Stable Diffusion: Is Video Coming Soon?

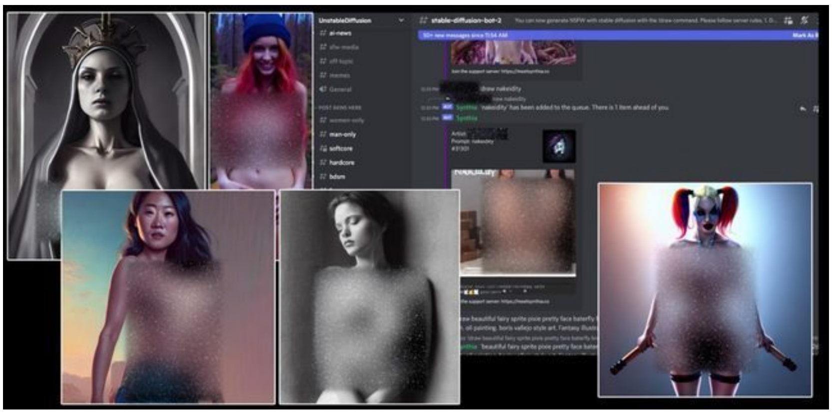
Originally published September 01, 2022 at https://metaphysic.ai/stable-diffusion-is-video-coming-soon/

For an excited public, many of whom consider diffusion-based image synthesis to be indistinguishable from magic, the open source release of Stable Diffusion seems certain to be quickly followed up by new and dazzling text-to-video frameworks – but the wait-time might be longer than they're expecting.

The recent and ongoing explosion of interest in AI-generated art reached a new peak last month, as stability.ai open sourced their Stable Diffusion image synthesis framework – a latent diffusion architecture similar to OpenAI's DALL-E 2 and Google's Imagen, and trained on millions of images scraped from the web.

Some of the SFW art emerging from the rapidly growing Stable Diffusion community. Source: Reddit and Discord

In a revolutionary and bold move, the model – which can create images on mid-range consumer video cards – was released with fully-trained weights, and the ability to easily remove the single line of code that prevents it creating pornographic or violent content.



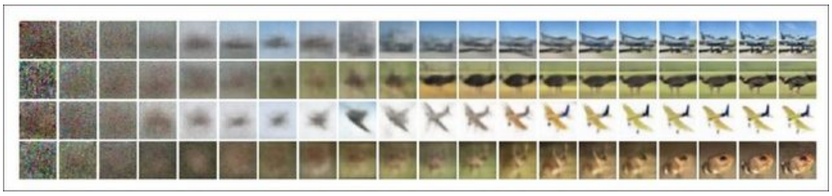
The Stable Diffusion bot 'Synthia' serves up AI porn on demand in one of the 'request areas' of a NSFW Discord community. Source: Discord

Unlike autoencoder-based deepfake content, or the human recreations that can be achieved by Neural Radiance Fields (NeRF) and Generative Adversarial Networks (GANs), diffusion-based systems learn to generate images by adding noise to existing source photos, which subsequently teaches the system how to make plausible and even photorealistic images solely from noise (or, as it transpires, practically anything else).

CLICK TO VIEW ANIMATED GIF

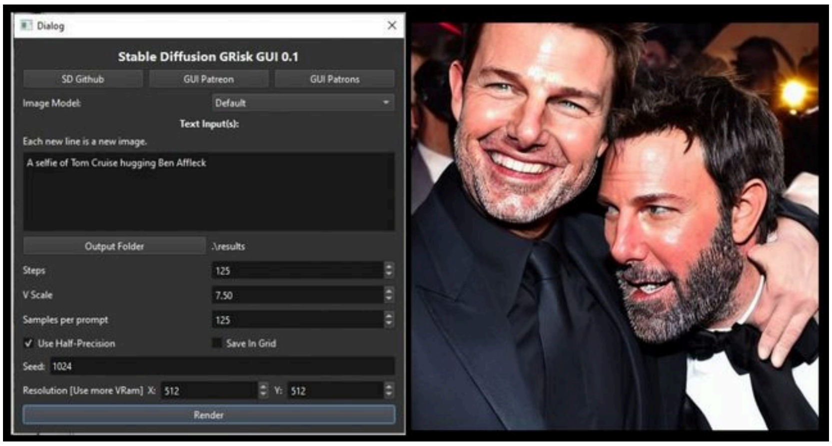
Diffusion-based models learn how to reconstruct photos from noise by adding noise to 'untainted' photos and noting the relationship between the unaffected and the tainted image as more noise is added. From this, the model begins to understand latent relationships between highly diffused sources and sharp, even high-resolution versions of those sources. Well-trained, a latent diffusion text-to-image model can then 'recover' images from base noise, using text prompts as guides for which content to recover. Source: https://ai.googleblog.com/2021/07/high-fidelity-image-generation-using.html

According to a 2021 paper from OpenAI, diffusion models have a clear advantage over GAN image synthesis in terms of accuracy and realism. Though this contention supports their own product (the DALL-E line), recent public interest in such systems seems to bear it out.



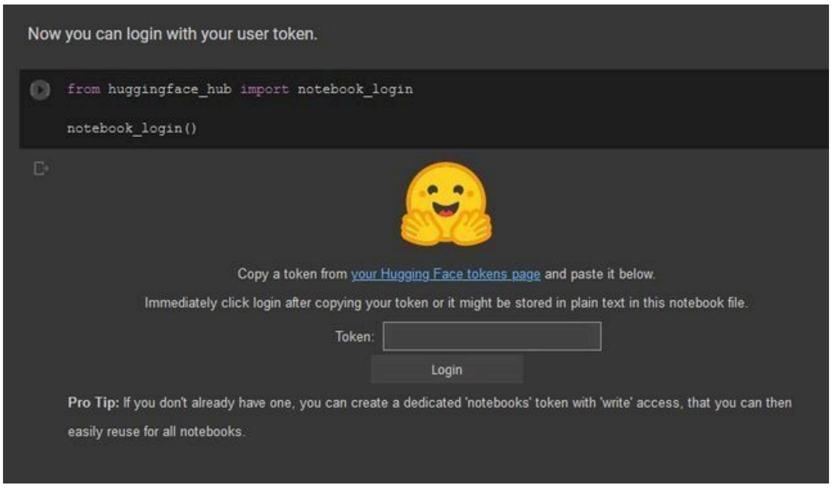
Photorealistic pictures emerge from pure static in diffusion models. Source: https://arxiv.org/pdf/2006.11239.pdf

The text-based annotations in the controversial LAION-400M image dataset on which Stable Diffusion was partially trained provide embedded connections between images and concepts (including real personalities, which are extensively represented in the dataset), enabled by the CLIP ViT-L/14 text encoder.



The first version of GRisk's free Stable Diffusion Windows executable provides most of the functionality available in Colab versions, and runs locally, using your own GPU. This image was rendered in about 12 seconds on a NVIDIA RTX 2070S with 8GB of VRAM, at 512x512px.

Within days of release, the open sourced Stable Diffusion code and weights were packaged into a free Windows executable (see image above) by a developer, obviating the one remaining limitation of the NSFW-filtered Hugging Face and monetized DreamStudio APIs – the need to obtain an authentication token to gain access to the model's weights (which constitute its entire practical value, at least at the moment).



