Toward Perpetual Full-Body Deepfake Video Generation

Originally published at https://www.unite.ai/toward-perpetual-full-body-deepfake-video-generation/

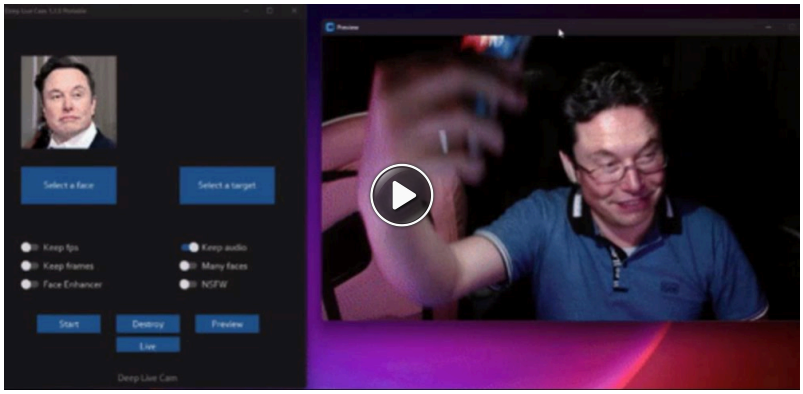
In a field where patience is required, a new system pushes us a little nearer to live-streamed human simulation without frustrating rendering rounds.

The state-of-the-art in full-length human simulation has come a long way since the possibility of full-body deepfakes first appeared in 2022. By now we have become habituated to the formidable and often-controversial ability of open source systems such as Wan-Animate, and closed-source systems such as Grok, to convert single or multiple images into a consistent, and often transformative video performance:

CLICK TO PLAY VIDEO ON SITE

***Click to play if necessary**. Examples of person replacement with WanAnimate-2.2. Please refer to source for better resolution.* Source

Additionally, the older autoencoder-based live facial deepfake framework DeepFaceLive has since been surpassed by more sophisticated frameworks such as Deep-Live-Cam, which leverages an orchestration of LivePortrait iterations, as well as legacy GAN and InsightFace modules, to create an effective real-time successor to the (now-abandoned) DFLive project:



CLICK TO PLAY VIDEO ON SITE

Click to play: *Elon Musk deepfaked in a live video session via Deep-Live-Cam. Please refer to source for better resolution.* Source

However, while frameworks such as Deep-Live-Cam can run forever and fake forever, and though they are better at generating tough facial angles than they used to be, the results are nonetheless constrained in terms of resolution and capability: you can obtain a particular face/identity, and it can do a lot of things, such as convincing facial expressions and lip-sync – but it's essentially a one-trick pony.

BRB…

Most of the current crop of AI human impersonation systems are likewise constrained and/or 'specialized'. One particular, recurrent constraint is that the best human simulation/impersonation systems are *offline* – which is to say, they are too resource-intensive to operate in real time, and instead need to go away, calculate the solution, and present the outcome to the user later.

Obviously such systems are unsuitable for live AI transmutation, such as transforming an entire range of body motion, like dance, in a live stream.

An example of this in recent years is the open-weights Wan2.2. Animate, which proved a hit with the hobbyist community and pro resellers – but once again, it's a 'generate and wait' scenario:

CLICK TO PLAY VIDEO ON SITE

Click to play. *From 2025, examples of Wan2.2-Animate's impressive capabilities – if you can have a little patience. Please refer to source for better resolution.* Source

So as it stands you can have it great, have it versatile, or have it now – pick two.

LiveAnimate

Into this Mexican stand-off comes a new offering from China, which effectively transforms Wan2.2 Animate into a live-driven animation framework operating, currently, at a respectable near-20fps, with impressive results across a range of scenarios:

CLICK TO PLAY VIDEO ON SITE

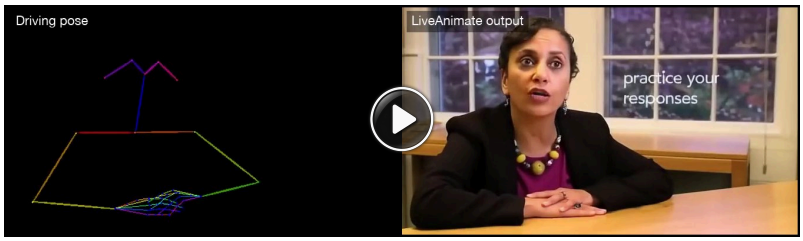
Click to play: *From the project site, LiveAnimate transfers driving poses across stage-dance, outdoor full-body and close-up portrait scenarios, reproducing whole-body, hand and facial motion. Please refer to source for better resolution.* Source

The new work extends the original system into a live-capable version by changing *how* video is generated: instead of processing past and future frames together, LiveAnimate generates each new segment as the action unfolds, using only a few steps, while retaining *selected earlier poses to keep the subject's appearance consistent over long sessions*.

