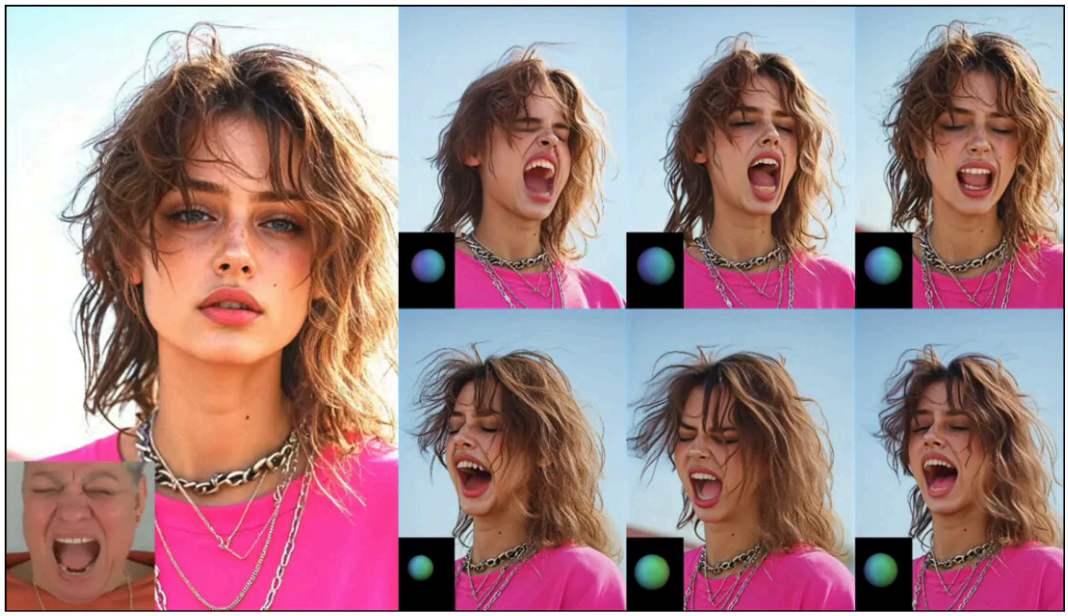
A Notable Advance in Human-Driven AI Video

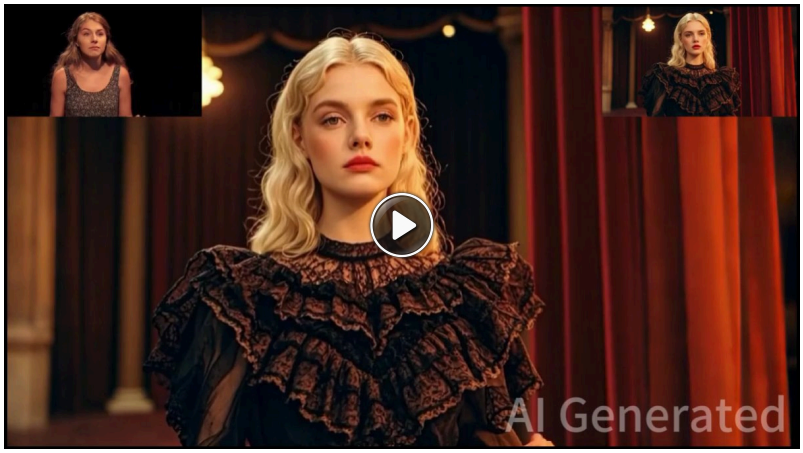
Originally published at https://www.unite.ai/a-notable-advance-in-human-driven-ai-video/



Note: *The project page for this work includes 33 autoplaying high-res videos totaling half a gigabyte, which destabilized my system on load. For this reason, I won't link to it directly. Readers can find the URL in the paper's abstract or PDF if they choose.*

One of the primary objectives in current video synthesis research is generating a complete AI-driven video performance from a single image. This week a new paper from Bytedance Intelligent Creation outlined what may be the most comprehensive system of this kind so far, capable of producing full- and semi-body animations that combine expressive facial detail with accurate large-scale motion, while also achieving improved identity consistency – an area where even leading commercial systems often fall short.

