# The Road to Realistic Full-Body Deepfakes

Originally published 22nd September 2022 at https://metaphysic.ai/the-road-to-realistic-full-body-deepfakes/

It's nearly five years since the advent of deepfakes released into the public realm the ability to alter people's facial identities; at first, in recorded video, and now even as a streaming implementation, with DeepFaceLive.

You still have to find someone who looks a bit like the person you're trying to imitate, though. The more they resemble the 'target' identity, the more convincing the illusion is going to be.

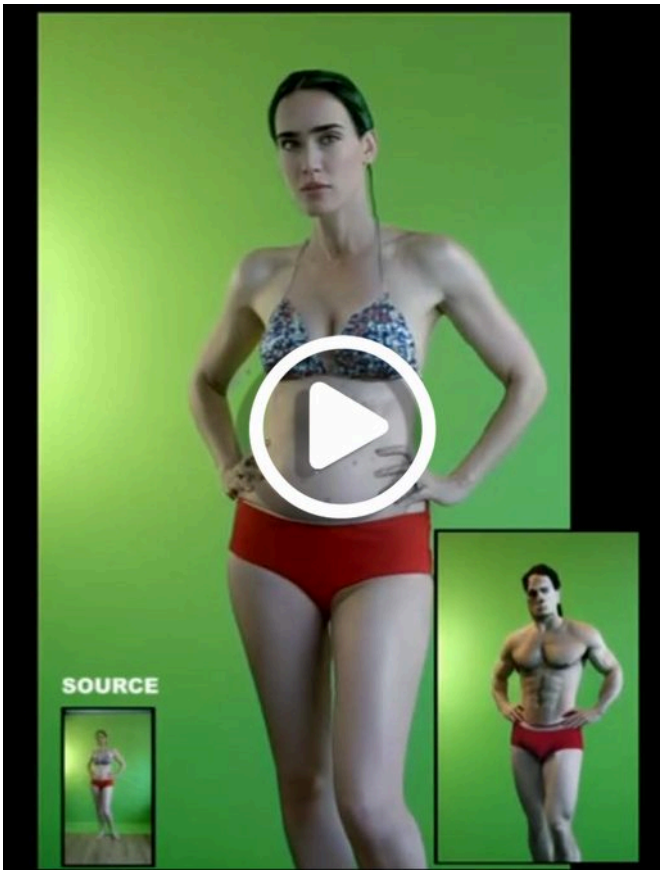
*If the face fits, wear it! Left, Miles Fisher is well-suited as a deepfake 'canvas' for Tom Cruise, while, right, Alexis Arquette proved an apt subject for Jerry Seinfeld in one deepfake parody, which can be seen at https://www.youtube.com/watch?v=S1MBVXkQbWU.*

Since autoencoder-based deepfake systems are trained at great length on a single and *relatively* similar 'opposite' identity, the authenticity of the subsequent model's recreation will suffer in accordance to how physically different the 'host' is from the personality being superimposed into a video clip.

Therefore it can be hard to find the right person to act as a 'canvas' for a deepfake rendition of a personality. Even if hair can be revised, and peripheral features such as ears, neck-height, and basic tone and physique (as well as age) are 'near enough', the chances of a full-face and full-body 'match' in one person are vanishingly small.

What if, instead, and in the complete absence of expensive and complex professional CGI techniques, you could recreate the *entire person* with machine learning?

CONTINUED ON NEXT PAGE




CLICK TO PLAY VIDEO ON SITE

*1990s-era Jennifer Connelly (and, inset, Henry Cavill) recreated through Stable Diffusion and EbSynth, based on the real movements of a female performer (lower left 'source' image). The actors' entire bodies here have been reinterpreted from that source footage, based on what Stable Diffusion knows about the face and physique of the two personalities recreated here – both of whom are well-represented in the database on which the model was trained. Predictably, the AI has an easier time transforming a woman into another woman than to a muscular man like Henry Cavill.*

Shortly, we'll take a look at the possibilities and *very* severe limitations of attempting photoreal, temporally coherent video with Stable Diffusion and the non-AI 'tweening' and style-transfer software EbSynth; and also (if you were wondering) why clothing represents such a formidable challenge in such attempts.

For now, we should consider at least a couple of reasons why a truly effective, 100% neural, Stable Diffusion-style text-to-video system for full-body deepfakes may be years or even decades away, rather than months, as many enthused Stable Diffusion fans now seem to believe.

## The (Slow) Future of 100% Neural Total-Body Deepfakes

Though the above video clip*, with all its rough edges, is just a cheap bag of tricks thrown together with open source software, this type of corporeal 'deepfake puppetry' is likely to be the earliest consumer-level incarnation of full-body deepfakes in the Metaverse, and in other potential virtual future environments and contexts (where the body movements and entire physical appearance of participants will eventually be capable of being transformed in real time).

This is because it's easy for a human (such as the performer 'powering' Jennifer Connelly and Henry Cavill in the above clip) to string together a series of concepts or instructions into a series of movements; and facial/body capture AI systems are already advanced enough to 'map' human movements in real time, so that image or video synthesis systems such as DeepFaceLive can 'overwrite' the original identity with very low latency.

But if you want to *describe* human activities in a text-to-video prompt (instead of using footage of real people as a guideline), and you're expecting convincing and photoreal results that last more than 2-3 seconds, the system in question is going to need an extraordinary, almost Akashic knowledge about many more things than Stable Diffusion (or any other existing or planned deepfake system) knows anything about.

These include anatomy, psychology, basic anthropology, probability, gravity, kinematics, inverse kinematics, and physics, to name but a few. Worse, the system will need *temporal* understanding of such events and concepts, rather than the fixed and time-static embeddings contained in Stable Diffusion, based on the 4.2 billion static images that it was trained on.