In one of the many Google Colab notebooks hastily derived from the released Stable Diffusion source code, the need to obtain authentication from the Hugging Face server remains an obstacle – unless you have downloaded one of the early executable packages (see above) or alternate Conda-based distributions, which may already include these weights. Source: Google Colab

Future versions, the author says, will include additional functionality such as Img2Img (see below), which allows visual prompting based on sketches or other photographs.

In terms of AI-facilitated image synthesis, the consequences of the Stable Diffusion release are immense. In a climate where DALL-E 2 operates under strict filtering (and semi-effective pre-training filtering) for violent and sexual content, and where Google is so cowed by the potential for the abuse of image synthesis that it seems it may never make Imagen available, stability.ai has effectively trashed all these barriers.

Custom-Made Image Synthesis Models

To boot, even if Stable Diffusion does not excel at every possible type of text-prompt (for example, in reproducing a particular genre, style of art, or celebrity that was under-represented in the original trained dataset), the weights of the model can be 'fine-tuned' by end-users, who can continue to train it solely on image collections of their choice.

At the time of writing, developers at both SFW and NSFW Discords are discussing collaborative efforts to systematize and rationalize this process. Additionally, images can be used either as scrappy guidelines for the higher-resolution of the model, or as incremental edits, using the Img2Img library, which is incorporated into the functionality of Stable Diffusion (though not available in all APIs or distributions at the moment).

A sketch-to-image paradigm, aided by text-prompts, can allow Stable Diffusion users to create extraordinary transformations using either crude information or existing images. Source: tinyurl.com/55vmb9t6 and tinyurl.com/23xw5pa3

The potential of using sketches as guidelines for sophisticated renders was first realized by nature-based initiatives such as NVIDIA's GauGAN project, with video-centered projects such as Google's Infinite Nature (which we'll look at later) adding 'hallucinated' animation capabilities.

The prompt for this integrated Img2Img/Stable Diffusion output was 'A man with golden armor, and mask, rises from the sands, a shiny golden magical staff in one hand, Artstation, Cinematic, Golden Hour, Sunlight, detailed, elegant, ornate, desert, rocky mountains, a big shiny white sun in the background, Illustration, by Weta Digital, sandstorm, Painting, Saturated, Sun rays'. Source: https://old.reddit.com/r/StableDiffusion/comments/wy7oa5/img2img_is_just_unreal_im_stunned/

Another facial improvement method gaining popularity with Stable Diffusion users is to exploit the rich facial relationships and embeddings trained into TenCent's GFPGAN Generative Adversarial Network.

Stable Diffusion's evenly-distributed concentration of effort across all aspects of the synthesized image can leave faces with artifacts, and a certain lack of distinction. Augmentation through GFPGAN can restore clarity focused on the face. Source: https://github.com/hlky/stable-diffusion-webui

Like Stable Diffusion, the GFPGAN network is trained on a huge number of faces, including many famous faces. Though it can't hope to have data on every possible celebrity, luminaries of the last 10-15 years are well-represented, as are 'classic' celebrities such as Marilyn Monroe, and the general algorithm will generically improve any faces that were not present in the training data.

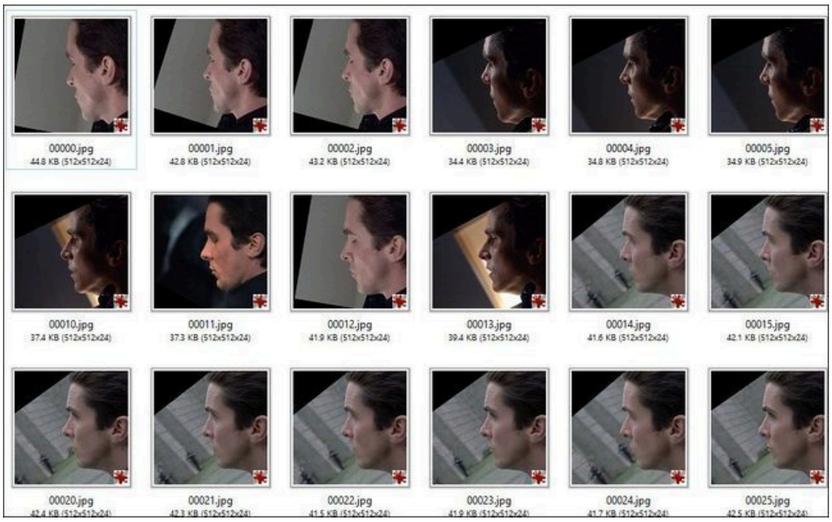
GFPGAN can make notable quality improvements to faces. Source: Discord

GFPGAN is now available in the Stable Diffusion web GUI, which currently requires installation through GIT, though the GRisk Windows executable is set to eventually incorporate GFPGAN as well, for a shallower learning curve.

Instant 'Deepfake' Face Sets

Exploiting the facial images trained into Stable Diffusion and GFPGA removes the need to gather and train thousands of source images of an individual – familiar drudgery for users of autoencoder-based deepfake systems.

CONTINUED ON NEXT PAGE



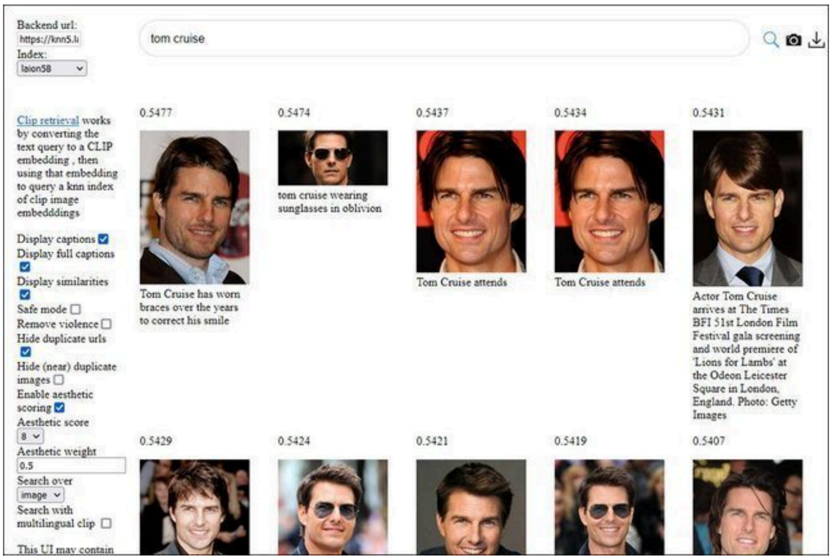
From a free and downloadable archive of face sets in the deepfake community, this Christian Bale dataset contains 7,670 images painstakingly curated and aligned, for training into a deepfake model. For the best results possible, it's also necessary for users to extract images directly from UHQ sources such as Blu-ray releases of movies.

The sprawling hyperscale datasets that inform the likes of DALL-E 2 and Stable Diffusion already include hundreds of thousands of celebrity images, as well as more general material to populate environments and provide full-body deepfake content.

Users can train underrepresented celebrities (or anyone else) into the system, if they want, even though the methods for doing so are currently nascent and rough-edged.