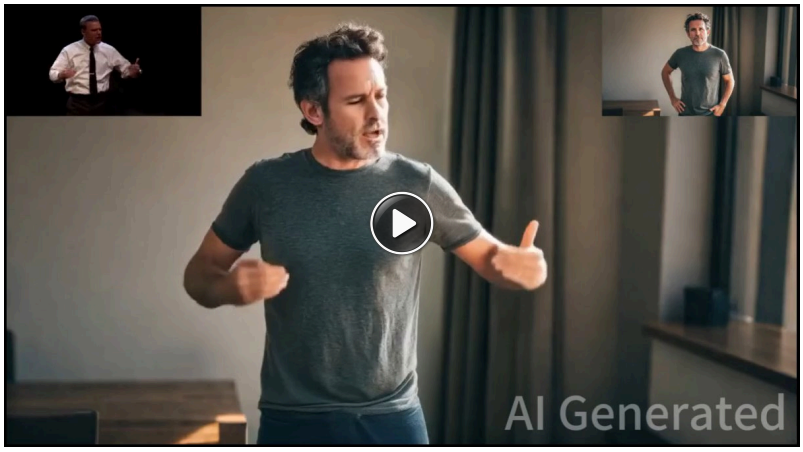
In the example below, we see a performance driven by an actor (top left) and derived from a single image (top right), that provides a remarkably flexible and dexterous rendering, with none of the usual issues around creating large movements or 'guessing' about occluded areas (i.e., parts of clothing and facial angles that must be inferred or invented because they are not visible in the sole source photo):



CLICK TO PLAY VIDEO ON SITE

AUDIO CONTENT. Click to play*. A performance is born from two sources, including lip-sync, which is normally the preserve of dedicated ancillary systems. This is a reduced version from the source site (see note at beginning of article – applies to all other embedded videos here).*

Though we can see some residual challenges regarding persistence of identity as each clip proceeds, this is the first system I have seen that excels in generally (though not always) maintaining ID over a sustained period without the use of LoRAs:

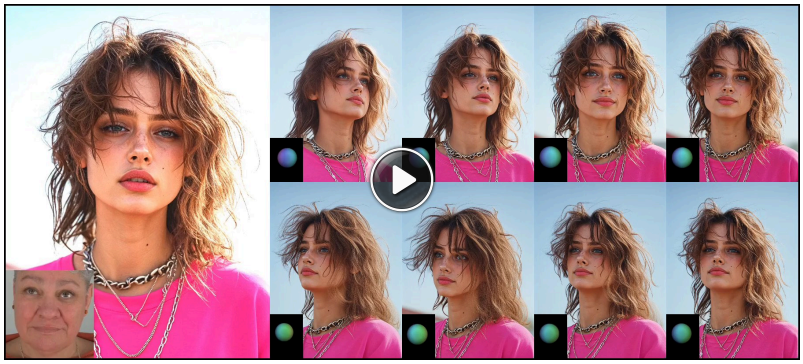


CLICK TO PLAY VIDEO ON SITE

***AUDIO CONTENT. Click to play**. Further examples from the DreamActor project.*

The new system, titled *DreamActor*, uses a three-part hybrid control system that gives dedicated attention to facial expression, head rotation and core skeleton design, thus accommodating AI-driven performances where neither the facial nor body aspect suffer at the expense of the other – a rare, arguably unknown capability among similar systems.

Below we see one of these facets, *head rotation*, in action. The colored ball in the corner of each thumbnail towards the right indicates a kind of virtual gimbal that defines head-orientation independently of facial movement and expression, which is here driven by an actor (lower left).



CLICK TO PLAY VIDEO ON SITE

***Click to play**. The multicolored ball visualized here represents the axis of rotation of the head of the avatar, while the expression is powered by a separate module and informed by an actor's performance (seen here lower left).*

One of the project's most interesting functionalities, which is not even included properly in the paper's tests, is its capacity to derive lip-sync movement directly from audio – a capability which works unusually well even without a driving actor-video.

The researchers have taken on the best incumbents in this pursuit, including the much-lauded Runway Act-One and LivePortrait, and report that DreamActor was able to achieve better quantitative results.

Since researchers can set their own criteria, quantitative results aren't necessarily an empirical standard; but the accompanying qualitative tests seem to support the authors' conclusions.

Unfortunately this system is not intended for public release, and the only value the community can potentially derive from the work is in potentially reproducing the methodologies outlined in the paper (as was done to notable effect for the equally closed-source Google Dreambooth in 2022).