Effectively, the system keeps earlier frames as a method of persistent memory to draw on as the video develops, so that the identity remains consistent – not entirely dissimilar to the way old-school CGI systems refer to texture maps.

Though a LoRA is involved at the processing stage, LiveAnimate is not just another LoRA system – its streaming generation, pose-memory/cache system, three-step inference and GPU optimizations represent additional mechanisms, on top of the diverse other technologies (some surprisingly old) in its repertoire.

The new system can cope not only with full-body driving scenarios such as those in the above examples, but also with upper-body movements, such as in interview scenarios:



CLICK TO PLAY VIDEO ON SITE

Click to play. *'Subtle head and hand motion over a static office scene'.*

To be fair regarding its limitations, the new paper concedes that two of the rival frameworks tested against LiveAnimate were able to preserve identity slightly better in certain specific circumstances; that said, not one of the contenders is an online rather than offline system, and the authors of the new work are clearly pushing the envelope in a plausible and optimistic direction.

Additionally, it may be worth noting that inference requires two H100 NVIDIA GPUs, for a total of 160GB of VRAM*.

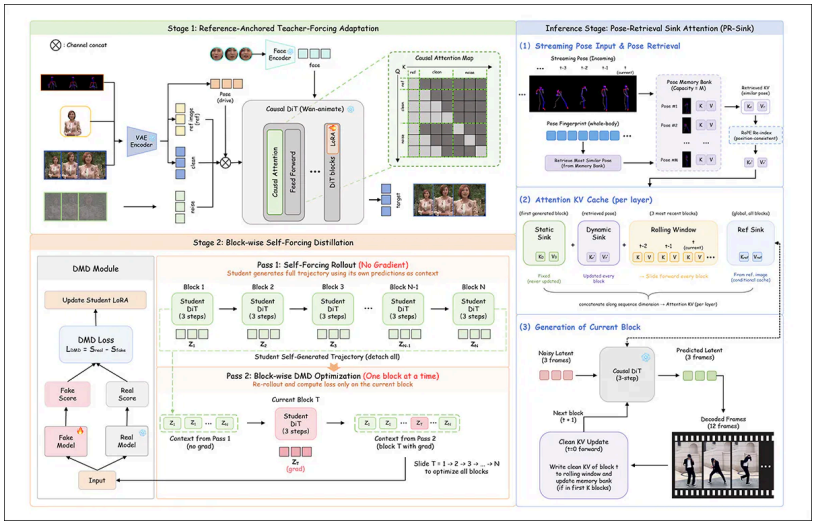
The paper states:

> *'LiveAnimate maintains nearly constant perceptual quality and identity from the first 30 seconds to the final [minute], while prior systems degrade substantially or require hours of offline computation for the same rollout.*
>
> *'These results establish a new operating point in quality, latency, and duration for interactive full-body animation.'*

The new paper is titled *LiveAnimate: Stable Long-Form Streaming Human Animation in Real-Time*, and comes from nine authors across The Chinese University of Hong Kong, Qwen Applications Business Group of Alibaba, and Liblib AI. A project site, replete with the videos also featured in this article, is also available, while code is 'coming soon', and weights…who knows?