CONTINUED ON NEXT PAGE

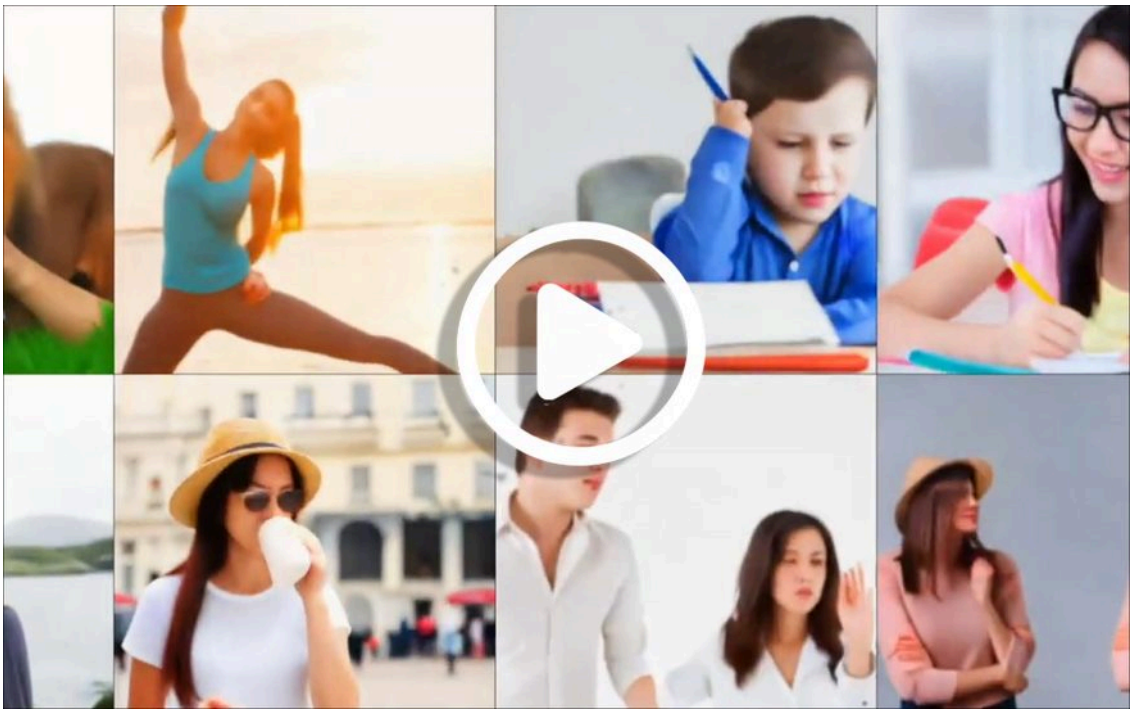
And that's before the hypothetical text-to-video system even starts *thinking* about what textures, lighting, geometries and other visible factors and facets might suit the scene, or how to generate an apposite accompanying soundtrack (another, almost equally complex database and adjunct model that would need to be developed).

## The Business Logistics of Text-To-Video Investment

Therefore it's going to be far more difficult to compile a truly comprehensive and versatile 'body movement' equivalent of the LAION database that powers Stable Diffusion, as well as for the academic and private research sectors to develop architectures and protocols that *cooperate* rather than compete in the race to achieve 'complete' deepfake capabilities for VFX and licensed entertainment applications.

It's one thing to share bare open source architectures on GitHub – quite another to release fully-trained models that cost millions to create, as Stability.ai has done with Stable Diffusion. From the standpoint of market share and general business logic, it's difficult to say whether such a generous occurrence is likely to ever happen again – and may depend on the extent to which the open-sourcing of Stable Diffusion ultimately undermines OpenAI's investment in DALL-E 2, and/or brings more financial acumen to Stability.ai than it would likely have achieved by gatekeeping its astounding product behind a commercial API.

In any case, the earliest such text-to-video system that has gained any ground is the 9-billion parameter transformer-based CogVideo architecture, which we covered in our recent article on the future of video in Stable Diffusion, and which launched in May of 2022.



## CogVideo Addresses Text-To-Video's Data Famine

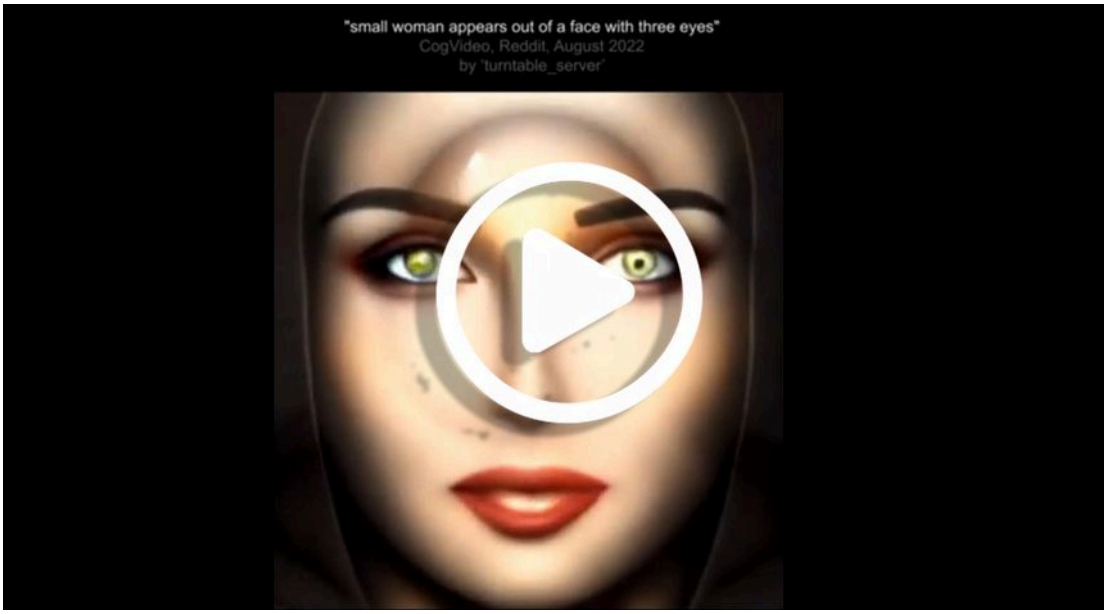
Though CogVideo is the premier text-to-video offering at the moment (and the only way that I am aware of to 'invent' fully neural and free-roaming, animated text-to-video humans, without any CGI involvement), the authors observe that it and similar systems are constrained by the cost and logistics of curating and training suitable movement-based datasets – just one factor that means Stable Diffusion fans may need to adjust their current expectations about hyper-realistic text-to-video a little.

As noted elsewhere, the largest current multilingual video description dataset (the movie clips need to be annotated with text descriptions, so that they have semantic meaning, a task which OpenAI's CLIP performs in Stable Diffusion's architecture) is VATEX, which contains a mere 41,250 videos supported by 825,000 captions.