The paper states*:

> *'Human image animation has possible social risks, like being misused to make fake videos. The proposed technology could be used to create fake videos of people, but existing detection tools [Demamba, Dormant] can spot these fakes.*
>
> *'To reduce these risks, clear ethical rules and responsible usage guidelines are necessary. We will strictly restrict access to our core models and codes to prevent misuse.'*

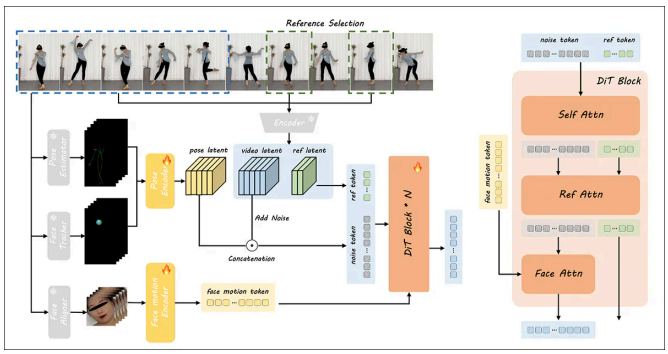
Naturally, ethical considerations of this kind are convenient from a commercial standpoint, since it provides a rationale for API-only access to the model, which can then be monetized. ByteDance has already done this once in 2025, by making the much-lauded OmniHuman available for paid credits on the Dreamina website. Therefore, since DreamActor is possibly an even stronger product, this seems the likely outcome. What remains to be seen is the extent to which its principles, as far as they are explained in the paper, can aid the open source community.

The new paper is titled *DreamActor-M1: Holistic, Expressive and Robust Human Image Animation with Hybrid Guidance*, and comes from six Bytedance researchers.

Method

The DreamActor system proposed in the paper aims to generate human animation from a reference image and a driving video, using a Diffusion Transformer (DiT) framework adapted for latent space (apparently some flavor of Stable Diffusion, though the paper cites only the 2022 landmark release publication).

Rather than relying on external modules to handle reference conditioning, the authors merge appearance and motion features directly inside the DiT backbone, allowing interaction across space and time through attention:



Schema for the new system: DreamActor encodes pose, facial motion, and appearance into separate latents, combining them with noised video latents produced by a 3D VAE. These signals are fused within a Diffusion Transformer using self- and cross-attention, with shared weights across branches. The model is supervised by comparing denoised outputs to clean video latents. Source: https://arxiv.org/pdf/2504.01724

To do this, the model uses a pretrained 3D variational autoencoder to encode both the input video and the reference image. These latents are patchified, concatenated, and fed into the DiT, which processes them jointly.

This architecture departs from the common practice of attaching a secondary network for reference injection, which was the approach for the influential Animate Anyone and Animate Anyone 2 projects.

Instead, DreamActor builds the fusion into the main model itself, simplifying the design while enhancing the flow of information between appearance and motion cues. The model is then trained using flow matching rather than the standard diffusion objective (Flow matching trains diffusion models by directly predicting velocity fields between data and noise, skipping score estimation).

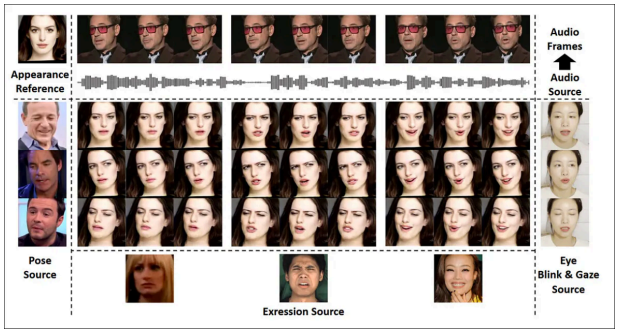
Hybrid Motion Guidance

The Hybrid Motion Guidance method that informs the neural renderings combines pose tokens derived from 3D body skeletons and head spheres; implicit facial representations extracted by a pretrained face encoder; and reference appearance tokens sampled from the source image.

These elements are integrated within the Diffusion Transformer using distinct attention mechanisms, allowing the system to coordinate global motion, facial expression, and visual identity throughout the generation process.

For the first of these, rather than relying on facial landmarks, DreamActor uses implicit facial representations to guide expression generation, apparently enabling finer control over facial dynamics while disentangling identity and head pose from expression.

