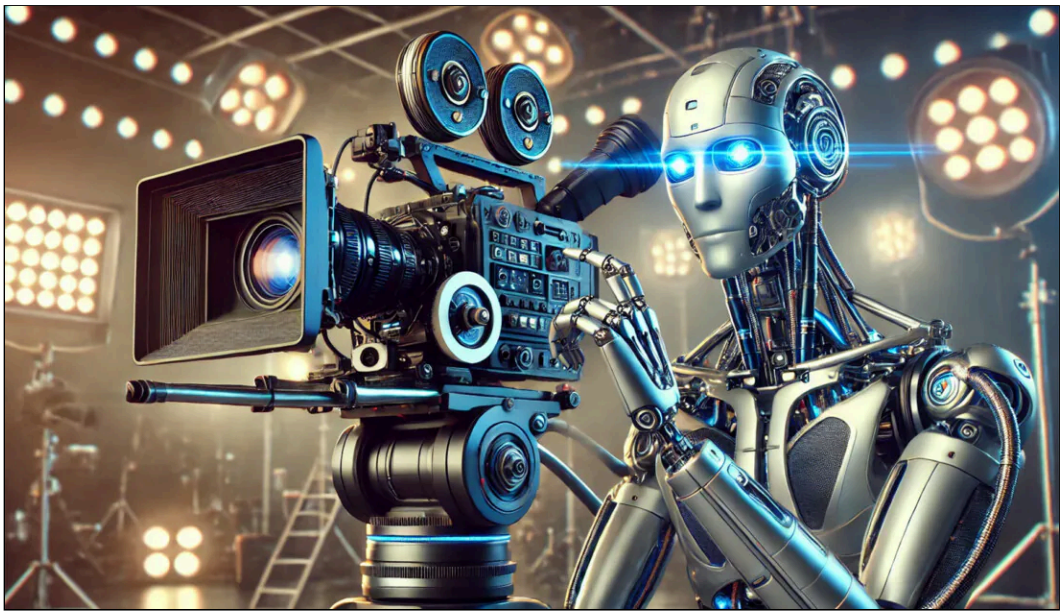
Why Can’t Generative Video Systems Make Complete Movies?

Originally published at https://www.unite.ai/why-cant-generative-video-systems-make-complete-movies/



The advent and progress of generative AI video has prompted many casual observers to predict that machine learning will prove the death of the movie industry as we know it – instead, single creators will be able to create Hollywood-style blockbusters at home, either on local or cloud-based GPU systems.

Is this possible? Even if it is possible, is it *imminent*, as so many believe?

That individuals will eventually be able to create movies, in the form that we know them, with consistent characters, narrative continuity and total photorealism, is quite possible – and perhaps even inevitable.

However there are several truly fundamental reasons why this is not likely to occur with video systems based on Latent Diffusion Models.

This last fact is important because, at the moment, that category includes *every* popular text-to-video (T2) and image-to-video (I2V) system available, including Minimax, Kling, Sora, Imagen, Luma, Amazon Video Generator, Runway ML, Kaiber (and, as far as we can discern, Adobe Firefly’s pending video functionality); among many others.

Here, we are considering the prospect of true *auteur* full-length gen-AI productions, created by individuals, with consistent characters, cinematography, and visual effects at least on a par with the current state of the art in Hollywood.

Let’s take a look at some of the biggest practical roadblocks to the challenges involved.

1: You Can’t Get an Accurate Follow-on Shot

Narrative inconsistency is the largest of these roadblocks. The fact is that no currently-available video generation system can make a truly accurate ‘follow on’ shot*.