Effectively this means that a task at least ten times more difficult to achieve than Stable Diffusion's generative power currently has at its disposal far less than a tenth of the necessary data.

To address this, CogVideo has adapted CogView 2, a Chinese *static* generative art transformer, to the task of text-to-video, and the resultant CogVideo dataset contains 5.4 million text/video pairings – still arguably scant data for the enormity of the task.

CONTINUED ON NEXT PAGE




However, if I seem a little dubious regarding the enormity of the challenges in creating a really good text-to-video framework without FAANG-level resources (with the inevitable ensuing commercialization and gate-keeping of the final product), my pessimism is not shared by Wenyi Hong, one of CogVideo's equally-contributing authors, with whom I recently had a chance to speak.

"I think it is not as expensive as you have indicated," Hong told me[†].

Though she concedes that a temporal video synthesis system equivalent to the generative power of Stable Diffusion or DALL-E 2 could possibly be perhaps five or ten times more costly to develop and train, she indicates that initial viral video synthesis clips are likely to be quite short, and to require less exorbitant resources.

Hong and her colleagues are developing CogVideo integration for a social media platform, and the earliest and widest dissemination of CogVideo output seems likely to come in the form of short videos lasting some seconds, which users might share, and which would not require hyperscale resources from the outset.

"It's OK to use a dataset much smaller than LAION," she says. "to train a video model like CogVideo, which normally generates videos lasting for several seconds. However, if we want to generate much more complicated videos, datasets of larger scale are needed."

Hong believes that, as with much of the entire machine learning research sector, the logistics of annotating videos and the availability of GPU memory (VRAM) represent the core challenges:

*"If we want to generate videos of high resolution, we have to trade off between the resolution, the frame-rate and the video's length. This is the biggest problem. We can generate video of any length if we have enough resources, enough memory.*

*"But there are also some problems that can crop up if the captions and video are not very closely related, or not extensive enough, or granular enough.*

*"Most captions would describe only one action in a video, like 'a person is holding a cup'. But if the video is very long, maybe a minute, the person won't hold the object for that long. Maybe they would put it down, or would start doing something else. The need to accommodate that level of complexity will make the whole training process difficult.*

*"However, we have already open-sourced CogVideo on GitHub, and I will try to develop an API where people can input their own sentences. For this, we're collaborating with Hugging Face, who've already created APIs for us."*

So it could transpire that the solution to the development of effective, powerful and *purely neural* text-to-video systems will be enabled by global participation, and perhaps by federated learning, in some implementation of SETI-style home-folding, now that Ethereum's move to proof-of-stake promises to free up GPU capacity and availability all over the world.

If we want text-to-video as badly as we seem to, such a system might be eagerly adopted by the growing army of image and video synthesis enthusiasts.

When I asked Wenyi how far we might be from a neural system that can effectively parse a script or a book into a movie, she responded: "Well, maybe ten to twenty years."

CONTINUED ON NEXT PAGE

## Beyond Stylized Transitions in Stable Diffusion Video

However, because Stable Diffusion is so powerful, and makes it so easy for anyone now to create amazing images; and because it captured the public imagination with only a small amount of warning from OpenAI's earlier and far more 'locked-down' DALL-E 2 product, there is a growing public expectation that text-prompted, hyper-realistic video, open to all and running at longer than just a few seconds, is likely to hit sooner than a decade or two.

In fact, AI VFX company Runway, a participant in the development of Stable Diffusion, is currently teasing the forthcoming release of a similar, prompt-powered video creation system.

*Runway's teaser for its text-to-video system previews some impressive functionality, but the only people in it appear to be from real source footage, and what remains to be seen is the extent to which neural humans may or may not feature in the system. Source: https://twitter.com/runwayml/status/1568220303808991232*

What's missing from the AI elements of the Runway teaser (and from any mature and usable current product) is *people* – the domain that we know most about, and the most challenging possible study for AI-based image and video synthesis: *whole*, walking, acting, interacting, running, tripping up, lying down, standing up, swimming, kissing, punching, slouching, laughing, crying, jumping, posing, talking *people*.

Though Stable Diffusion can generate static people and humanoid creatures very convincingly, and even photorealistically, most of the videos that are emerging from the frantic efforts of the SD community are either stylized (i.e. cartoon-like, often via inverting the noise-based pipeline in Stable Diffusion), 'psychedelic' (often done with Stable WarpFusion or Deforum), or display very limited movement (often done with EbSynth, which we'll shortly examine).

*Some of the more stylized or even trippy implementations of movement in Stable Diffusion. Sources (clockwise): https://www.youtube.com/watch?v=pkEQAKmDMa8 | https://old.reddit.com/r/StableDiffusion/comments/xeuuef/dance_like_no_one_is_watching/ | https://www.youtube.com/watch?v=_MDsKJYqaoY | https://old.reddit.com/r/StableDiffusion/comments/xev31d/stable_diffusion_experiment_ai_img2img_julie/*

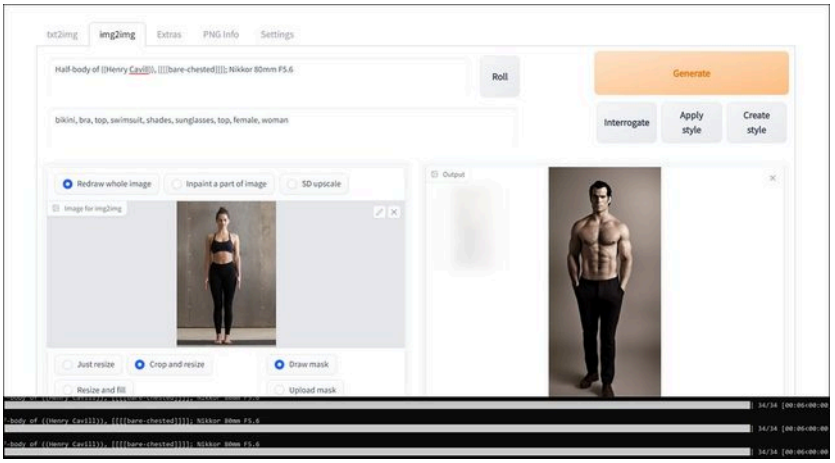
As we'll see, using EbSynth to animate Stable Diffusion output can produce much more realistic images; however, there are implicit limitations in both Stable Diffusion and EbSynth that curtail the ability of any realistic human (or humanoid) creatures to move about very much – which can too

easily put such simulations in the limited class of *'let's animate that static head a little bit'* that typifies the output of DeepNostalgia, and a huge raft of scientific attempts over the last 4-5 years to give 'limited life' to static human representations:

