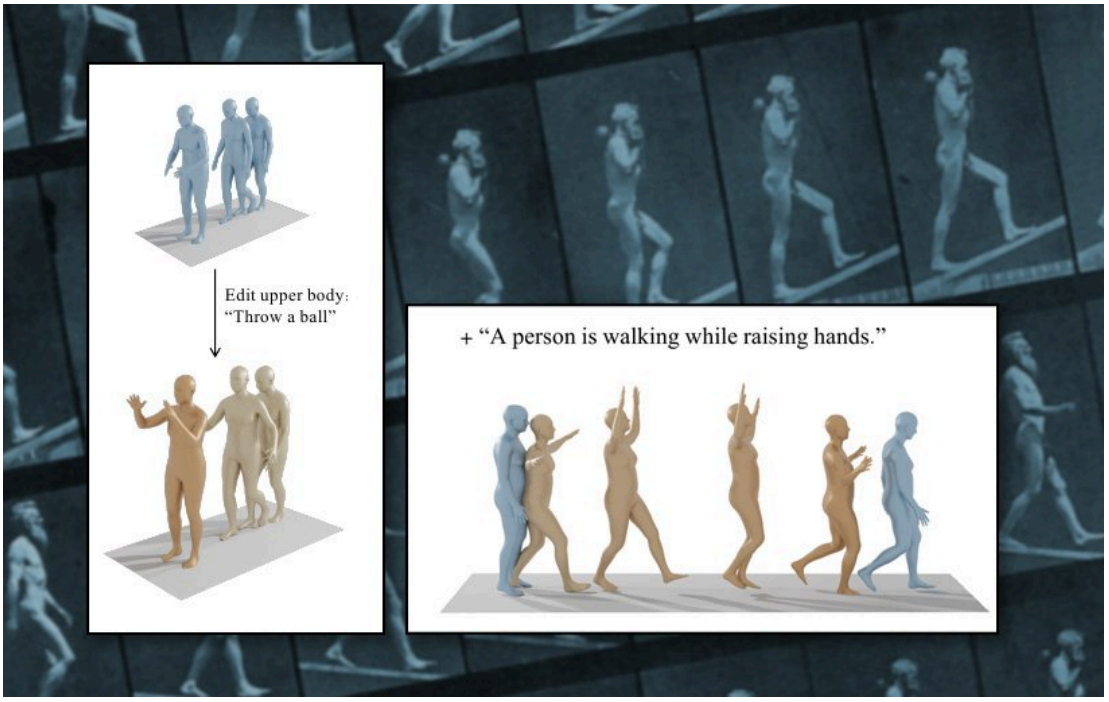
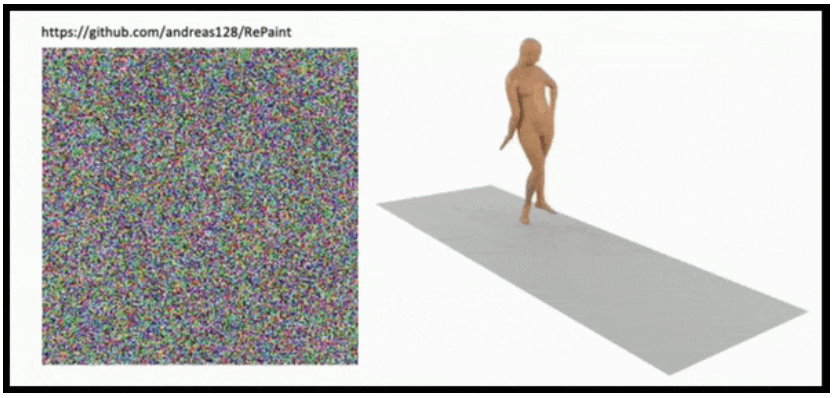
Creating Authentic Human Motion Synthesis via Diffusion

Originally published at https://metaphysic.ai/creating-authentic-human-motion-synthesis-via-diffusion/

October 10, 2022



New research from Tel Aviv University may prove capable of bringing authentic human motion to text-to-video synthesis, videogames, motion capture architectures in VFX pipelines, and also function as a synthetic data generator for downstream research initiatives, among a myriad of other potential applications – all thanks to what has turned out to be the most sensational AI technology of 2022 – diffusion models.