Celebrity images: raw output from Stable Diffusion with the prompt "Award-winning color studio portrait of [CELEBRITY NAME], high detail, studio lighting, 8k, colour, color'. All that painstaking data-gathering associated with traditional deepfakes has already been done within the extensive library of images in the LAION dataset (and the FFHQ dataset and synthetic data, if you augment the images with GFPGAN). Many other libraries can be incorporated into the Stable Diffusion installation to fine-tune the quality of the faces.

Stable Diffusion was trained primarily on [LAION-Aesthetics](), a collection of subsets of the LAION 5B dataset, which is itself a subset of LAION-400M. LAION-aesthetics [can be explored]() by ticking 'enable aesthetic scoring' in a [searchable and browsable GUI]().



The LAION-aesthetics subset can be browsed and searched interactively online. Source: https://rom1504.github.io/clip-retrieval/?back=https%3A%2F%2Fknn5.laion.ai&index=laion5B&useMclip=false&query=tom+cruise

The searchable LAION database also allows Stable Diffusion users to get a better understanding of the way the model 'thinks', by exploring which labels are most strongly associated with concepts or human movement 8s that the user might like to reproduce in a prompt.

Searching LAION 5b for 'hugging' gives an insight into the labels that may best reproduce (in this case) a 'hug' pose.

By examining the strongest correlations between labels and the desired pose/movement, Stable Diffusion users can incorporate apposite prompts designed to exploit the semantic relationships between words and images that are trained into the model.

Since the searchable database is ordered not by labels but by the innate aesthetic scores for images and groups of images, it's remarkably easy to see what the dataset considers to be the 'nearest neighbor' for any particular face, concept, or object, simply by scrolling down until the 'target' results begin to trail off:

The 'nearest-scored neighbors' for actor Gene Hackman, in LAION-aesthetics. Source: https://rom1504.github.io/clip-retrieval/?back=https%3A%2F%2Fknn5.laion.ai&index=laion5B&useMclip=false&query=kissing

This kind of relationship-mapping is not possible for celebrities represented at volume in the database (for instance, there are more 'Jennifer Lawrence' pictures in LAION-aesthetics than a search will disclose), but can provide useful insight into the way Stable Diffusion has formed associations.

Composite Understanding

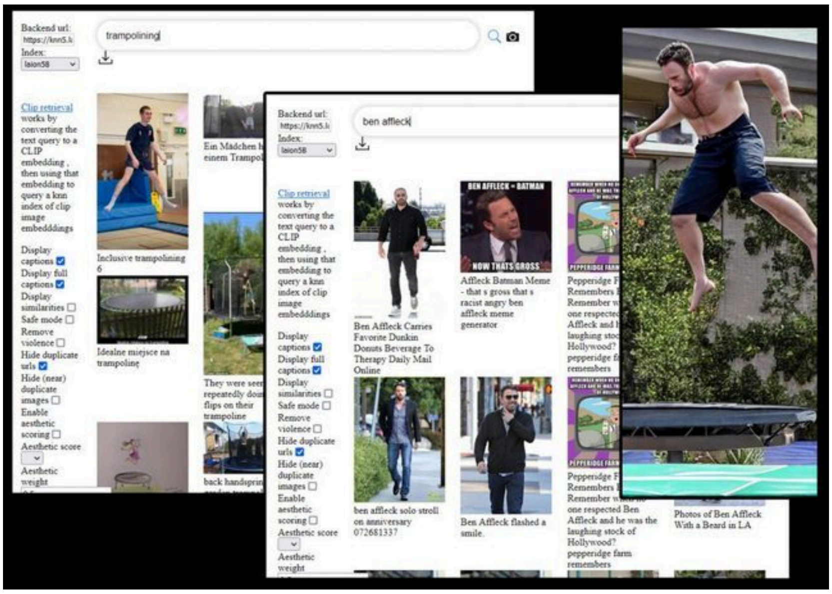
It is not easy to conceptualize what's happening when the training of an image synthesis system (such as a GAN or a latent diffusion architecture) 'generalizes' classes, priors and entities from a group of images of a particular individual.

One recent medical paper, studying the use of expression recognition in diagnosing mental disorders, perhaps illustrated this best, by providing two 'composite' photos of two groups of multiple people (*'unaffected'* and *'affected by mental illness'*), to offer a visual 'average' or mean representation of all the faces in each group:

These pictures represent multiple, combined identities from two groups of people, and neither image is accurate to any individual identity. Source: https://arxiv.org/abs/2208.01369

Likewise the training of a model such as Stable Diffusion creates an internal 'palette' representing any one person whose images have been correctly and consistently labeled in the training dataset (i.e. *'image of Ben Affleck*, *'photo of Ben Affleck attending an event in Los Angeles'*, etc.) In effect, the model develops an ability to reproduce any well-represented identity across a range of poses. This transformative capability is very similar to the traditional deepfake functionality that emerged in 2017 (and which, for most people, still defines the term 'deepfakes').

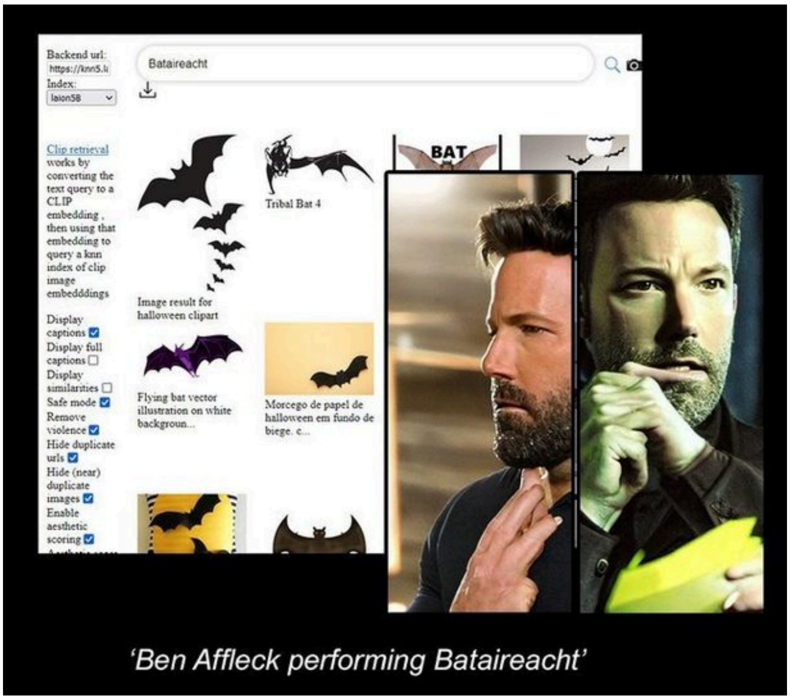
The difference is that the Stable Diffusion model has also ingested vast amounts of environmental and incidental data, as well as data related to the general physical characteristics of celebrities (i.e., not just their inner faces), and can accurately reproduce people in their entirety, and in the context of the user's choice (such as 'studio setting').