*Some of the GAN-based approaches that can give constrained movement and limited vivacity to human faces. Sources, clockwise: https://www.youtube.com/watch?v=uoftpl3Bj6w | https://www.myheritage.com/deep-nostalgia | https://studios.disneyresearch.com/2021/11/30/rendering-with-style-combining-traditional-and-neural-approaches-for-high-quality-face-rendering/*

Many of these systems rely on interpreting existing, real-world human movement, and using that motion information to power transformations, rather than relying on a database of distilled knowledge about human movement, as CogVideo does.

For instance, for the Connelly/Cavill full-body deepfakes featured earlier, I used the Img2Img function of Stable Diffusion to transform footage that I took of a performer into the two personalities. With Img2Img, you provide Stable Diffusion with a source image (anything from a crude sketch to a regular photo), and also provide a text-prompt that suggests to the system the way in which it should alter the image (such as *'Jennifer Connelly in the 1990s'*, or *'Henry Cavill, bare-chested*).

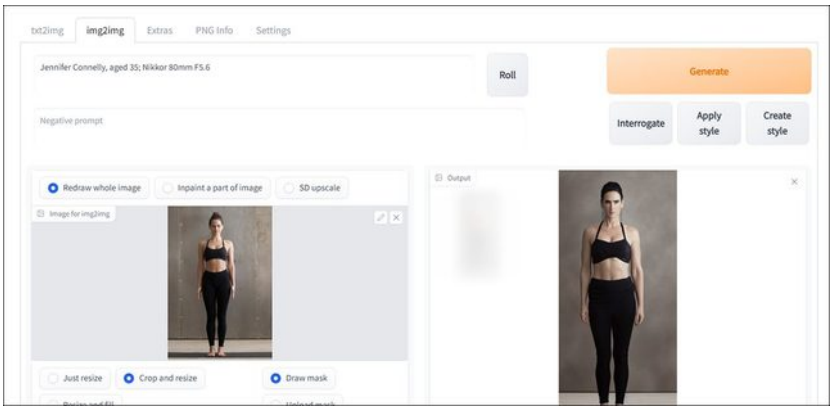



*A source picture and some text instructions (with negative instructions in the box below) lead to a fairly accurate Img2Img transformation of a woman into the actor Henry Cavill, in the highly popular AUTOMATIC1111 distribution of Stable Diffusion.*

As with autoencoder-based deepfakes (i.e., the open source system that has been used for making viral deepfake videos for the last five years), it is a lot easier for the machine learning system to effect a transformation when the source and the target have more in common – for instance, in the above image, Henry Cavill has his hands in his pockets, which does not exactly reflect the source pose.

By contrast, the image below shows that Stable Diffusion can transform the yoga woman source picture into a more accurately-posed approximation of Jennifer Connelly:

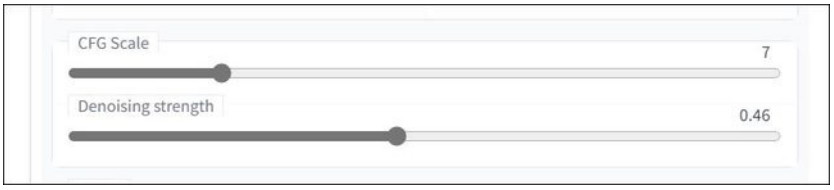
CONTINUED ON NEXT PAGE




*Even with lower settings, Stable Diffusion has a far easier time transforming the woman in the source picture into another woman, rather than a man – in this case, a representation of actress Jennifer Connelly as she appeared in the late 1990s.*

## Controlling 'Power' and 'Restraint' in Stable Diffusion

The two defining forces in a Stable Diffusion transformation are *CFG scale* and *Denoising Strength*.

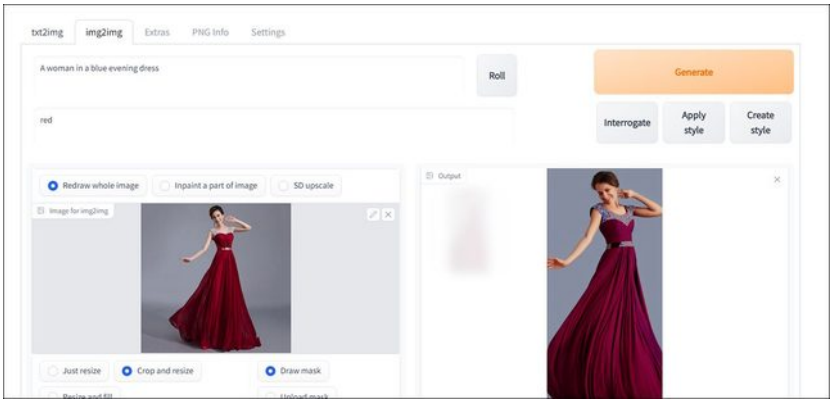



CFG stands for classifier free guidance. The higher you set this scale, the more strictly the system is instructed to follow the instructions in your prompt, even though that can lead to artifacts and other visual anomalies.

In many cases, the LAION dataset on which the model was trained is so authoritative that even short and simple adjunct Img2Img instructions can lead to effective results, removing the need to turn this setting up very high.

But if you're trying to make something happen that Stable Diffusion has *no prior knowledge of*; subtract something from the output which is challenging to isolate in the source image you provided; or to conjoin things, people or concepts that are very difficult to assemble coherently; then you may have to turn up either the CFG scale or the Denoising Strength, which will force Stable Diffusion to act more 'imaginatively' – though usually at the cost of some aspect of image quality.

For instance, though Stable Diffusion can turn a slender woman into a well-generalized muscular man such as Henry Cavill, it has *extraordinary* difficulty simply changing the color of a dress (part of Stable Diffusion's general issues with clothing, which we'll look at later).