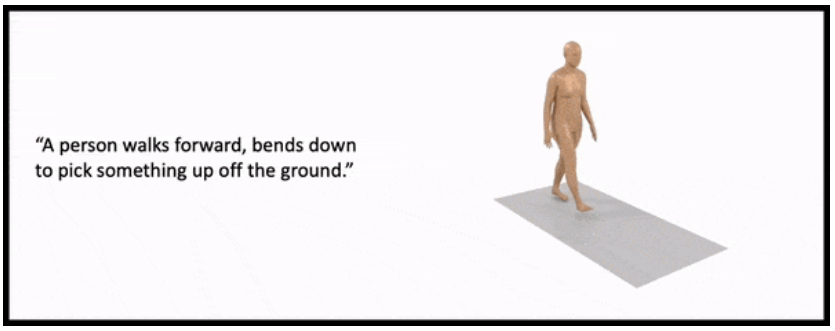
From sheer static to cohesive and realistic movement, via noise-based diffusion. On the left, by analogy, we see a Stable Diffusion-style image forming; on the right, we see a representative idea of the way that chaotic motion becomes authoritative and convincing as the prior information from the trained dataset likewise concretizes. Source: https://guytevet.github.io/mdm-page/

As with Stable Diffusion, the new system creates order out of chaos, by devolving base noise into coherent output (which it has learned to do by reversing this process for thousands of data samples during training).

Unlike Stable Diffusion, what emerges is not an accurate picture, but an accurate *movement*, informed by comprehensively labeled upstream motion-based datasets.

The new model, called *Motion Diffusion Model* (MDM), leverages Transformers to interpret priors (extracted features from snippets of movement data, collected into voluminous datasets) and then reproduce them according to several possible input methods; notably text-prompts, which most of us have become well aware of since the launch of OpenAI's DALL-E 2 earlier this year, and then Stability.ai's sensational open source release of the Stable Diffusion model and weights in August.

Below are some examples of prompts that can produce realistic motion data under MDM:

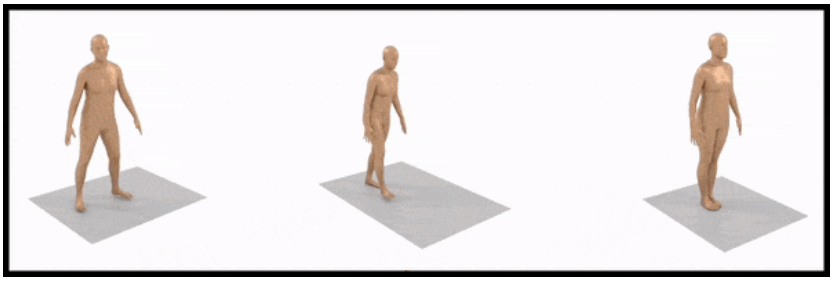


Prompts interpreted through MDM.

The paper is titled *Human Motion Diffusion Model*, and comes from six authors at Tel Aviv University. The lead author, Guy Tevet, from the university's Visual Computing Lab, spoke to us about the challenges and potential of the new system, and also considered its possible applications across a range of sectors.

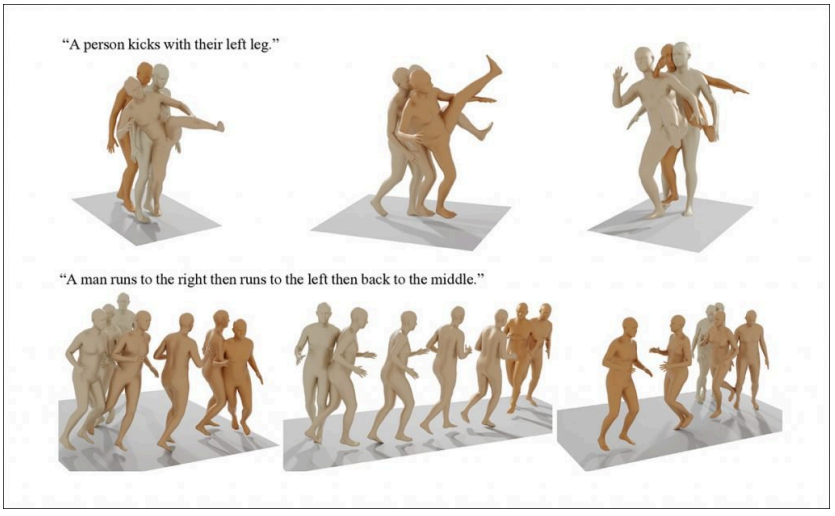
Ideas into Motion

"What I mostly have in mind," he says. "is that in video games, and especially now in the Metaverse, you may need to animate a lot of figures. For example, you want to animate a crowd at a soccer game – a crowd of thousands of people, where some are 'real', but some of them are not – and you want to animate them in an authentic way; you don't want all of them to do the same thing."
"In this case," he continues. "diffusion models are fantastic. For instance, a prompt to *'kick'* can mean many things, and being able to clarify which type of kick you mean, via a text prompt, can save a lot of time in obtaining a movement that's correct for the context.



From left to right, three types of 'kick' interpreted through the new model, from rough-housing to martial arts.

