# Generating Temporally Coherent High Resolution Video With Stable Diffusion

Originally published at https://blog.metaphysic.ai/generating-temporally-coherent-high-resolution-video-with-stable-diffusion/ · April 21, 2023

A new collaboration between NVIDIA and several academic institutions across Germany and Canada offers a novel solution to what has become a motivating goal among image synthesis researchers – a method to make text-to-image generative latent diffusion systems, such as Stable Diffusion, output *temporally coherent* video instead of just static images.

*Examples from the new video synthesis system. Indicative samples only in this article: please refer to the linked or cited sources for more examples at better resolution and frame rates. Source: https://research.nvidia.com/labs/toronto-ai/VideoLDM/samples.html*

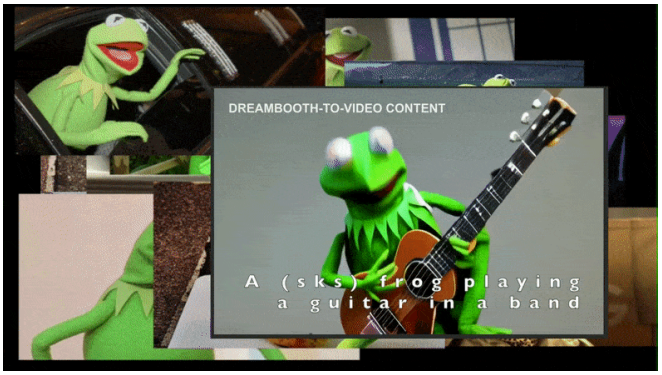
The new system comes on the heels of much progress in this regard over the last several months, and is not only capable of generating several seconds of fairly convincing video from simple text prompts, but also of producing POV driving videos at lengths of up to five minutes – and, thanks to innovative use of upsampling, of producing footage of all kinds at truly HQ (1280 × 2048) resolution.

However, the system's more abstract generative text-to-video capabilities are likely to be of most interest to both consumers and VFX professionals:


*Though we're not able to reproduce the high frame rates and full resolution of the supplementary source videos for the project, they can be casually browsed (including this example) at https://drive.google.com/drive/folders/1ENd9_9lzN6mI3E_HAjP52_KMWFdd0c8u*

The nature of the samples produced for the release of the system makes them difficult to reproduce faithfully here, and we refer the reader to the project site, where a multitude of videos can be casually (and simultaneously) viewed; and to the supplementary videos that have been made available on Google Drive.

The new approach – loosely termed 'Video Latent Diffusion Model' (Video LDM, here *VLDM*) – can not only generate videos from the default trained models that come with Stable Diffusion, but can also realistically animate models that have been fine-tuned by users of DreamBooth – a method that can 'inject' a specific object or person into a publicly-available trained model, so that a user can, for instance, create images (and now videos) of themselves:




*In the background, the handful of images that taught DreamBooth how to reproduce Kermit the Frog; in the foreground, the new system successfully reproduces plausible motion from the fine-tuned DreamBooth model, based on a user's text-prompt. Here the 'Kermit entity' has been named 'sks' – by now a traditional moniker for subjects that are trained into DreamBooth, since this sequence of consecutive characters appears nowhere in the database on which Stable Diffusion was trained, and therefore is guaranteed to produce a result uncluttered by 'competing' terms.*

In the background of the illustrated example above, we see that DreamBooth has obtained a comprehensive visual understanding of the popular Muppet character Kermit the Frog, which the new system is able to leverage into actual motion, based on what it has learned about 'playing a guitar' from massive datasets of short video clips.

As the paper notes towards its conclusion, this is a powerful capability with much potential for abuse, since it effectively takes the controversial ability to inject random celebrities (or ex-partners, enemies, etc.) into a powerful generative model and, now, actually output full-motion, high-resolution video running at up to 30FPS.

The new paper is titled *Align your Latents: High-Resolution Video Synthesis with Latent Diffusion Models*, and comes from seven researchers variously associated with NVIDIA, the Ludwig Maximilian University of Munich (LMU), the Vector Institute for Artificial Intelligence at Toronto, the University of Toronto, and the University of Waterloo. Let's take a look at their work.

## The Gift of Time

Since Stable Diffusion took the AI scene by storm when it was open sourced in August of 2023, the world has been waiting for it, or a similar system, to provide users with the capability to generate temporally stable and photorealistic *video* footage.