*Even with CFG at an above-average 13.5 and Denoising Strength at a racy 0.58, and even with 'red' as a banned (negative) word, the dress will not change color.*

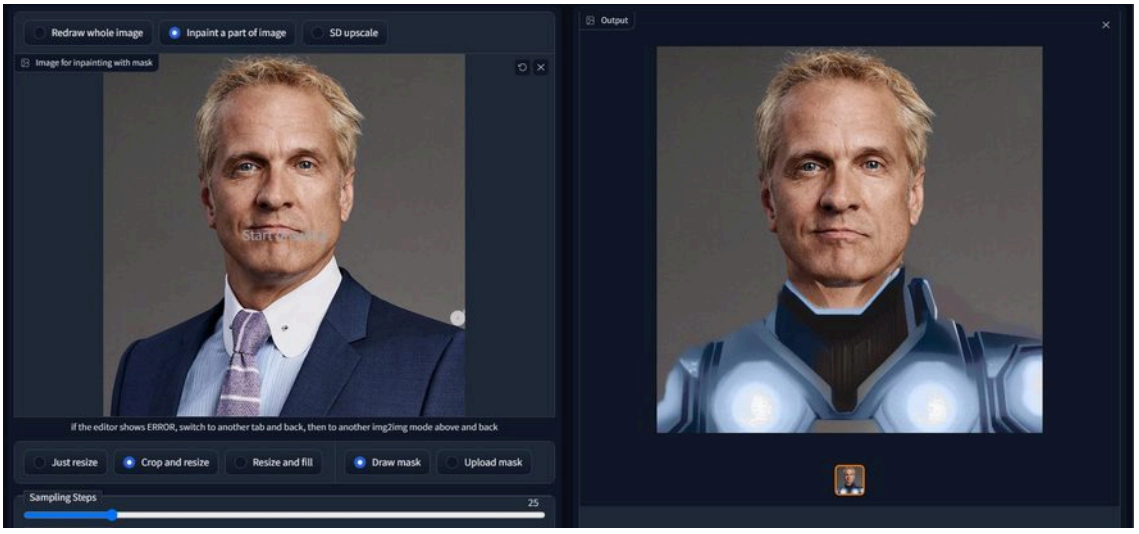
In one experiment, I attempted to change the color of the dress that the female performer was wearing in the source shoots, intended for a Stable Diffusion transformation to the actress Salma Hayek.

However, I found that no combination of settings, plugins or other chicanery could accomplish this apparently more minor task. In the end, it was necessary to set both CFG and Denoising almost to maximum settings before Stable Diffusion would transform the dress color – and in the process, 90-95% of the transformation's pose fidelity, style and coherence was lost:

CONTINUED ON NEXT PAGE


In general, similar to the way that traditional autoencoder deepfakes tend to use hosts that resemble the identity they want to impose, it's often easier to use source material that's at least a little closer to what you ultimately want to render (i.e. ask your performer to just wear a red dress in the first place).

Though there is at least one supplementary script that can use CLIP to recognize, mask and change a specific element, such as an item of clothing, it's too inconsistent to generate temporally coherent video for full-body deepfakes.




*The Txt2Mask addon for Stable Diffusion can isolate and change clothing, but, characteristic of many of Stable Diffusion's most 'bleeding edge' features, it's currently a hit-and-miss affair. Source: https://github.com/ThereforeGames/txt2mask*

## Fashion Chaos in Stable Diffusion

In case you're wondering why there's so much bare skin in some of these examples, it's not least because Stable Diffusion has additional issues with clothes and body adornment.

Surprisingly, there is very little 'famous clothing' that has become so resolutely generalized into the LAION-based Stable Diffusion model that you can rely on it to appear, consistently, across a series of sequential rendered frames.

Even Levi 501 jeans (of which there are numerous examples in the LAION database), which were voted in 2020 as the most iconic item of clothing of all time, can't be depended on to render consistently for an Img2Img full-body deepfake sequence in Stable Diffusion.

CONTINUED ON NEXT PAGE

*In terms of temporal coherence, and with a fixed seed (i.e., Stable Diffusion will not 'randomize' how it represents the jeans, but will stick to the settings of a good render that you chose earlier) the most recognizable item of clothing in the world performs way above average – but there are still random rips and glitches.*

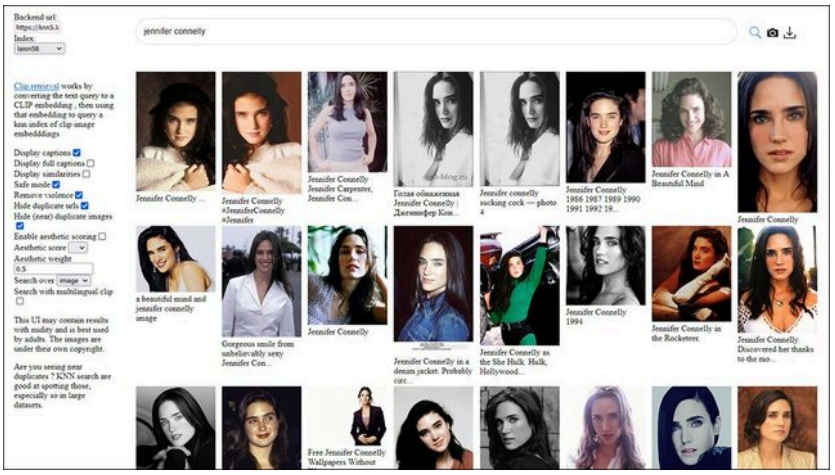
Jennifer Connelly's face and body? Fine – LAION-trained Stable Diffusion has assimilated nearly forty years of pictures of Connelly – event pictures, paparazzi beach grabs, publicity stills, extracted video frames, and many other sources that have enabled the system to generalize the actress's core identity, face and physique across a range of ages.

To boot, Connelly's hair styling is relatively consistent over the years, which is not always the case with women (because of fashion and aging) or men (because of fashion, ageing, and male pattern baldness).