"In video games particularly, such as the GTA series, even today, movements of this type are not as realistic as they could be. That's one area where motion diffusion could really have an impact."
Tevet also believes that MDM, and similar systems, have potential in a VFX workflow, in the context of motion capture and authentic movement reproduction – one aspect of traditional CGI that has remained subject to criticism as these older technologies have improved in other areas over the last thirty years.
Though the VFX industry has certainly not dismissed AI, many practitioners retain a skeptical standpoint in regard to the level of editability for many of the new technologies that are so dazzling to the general public, such as autoencoder deepfakes and latent diffusion image generation. To what extent, then, could VFX professionals intervene at a truly *granular* level in a pipeline that uses motion diffusion?
"Yes, it's possible to address these movements at the joint level," Tevet observes. "You can select specific joints. If you want to take an existing motion and edit just one of the hands."



Examples of MDM (literally) in action. Darker colors indicate later frames in the sequence being depicted. Source: https://arxiv.org/pdf/2209.14916.pdf

The new architecture, Tevet agrees, is also amenable to third-party technologies that could allow for direct (i.e. manipulated) interaction via interfaces based around SDF and/or 3DMM interfaces, amongst other possible ways that researchers have been applying recently to obtain a finer-grained access to content that's resident in the latent space of a Generative Adversarial Network (GAN), or of Neural Radiance Fields (NeRF) – or of a latent diffusion network such as Stable Diffusion.

Commanding MDM

Currently, however, there are three primary methods by which the end-user can obtain motion from MDM: *text-to-motion*; *action-to-motion*; and *unconditioned generation*.
Text-to-motion, where MDM has achieved state-of-the-art performance in the HumanML3D and KIT benchmarks, is most familiar to us, also from the above examples, in that a user text-prompt is fed to the model, which constructs a semantic interpretation and then a corresponding action.
In action-to-motion, a model is tasked with generating diverse types of movement within a specified action, and creating rational action transfer when some of the central facets of the reconstruction have changed, such as a change in body-type. Here too, MDM has outperformed prior efforts, in tests published in the new paper, this time in the HumanAct12 and UESTC benchmarks.
In unconditioned generation, character movement is required to conform to natural and realistic motion in a situation where it may have little prior instruction, except to *follow a certain path*; it's one of the most challenging and least-studied aspects of motion synthesis, and effectively demands character movement more from 'motivation' or scant and 'unexplained' goals than through direct instruction of any kind.

From a 2016 paper led by the University of Edinburgh, a character's movements are synthesized automatically along certain trajectories, adapting to ad hoc conditions. Source: https://www.pure.ed.ac.uk/ws/files/25269752/motionsynthesis.pdf

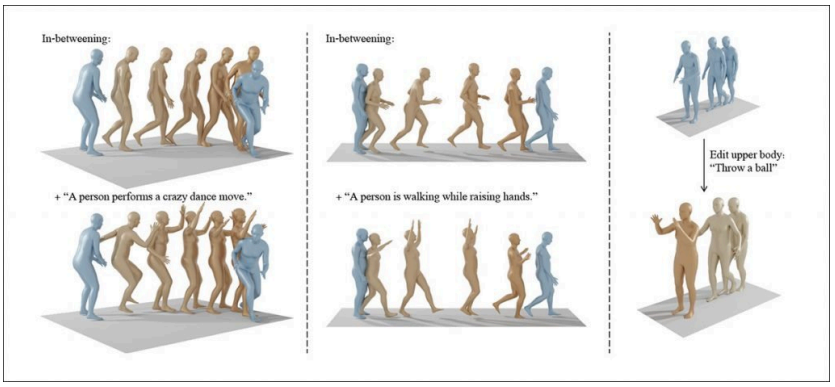
From a certain perspective, the task is analogous to the autonomous behavior of AI-driven videogame characters, and also to reinforcement learning, where an AI entity may be regularly challenged to adapt to changing conditions.

Into the Wild

Our talk took place on the day that MDM was released to open source – including the checkpoints: the trained model, that is usually so expensive to develop, and which (except for a rare case such as Stable Diffusion) is not always made available in its entirety, or at all, except through metered APIs and other limited measures.

Those downloading and running the MDM GitHub framework receive a visual indicator of the generated motion in the form of an .mp4 file that presents a stick figure performing the motion – but the obvious intent is that this data will be actively incorporated into secondary frameworks of various types. Source: https://github.com/GuyTevet/motion-diffusion-model

At the time of writing, the MDM authors were in discussion with AI proponent and prolific computer vision Twitter influencer Ahsen Khaliq, who joined Hugging Face towards the end of last year. Khaliq, Tevet told us, is currently arranging a Gradio demonstration of MDM. Nonetheless, unusually, the full release has preceded the 'trailer', as it were.
"Yes," Tevet said. "I'm releasing the full code of the model, including the checkpoints, and people will be able to fine tune it to their needs."



