A New System for Temporally Consistent Stable Diffusion Video Characters

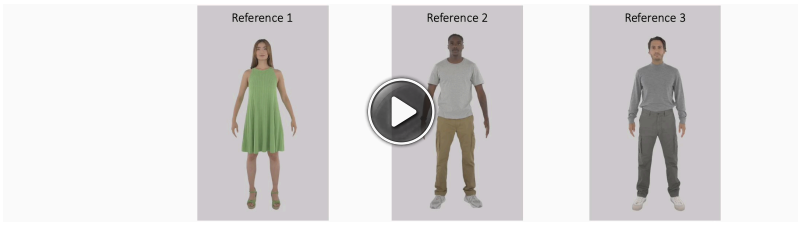
Originally published at https://www.unite.ai/a-new-system-for-temporally-consistent-stable-diffusion-video-characters/

A new initiative from the Alibaba Group offers one of the best methods I have seen for generating full-body human avatars from a Stable Diffusion-based foundation model.

Titled *MIMO* (**MIM**icking with **O**bject Interactions), the system uses a range of popular technologies and modules, including CGI-based human models and AnimateDiff, to enable temporally consistent character replacement in videos – or else to drive a character with a user-defined skeletal pose.

Here we see characters interpolated from a single image source, and driven by a predefined motion:

[Click video below to play]



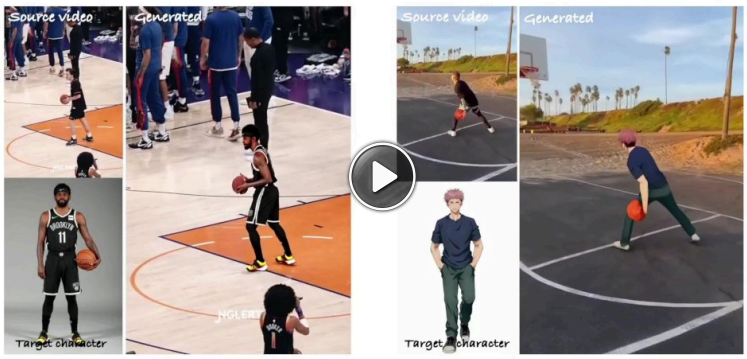
CLICK TO PLAY VIDEO ON SITE

From single source images, three diverse characters are driven by a 3D pose sequence (far left) using the MIMO system. See the project website and the accompanying YouTube video (embedded at the end of this article) for more examples and superior resolution. Source: https://menyifang.github.io/projects/MIMO/index.html

Generated characters, which can also be sourced from frames in videos and in diverse other ways, can be integrated into real-world footage.

MIMO offers a novel system which generates three discrete encodings, each for character, scene, and occlusion (i.e., matting, when some object or person passes in front of the character being depicted). These encodings are integrated at inference time.

[Click video below to play]



CLICK TO PLAY VIDEO ON SITE

MIMO can replace original characters with photorealistic or stylized characters that follow the motion from the target video. See the project website and the accompanying YouTube video (embedded at the end of this article) for more examples and superior resolution.

The system is trained over the Stable Diffusion V1.5 model, using a custom dataset curated by the researchers, and composed equally of real-world and simulated videos.

The great bugbear of diffusion-based video is temporal stability, where the content of the video either flickers or 'evolves' in ways that are not desired for consistent character representation.

MIMO, instead, effectively uses a single image as a map for consistent guidance, which can be orchestrated and constrained by the interstitial SMPL CGI model.

Since the source reference is consistent, and the base model over which the system is trained has been enhanced with adequate representative motion examples, the system's capabilities for temporally consistent output are well above the general standard for diffusion-based avatars.

[Click video below to play]

CLICK TO PLAY VIDEO ON SITE

Further examples of pose-driven MIMO characters. See the project website and the accompanying YouTube video (embedded at the end of this article) for more examples and superior resolution.