Data trained into Stable Diffusion from the LAION database, which has generalized well enough on 'trampolining' and 'Ben Affleck' to be able to combine these two concepts. Prompt: 'Ben Affleck jumping on a trampoline, award-winning 8K photo, high resolution, high details'

Therefore Stable Diffusion synthesizes images in a similar way to how we internally imagine scenes and events. If you've ever seen anyone trampolining, and have seen one or two Ben Affleck movies, it isn't a stretch to imagine the actor trampolining: you've 'internalized' the high-level concept of trampolining, and you've also seen enough video and still images of the actor to conjoin these two concepts.

Unless you're familiar with *Bataireacht*, the obscure Irish martial art of stick-fighting, you won't be able to imagine Affleck doing this, and, as it turns out, neither will Stable Diffusion, because this activity did not make it into the LAION sub-set.

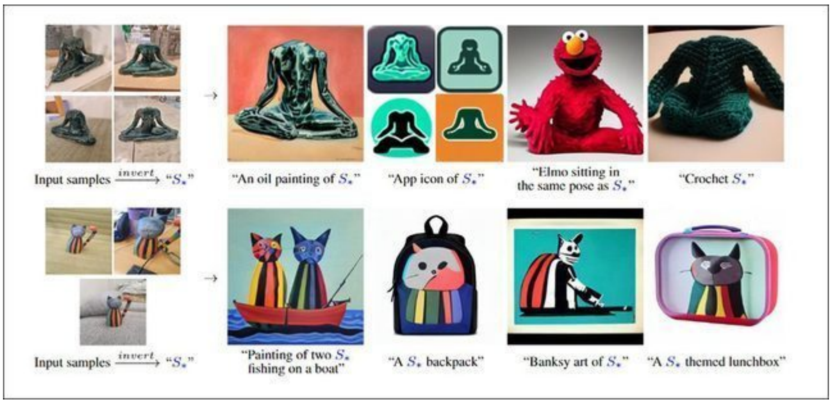


"Did you mean 'Batman'?". Stable Diffusion's got nothing regarding the obscure Irish martial art, and defaults to basic renditions of Affleck. In the rear image, we see the online searchable version of the database used to train Stable Diffusion (more details below).

For this reason, the Stable Diffusion fan-base is currently very interested in methods which could allow their projects to exceed the limits of the database's knowledge.

Textual Inversion

One such method comes via a novel technique titled *textual inversion*, recently published by Tel Aviv University. Textual inversion offers a way to encapsulate a desired concept or visual that the user wishes to retain in an image generation, and impose it onto subsequent generations, without the need for direct and detailed image prompts on each occasion:



Why rely on the embeddings from some out-of-date image corpus on which your image synthesis framework was trained? By embodying, representing and defining your own image source through textual inversion, you can then apply the essential characteristics of it to any text prompt (represented by 'S' in the above examples). Source: https://arxiv.org/pdf/2208.01618.pdf

The idea behind textual inversion is that the user trains a small number of images into the existing model, together with labels that cohere the concept.

This single entity can then be referenced by a token (i.e. *'S'*) when the post-trained model is use for inference.

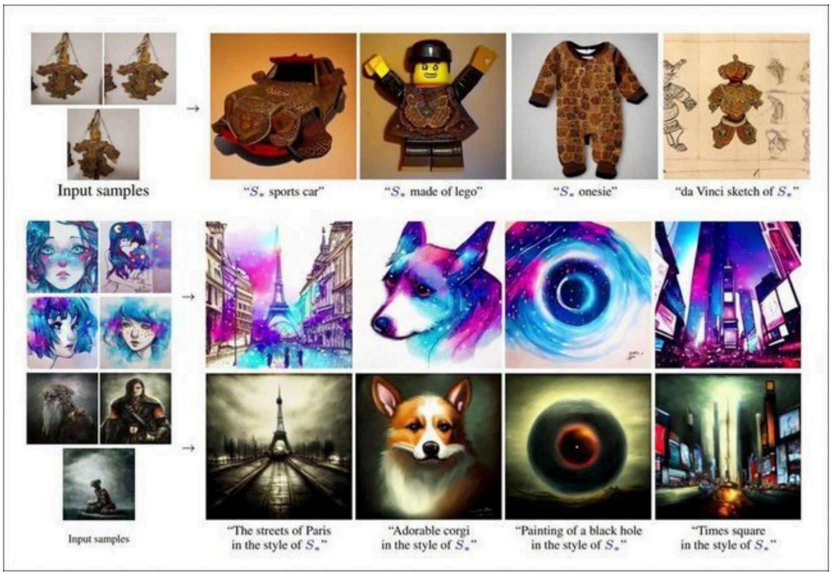
One word will suffice, when images and concepts have been 'baked' into a single reference term with textual inversion:

Additionally, at the time of writing, Google Research has just released a similar system called [DreamBooth](https://dreambooth.github.io/), which likewise 'tokenizes' a desired element into a distinct 'object'. Though the object retains some interpretive freedom of movement (pose, lighting, interaction, etc.) when inserted into 'alien' contexts, it is able to maintain its semantic integrity:

DreamBooth allows for the creation of distinct and resilient objects, creatures and people that are able to 'remain themselves' in challenging virtual circumstances without looking like they've been 'cut and pasted' into a scene. Source: https://dreambooth.github.io/

As I write, the first of the textual inversion approaches is being shoe-horned into Stable Diffusion, albeit with the severe limitation of needing a graphics card with a minimum of 20GB of VRAM, which may limit its widespread use to Colab environments.

More sophisticated and user-friendly integrations are being discussed and evolved, though it may prove difficult to lower the hardware requirements significantly.



Textual Inversion integrated, at some expense of hardware requirements, into Stable Diffusion. Source: https://old.reddit.com/r/StableDiffusion/comments/wvzr7s/tutorial_fine_tuning_stable_diffusion_using_only/

These two approaches could help diffusion models such as Stable Diffusion, DALL-E 2 and Midjourney to create 'resilient' entities – including the identities of people – that are less inclined to becoming indistinct or distorted through interaction with other entities in a projected fictitious space, in situations that depict close physical interaction or complex self-poses (see 'Out on a Limb' below).

Eventually, the evolution of textual inversion could even remove the need to train under-represented ideas, concepts, and visual lexicons into existing Stable Diffusion weights, since this kind of 'fine-tuning' can have an adverse affect on the overall quality of output from the model, and set it back to a less developed phase in its training, without ever incorporating the 'supplemental material' as fully as if it had been present from the start of training.

Nonetheless, despite excitement in the Stable Diffusion community about the potential of such approaches, the authors of the original paper note that textual inversion might be more effective for style-transfer, rather than 'object transfer'; and as several others have observed, the very small number (about five) of images recommended to be trained into the model *post facto* are not likely to produce a very flexible or exploitable image entity.

Out on a Limb

As we'll see, the way that machine learning models internalize and embody movement-based and pose-based human concepts, such as 'hugging', is essential not only to accurate reproduction of *still* images, but also to the far more complex interpretation necessary to recreate video movement from synthesized images.

In truth, complex poses can be a hit-and-miss affair in latent diffusion image generators, most particularly in regard to human extremities, and the general disposition of human limbs, including hands. Close examination of Tom Cruise's hand in our earlier Stable Diffusion-prompted image reveals some problems:

CONTINUED ON NEXT PAGE

Tom Cruise's hand is not in great shape, as Stable Diffusion has difficulty eliciting and representing complex poses and disposition of human limbs.