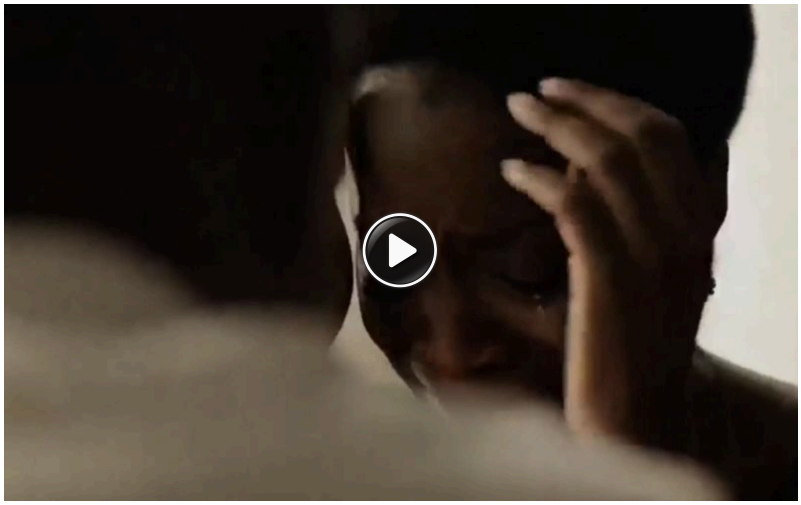
This is because the denoising diffusion model at the heart of these systems relies on random noise, and this core principle is not amenable to reinterpreting exactly the same content twice (i.e., from different angles, or by developing the previous shot into a follow-on shot which maintains consistency with the previous shot).

Where text prompts are used, alone or together with uploaded ‘seed’ images (multimodal input), the tokens derived from the prompt will elicit semantically-appropriate content from the trained latent space of the model.

However, further hindered by the ‘random noise’ factor, it will *never do it the same way twice*.

This means that the identities of people in the video will tend to shift, and objects and environments will not match the initial shot.

This is why viral clips depicting extraordinary visuals and Hollywood-level output tend to be either single shots, or a 'showcase montage' of the system's capabilities, where each shot features different characters and environments.



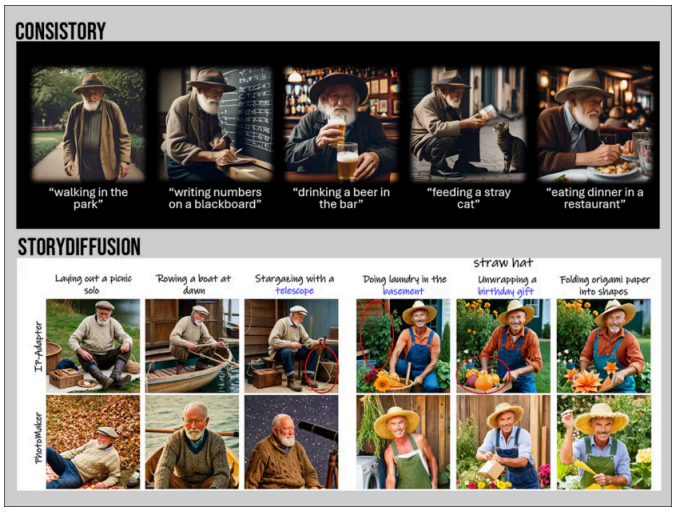
CLICK TO PLAY VIDEO ON SITE

Excerpts from a generative AI montage from Marco van Hylckama Vlieg – source: https://www.linkedin.com/posts/marcovhv_thanks-to-generative-ai-we-are-all-filmmakers-activity-7240024800906076160-nEXZ/

The implication in these collections of *ad hoc* video generations (which may be disingenuous in the case of commercial systems) is that the underlying system *can* create contiguous and consistent narratives.

The analogy being exploited here is a movie trailer, which features only a minute or two of footage from the film, but gives the audience reason to believe that the entire film exists.

The only systems which currently offer narrative consistency in a diffusion model are those that produce still images. These include NVIDIA's ConsiStory, and diverse projects in the scientific literature, such as TheaterGen, DreamStory, and StoryDiffusion.