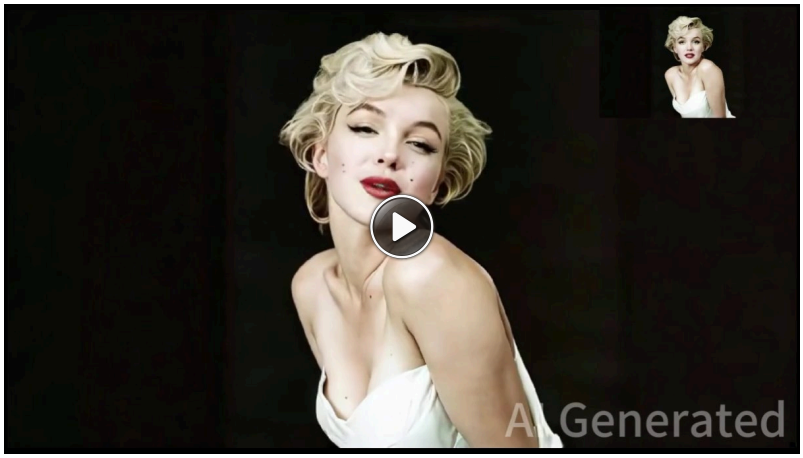
To create these representations, the pipeline first detects and crops the face region in each frame of the driving video, resizing it to 224×224. The cropped faces are processed by a face motion encoder pretrained on the PD-FGC dataset, which is then conditioned by an MLP layer.



PD-FGC, employed in DreamActor, generates a talking head from a reference image with disentangled control of lip sync (from audio), head pose, eye movement, and expression (from separate videos), allowing precise, independent manipulation of each. Source: https://arxiv.org/pdf/2211.14506

The result is a sequence of face motion tokens, which are injected into the Diffusion Transformer through a cross-attention layer.

The same framework also supports an *audio-driven* variant, wherein a separate encoder is trained that maps speech input directly to face motion tokens. This makes it possible to generate synchronized facial animation – including lip movements – without a driving video.

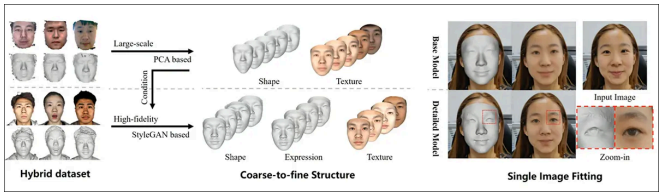


CLICK TO PLAY VIDEO ON SITE

***AUDIO CONTENT. Click to play**. Lip-sync derived purely from audio, without a driving actor reference. The sole character input is the static photo seen upper-right.*

Secondly, to control head pose independently of facial expression, the system introduces a 3D head sphere representation (see video embedded earlier in this article), which decouples facial dynamics from global head movement, improving precision and flexibility during animation.

Head spheres are generated by extracting 3D facial parameters – such as rotation and camera pose – from the driving video using the FaceVerse tracking method.



Schema for the FaceVerse project. Source: https://www.liuyebin.com/faceverse/faceverse.html

These parameters are used to render a color sphere projected onto the 2D image plane, spatially aligned with the driving head. The sphere's size matches the reference head, and its color reflects the head's orientation. This abstraction reduces the complexity of learning 3D head motion, helping to preserve stylized or exaggerated head shapes in characters drawn from animation.

Visualization of the control sphere influencing head orientation.

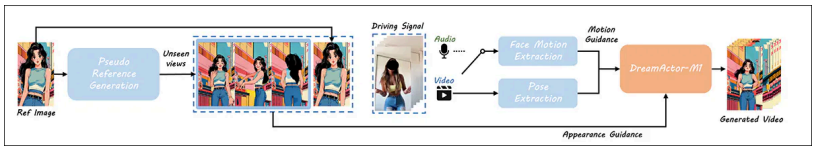
Finally, to guide full-body motion, the system uses 3D body skeletons with adaptive bone length normalization. Body and hand parameters are estimated using 4DHumans and the hand-focused HaMeR, both of which operate on the SMPL-X body model.

SMPL-X applies a parametric mesh over the full human body in an image, aligning with estimated pose and expression to enable pose-aware manipulation using the mesh as a volumetric guide. Source: https://arxiv.org/pdf/1904.05866

From these outputs, key joints are selected, projected into 2D, and connected into line-based skeleton maps. Unlike methods such as Champ, that render full-body meshes, this approach avoids imposing predefined shape priors, and by relying solely on skeletal structure, the model is thus encouraged to infer body shape and appearance directly from the reference images, reducing bias toward fixed body types, and improving generalization across a range of poses and builds.