As with GANs and autoencoder-based deepfakes, the closer two subjects are together, the more prone diffusion models are to become confused. Here are some Stable Diffusion examples featuring 1997-era Kate Winslet and Leonardo DiCaprio trying for a romantic embrace in James Cameron's *Titanic*.

In the first two images, the Winslet character's arm seems to be penetrating DiCaprio's shoulder (and even duplicating itself, in the second image), in much the same way clipping errors have amused videogame enthusiasts over the years. The problem is also very familiar to CGI practitioners, and falls into the realm of collision detection. In the third image, both the forearm and upper arm are overly extended, and not completely straight. But it certainly doesn't take two to trigger such defects: trying to get 1990s Kate Winslet to perform the handstand scorpion yoga position in Stable Diffusion is a doomed endeavor, while even the simple lotus position produces some Cronenberg-style body horror.

For this reason, in much the same way that practitioners of autoencoder-based deepfake systems tend to avoid close quarters between two characters, the best of the current dazzling crop of human-centric Stable Diffusion prompt outputs is characterized by sole figures beautifully rendered in styles drawn wholesale from art communities such as ArtStation.

The best Stable Diffusion characters tend to go it alone.

Not only does this limit a lot of the finest output of the system to the realm of advanced production art (at least, in terms of film and TV production), but 'concept art' is a term frequently included in prompts in the Stable Diffusion community.

Limb entanglement is a fundamental problem in diffusion-based image synthesis systems. Despite the superior rendering quality and finish of output from DALL-E 2, which was trained at greater length and expense than Stable Diffusion has been to date, OpenAI's current flagship product replicates such goofs frequently:

Above, DALL-E 2's four attempts (from a single prompt request) to reproduce 'A woman embracing a man', produces only one image free of limb-confusion – the same success rate as Stable Diffusion (images below) is able to achieve. These are not cherry-picked images.

It could be that increased and more varied data could help address this kind of limb-based confusion. There are so many possible configurations and dispositions of limbs, particularly between multiple individuals within the same generated photo, that the process of generalization may only have 2-3 images for each pose at its disposal, where (for instance) hands are in different positions, leading to a 'fusion' of 'multiple hands'.

Grass-Roots, Global AI Training?

Ultimately, increased training as well as higher volumes of images would be likely to alleviate the problem. There is currently talk in the Stable Diffusion developer community regarding the possibility of exploiting increased consumer-level GPU availability (brought about by Ethereum's switch to proof-of-stake, and an unexpected surfeit of NVIDIA GPU inventory) into a collective system that would distribute training tasks across a global community of contributing users.
This would allow perpetual networked training of the common models that underpin Stable Diffusion (rather than fine-tuning, which involves one user adding some pictures and continuing to train the public model so that it incorporates those images, which is likely to compromise the original rendering quality of the model).

In the meantime, it seems likely that proponents of Stable Diffusion output will develop their own set of 'situations to avoid', just as the autoencoder deepfake community has done.

Square Dealing and Decapitation

Not all the weird contortions and nightmare-style limb-manglings that Stable Diffusion can unwittingly generate are centered around poor training or inadequate data. Sometimes the system has to make unappealing compromises for prompts that are at odds with other factors.

The system has been consistently trained in a square aspect ratio, notwithstanding the (often differing) dimensions of the input training images. Because this can lead either to cropping or padding of the input image (depending on decisions that are automatically made at training time), this could be one of the reasons for several different types of rendering issue – most famously, the 'extra anatomy' that can appear when the source data was primarily 'square', but the user has set their resolution to a non-square mode such as 320×768:

Emma Watson gets some additional anatomy, as Stable Diffusion was 'thinking square' but required to output to portrait mode. Source: https://boards.4channel.org/g//thread/88281828/sdg-stable-diffusion-general

Such duplications can also occur when users repeat words in their prompt that they want the model to pay attention to while rendering, such as typing ‘hands’ several times (i.e., when the model is refusing to render hands for a given prompt).

In such a case, Stable Diffusion will sometimes intelligently interpret the user’s hint that the picture should have *some* hands on show; other times, not understanding this, it just goes to town with ‘extra’ hands (or whatever the repeated word was). A far, far more frequent side-effect of Stable Diffusion’s uncertain image cropping is that it will cut the heads off of depicted subjects, like a bad photographer. This happens most frequently in NSFW generations, where the text prompts are likely to emphasize below-the-neck anatomy, or the model is drawing on ‘anonymous’ pornographic content in the training data, which did not include faces:

This, again, appears to be either a potential mismatch in the aspect ratio most strongly associated with the concept that the user is eliciting (which may be at odds with the dimensions they have asked for); or else can occur because the prompt has focused Stable Diffusion’s attention excessively on the body area. The latter case can sometimes be remedied by adding face-based text content to the prompt.

Stable Diffusion can also ‘fix’ aspect ratio mismatches of this type by producing extra images inside the same image. In the case below, the prompt has elicited associations with primarily portrait-style ratios, but the user has set the output to the standard 512x512px. Unwilling to leave large bars of black on the side, or to inpaint the environment extensively, Stable Diffusion has instead produced three iterations of the same prompt within the same generated image:

Three similar iterations from a single prompt in one image.

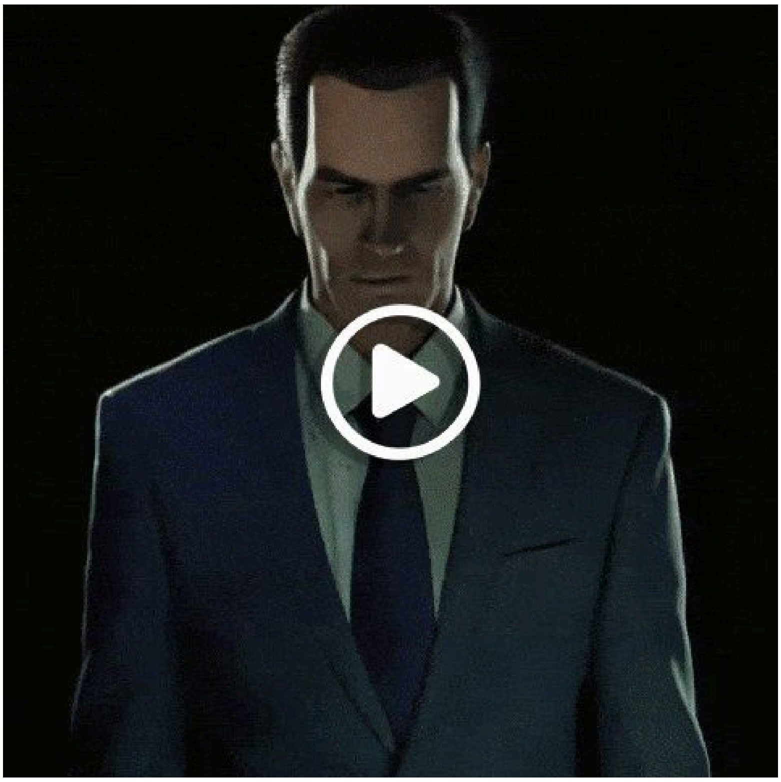
However, this is not predictable or reliable behavior:

Forward Movement

In terms of public access to unfettered facial image synthesis, this is without doubt the most significant moment since Reddit promptly shut down the first headline-grabbing autoencoder deepfake subreddits in 2017. Mirroring that momentous time, four emergent NSFW Stable Diffusion subreddits (r/UnstableDiffusion, r/stablediffusionnsfw r/PornDiffusion, and r/HentaiDiffusion) were quickly banned by Reddit shortly after the release of the model and weights.

