John Brownell - This Machine

Discussion between Kirill Zubovsky (Q) and John Brownell (A) about the making of his AI-powered video, This Machine. If you enjoyed this discussion, please subscribe to John's YouTube channel, and Kirill's newsletter The Novice.

—

Q: It seems that you trained a model to convert images of people into images of robots, and then interpolated between the images to create an illusion of flow. Then, perhaps in an entirely different application, you merged it all together. Could you explain the details of how the background video is done? Please be as technical as you like, or not.

A: Yeah, so I used an extension on top of Stable Diffusion called Deforum. The video's intro (where people turn into robots) and the background animations throughout the video were created using Deforum. I don't really know a lot of the deep technicalities about how it works, but I'll try to give a general explanation.

At the core is Stable Diffusion - a deep learning "text-to-image" model created by Stability AI that was released and open sourced earlier this year. The models and the scripts they have released are very powerful and once the system was open sourced, it really started a tidal wave of new tech, new models, and new apps built on top of it. I got Stable Diffusion running on a computer at my house on day one and was immediately blown away and have been pretty much obsessed with it ever since.

One of the core capabilities of Stable Diffusion is "image-to-image" rendering. Meaning that instead of just describing an image with a text prompt, you can give it an existing image along with the prompt and it will generate a new image by combining the two - and the user specifies how much weight to give the image vs the prompt during rendering

Deforum generates videos by using SD to generate one frame at a time. In simple terms, it generates each new frame using the image-to-image functionality, passing in the previous frame along with a text prompt. On top of that, the system adds some animation parameters that allow you to "move the camera" between frames as well. Put all of this together and you can make some pretty amazing videos.

I spent a lot of time experimenting with the various settings available in Deforum, and at the same time development has been moving very fast. So almost every week they are adding new features and new parameters that can be animated. One of the tests I did involved animating between 2 prompts that only change slightly and this is how I ended up building the intro to the This Machine video.

For example, imagine it is configured with the following prompts:

Frame 0: “A portrait of a suburban family standing in front a suburban house”
Frame 50: A portrait of a robot family standing in front of a suburban house”

At frame 0, Deforum generates a new image from stable Diffusion using the first prompt above, and then each subsequent frame combines the image from the previous frame with the current prompt (image-to-image). When it gets to frame 50, the prompt has changed and image-to-image will start transforming the previous human family into a robot family. By experimenting with the parameter that specifies how much weight to give the image vs the prompt, I was able to eventually get it to create what you see there. The hard part is getting the animations to maintain some coherence from frame to frame.

Using Deforum, the prompts are set up in a configuration file. So I would set up the prompts and let the model generate the images. My only control was reviewing the results and deciding whether there was something usable or to try again.

One of the parameters that you can play with is the "denoising strength". This is a value that determines how much weight is given to the previous image and how much weight is given to the prompt. So I spend a lot of time messing around with that value to get results that I like.

It's all a little hard to describe... but when you want the animation to change more - move closer to a new prompt - you can reduce the value of the denoising strength for a few frames to force it. And when you want to maintain consistency from frame to frame, you specify a higher denoising strength. By changing this parameter over time you can have some decent control over the animation - BUT the actual images are created by the model based upon prompts without direct input by me.

Part of the fun of this process is realizing that you are at least partially at the whim of the machines!

Q: Likewise, what is the magic behind the robot on the foreground? That seems like a Stable Diffusion depth guided model, but it’s remarkably good at following your lips, and may have been done before SD2 was even available. How did you make it work? Is the publicly available AI really this good now, or did you make your own tweaks that make it so much better?

A: The singing robot uses a method called “Thin-Plate Spline Motion Model for Image Animation” [https://arxiv.org/abs/2203.14367] . A paper describing this method of animation was released in March 2022 and then an implementation was release on Github [https://github.com/yoyo-nb/Thin-Plate-Spline-Motion-Model] shortly afterwards. They trained a few different models on videos of talking heads or torsos (I believe the datasets were “talking head” videos of celebrities as well as Ted Talks). And created a system that uses those models to map an image to a source video. The result is that you can animate a static image using a source video. It’s amazing!