Method

LiveAnimate constitutes a two-stage training process designed to make Wan2.2-Animate generate continuously and quickly, followed by a memory system that retrieves useful earlier poses as each new block of video is produced:



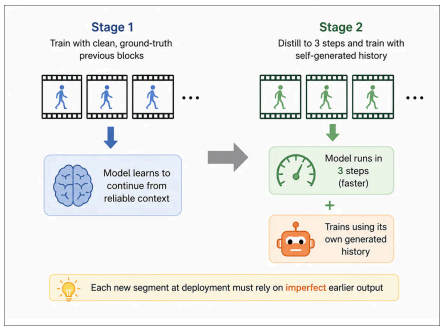
An overview of the LiveAnimate pipeline, showing its two-stage training process on the left and live generation on the right, where incoming poses are matched ag combined with recent frames to generate each new block of video in three steps. Source

The original Wan2.2-Animate is designed to consider an entire sequence rather than generate an indefinitely extending video. Therefore to adapt it, the authors divided training videos into successive blocks, with each generated using earlier blocks as context/ground truth. This initially teaches the model to continue from reliable previous material, before having to cope with errors accumulated from its own output.

Forget Me Not

Throughout this process, the original reference image is kept permanently available through a 'Ref Sink', providing a fixed reminder of the person's identity and appearance. Once a block has been completed, a 'Clean KV Update' converts information obtained from it into historical context for subsequent blocks (though this does not repair existing errors).

This first training stage still requires 50 denoising steps per block, making live operation impractical. The second stage therefore distills the process down to three steps, *while exposing the model to its own generated history*, since deployment requires every new segment to depend on imperfect earlier output:



The two training stages† used to prepare LiveAnimate for deployment, first learning from clean previous video blocks – then reducing generation to three steps, while training on the model's own imperfect output.

Training against these self-generated sequences presents another problem, since retaining an entire video's computational history would be prohibitively expensive. Instead, one complete practice run is made, then revisited, one block at a time, for training. Each block can therefore receive an update *without keeping the entire sequence computationally 'live'*.

Combined with LoRA adaptation, this allows the 14-billion-parameter model to be distilled on a single node containing eight 80GB H100 GPUs (noting that this cluster is for training, not runtime inference).

Don't Stop Now

For indefinite generation, LiveAnimate must also decide what is *worth* remembering. Keeping *everything* would make processing requirements balloon out-of-control; but retaining only recent frames might also discard useful earlier views.

Therefore *Pose-Retrieval Sink Attention* (PR-Sink) addresses this, by keeping the first generated block as a permanent 'Static Sink' – a relevant earlier pose, operating as a replaceable 'Dynamic Sink', and a rolling window containing the current and two preceding blocks. The reference image remains separately available through the Ref Sink, while fixed storage sizes prevent costs from growing as the video length increases.

Block Wars

To choose older poses for this rather limited memory, ViTPose reduces body and hand positions across three frames into a compact representation. The memory-bank holds five such representative poses, and is populated during the first 20 blocks, favoring varied poses over near-duplicates.

During generation, the incoming pose is compared with these representatives and the closest match retrieved, providing earlier evidence of how the person looked in a similar position. However, the immediately-preceding block is *excluded*, because it already exists in the rolling window, leaving the Dynamic Sink free to retrieve information from further back.

Finally, the runtime itself is optimized for speed by distributing attention processing across two H100 GPUs, overlapping communication with computation, and reusing cached information wherever possible.

Data and Tests

LiveAnimate was trained on 40,000 talking videos from *AVSpeech*, and 20,000 human-motion videos from *TikTok dataset* and *HumanVid*, with training and inference supporting resolutions of 480×480; 384×672; and 672×384 pixels.

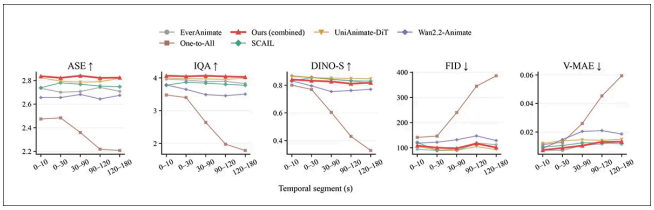
Training began from the Wan2.2-Animate-14B base checkpoint, using LoRA with rank 128, and proceeded on eight NVIDIA H100 GPUs for 10,000 steps in the first stage, and 20,000 in the second.

Video was generated in blocks of three latent frames (the model's compressed internal representation), corresponding to 12 RGB frames, while inference was performed on the two H100s.

The evaluation compared LiveAnimate with existing pose-driven human-animation methods across both short and long sequences, using reference images and driving poses under consistent settings. Long-form tests extended generation to three minutes to examine whether quality and identity remained stable over time.