The release has led to a storm of interest in image synthesis systems, and, perhaps, some unrealistic expectations (and fears) as to what's directly round the corner, such as open source text-to-video implementations of sophisticated synthesis frameworks like DALL-E 2, Stable Diffusion, Midjourney, and Stable Diffusion.

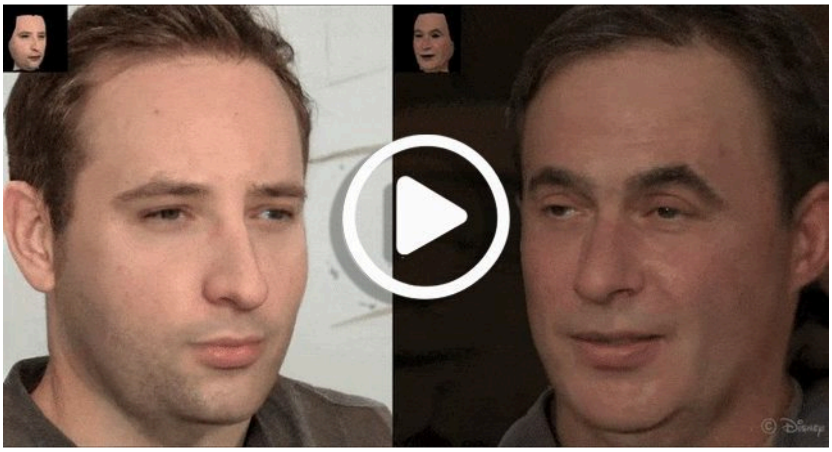
However, in this regard, the fabulous pace of evolution in facial synthesis comes up against some notable problems concerning the nascent state-of-the-art in synthesizing *human* movement.



CLICK TO VIEW ANIMATED GIF

A slightly above-average rendition of a human movement created with CogVideo (see below). Source: https://twitter.com/LucasCoolSouza/status/1554844834388647937

The current crop of available video synthesis frameworks include attempts to create video via Generative Adversarial Networks (GANs), which are usually hindered by the inherent instabilities and possibly insuperable obstacles that arise in such cases (discussed in our recent article on GANs).



CLICK TO VIEW ANIMATED GIF

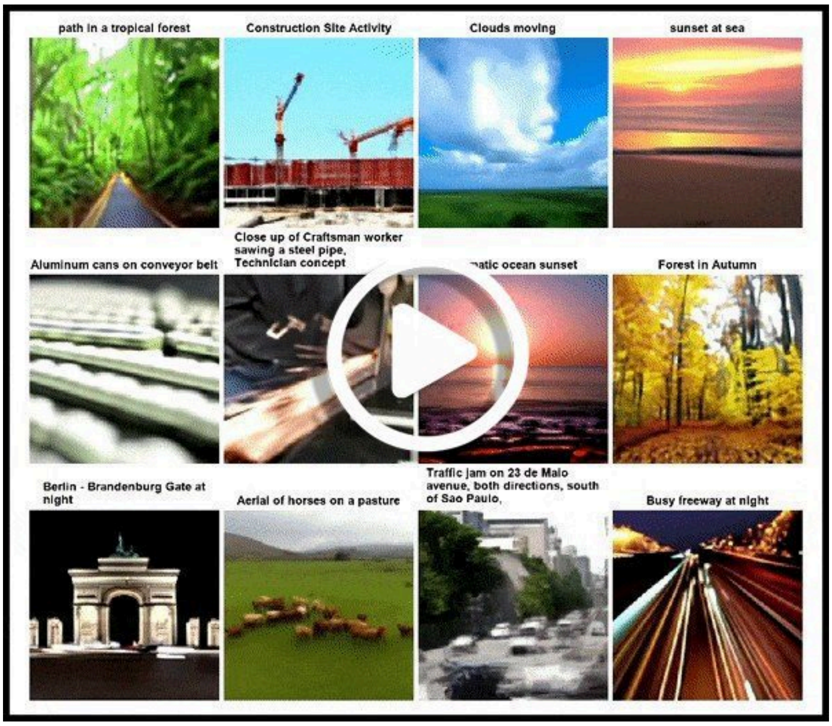
Rather stilted facial synthesis output from a heavily-modified GAN architecture in the 2019 Disney paper 'Rendering with Style'. Source: https://www.youtube.com/watch?v=TwpLqTmvqVk

Architectures such as InsetGAN produce only static images, or 1980s-style 'morph' sequences; MVCGAN, along with a host of similar approaches, can produce repetitive loops (which are in effect 'walkarounds' of otherwise still images); VGAN (from 2016, the first to leverage GANs for this kind of research) produces tiny movements at only 64x64px; TGAN (2017) glues together a temporal and an image generator, which results primarily in yet another lip-synthesis layer in one of many such limited re-composer systems; and DIGAN can produce slightly longer videos, but with a distinctly hallucinatory quality.

CLICK TO VIEW ANIMATED GIF

A 128x128px generation that attempts to animate a Tai Chi movement, from DIGAN. Source: https://sihyun.me/digan/

More recently T2V systems based around video diffusion and autoregressive transformers have offered a rationale as to why they could potentially improve in the future without excessively relying on CGI interfaces such as 3DMM models, though current results from most of them are not much better than their GAN-based predecessors, and rigorously avoid depicting humans or animals.



CLICK TO VIEW ANIMATED GIF

Examples of looped synthesized video created with diffusion-based video models. Source: https://video-diffusion.github.io/

Google Research has just released details of its Transframer architecture, which can predict the next frame in a sequence based on the current frame, albeit currently at extremely low resolution.

CLICK TO VIEW ANIMATED GIF

This example of Transframer output is enlarged 300% from the source GIF provided by DeepMind. Source: https://twitter.com/DeepMind/status/1559178172280840196

Though prompted by an initial (real) starting image, in effect, Transframer does not know what would be round the corner in the real world, since it is not performing interpolation between two A/B images, but using a sole A image as a departure for a short 'hallucinated' journey that will become more and more fictitious as time progresses.

In this sense it's essentially an urbanized version of the company's prior work Infinite Nature, which can likewise 'invent' a journey from a real starting image, albeit that Transframer offers a host of new modules and approaches.