Two examples of 'static' narrative continuity, from recent models:: Sources: https://research.nvidia.com/labs/par/consistory/ and https://arxiv.org/pdf/2405.01434

In theory, one could use a better version of such systems (none of the above are truly consistent) to create a series of image-to-video shots, which could be strung together into a sequence.

At the current state of the art, this approach does not produce plausible follow-on shots; and, in any case, we have already departed from the *auteur* dream by adding a layer of complexity.

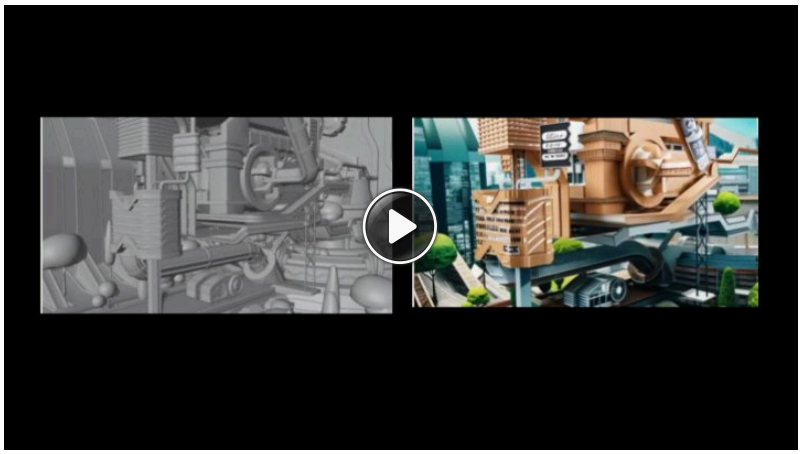
We can, additionally, use Low Rank Adaptation (LoRA) models, specifically trained on characters, things or environments, to maintain better consistency across shots.

However, if a character wishes to appear in a new costume, an entirely new LoRA will usually need to be trained that embodies the character dressed in that fashion (although sub-concepts such as 'red dress' can be trained into individual LoRAs, together with apposite images, they are not always easy to work with).

This adds considerable complexity, even to an opening scene in a movie, where a person gets out of bed, puts on a dressing gown, yawns, looks out the bedroom window, and goes to the bathroom to brush their teeth.

Such a scene, containing roughly 4-8 shots, can be filmed in one morning by conventional film-making procedures; at the current state of the art in generative AI, it potentially represents weeks of work, multiple trained LoRAs (or other adjunct systems), and a considerable amount of post-processing

Alternatively, video-to-video can be used, where mundane or CGI footage is transformed through text-prompts into alternative interpretations. Runway offers such a system, for instance.



CLICK TO PLAY VIDEO ON SITE

CGI (left) from Blender, interpreted in a text-aided Runway video-to-video experiment by Mathieu Visnjevec – Source: https://www.linkedin.com/feed/update/urn:li:activity:7240525965309726721/

There are two problems here: you are already having to create the core footage, so you're already making the movie *twice*, even if you're using a synthetic system such as UnReal's MetaHuman.

If you create CGI models (as in the clip above) and use these in a video-to-image transformation, their consistency across shots cannot be relied upon.

This is because video diffusion models do not see the 'big picture' – rather, they create a new frame based on previous frame/s, and, in some cases, consider a nearby future frame; but, to compare the process to a chess game, they cannot think 'ten moves ahead', and cannot remember ten moves behind.

Secondly, a diffusion model will still struggle to maintain a consistent appearance across the shots, even if you include multiple LoRAs for character, environment, and lighting style, for reasons mentioned at the start of this section.

2: You Can't Edit a Shot Easily

If you depict a character walking down a street using old-school CGI methods, and you decide that you want to change some aspect of the shot, you can adjust the model and render it again.

If it's a real-life shoot, you just reset and shoot it again, with the apposite changes.

However, if you produce a gen-AI video shot that you love, but want to change *one aspect* of it, you can only achieve this by painstaking post-production methods developed over the last 30-40 years: CGI, rotoscoping, modeling and matting – all labor-intensive and expensive, time-consuming procedures.