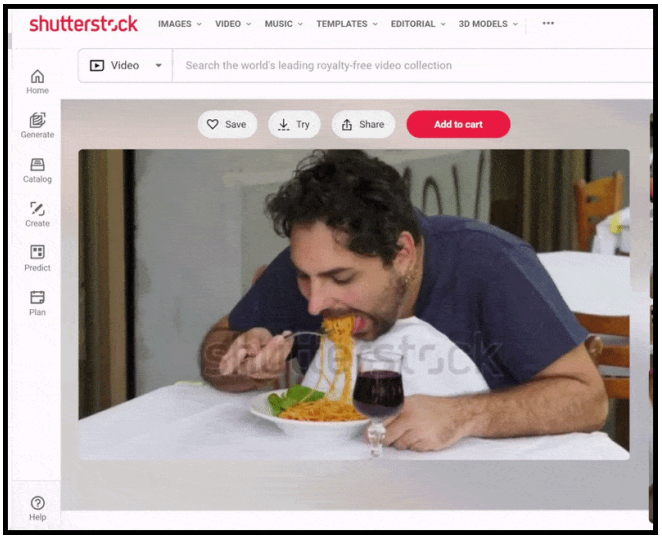
However, this is no trivial endeavor, since, despite their profound and even shocking ability to reproduce real and fantastic images, latent diffusion models such as Stable Diffusion have absolutely no understanding of time, and no native central mechanisms that could be easily exploited to make up for this lack.

Therefore, since its launch, a series of projects have been initiated to provide Stable Diffusion with video-generating capabilities. Systems such as Deforum allow the generation of 'trippy' or abstract animation sequences, while the (non-AI) interpolation application EbSynth can create interstitial transitions between user-supplied keyframes that have been generated or altered by Stable Diffusion.

Though such results can be very smooth, and give the impression of temporal coherence, there is no abstract 'intelligence' or prior knowledge about movement dynamics informing this kind of output. Instead, the systems are simply 'morphing' blindly between frames, either based on the pixels in the frames, or, optionally, guided by 'example' videos.

The future of generative video lies in obtaining and using *priors* – video-based, temporally rich features that have been generalized by training systems on large collections of short videos.

This means that one can take a concept that a latent diffusion system knows about – such as the actor Will Smith – and superimpose temporal knowledge about an action that featured in the video dataset on which a video generative system was trained – such as 'eating spaghetti'.



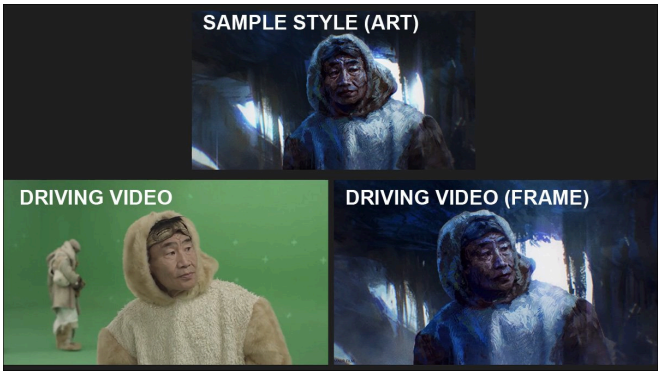

*One of many Shutterstock stock videos depicting a person eating spaghetti – the presence of the Shutterstock logo (see Will Smith video below) in nearly all recent video synthesis outings indicate that such videos are the direct source for the priors that inform video synthesis. Source: https://www.shutterstock.com/video/clip-1098199103-food-spaghetti---hungry-young-man-eats*

Though the results are certainly not perfect, you can indeed, by this crude assembly of two priors (*'Will Smith'* + *'Eating Spaghetti'*), make Will Smith eat copious amounts of spaghetti.

*Will Smith has the brute force of a brief 'stock footage' spaghetti-eating clip copied and pasted into temporal space until his appetite seems insatiable. Source: https://www.reddit.com/r/StableDiffusion/comments/1244h2c/will_smith_eating_spaghetti/*

In much the same way as deepfake models generalize to a versatile and flexible model of an identity, based on the input from thousands of source images of one person, a single spaghetti-eating video (such as the Shutterstock one featured above) will not be adequate to provide priors for a realistic novel synthesis. Therefore, the more available examples there are of any one particular action, the more authentic and adaptable that action will be to new scenarios and syntheses.