CLICK TO VIEW ANIMATED GIF

Generating and flying over 'imagined' landscapes from a real starting image. Source: https://infinite-nature.github.io/

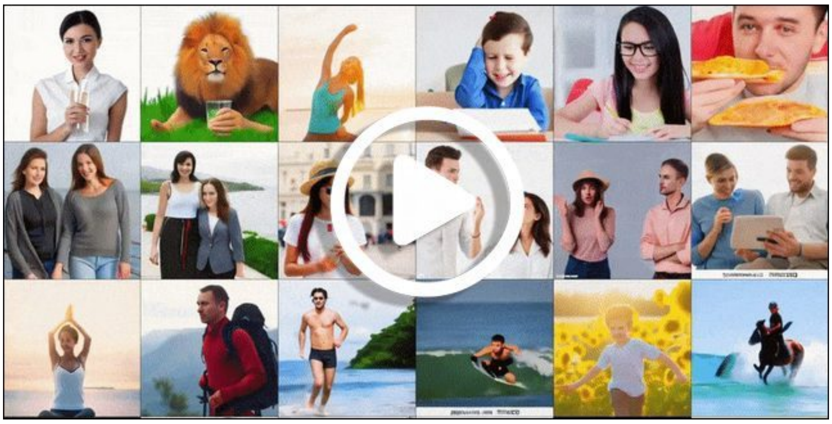
Notably absent from both projects is any footage of humans, because humans move, and act, and interact, and collapse down, and expand out to confusing configurations, and exhibit inverse kinematics, which are difficult to model even with mature CGI systems; because we're attuned to even the smallest visual inconsistency in human appearance, making it an unforgiving domain in which to perform video synthesis; and because depicting people moving convincingly requires extraordinary volumes of data, or else new breakthroughs in how we can quantize and capture the movement of people in a way that's interpretable in systems such as Stable Diffusion.

CogVideo

Because the advent of the open source Stable Diffusion release has been such a culture shock, many commenters and writers have postulated that text-to-video is the next logical step, with the possibility of a video synthesis system capable of parsing long texts directly into movies. However, the gulf between generating convincing single images and convincing temporal representations is quite severe – arguably more than several orders of magnitude.

Part of the reason for this is the scarcity of good-quality datasets; but the primary challenge is to embody 'movement concepts' as prior paths into which other elements can be injected, because this entails significant understanding about the way people move, change their expressions, and most particularly about how they organize those confusing messes of limbs into coherent ambulatory motion, and hundreds of other types of motion.

The video synthesis research sector is, of necessity, taking baby steps towards extrapolating convincing movement from an initial starting frame, and the current front-runner in the field is Tsinghua University's CogVideo, which is the subject of much interest in the SFW and NSFW Stable Diffusion communities at the moment.



CLICK TO VIEW ANIMATED GIF

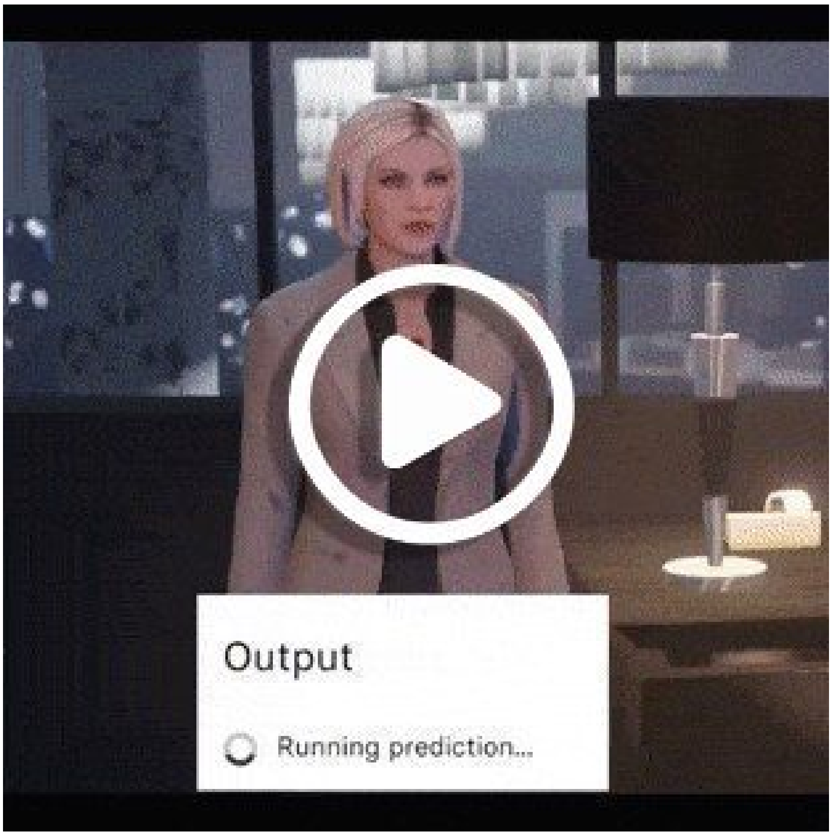
Official samples from the CogVideo project, where the pre-trained CogView 2 model has been extended into temporal space, so that some limited human movement can be extrapolated from a seed starting frame. Please see the source site for more examples at better quality. Source: https://github.com/THUDM/CogVideo

CogVideo is a 9-billion parameter open source model that extends the text-to-image architecture CogView 2 via a novel approach called multi-frame hierarchical training. It's capable of producing a few seconds of forward movement from an initial starting point, without the need for a 'target' frame.

In other words, it differs from many such interpretive systems, including DAIN and other upscaling interpretive architectures, in that it is not just creating 'tween' frames between two real (or even unreal) images, but actually creating movement based on learned priors that are trained into the model from a dataset of 5.4 million captioned videos.

But even this enormous volume of training data can only capture limited instances of the wide range of human activities that might need to be represented in a video, which currently makes it impossible to seed complex motions (i.e. 'woman stands up, stretches, walks towards camera, picks up magazine, returns to chair, sits down, reads magazine') from a single image.

Therefore the standard method currently used by CogVideo enthusiasts, including in the Stable Diffusion communities, is to use the architecture to 'fill in' keyframes generated either by CogVideo itself, or 'anchor frames' that have been generated by static image synthesis systems such as Stable Diffusion and DALL-E 2.



CLICK TO VIEW ANIMATED GIF

A typical example of multiple inputs 'tent-poling' a more complex, albeit fairly repetitive animation, via CogVideo. Source: https://twitter.com/LucasCoolSouza/status/1555288351971983365

Can You Repeat That?

It's well-known that porn is a primary driver of new technologies, and has been for a long time; so whatever your view may be on the growing corpus of NSFW material that's emerging from the new Stable Diffusion communities, NSFW 'user enthusiasm' seems as likely as it ever was to come up with new and innovative techniques for driving the technology forward into more mainstream applications.