So I took a video of my head against a green screen as I lip synced to my own song. Then, I took a frame from that video and used Stable Diffusion image-to-image to create images of robot heads. I then used the static robot head image with the source video to generate the animation. I generated videos of over 70 different singing robots and picked the best one. Then I used Adobe Premiere to overlay that video onto the Deforum animations I described above.

Q: What technology are you using for this? Is this a home-made model, or something you are getting off the shelf, and modifying to your needs? How much of this is relying on APIs, vs. say your own tech running on a cloud and/or a computer in your closet.

A: As mentioned earlier, most of this work is using Stable Diffusion, an open source machine learning image generation model. I did almost immediately upgrade my computer and bought a 24GB video card to make it work faster and allow for generating images at higher resolutions. I mostly use the scripts as provided, but I have made some scripts of my own to automate certain parts of the process. For example, if I am working on a specific section of a song, I have a script that allows me to specify all of the parameters and it will loop through multiple renders. That way I can go to bed and when I wake up I'll have 20 or 30 videos for that section. I can then review the videos, pick the best, and move on to the next section.

On my primary machine, depending on the parameters, I can generate an image in about 5-7 seconds. So a 10 second video would take about a minute.

I have a Windows PC with an RTX 3090. I upgraded my machine specifically to run Stable Diffusion locally!

I have done some of this work in the cloud using Google Colab, but I just like the control of being able to use my own GPU on my own network. Plus, once you get everything installed and working correctly you can just get to work whenever you want - no configuring or downloading dependencies. And at some point I hope to find some time to game on it. :)

Q: How much work did you put into this? What was the planning and making process like, and what was the timeline for creating the music, and the AI stuff, and then putting it all together.

A: Man … a lot! The tech is still so new but also moving so fast. So much of the time I've spent has just been experimenting and learning what *doesn't* work. In general, I start with the music. I will import the song file into Adobe Premiere and write down the frame numbers for each section in the song - anyplace in the song where I know I want to make prompt or animation changes in the video. Because of this, it's important to choose the frame rate of the video at the beginning, because it is a pain to change keyframes later if you need to change the FPS. Then, I start testing prompts to generate imagery that works for each section of the song. I'll start building the configuration in Deforum based upon these prompts and then experiment with animating the results. It is a LOT of trial and error!

Q: What drives the choice of the frame rate?

A: Yeah, this is a huge choice. The lower the framerate, the faster you can make a video. But the higher the framerate, the smoother the animations are. For these music videos, I need to make the decision up front because the first step is to go through the song and figure out where the parameter and prompt changes will happen. When I started I was working at 12fps (which is the default in Deforum), but once I tried 24 fps it just looked so much better that I won't go lower than that.

But the frame rate also effects animating the parameters! If you use the same denoising strength at 12 fps and at 24 fps, you will find the animations change twice as much in the same period of time - because there are twice as many frames getting generated. It takes a lot of experimentation!

There have been several optimizations to get it to work with lower vram but I think the core of the models work the same regardless. I have heard of some folks getting Stable Diffusion working on 8GB but my guess is that would be very slow.

Q: How much does it cost to make a video like this?

A: Well, just the cost of the computer and the power required to run the computer. To run Stable Diffusion locally, you need an Nvidia video card with a lot of video RAM. At the high end, a 24GB card will likely cost over $1,000. But I believe it can run on cards with as little as 8GB - it's just going to be much slower and constrained to lower resolutions.

Q: How did you even come up with the idea of making AI videos? From what I understand, you were actually very very early to start this, before AI was hot, or anyone was even considering doing video with it.