Frameworks tested were EverAnimate; One-to-All; SCAIL††; UniAnimate-DiT; and Wan2.2-Animate.

Metrics used covered visual quality, identity consistency, distributional quality and temporal representation error. *Aesthetic Score* (ASE) and *no-reference Image-Quality Assessment* (IQA) measured frame-level visual quality; DINO, similarity (DINO-S), and appearance consistency with the reference identity; Fréchet Inception Distance (FID), distributional quality; and VideoMAE feature distance (V-MAE), how well the generated video preserved motion over time:



Test results pertaining to quality and identity across three minutes of continuous generation, with LiveAnimate remaining comparatively stable across all five metrics as the sequence progresses. Several competing methods show greater degradation over time. Higher readings are better for IQA, ASE, and DINO-S, with lower readings better for FID and V-MAE.

As shown in the initial results graph above, LiveAnimate achieved the highest initial ASE of 2.823, and IQA of 4.047, over the first 30 seconds. The authors contend that this indicates three-step distillation preserves perceptual quality, despite the reduced inference budget.

More significantly, LiveAnimate showed little deterioration during three minutes of continuous generation, with its visual-quality and identity-consistency scores remaining almost unchanged.

Conversely, One-to-All deteriorated substantially over time, with lower visual quality and identity consistency and higher distributional error; and base Wan2.2-Animate also showed some loss of identity consistency.

The authors state:

> *'These trends support our central claim that explicitly managing long-range context is important for streaming animation.'*

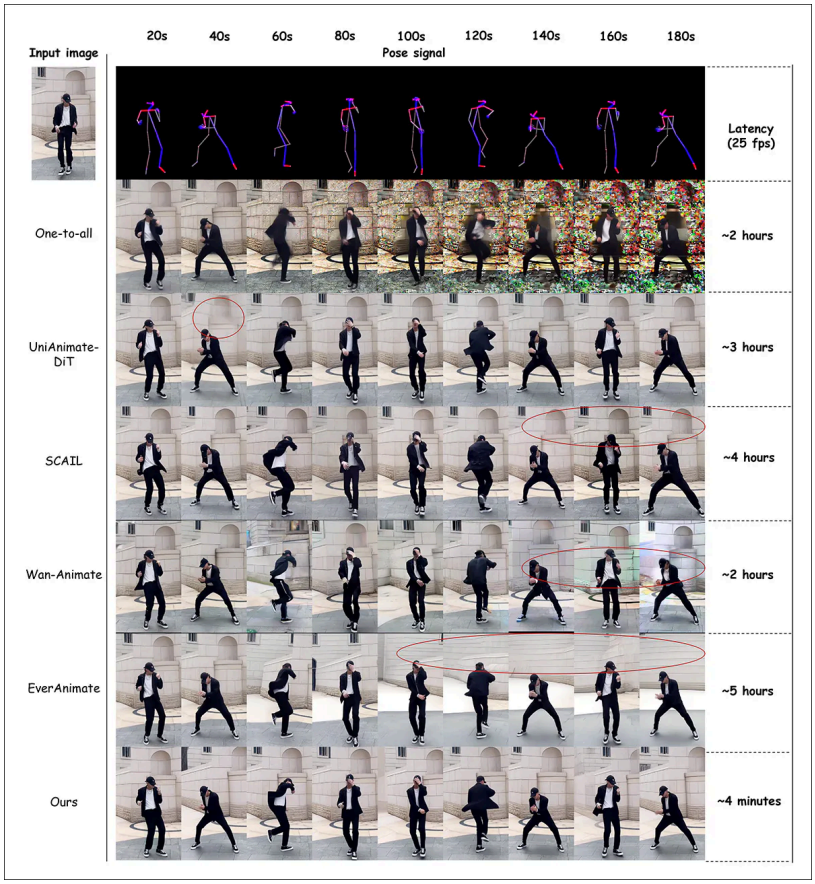
LiveAnimate's scaling efficiency was also tested across GPUs. As shown below, 12.41 FPS was achieved with one H100; 19.63 FPS with two; and 22.13 FPS with four, with the gains increasingly limited by communication overhead. Two H100s were therefore selected as the preferred configuration:

GPUs	Latency (ms)	FPS	Speedup	Efficiency
1	966.96	12.41	1.00×	100.0%
2	611.28	19.63	1.58×	79.1%
4	542.25	22.13	1.78×	44.6%

Scaling efficiency across one, two and four H100 GPUs at 480×480 resolution, showing latency, frame rate, speedup and efficiency. Two GPUs achieved 19.63 FPS, while further scaling to four produced only a modest increase to 22.13 FPS.