If you want to get a better idea of how such priors work, the AI-free EbSynth workflow process is quite illustrative, in that it requires an 'example' video that is used as a 'template' for the synthesized video:




*EbSynth obtains movement vectors from a user-selected video, which drives the style-transferred synthesis. In the case of the new wave of text-to-video systems, such 'priors' have been obtained instead from massive collections of video clips, and the resulting temporal features can be elicited with a mere text prompt. Source: https://www.youtube.com/watch?v=0RLtHuu5jV4*

The new wave of text-to-video systems operate on exactly the same principle, except that the movement data has already been trained into distilled and highly compressed temporal features, and can be summoned up via text prompts in much the same way as text-to-image systems; and, crucially, the user has access to thousands, or even potentially millions of such movements, without needing to find an explicit example for the desired synthesis to follow.

In terms of quality, this kind of video synthesis (from systems such as Stable Diffusion Videos) is currently on a par with the emergence a few years ago of the earliest publicly-usable image synthesis systems, such as Aleph2Image and BigSleep; and users are now expecting the same quantum leaps in quality that distinguish those bizarre initial outings from the sophisticated generative powers of DALL-E 2 and Stable Diffusion today.

## Brief Encounter

Many have observed that the new text-to-video (T2V) systems produce only very short clips (or else longer videos that are clearly 'episodic', and composed of concatenated instances of 'shorter' data), more in the way of 'meme'-style GIF animations than actual video footage.

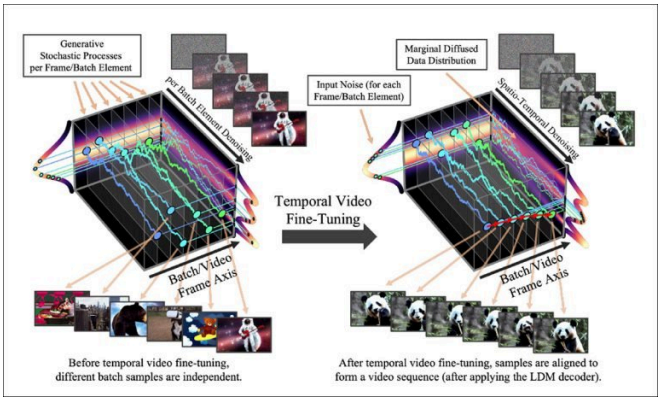
Though current hardware limitations for the generative systems themselves are one contributing factor in this regard, the main reason is that all the available video datasets feature *very short clips*, usually of less than five seconds. Since the value of such datasets lies more in breadth and diversity than in 'per-clip' richness and length, this logistical economy at the curation level dictates what can currently be output by T2V frameworks.

Nothing impedes researchers from creating lengthier video-clips in new datasets, featuring more intricate and complex sets of movements (the researchers of the current paper have done so, for their 'POV driving' tests); except that it costs a lot of money and effort to gather and annotate that kind of material, whereas the existing available data can help establish the viability of T2V systems right now.

Therefore these early efforts are likely to continue to obtain continuous video by either repeating movements or transitioning between different types of movement.

## VLDM: Approach

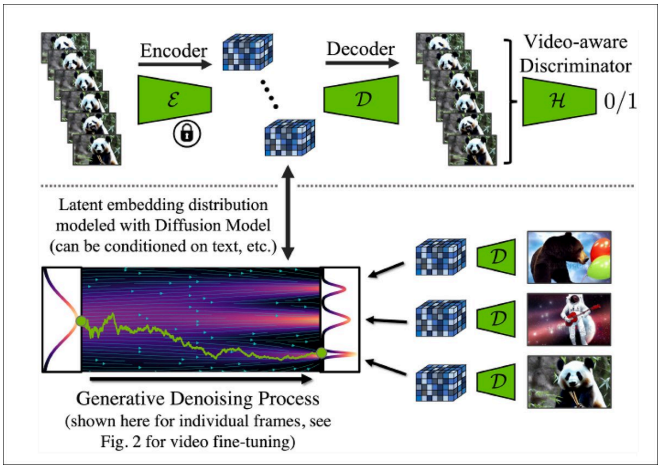
VLDM imposes a framework of temporal layers onto an established text-to-image system. Though this is applicable in theory to many brands of latent diffusion model, and even potentially to other architectures, the researchers have used Stable Diffusion – not least, perhaps, because its open source nature allows for potential commercial deployment, and for free experimentation.