The way that diffusion models work, simply changing one aspect of a text-prompt (even in a multimodal prompt, where you provide a complete source seed image) will change *multiple aspects* of the generated output, leading to a game of prompting 'whack-a-mole'.

3: You Can't Rely on the Laws of Physics

Traditional CGI methods offer a variety of algorithmic physics-based models that can simulate things such as fluid dynamics, gaseous movement, inverse kinematics (the accurate modeling of human movement), cloth dynamics, explosions, and diverse other real-world phenomena.

However, diffusion-based methods, as we have seen, have short memories, and also a limited range of motion priors (examples of such actions, included in the training dataset) to draw on.

In an earlier version of OpenAI's landing page for the acclaimed Sora generative system, the company conceded that Sora has limitations in this regard (though this text has since been removed):

> *'[Sora] may struggle to simulate the physics of a complex scene, and may not comprehend specific instances of cause and effect (for example: a cookie might not show a mark after a character bites it).*
>
> *'The model may also confuse spatial details included in a prompt, such as discerning left from right, or struggle with precise descriptions of events that unfold over time, like specific camera trajectories.'*

The practical use of various API-based generative video systems reveals similar limitations in depicting accurate physics. However, certain common physical phenomena, like explosions, appear to be better represented in their training datasets.

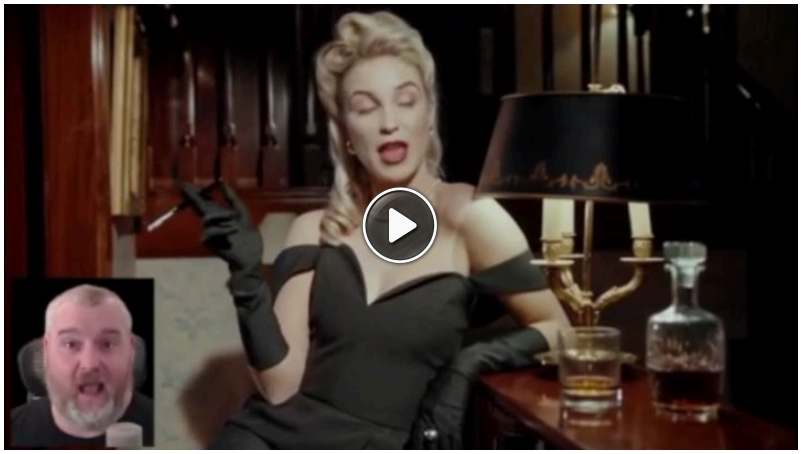
Some motion prior embeddings, either trained into the generative model or fed in from a source video, take a while to complete (such as a person performing a complex and non-repetitive dance sequence in an elaborate costume) and, once again, the diffusion model's myopic window of attention is likely to transform the content (facial ID, costume details, etc.) by the time the motion has played out. However, LoRAs can mitigate this, to an extent.

Fixing It in Post

There are other shortcomings to pure 'single user' AI video generation, such as the difficulty they have in depicting rapid movements, and the general and far more pressing problem of obtaining temporal consistency in output video.

Additionally, creating specific facial performances is pretty much a matter of luck in generative video, as is lip-sync for dialogue.

In both cases, the use of ancillary systems such as LivePortrait and AnimateDiff is becoming very popular in the VFX community, since this allows the transposition of at least broad facial expression and lip-sync to existing generated output.