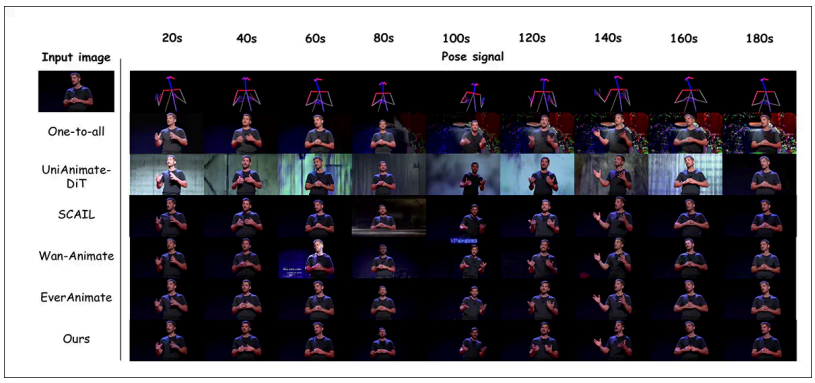
At 19.63 FPS, each 12-frame block was generated in 0.611 seconds. By comparison, approximately 2–5 hours were required by the competing systems to generate the same three-minute sequence.

Qualitative tests brought up similar differences: in the full-body test shown below, severe deterioration was observed in One-to-All; flickering was produced by UniAnimate-DiT and SCAIL; and later color or background drift became evident with Wan-Animate and EverAnimate.



Test results comparing full-body animation over three minutes. Frames were sampled every 20 seconds, with red annotations highlighting long-term deterioration i baselines required approximately 2–5 hours to generate the sequence, compared with approximately four minutes for LiveAnimate. Please refer to source paper fo

LiveAnimate maintained the subject's appearance and surrounding scene throughout the three-minute sequence. In the less demanding upper-body test shown below, comparable long-term stability was achieved by EverAnimate and LiveAnimate (though real-time generation was provided only by LiveAnimate):



Test results comparing upper-body animation over three minutes. Frames sampled every 20 seconds show LiveAnimate maintaining the subject's identity, clothing consistently than the competing methods.

The authors conclude:

> *'On the three-minute benchmark, LiveAnimate sustains nearly constant perceptual quality and identity at 19.63 FPS on two H100 GPUs, with memory and latency independent of stream duration, whereas offline baselines degrade visibly or require hours of computation.*
>
> *'These results bring billion-scale video diffusion models within reach of interactive applications such as live streaming, telepresence, and virtual avatars.'*

Conclusion

It is encouraging to see a driven-generation framework take a novel approach to the persistent problems of memory and identity that plague generative video. There are already many generative frameworks that can reference fixed imagery supplied by the user, without the need to 'paste' the reference image into essential image-to-video content.

However, this only solves some of the problems of generating a single video-clip, such as retaining the identity (including clothing and hairstyle) of a person who reenters frame, or has become momentarily obscured, and then is seen once more. To date, only the heavy-duty and annoyingly temporary LoRA approach has made any headway with these issues, albeit limited and provisional.

For video, and, in fact, for the entire current AI revolution, the development of *effective* persistent memory is a critical, existential issue – one that will likely define the difference between the advent of AGI, or a third AI winter.

** Is this a 'gotcha' anymore? The current thinking is that smaller specialized models may eventually run locally, rent-free, while higher-end needs are served by a competitive and hopefully balkanized GPU rental market. This assumes the frontier companies fail to EEE the competition, and that substantial GPU compute remains available regardless. On that basis, I no longer consider egregious GPU requirements a demerit, but hope that access becomes increasingly democratic, and that later optimizations reduce the initial resource demands.*

† *AI-generated and checked by myself.*

†† *For long-video testing, SCAIL and UniAnimate-DiT were extended with the same training-free sliding-window procedure because neither natively supports long-form generation. EverAnimate and One-to-All were evaluated using their own long-video generation methods.*

First published Thursday, August 13, 2026

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