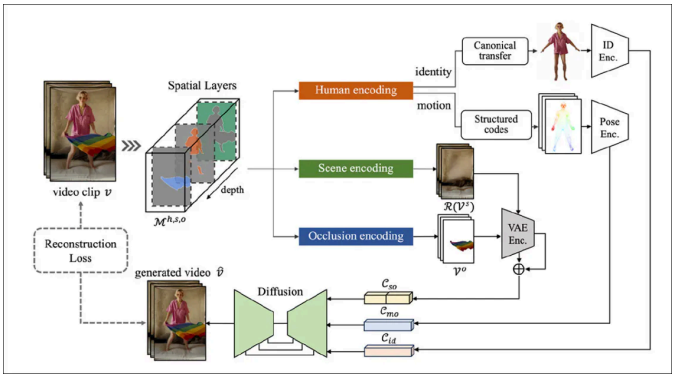
It is becoming more common for single images to be used as a source for effective neural representations, either by themselves, or in a multimodal way, combined with text prompts. For example, the popular LivePortrait facial-transfer system can also generate highly plausible deepfaked faces from single face images.

The researchers believe that the principles used in the MIMO system can be extended into other and novel types of generative systems and frameworks.

The new paper is titled *MIMO: Controllable Character Video Synthesis with Spatial Decomposed Modeling*, and comes from four researchers at Alibaba Group's Institute for Intelligent Computing. The work has a video-laden project page and an accompanying YouTube video, which is also embedded at the bottom of this article.

Method

MIMO achieves automatic and unsupervised separation of the aforementioned three spatial components, in an end-to-end architecture (i.e., all the sub-processes are integrated into the system, and the user need only provide the input material).



The conceptual schema for MIMO. Source: https://arxiv.org/pdf/2409.16160

Objects in source videos are translated from 2D to 3D, initially using the monocular depth estimator Depth Anything. The human element in any frame is extracted with methods adapted from the Tune-A-Video project.

These features are then translated into video-based volumetric facets via Facebook Research's Segment Anything 2 architecture.

The scene layer itself is obtained by removing objects detected in the other two layers, effectively providing a rotoscope-style mask automatically.

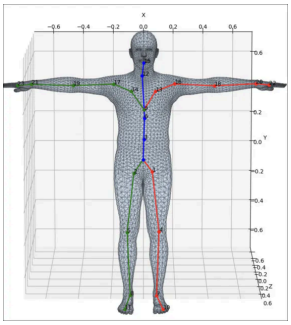
For the motion, a set of extracted latent codes for the human element are anchored to a default human CGI-based SMPL model, whose movements provide the context for the rendered human content.

A 2D feature map for the human content is obtained by a differentiable rasterizer derived from a 2020 initiative from NVIDIA .

Combining the obtained 3D data from SMPL with the 2D data obtained by the NVIDIA method, the latent codes representing the 'neural person' have a solid correspondence to their eventual context.

At this point, it is necessary to establish a reference commonly needed in architectures that use SMPL – a canonical pose. This is broadly similar to Da Vinci's 'Vitruvian man', in that it represents a zero-pose template which can accept content and then be deformed, bringing the (effectively) texture-mapped content with it.

These deformations, or 'deviations from the norm', represent human movement, while the SMPL model preserves the latent codes that constitute the human identity that has been extracted, and thus represents the resulting avatar correctly in terms of pose and texture.



An example of a canonical pose in an SMPL figure. Source: https://www.researchgate.net/figure/Layout-of-23-joints-in-the-SMPL-models_fig2_351179264

Regarding the issue of entanglement (the extent to which trained data can turn out to be inflexible when you stretch it beyond its trained confines and associations), the authors state*:

> *'To fully disentangle the appearance from posed video frames, an ideal solution is to learn the dynamic human representation from the monocular video and transform it from the posed space to the canonical space.*
>
> *'Considering the efficiency, we employ a simplified method that directly transforms the posed human image to the canonical result in standard A-pose using a pretrained human repose model. The synthesized canonical appearance image is fed to ID encoders to obtain the identity .*
>
> *'This simple design enables full disentanglement of identity and motion attributes. Following [Animate Anyone], the ID encoders include a CLIP image encoder and a reference-net architecture to embed for the global and local feature, [respectively].'*

For the scene and occlusion aspects, a shared and fixed Variational Autoencoder (VAE – in this case derived from a 2013 publication) is used to embed the scene and occlusion elements into the latent space. Incongruities are handled by an inpainting method from the 2023 ProPainter project.

