose latent space may be applied to the training of numerous generative models as well as other downstream uses like single-image CLIP-guided synthesis.

3.1 Latent diffusion model

The latent diffusion model builds on the diffusion model and allows the diffusion operation to be performed in a latent space. With the use of trained perceptual compression models consisting of $\mathcal{E}$ and $\mathcal{D}$, the model is able to operate in a lower-dimensional, computationally efficient latent space in which high-frequency, imperceptible details are abstracted away. This makes the model more suitable for likelihood-based generative models, as they can focus on the important, semantic bits of the data and train in a more efficient space.

3.2 Conditioning mechanism

Diffusion models, like other generative models, may theoretically simulate conditional distributions of the form $p(aly)$. This opens the door to regulating the synthesis process through inputs y like text, semantic maps, or other image-to-image translation jobs, and may be accomplished with a conditional denoising autoencoder. Nevertheless, integrating the generating potential of DMs with other forms of conditionings besides class-labels or blurred variations of the input picture is still a relatively unexplored field of research in the context of image synthesis. By adding the cross-attention mechanism to the basic UNet backbone of DMs, which is useful for learning attention-based models of varied input modallities, we make DMs become more adaptable conditional image generators.

4 Expected evaluation metrics

We dicide to follow the work CogVideo by reporting Frechet Video Distance (FVD) and Inception Score (IS) in our future experiments.

4.1 Frechet Video Distance

A crucial step towards a more accurate evaluation of models for producing videos is the Frechet Video Distance (FVD) evaluation measure for generative models of video. FVD may be applied in circumstances when this is not the case by design, for as when creating unconditional videos using generative adversarial networks. When reviewing movies that have been altered to contain static noise and temporal noise, FVD is reliable. The fact that FVD regularly beats SSIM and PSNR in agreeing with human judgment is more significant than the results of a large-scale human investigation conducted on produced movies from various current generative models.

4.2 Inception Score

The Inception score gives us a basis for comparing the quality of these models. The reasearchers have applied the technique to the problem of semi-supervised learning, achieving state-of-the-art results on a number of different data sets in computer vision.

5 Reference

[1] Reed S, Akata Z, Yan X, et al. Generative adversarial text to image synthesis[C] International conference on machine learning. PMLR, 2016: 1060-1069.
[2] Hong S, Yang D, Choi J, et al. Inferring semantic layout for hierarchical text-to-image synthesis[C] Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 7986-7994.
[3] Ramesh A, Dhariwal P, Nichol A, et al. Hierarchical text-conditional image generation with clip latents[J]. arXiv preprint arXiv:2204.06125, 2022.
[4] Gafni O, Polyak A, Ashual O, et al. Make-a-scene: Scene-based text-to-image generation with human priors[C] Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XV. Cham: Springer Nature Switzerland, 2022: 89-106.
[5] Yu J, Xu Y, Koh J Y, et al. Scaling autoregressive models for content-rich text-to-image generation[J]. arXiv preprint arXiv:2206.10789, 2022.
[6] Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models[J]. Advances in Neural Information Processing Systems, 2020, 33: 6840-6851.
[7] Nichol A, Dhariwal P, Ramesh A, et al. Glide: Towards photorealistic image generation and editing with text-guided diffusion models[J]. arXiv preprint arXiv:2112.10741, 2021.
[8] Gu S, Chen D, Bao J, et al. Vector quantized diffusion model for text-to-image synthesis[C] Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 10696-10706.
[9] Pan Y, Qiu Z, Yao T, et al. To create what you tell: Generating videos from captions[C] Proceedings of the 25th ACM international conference on Multimedia. 2017: 1789-1798.
[10] Mittal G, Marwah T, Balasubramanian V N. Sync-draw: Automatic video generation using deep recurrent attentive architectures[C] Proceedings of the 25th ACM international conference on Multimedia. 2017: 1096-1104.
[11] Li Y, Min M, Shen D, et al. Video generation from text[C] Proceedings of the AAAI conference on artificial intelligence. 2018, 32(1).
[12] Wu C, Huang L, Zhang Q, et al. Godiva: Generating open-domain videos from natural descriptions[J]. arXiv preprint arXiv:2104.14806, 2021.
[13] Wu C, Liang J, Ji L, et al. Nüwa: Visual synthesis pre-training for neural visual world creation[C] Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVI. Cham: Springer Nature Switzerland, 2022: 720-736.
[14] Ho J, Salimans T, Gritsenko A, et al. Video diffusion models[J]. arXiv preprint arXiv:2204.03458, 2022.

[15] Hong W, Ding M, Zheng W, et al. Cogvideo: Large-scale pretraining for text-to-video generation via transformers[J]. arXiv preprint arXiv:2205.15868, 2022.
[16] Unterthiner T, Van Steenkiste S, Kurach K, et al. Towards accurate generative models of video: A new metric challenges[J]. arXiv preprint arXiv:1812.01717, 2018.
[17] Salimans T, Goodfellow I, Zaremba W, et al. Improved techniques for training gans[J]. Advances in neural information processing systems, 2016, 29.