*Conceptual workflow for VLDM. Source: https://arxiv.org/pdf/2304.08818.pdf*

The temporal neural network layers imposed onto Stable Diffusion by VLDM learn to align the individual frames output by Stable Diffusion in a temporally consistent way. A frozen encoder (i.e., an encoder that will not learn anything new from the workflow) processes frames independently, while the video-aware layers impose a consistency that's not dissimilar to optical flow (where a video is 'flattened' out into a single entity that can be examined like a painting).

A video-aware discriminator adds an additional layer of oversight:



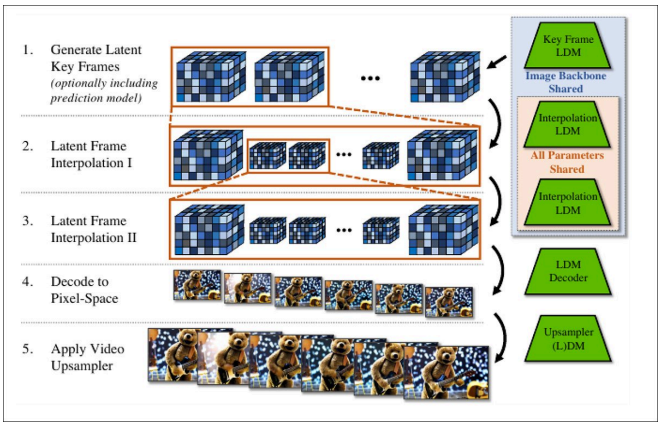

*Above, the processed output from the imposed layers are passed to a video-aware discriminator; below, a generative denoising process is applied to individual frames, which are then passed to the video fine-tuning layer (see image above this one).*

After a 'brief' fine-tuning (only the parameters affecting the temporal layers are trained – see 'Data and Tests', below), during optimization, the core Stable Diffusion architecture is not directly affected – a 'passive' approach that more or less treats Stable Diffusion as a 'read only' process, much like Adobe's DreamBooth clone InstantBooth.

Two kinds of temporal layer are used in this process: temporal attention, and residual blocks based on 3D convolutions. Sinusoidal embeddings (essentially, Transformers) are used to provide a temporal positional encoding.

There is a degree of in-betweening occurring in this process, in that VLDM initially generates sparse frames that in themselves would not be adequate for full-motion video. Interpolation is carried out between these frames, in much the same way that the far humbler EbSynth generates interstitial frames

from keyframes; and, optionally, the output is passed to upsampling layers that achieve the UHQ resolution demonstrated in the published results (a now-common upsampling hierarchy that Google Research uses extensively in its own generative systems, such as GigaGAN).




*The interpolation stack in VLDM. Prediction models can also be used in this process, using priors not just to interpolate current frames, but also to anticipate how the rendered movement should develop.*

The most critical factor for the quality of results for VLDM may be the temporal autoencoder finetuning stage, which eliminates the 'flicker' that has tended to plague attempts to use Stable Diffusion for generating video – with most examples inevitably being compared to the 'psychedelic' rotoscoping effects in the 2006 film adaptation of Philip K. Dick's *A Scanner Darkly*.

Regarding this, the authors comment:

*'Our video models build on pre-trained image LDMs. While this increases efficiency, the autoencoder of the LDM is trained on images only, causing flickering artifacts when encoding and decoding a temporally coherent sequence of images. To counteract this, we introduce additional temporal layers for the autoencoder's decoder, which we finetune on video data with a (patch-wise) temporal discriminator built from 3D [convolutions].'*

VLDM's methods for prediction and interpolation, the authors acknowledge, build on the masking techniques used in recent projects such as *Flexible Diffusion Modeling of Long Videos*, *Diffusion Models for Video Prediction and Infilling*, and MCVD.

## Data and Tests

Experiments for VLDM were divided between two scopes: driving videos (i.e., generating videos of roads passing by from the POV of a driver), and abstract generative text-to-video output.

However, since radically different dataset sources and notably diverse testing methodologies were used for these, it could be argued that the latter may have been a more appropriate and discrete focus for the paper. Therefore here we will concentrate primarily on the shorter T2V output that dominates the paper, project page, and supplementary video examples.

The dataset for the driving videos were generated internally, according to the researchers, and consisted of 683,000 videos, each lasting eight seconds, and at a resolution of 512 x 1024. The videos feature both day and night scenarios, and a subset that includes bounding boxes (i.e., 'recognized' vehicles).

For the abstract T2V tests, the researchers used the WebVid-10M dataset to transform the publicly available Stable Diffusion architecture.