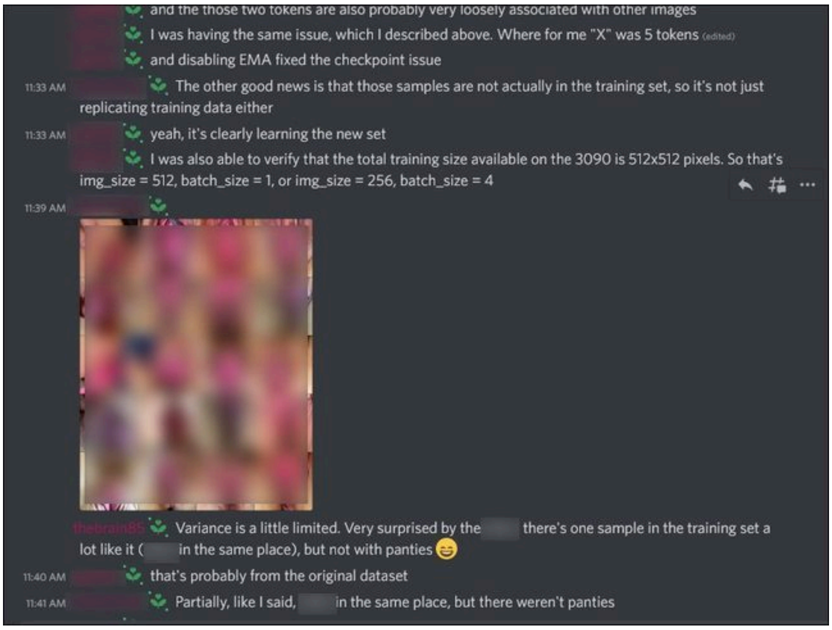
A number of Stable Diffusion/CogVideo creations have been posted at the primary CogVideo subreddit, none of which we can feature or link to here, and several of which involve same-gender sexual activity created with pornographic intent. The results are crude and semi-hallucinatory, and very, very far from photorealistic, but exhibit far more temporal cohesiveness and narrative clarity than the DeepDream craze of several years ago (which LSD-laced style has also lately been adopted by Stable Diffusion users).

To develop these complex videos, the users have effectively extended the A>B tweening scenario into an A>B>C>D etc. paradigm, where CogVideo begins a new interpretation where the previous segment finished up. This allows for a non-linear and non-repetitive video that can develop in ways that are unpredictable to the viewer.

However, the primary reason (besides sheer intensity of interest) that the earliest hyper-realistic AI videos are likely to be pornographic is that nearly all of the acts depicted in pornography are repetitious by nature, and as such already ideally suited to both A>B 'tweening' via AI, and the creation of CogVideo 'loops' that amount to the same effect.

Here, as with LAION, the possible movements that can be depicted are limited to whatever was included in the dataset on which CogVideo was trained. To the best of my knowledge, none of the 5.4 million videos on which CogVideo was trained are NSFW, or depict sexual activity, though presumably some common human activities could be adapted to this cause.

Additionally, there is nothing to stop NSFW Stable Diffusion communities from developing exclusively pornographic video datasets that would be far more efficient at generating this kind of material after training. They are already doing so for static porn datasets:



What this effectively means is that some of the earliest Stable Diffusion videos are likely to consist of such NSFW loops, and will almost certainly be short enough to be shared as animated GIFs.

'Racy' GIFS are nothing new, and neither is deepfake porn; the difference here is that such videos, including videos of celebrities, will at last be genuinely easy to produce (as many believe – wrongly – deepfake porn currently is).

Other Paths to Stable Diffusion Video Synthesis

Stable Diffusion's interpretive powers can be leveraged in other ways to create video. One such approach is 'deepfake puppetry'.

CONTINUED ON NEXT PAGE

CLICK TO VIEW ANIMATED GIF

Far from real-time, but here Stable Diffusion is run over each individual extracted frame in a real-world face sequence. Source: https://old.reddit.com/r/StableDiffusion/comments/wyeoqq/turning_img2img_into_vid2vid/

In the example above, the user has fed each extracted frame into Stable Diffusion, together with a prompt that identifies the celebrity that should be superimposed. The same seed value has been retained for each generation, and the code modified so that the same noise tensor is passed to the *stochastic_encode()* parameter every time – a trick that could easily be integrated into Stable Diffusion.

In contrast to real-time deepfake streaming systems such as DeepFaceLive, which only replace the internal facial features in videos which are otherwise genuine, puppetry allows real-world movement to control *completely synthetic* video output, a paradigm that Neural Radiance Fields has implemented as well, notably in RigNeRF.

It remains to be seen to what extent code alterations could produce temporally and semantically consistent transformations, without that characteristic hallucinatory shimmer that currently hallmarks AI-generated synthetic content (unless it employs some kind of CGI routine to stabilize the output and dial down the level of *ad hoc* frame-based interpretation).

If these problems are solved, the path could eventually open up for authentic and 100% synthesized deepfake puppetry via diffusion-based image generators. For longer video clips, the process would need to become quite industrialized, but would represent a notable usability improvement over current explorations of the GAN latent space, and could pave the way for full-body deepfakes with photorealistic faces that correspond to people in the real world.

There are also a number of third-party applications and APIs that can add repetitive or simple movement to any input image – mostly circular, repetitive movement. While there’s nothing to stop Stable Diffusion users feeding their output (SFW or otherwise) into such systems, the results are limited, if momentarily engaging.
For instance, MyHeritage, an avid adopter of image synthesis frameworks, can bring static Stable Diffusion renderings to life via the DeepNostalgia architecture:

CLICK TO VIEW ANIMATED GIF

A Stable Diffusion-generated image of Tom Cruise, animated by DeepNostalgia. Source:myheritage.com

There are also downloadable applications such as Pixbim Animate Photos AI, mobile apps such as TokkingHeads, and a notable number of GitHub repositories and Colabs that can also take static photos for a brief spin into the temporal realm, without really offering any flexibility or utility in a genuine text-to-video pipeline.

The Future of Stable Diffusion

To answer the question posed by the title of this feature, yes – Stable Diffusion is *already* producing video content, mainly via third-party architectures – but it's pretty terrible; and it's going to take more time than the general public probably realizes to create 'pure' video synthesis systems based around diffusion architectures, which by default have *zero* capabilities for temporal analysis and reproduction.

CLICK TO VIEW ANIMATED GIF

'Anime girl says goodbye, she is sad because she needs to leave' – Stable Diffusion output manipulated and extended by CogVideo. Source: https://twitter.com/FabianMosele/status/1552412134251892737

The difficulty of maintaining a consistent appearance and style across multiple image generations, even by hacking the code, means that effective deepfake-style puppetry systems for frameworks such as Stable Diffusion will be a challenge to implement. Nonetheless, they may currently represent the best hope for hyper-realistic diffusion-based video in the near-term.
If history is anything to go by (particularly the history of GANs), we're in for 18 months of new research papers claiming new incremental victories in making diffusion-based image synthesis systems amenable to video generation, simply by exploring and manipulating their latent space.

After this, there'll be another 18 months of further research papers that have abandoned that particular hope, and which instead offer frameworks that use parametric, CGI-style interstitial 'bridges' capable of bringing some kind of temporal consistency to diffusion-based video.

In the meantime, a new wave of SFW and NSFW implementations and APIs of Stable Diffusion are likely to emerge in the next six months, many of which will seek to differentiate themselves by offering 'fine-tuned' models that have added domains (certainly not excluding porn, in some cases) that are otherwise under-represented or absent in the official sources used by Stability.ai – or which offer facile training or fine-tuning, so that users can develop their own 'specialized' models – for a price (likely a subscription-based price).

The next thing to watch for with Stable Diffusion is not necessarily video, but new and better implementations of textual inversion than are currently available, which could obviate the need for further training of the core model, and which could represent a quantum leap forward in image synthesis – if hardware requirements, data issues, and other obstacles can be overcome.