During training, the 3D body skeletons are concatenated with head spheres and passed through a pose encoder, which outputs features that are then combined with noised video latents to produce the noise tokens used by the Diffusion Transformer.

At inference time, the system accounts for skeletal differences between subjects by normalizing bone lengths. The SeedEdit pretrained image editing model transforms both reference and driving images into a standard canonical configuration. RTMPose is then used to extract skeletal proportions, which are used to adjust the driving skeleton to match the anatomy of the reference subject.



Overview of the inference pipeline. Pseudo-references may be generated to enrich appearance cues, while hybrid control signals – implicit facial motion and explic body skeletons – are extracted from the driving video. These are then fed into a DiT model to produce animated output, with facial motion decoupled from body po as a driver.

Appearance Guidance

To enhance appearance fidelity, particularly in occluded or rarely visible areas, the system supplements the primary reference image with pseudo-references sampled from the input video.

CLICK TO PLAY VIDEO ON SITE

***Click to play**. The system anticipates the need to accurately and consistently render occluded regions. This is about as close as I have seen, in a project of this kind, to a CGI-style bitmap-texture approach.*

These additional frames are chosen for pose diversity using RTMPose, and filtered using CLIP-based similarity to ensure they remain consistent with the subject's identity.

All reference frames (primary and pseudo) are encoded by the same visual encoder and fused through a self-attention mechanism, allowing the model to access complementary appearance cues. This setup improves coverage of details such as profile views or limb textures. Pseudo-references are always used during training and optionally during inference.

Training

DreamActor was trained in three stages to gradually introduce complexity and improve stability.

In the first stage, only 3D body skeletons and 3D head spheres were used as control signals, excluding facial representations. This allowed the base video generation model, initialized from MMDiT, to adapt to human animation without being overwhelmed by fine-grained controls.

In the second stage, implicit facial representations were added, but all other parameters frozen. Only the face motion encoder and face attention layers were trained at this point, enabling the model to learn expressive detail in isolation.

In the final stage, all parameters were unfrozen for joint optimization across appearance, pose, and facial dynamics.

Data and Tests

For the testing phase, the model is initialized from a pretrained image-to-video DiT checkpoint† and trained in three stages: 20,000 steps for each of the first two stages and 30,000 steps for the third.

To improve generalization across different durations and resolutions, video clips were randomly sampled with lengths between 25 and 121 frames. These were then resized to 960x640px, while preserving aspect ratio.

Training was performed on eight (China-focused) NVIDIA H20 GPUs, each with 96GB of VRAM, using the AdamW optimizer with a (tolerably high) learning rate of 5e−6.

At inference, each video segment contained 73 frames. To maintain consistency across segments, the final latent from one segment was reused as the initial latent for the next, which contextualizes the task as sequential image-to-video generation.

Classifier-free guidance was applied with a weight of 2.5 for both reference images and motion control signals.

The authors constructed a training dataset (no sources are stated in the paper) comprising 500 hours of video sourced from diverse domains, featuring instances of (among others) dance, sports, film, and public speaking. The dataset was designed to capture a broad spectrum of human motion and expression, with an even distribution between full-body and half-body shots.

To enhance facial synthesis quality, Nersemble was incorporated in the data preparation process.

Examples from the Nersemble dataset, used to augment the data for DreamActor. Source: https://www.youtube.com/watch?v=a-OAWqBzldU

For evaluation, the researchers used their dataset also as a benchmark to assess generalization across various scenarios.

The model's performance was measured using standard metrics from prior work: Fréchet Inception Distance (FID); Structural Similarity Index (SSIM); Learned Perceptual Image Patch Similarity (LPIPS); and Peak Signal-to-Noise Ratio (PSNR) for frame-level quality. Fréchet Video Distance (FVD) was used for assessing temporal coherence and overall video fidelity.

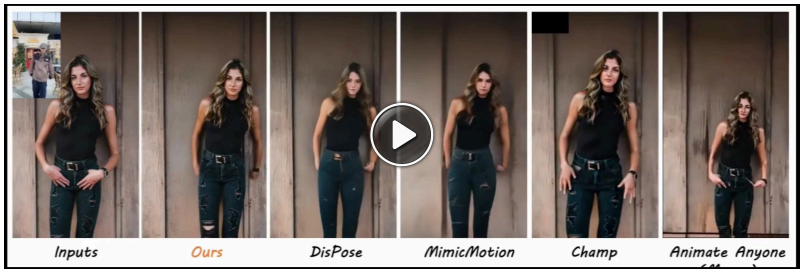
The authors conducted experiments on both body animation and portrait animation tasks, all employing a single (target) reference image.