*Despite only fairly recent stardom, Stable Diffusion has internalized a wide range of hairstyles for actress Margot Robbie, many of which can be quite resistant to prompt-based attempts to stabilize them for purposes of coherent temporal video.*

However, not least because of the sheer volume of material in the database, Connelly is wearing something different in nearly all of her LAION photos:

CONTINUED ON NEXT PAGE




*A diversity of sartorial choices for Jennifer Connelly over the years, as represented in the LAION database, and subsequently trained into Stable Diffusion. Source: https://rom1504.github.io/clip-retrieval/?index=laion5B&useMclip=false&query=jennifer+connelly*

So if you ask Stable Diffusion for a picture of *'Jennifer Connelly'*, does it choose a particular outfit that's above-averagely represented in her stable of LAION pictures? Might it generalize every single outfit she wears in LAION into something 'representatively generic'? Does it pick from a range of outfits with the highest LAION aesthetic score? And to what extent will the prompt itself affect the choice or continuity of the clothing that gets depicted in a range of rendered frames?


*Diverse renders from Stable Diffusion prompts related to 'Jennifer Connelly', show a largely random range of attire.*

Stable Diffusion was only released to open source little more than a month ago, and these are among the many questions that are yet to be answered; but in practice, even with a *fixed seed* (which we'll look at in a moment), it's hard to obtain temporally consistent clothing in full-body deepfake video derived from a latent diffusion model such as Stable Diffusion or DALL-E 2 – unless the clothing in question is distinct, unchanging over the years, and already well-represented in the model's training database.

## Potential Full-Body Deepfake Consistency Through Textual Inversion

One solution for consistent clothing in this scenario could be the use of Textual Inversion models – small scraps of adjunct code that encapsulate the look and semantic meaning of a custom object, person or entity, via the short training of a limited number of annotated photos.

Textual inversions can be created by users and placed 'adjacent' to the standard trained model at inference time, and can effectively act as if they had originally been trained into the system.

In this way, in theory, it would be possible to summon up a pair of Levi 501s (or a specific hairstyle) with a consistent enough appearance to support temporal video; and also to create truly 'stable' models of more obscure items of clothing.

If this became an established solution, it would be a little like the early heyday of the Renderosity marketplace, where users still trade or sell outfits and 'mods' for the CGI-based virtual humans in Poser and Daz 3D.

Ultimately, Textual Inversion might represent the only rational way to gain temporal consistency for objects in Stable Diffusion, and to easily insert 'unknown' people into the system, for the purpose of creating full-body deepfakes via a latent diffusion system. Some Reddit users are currently putting themselves (and some more obscure public figures) into Stable Diffusion via this route:

*None of these people are in your copy of Stable Diffusion, either at all, or at this resolution - but rather have been added by enthusiasts using Textual Inversion. The top images are self-portraits of Reddit user 'Dalle2Pictures', who, despite his username, in this case used Textual Inversion with Stable Diffusion; the middle row is another Reddit user, sEi_, who likewise used Textual Inversion to insert his own likeness into the system; the bottom row are Stable Diffusion renders of Former United States Representative Tulsi Gabbard, not featured at this level of detail in a standard Stable Diffusion distribution; in this case Reddit user Visual-Ad-8655 reportedly took just two hours to generate a Textual Inversion for Gabbard. Sources, top to bottom: https://old.reddit.com/r/StableDiffusion/comments/xjl49b/i_used_textual_inversion_with_stabledifussion_to/ | https://old.reddit.com/r/StableDiffusion/comments/x9uol8/adding_new_objects_to_the_model_added_my_face_so/ | https://old.reddit.com/r/StableDiffusion/comments/xdl48y/textual_inversion_test_of_tulsi_gabbard/*

Though the creation process currently has high hardware demands, users can create custom Textual Inversion files via web-based Google Colabs and Hugging Face APIs.

Additionally, the rapid pace of development and optimization in the Stable Diffusion developer community means that it might become easier to put yourself (or any celebrity that's absent or under-represented in LAION) into the world of Stable Diffusion via a local, consumer-level video card.

*(For more about Textual Inversion, check out our August feature on the future of 'general' video synthesis in Stable Diffusion).*

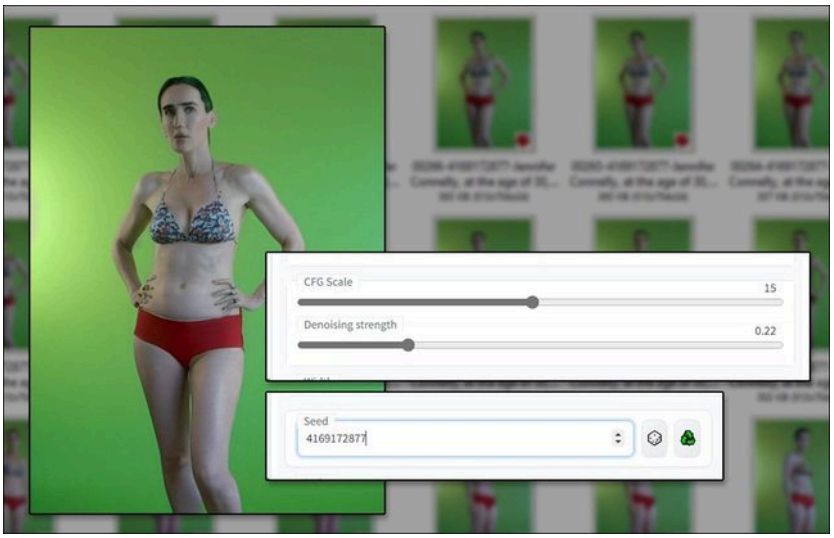
## Full-Body Video Deepfakes with Stable Diffusion and EbSynth

To create the Jennifer Connelly and Henry Cavill Stable Diffusion-based full-body deepfakes shown at the start of the article, I took a brief section of footage from a custom filming session with a performer, and extracted the short video into its constituent frames.

As is clear from the clip, the same brief footage is used for both transformations.

I then made some tests of some of the original source frames, and eventually found one combination of settings (in this example, for Jennifer Connelly), that seemed to produce a good result.

CONTINUED ON NEXT PAGE




*Settings that more or less worked to effect a transformation from the real-world model into the target personality.*

We've already seen how 'random' Stable Diffusion's interpretations of an Img2Img text prompt can be. In fact, in order to produce novel and diverse results, the system filters the text prompt through a random seed for each individual image generation – a single, unique route through the latent space of Stable Diffusion, represented by a hash number. Without this functionality, it would be difficult to explore the potential of the software, or generate variations on a prompt.