A: Well, it really goes back to GPT-3, a large language text-generation model created by OpenAI. I received access to the GPT-3 beta last year. I started building an app that I could use to help with writing song lyrics. I even generated a model that was fine-tuned on my own lyrics! I started working on several new songs using that system and I found myself using AI to write songs about AI. It was like an AI inception. So I decided I was going to record a whole album along those lines.

I would guess about 10% of the lyrics my new album were written by or assisted by the songwriting app I wrote. I wish I had been better about keeping track of exactly how much I used! It's hard though because often the model will spit out a neat idea that I use but in a modified form. It's not unlike reading a line of a poem or book that inspires you to write something similar but from a slightly different angle. There are a couple songs that use some full lines, or even a verse or two, that were completely written by my fine-tuned lyric model.

As I was working on the album, DALL-E (an image generation model from OpenAI that is similar to Stable Diffusion) was released and started making a lot of noise. I immediately knew that I

needed to incorporate some AI generated imagery into my AI themed album. So once I got beta access, I started working on that - using GPT-3 to help me generate image descriptions that I would then use as prompts in DALL-E. Think of that - I was using AI to assist with writing lyrics, then using AI to interpret those lyrics to generate a text description of an image, and then using another AI to generate the actual images!

But it wasn't long before Stable Diffusion came along, was open sourced and blew everything wide open. Deforum was released shortly after and I knew I needed to incorporate video into the project as well. This is my biggest problem anymore - it all is moving too fast! It seems like every day some new ridiculously cool feature, model, application, or paper, is released. For example, in the middle of making one of my videos a new fine-tuning method was discovered that allows you to add new things into the Stable Diffusion models. So I had to stop everything so I could figure out how to train it on pictures of myself so I could put my own likeness in the video as well.

And of course, in the meantime, I've still been trying to finish recording my album so I can finally release this thing!

Q: Can a non-technical artist attempt to make something like your video today, and whether yes or no, where do you recommend they start? Also, do you plan to offer these videos as a service?

A: It definitely helps if you are familiar with the command line, python, Github, and such. But anyone with some patience and the right hardware (an nvidia video card with a decent amount of vram) should be able to do this. It's not exactly simple, but there are a ton of Youtube videos and tutorials out there that can help a person get started.

I'd recommend starting by getting Stable Diffusion running inside of the Automatic1111 web interface. This provides a GUI on top of Stable Diffusion. Once you have that working, Deforum can now be installed as an extension from right within the GUI:

Here is a decent guide to installing in Windows:
https://www.youtube.com/watch?v=sTcTYnv2o74

And here is a guide to installing Deforum within the GUI:
https://www.youtube.com/watch?v=R52hxnpNews

Also, join the discord communities:
Stable Diffusion: https://discord.com/invite/stablediffusion
Deforum: https://discord.gg/upmXXsrwZc

Q: It seems that a lot of artists on the internet are frustrated with AI, fearing it will take away something unique from art. Meanwhile, AI also seems to be empowering a whole

new group of greators. Where do you think AI art is going? How do you think AI is going to impact us in general in the short and long term?

A: I am sympathetic to artists that are worried and frustrated. I also see a lot to be worried and frustrated about! I have generated thousands of images - many of them of a quality that is indistinguishable from human generated art. My computer can create, in seconds, something that would take a human artist hours, weeks, even months to make by hand. I do think something will be lost - it's inevitable. But I also think it is revealing a vast, exciting, new world. Artists who get started now have the chance to be on the cutting edge of something entirely unexplored.

I don't think we can comprehend where this is going. Even in the near term! Emad from Stability says they are on the verge of having the capability to generate 30 frames per second… real time AI-generated video! There are already very smart people working on AI-generated virtual reality environments. Imagine that. We are potentially on the verge of technology very similar to the holodeck from Star Trek. A dream come true!

So I can barely comprehend where this is going in the next 3-5 years, much less any kind of long-term outlook. I just hope when the AIs wake up and become conscious that they like my album. 🙂