For body animation, DreamActor-M1 was compared against Animate Anyone; Champ; MimicMotion, and DisPose.

	FID↓	SSIM↑	PSNR↑	LPIPS↓	FVD↓
AA [14]	36.72	0.791	21.74	0.266	158.3
Champ [61]	40.21	0.732	20.18	0.281	171.2
MimicMotion [57]	35.90	0.799	22.25	0.253	149.9
DisPose [20]	33.01	0.804	21.99	0.248	144.7
Ours	**27.27**	**0.821**	**23.93**	**0.206**	**122.0**

Quantitative comparisons against rival frameworks.

Though the PDF provides a static image as a visual comparison, one of the videos from the project site may highlight the differences more clearly:



CLICK TO PLAY VIDEO ON SITE

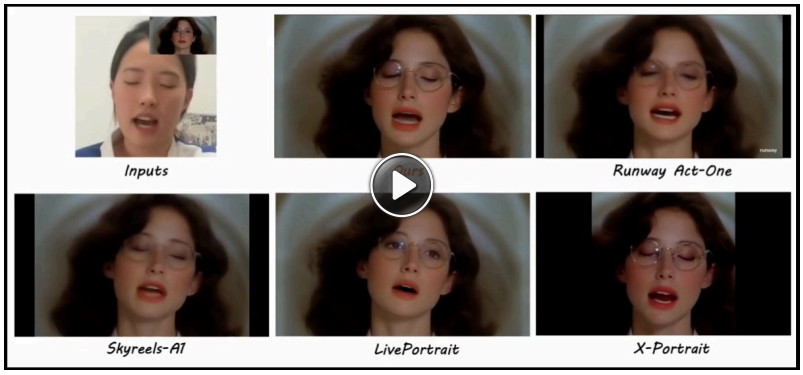
***AUDIO CONTENT. Click to play**. A visual comparison across the challenger frameworks. The driving video is seen top-left, and the authors' conclusion that DreamActor produces the best results seems reasonable.*

For portrait animation tests, the model was evaluated against LivePortrait; X-Portrait; SkyReels-A1; and Act-One.

	FID↓	SSIM↑	PSNR↑	LPIPS↓	FVD↓
LivePortrait [12]	31.72	0.809	24.25	0.270	147.1
X-Portrait [49]	30.09	0.774	22.98	0.281	150.9
SkyReels-A1 [30]	30.66	0.811	24.11	0.262	133.8
Act-One [33]	29.84	0.817	25.07	0.259	135.2
Ours	**25.70**	**0.823**	**28.44**	**0.238**	**110.3**

Quantitative comparisons for portrait animation.

The authors note that their method wins out in quantitative tests, and contend that it is also superior qualitatively.



CLICK TO PLAY VIDEO ON SITE

***AUDIO CONTENT. Click to play**. Examples of portrait animation comparisons.*

Arguably the third and final of the clips shown in the video above exhibits a less convincing lip-sync compared to a couple of the rival frameworks, though the general quality is remarkably high.

Conclusion

In anticipating the need for textures that are implied but not actually present in the sole target image fueling these recreations, ByteDance has addressed one of the biggest challenges facing diffusion-based video generation – consistent, persistent textures. The next logical step after perfecting such an approach would be to somehow create a reference atlas from the initial generated clip that could be applied to subsequent, different generations, to maintain appearance without LoRAs.

Though such an approach would effectively still be an external reference, this is no different from texture-mapping in traditional CGI techniques, and the quality of realism and plausibility is far higher than those older methods can obtain.

That said, the most impressive aspect of DreamActor is the combined three-part guidance system, which bridges the traditional divide between face-focused and body-focused human synthesis in an ingenious way.

It only remains to be seen if some of these core principles can be leveraged in more accessible offerings; as it stands, DreamActor seems destined to become yet another synthesis-as-a-service offering, severely bound by restrictions on usage, and by the impracticality of experimenting extensively with a commercial architecture.

* *My substitution of hyperlinks for the authors; inline citations*

† *As mentioned earlier, it is not clear with flavor of Stable Diffusion was used in this project.*

First published Friday, April 4, 2025

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