Once assembled and retouched in this way, both the background and any occluding objects in the video will provide a matte for the moving human avatar.

These decomposed attributes are then fed into a U-Net backbone based on the Stable Diffusion V1.5 architecture. The complete scene code is concatenated with the host system's native latent noise. The human component is integrated via self-attention and cross-attention layers, respectively.

Then, the denoised result is output via the VAE decoder.

Data and Tests

For training, the researchers created human video dataset titled HUD-7K, which consisted of 5,000 real character videos and 2,000 synthetic animations created by the En3D system. The real videos required no annotation, due to the non-semantic nature of the figure extraction procedures in MIMO's architecture. The synthetic data was fully annotated.

The model was trained on eight NVIDIA A100 GPUs (though the paper does not specify whether these were the 40GB or 80GB VRAM models), for 50 iterations, using 24 video frames and a batch size of four, until convergence.

The motion module for the system was trained on the weights of AnimateDiff. During the training process, the weights of the VAE encoder/decoder, and the CLIP image encoder were frozen (in contrast to full fine-tuning, which will have a much broader effect on a foundation model).

Though MIMO was not trialed against analogous systems, the researchers tested it on difficult out-of-distribution motion sequence sourced from AMASS and Mixamo. These movements included climbing, playing, and dancing.

They also tested the system on in-the-wild human videos. In both cases, the paper reports 'high robustness' for these unseen 3D motions, from different viewpoints.

Though the paper offers multiple static image results demonstrating the effectiveness of the system, the true performance of MIMO is best assessed with the extensive video results provided at the project page, and in the YouTube video embedded below (from which the videos at the start of this article have been derived).

The authors conclude:

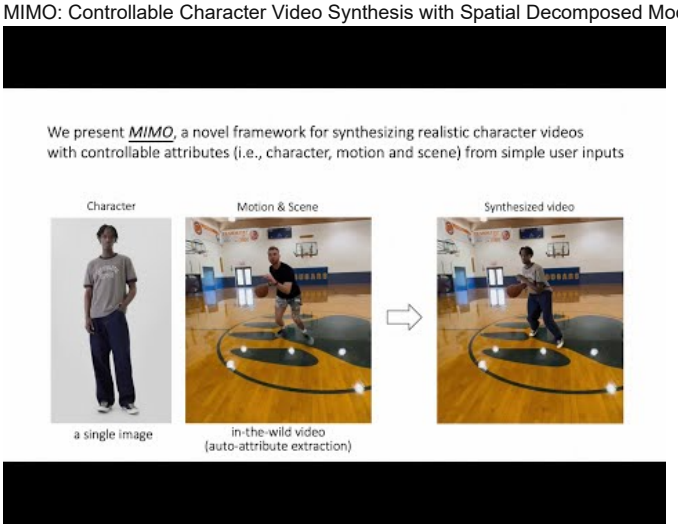
> *'Experimental results [demonstrate] that our method enables not only flexible character, motion and scene control, but also advanced scalability to arbitrary characters, generality to novel 3D motions, and applicability to interactive scenes.*
>
> *'We also [believe] that our solution, which considers inherent 3D nature and automatically encodes the 2D video to hierarchical spatial components could inspire future researches for 3D-aware video synthesis.*
>
> *'Furthermore, our framework is not only well suited to generate character videos but also can be potentially adapted to other controllable video synthesis tasks.'*

Conclusion

It's refreshing to see an avatar system based on Stable Diffusion that appears capable of such temporal stability – not least because Gaussian Avatars seem to be gaining the high ground in this particular research sector.

The stylized avatars represented in the results are effective, and while the level of photorealism that MIMO can produce is not currently equal to what Gaussian Splatting is capable of, the diverse advantages of creating temporally consistent humans in a semantically-based Latent Diffusion Network (LDM) are considerable.

MIMO: Controllable Character Video Synthesis with Spatial Decomposed Modeling



** My conversion of the authors' inline citations to hyperlinks, and where necessary, external explanatory hyperlinks.*

First published Wednesday, September 25, 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