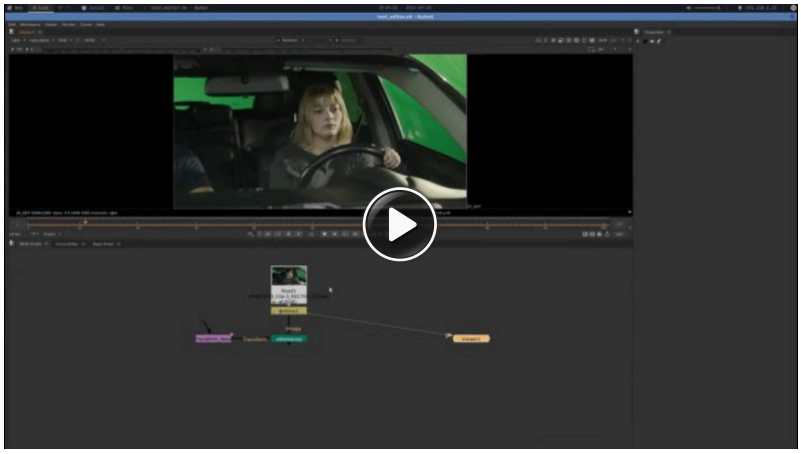


CLICK TO PLAY VIDEO ON SITE

An example of expression transfer (driving video in lower left) being imposed on a target video with LivePortrait. The video is from Generative Z TunisiaGenerative. See the full-length version in better quality at https://www.linkedin.com/posts/genz-tunisia_digitalcreation-liveportrait-aianimation-activity-7240776811737972736-uxiB/?

Further, a myriad of complex solutions, incorporating tools such as the Stable Diffusion GUI ComfyUI and the professional compositing and manipulation application Nuke, as well as latent space manipulation, allow AI VFX practitioners to gain greater control over facial expression and disposition.

Though he describes the process of facial animation in ComfyUI as 'torture', VFX professional Francisco Contreras has developed such a procedure, which allows the imposition of lip phonemes and other aspects of facial/head depiction"



CLICK TO PLAY VIDEO ON SITE

Stable Diffusion, helped by a Nuke-powered ComfyUI workflow, allowed VFX pro Francisco Contreras to gain unusual control over facial aspects. For the full video, at better resolution, go to https://www.linkedin.com/feed/update/urn:li:activity:7243056650012495872/

Conclusion

None of this is promising for the prospect of a single user generating coherent and photorealistic blockbuster-style full-length movies, with realistic dialogue, lip-sync, performances, environments and continuity.

Furthermore, the obstacles described here, at least in relation to diffusion-based generative video models, are not necessarily solvable 'any minute' now, despite forum comments and media attention that make this case. The constraints described seem to be intrinsic to the architecture.

In AI synthesis research, as in all scientific research, brilliant ideas periodically dazzle us with their potential, only for further research to unearth their fundamental limitations.

In the generative/synthesis space, this has already happened with Generative Adversarial Networks (GANs) and Neural Radiance Fields (NeRF), both of which ultimately proved very difficult to instrumentalize into performant commercial systems, despite years of academic research towards that goal. These technologies now show up most frequently as adjunct components in alternative architectures.

Much as movie studios may hope that training on legitimately-licensed movie catalogs could eliminate VFX artists, AI is actually *adding* roles to the workforce at the present time.

Whether diffusion-based video systems can really be transformed into narratively-consistent and photorealistic movie generators, or whether the whole business is just another alchemic pursuit, should become apparent over the next 12 months.

It may be that we need an entirely new approach; or it may be that Gaussian Splatting (GSplat), which was developed in the early 1990s and has recently taken off in the image synthesis space, represents a potential alternative to diffusion-based video generation.

Since GSplat took 34 years to come to the fore, it's possible too that older contenders such as NeRF and GANs – and even latent diffusion models – are yet to have their day.

** Though Kaiber's AI Storyboard feature offers this kind of functionality, the results I have seen are not production quality.*

Martin Anderson is the former head of scientific research content at metaphysic.ai
First published Monday, September 23, 2024

Writer on machine learning, domain specialist in human image synthesis. Former head of research content at Metaphysic.ai, until its dissolution into DNEG's Brahma.ai.
Portfolio site: martinanderson.ai
Contact: martin@martinanderson.ai

Discover More