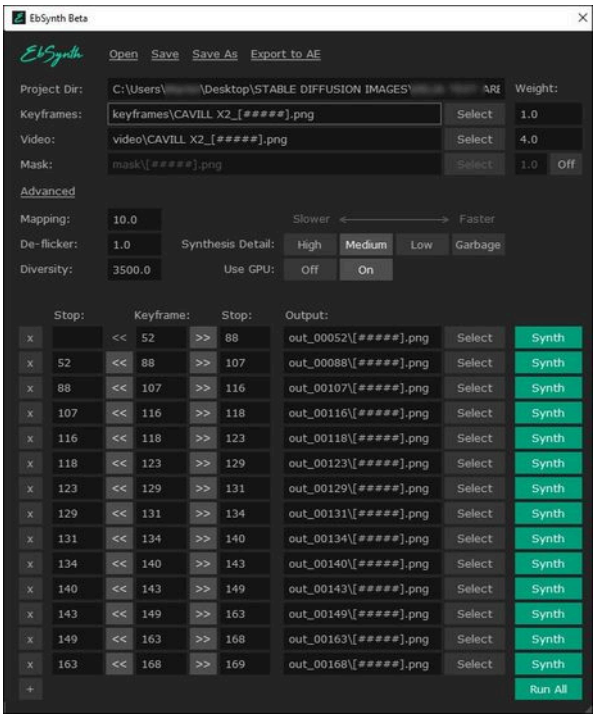
All distributions of Stable Diffusion let you 'freeze' this seed, if you find one that really works well – and this ability is absolutely essential for any hope of temporal coherence when working with a contiguous sequence of images, as in this scenario.

However, if the subject moves around a lot in the video, the seed, which operated so effectively on a single frame, is not likely to work as well on frames that are a little different:


*The seed that produced the first image transformation proved very effective, and was chosen as the 'fixed seed' for the entire sequence. But it's not as applicable to the second image, which is also part of the video sequence. Here the difference in quality is exaggerated for illustrative purposes – though it can be even worse, depending on how 'mobile' the performer is in the clip.*

As I write, a new Stable Diffusion script has been developed that can ostensibly 'morph' between two optimal seeds in a rendered sequence. Though such a solution wouldn't solve all the problems of 'seed shift', it could allow performers to move about a bit more in Stable Diffusion/EbSynth transformations, as most 'photorealistic' examples of SD/EbSynth video clips at the moment are characterised by very limited character movement.

Returning to the celebrity transformations: enter the aforementioned EbSynth – an innovative, obscure, and scantly-documented non-AI application that was originally designed to apply painterly styles to short video clips, but which is gaining popularity as a 'tweening' tool for videos that use Stable Diffusion output.




*EbSynth in action.*

To give some idea of the increased smoothness that EbSynth can offer a full-body Stable Diffusion deepfake, compare the original, raw Jennifer Connelly transforms produced by Stable Diffusion, on the left of the video below, to the EbSynth version on the right. A smoother video has been created by EbSynth, by 'morphing' between a handful of carefully-selected keyframes, and using only these (24 is apparently the maximum allowed per clip) to recreate the full, though inevitably short runtime of the video:


*On the left, the raw output from Stable Diffusion 'sizzles', because even with a fixed seed, temporal consistency is hard to achieve by just gluing the raw output frames together. On the right, we see better temporal consistency obtained by EbSynth, which has converted a mere 24 frames (out of the original 200 frames in the clip) into a smoother reconstruction. To improve the facial quality of the final video on the right, a publicly-shared autoencoder was used – though in fact better results can be obtained by zooming in on the face and re-rendering it entirely in Stable Diffusion (a process which is currently rather more time-consuming).*

Despite its brilliance, EbSynth is a frustrating tool to use for this purpose, due to a number of confounding interface quirks, the lack of cohesive or centralized documentation, a minimal and restrictive Reddit presence, and conflicting opinions about what the crucial settings in the application's 'advanced' section actually do, either for the original intent of style transfer, or for this jury-rigged purpose.

Additionally, the very small amount of keyframes you are allowed to set in EbSynth means that a) the clip will probably need to be very short, and b) the person in the clip will probably need to limit any sudden movements, because every additional movement eats up that precious allotment of keyframes.

As a principle workflow, however, the basic tenets and functionalities of EbSynth could possibly be adapted into new software with greater keyframe capacity, some ability to detect where additional keyframes should be assigned (in EbSynth, you need to curate them quite carefully), and more transparent instrumentality for controlling the interpolation settings.

## Other Routes to Full-Body Deepfakes

Besides the tortuous and challenging path to effective CogVideo-style neural text-to-video systems, and these kind of extremely limited 'hacks' for temporally coherent Img2Img Stable Diffusion full-body deepfakes, there are certainly other roads forward for extending identity transformation beyond the facial area.

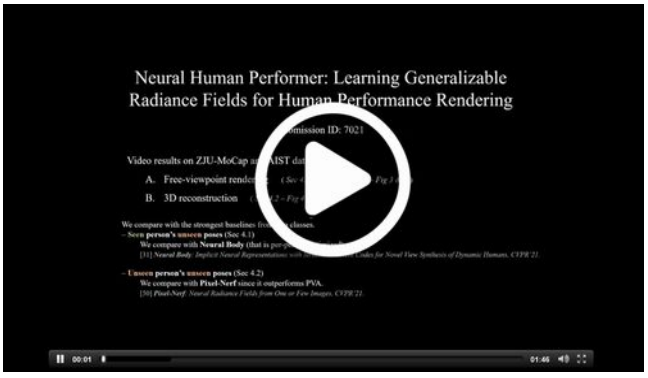
I have covered most of these alternative methods quite extensively in previous features here at Metaphysic, including each approach's capacity and potential to generate full-body deepfakes. Therefore I refer you to those features, on The Future of Autoencoder-Based Deepfakes; the possibility that Neural Radiance Fields (NeRF) might become an eventual successor to autoencoders; and the future of Generative Adversarial Networks (GANs) in regard to deepfakes.

Nonetheless, let's briefly review these alternative options.