*Samples from the project page of WebVid-10M. Source: https://m-bain.github.io/webvid-dataset/*

WebVid-10M contains 10.7 million video-caption pairs, running to 52,000 hours of videos, which the researchers were obliged to resize to 320×512 – in itself, a formidable act of data pre-processing.

Metrics used for the tests were Fréchet Inception Distance (FID), and Fréchet Video Distance (FVD). Exclusively for the text-to-video tests, the authors also used the CLIP Similarity (CLIP-SIM) metric introduced by Microsoft's GODIVA generative system. Noting that CLIP-SIM has been demonstrated to be less than entirely reliable, human evaluation tests were also conducted, and (video) inception scores used.

The Stable Diffusion denoising model was undertaken with *Denoising Diffusion Implicit Models* (DDIM) across all experiments.

The authors describe the initial approach for the Stable Diffusion T2V tests:

*'[We] train a temporally aligned version of Stable Diffusion for text-conditioned video synthesis. We briefly fine-tune Stable Diffusion's spatial layers on frames from WebVid, and then insert the temporal alignment layers and train them (at resolution 320 × 512). We also add text-conditioning in those alignment layers.*

*'Moreover, we further video fine-tune the publicly available latent Stable Diffusion upsampler, which enables 4× upscaling and allows us to generate videos at resolution [1280×2048]. We generate videos consisting of 113 frames, which we can render, for instance, into clips of 4.7 seconds length at 24 fps or into clips of 3.8 seconds length at 30 fps.'*

The authors note that the Stable Diffusion backbone 'readily translates' to video generation under VLDM, despite the fact that the source dataset used for training, and the system competently combines the expressiveness of Stable Diffusion with the temporal coherence gleaned from the WebVid-10M sources.

Competing frameworks used in the tests were: CogVideo (both Chinese and English implementations); ByteDance's MagicVideo; and MetaAI's Make-A-Video. The base models used were Stable Diffusion versions 1.4 and 2.1. For zero-shot generation, the authors used UCF101 and Microsoft's MSR-VTT.

VLDM performed comparably to the rival frameworks, and outperformed them in certain scenarios:

*Results for the T2V tests.*

The authors note, regarding the MSR-VTT tests, that the Make-A-Video framework is trained with a far higher amount of data, using the extensive HD-VILA-100M dataset. They comment:

*'We significantly outperform all baselines except [Make-A-Video], which we still surpass in IS on UCF-101. However, Make-A-Video is concurrent work, focuses entirely on text-to-video and trains with more video data than we do. We use only WebVid-10M.'*

It was not possible to conduct comparative experiments on DreamBooth video synthesis, since, as far as the authors know, VLDM is the only framework to currently facilitate this.

## Conclusion

The new paper runs to a dense and comprehensive 44 pages, with extensive supplementary material, and we can only superficially review some of the technologies, approaches and tests conducted for the work. We therefore encourage the reader to refer to the full paper, and most particularly to the video samples provided by the researchers, which arguably demonstrate the best temporal coherence of any recent similar project, and which would appear to address a long-felt need in the Stable Diffusion community (though the paper suggests, by inference, that this code will not be publicly released).

The two most notable take-aways from this paper are that the system described can create realistic synthesized video from user-generated models via DreamBooth; and that the authors acknowledge, at the paper's conclusion, the growing retrenchment around data availability for such systems – a new protectionism that seems set to slow down the video and image synthesis scene notably.

From the paper:

*'Our synthesized videos are not indistinguishable from real content yet. However, enhanced versions of our model may in the future reach an even higher quality, potentially being able to generate videos that appear to be deceptively real. This has important ethical and safety implications, as state-of-the-art deep generative models can also be used for malicious purposes, and therefore generative models like ours generally need to be applied with an abundance of caution.*

*'Moreover, the data sources cited in this paper are for research purposes only and not intended for commercial application or use, and the text-to-image backbone LDMs used in this research project have been trained on large amounts of internet data.*

*'Consequently, our model is not suitable for productization. An important direction for future work is training large-scale generative models with ethically sourced, commercially viable data.'*

This reiterates the point we made earlier this month regarding Adobe's growing interest in generative systems, which is underpinned by its massive investment in stock images and video: that the 'data party' may indeed be coming to an end, as freely web-scraped LAION-scale datasets come under increasing legal scrutiny, and business investors in generative AI begin seem likely to seek out 'safer' systems, using IP-safe data that is less likely to face legal challenges – even if smaller 'legitimate' collections cannot compete with the unfettered priors that can be obtained by liberally exploiting the entire internet to populate datasets.