### Neural Radiance Fields (NeRF)

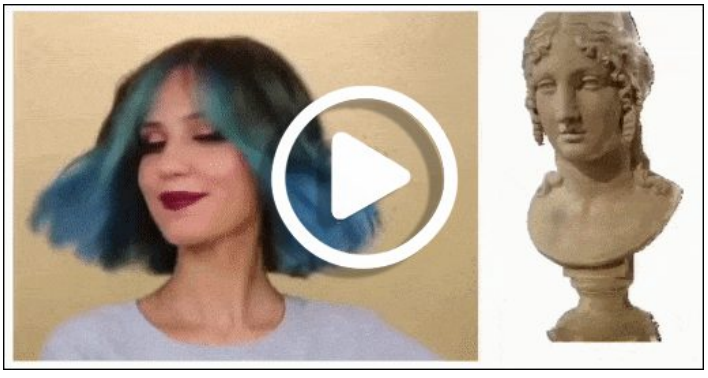
There is no video synthesis technology that deals more extensively with full-body neural representations of people than Neural Radiance Fields. NeRF can recreate temporally-accurate video as well as 'frozen', explorable 3D representations, by training images and videos into neural scene and object representations.

For instance, Neural Human Performer can perform a style of deepfake puppetry, albeit currently at very low resolution (a limitation common to most NeRF initiatives):




As I have mentioned in the previous NeRF article, other projects that deal directly with neural humans in NeRF are numerous, and include MirrorNeRF, A-NeRF, Animatable Neural Radiance Fields, Neural Actor, DFA-NeRF, Portrait NeRF, DD-NeRF, H-NeRF, and Surface-Aligned Neural Radiance Fields.

A further example of NeRF-based deepfake puppetry is NeRF-Editing, which uses Signed Distance Functions/Fields (SDF) as an interpretive layer between a human performer (or, in theory, priors taken from a CogVideo-style database) and the usually inaccessible parameters of a NeRF object – or, potentially, a different identity:



*Deepfake puppetry with NeRF-editing. Source: Source: http://geometrylearning.com/NeRFEditing/*

Some human synthesis projects are beginning to integrate NeRF into a wider and more complex workflow that includes elements of traditional CGI, such as texturing, including Disney Research's Morphable Radiance Fields, or else are beginning to use NeRF to swap faces rather than render entire bodies. An example of the latter case is RigNeRF, a NeRF-based face-swapping method that offers deepfake puppetry very similar to DeepFaceLive, though it's not remotely as mature in implementation.

I could go on all day, because this is a fertile and well-funded strand of video synthesis research. The commercial and academic sectors are very keen to develop neural humans using this technology, while NVIDIA's recent foray into more efficient NeRF generation has re-invigorated industry interest.

Nonetheless, NeRF's challenges and inherent constraints are formidable: Neural Radiance Fields are very difficult to edit, and usually expensive and time-consuming to train, while NeRF-based neural humans are characterized by limited resolution, which tends to undermine the potential of

the system to leverage real-world images and videos to create, potentially, the most authentic neural humans possible.

Nonetheless, as Stable Diffusion has proved, and DALL-E 2 has presaged, quantum leaps in image synthesis technologies tend to take us by surprise, so NeRF may yet improve, at one sweep, its current, struggling position as a practicable method of simulating full-body humans.

## Autoencoders (Deepfakes)

Autoencoder-based open source repositories such as DeepFaceLab and FaceSwap (both based on the controversial 2017 code that premiered sensationally on Reddit) are what most of us think of when we hear the term 'deepfakes' – models which are trained on thousands of images of celebrities, and which can subsequently impose those learned faces into the central facial area of other people, effectively changing that person's facial identity.


Autoencoder deepfake systems only swap faces, not bodies. Notwithstanding, there is occasional speculation among fans and developers that an autoencoder system modeled along the same lines could be devised that uses the kind of full-body motion capture software that can create deepfaked dancers, therefore enabling full-body deepfakes.

However, even if such a system could be devised, it would face many of the same problems with clothing that Stable Diffusion does when producing temporal deepfake content, making the creation of a usable training dataset practically impossible.

Unless, of course, clothes don't enter into the equation, and the putative system were to be trained on images of naked bodies, and intended to produce full-body deepfake porn.

However, whose face would be in those training pictures? If it were a particular celebrity, practically the entire contents of the dataset would need to be synthetic, i.e. Photoshopped; and, after training, the difficulty of finding a 'body match' would be effectively doubled. In a best-case scenario, a porn deepfake would now have to be processed *twice*, with two different frameworks, at double the preliminary effort, and for relatively little gain, compared to what is currently achievable.

Additionally, truly distinct physiques are relatively rare, and, given the demands and relatively low standards of the deepfake porn community, this kind of effort would arguably qualify as 'overkill'.

Taking into consideration those factors, and how improbable it is that any substantial corporate entity would fund such an effort, there seems no obvious road ahead for autoencoders in the production of full-body deepfakes, except as a possible adjunct technology, retaining a focus on face-swapping in the wider context of full-body deepfakes produced by other methods (assuming such methods cannot handle the task at least as well, if not better).

## Generative Adversarial Networks (GANs)

The primary use of Generative Adversarial Networks in full-body deepfake initiatives comes in the form of well-funded industry interest in *fashion-based* body and clothing synthesis – especially in regard to systems that could allow 'virtual try-ons', primarily in the women's clothing market.

Though projects such as InsetGAN and StyleGAN-Human (see video below) are keen to develop commercial applications of this nature, the resulting renders are always either static, or nearly-static:

CONTINUED ON NEXT PAGE


Though GANs have gained public acclaim and notoriety over the past five years for their ability to produce the most realistic faces of any image synthesis system (including DALL-E 2 and Stable Diffusion), a Generative Adversarial Network lacks any temporal architecture or instrumentality that might suit it for the production of full-body deepfakes.

After years of near-fruitless exploration into the possibility of realistically animating faces in the GAN's latent space via purely neural methods, the research sector, exemplified by Disney Research's efforts, is increasingly coming to accept that GANs may only be useful as texture generators that are powered by entirely different, often older technologies based around CGI, such as 3D morphable models (3DMM).

If there is any real 'race' to develop effective and versatile full-body deepfakes, Generative Adversarial Networks appear, currently at least, to be stuck at the starting line.

** Though the author is a regular freelance contributor to the Metaphysic blog, he is not an employee of Metaphysic. The original full-body deepfake examples in this feature are the author's own experiments, and entirely unrelated to the work, technologies and output of Metaphysic.*

† *In a Zoom conversation on 10th August 2022.*