'In-betweening' with MDM. The blue figures represent the 'key-frames'.

One of the central achievements of the model is its very meagre hardware requirements.
"So that that's a thing that we are very proud of," Tevet says. "It's not so resource demanding. You need only a single GPU – a standard GPU. Besides fine-tuning it, you can even conceivably train this model from scratch, which would take about three days on as little as 8GB of VRAM."
However, Stable Diffusion fans that are hoping such a system could play a factor in developing human motion in text-to-video SD output should manage their expectations. Stable Diffusion, like any similar latent diffusion architecture, has no intrinsic mechanisms that could facilitate a temporal architecture, and therefore no direct 'hook' to which motion data could conceivably be attached. Even if it were possible to create an interpretive mechanism, temporal consistency of objects, characters and environments remains an unsolved (though eagerly pursued) problem in SD.

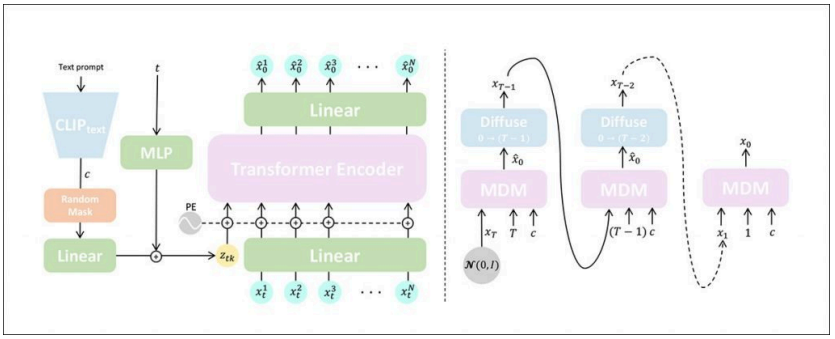
'Land like Spider-Man'

It's well-known in motion synthesis research that data volume, availability and accuracy of labeling are central bottlenecks to progress. Nascent text-to-video systems such as CogVideo have had to bootstrap their temporal architectures on the back of existing text-to-*image* frameworks (see our chat with one of CogVideo's authors here); even the stunning recent revelation of Google's Imagen Video framework is built in a similar way, by extending the original Imagen text-to-video system into the temporal domain, and training jointly on video and images in a 7-tier array of diverse video diffusion models.
"My main limitation now as a researcher," says Tevet. "is the data. We can we show that we can train very nice models with low resources. But I know for sure that if I could get data that uses more specific language, and that more specifically describes a motion, and a greater diversity and range of motion, we could get better results. That's the main limitation today, in my opinion."

https://www.youtube.com/watch?v=cceRrlnTCEs

https://www.youtube.com/watch?v=cceRrlnTCEs

Contributing databases to MDM include the aforementioned HumanML3D, which advances the textual annotations of the AMASS (see video above) and HumanAct12 collection (again, see links above).



The MDM pipeline converts a CLIP-annotated motion sequence into linear embeddings via Transformers. At each sampling step, the model makes a new prediction, iterating through the sequence until a loss equating to 'clean' movement is obtained.

The annotation problem is not just a logistical one, but also has a semantic and even cultural character.
"Currently," says Tevet. "you can't tell MDM to make a figure *'land like Spider-Man'*, even though there may be suitable motions in the upstream data that would fit a description like that.
"The annotations tend to be more prosaic, such as *'lands to crouching position, supported by left hand'*. That's the same movement, really, but annotated in a much drier and less interpretive way. So whether or not we'll eventually be able to increase the lexicon and interpretability of systems such as MDM may depend on future initiatives in re-labeling existing data, not just the development of new datasets.
"In the future, I want to give people the ability to really control motion with free language, and to be able to generate Spider-Man, if they want."

Corporeal Expression

The current MDM model is limited to joint-based movement. Therefore, any attempt to convey facial expressions is limited to coarse actions such as tilting the head (*'curiosity'*) or rapidly turning it (*'surprise'*). Tevet does not envision incorporating facial motility into the full-body data generated by MDM, at least in the near future.
"It's an interesting idea," Tevet comments. "but I'm not exactly sure how we could implement it. One thing I'm fairly sure about is that we would hit a data problem quite quickly if we did try, because motion capture is expensive, and facial expression datasets tend to be more fragmented and

project-specific than the larger corpora that we use for MDM. That would make it a challenge, just to incorporate that kind of upstream data.
“I suppose that if you wanted a facially expressive solution that’s more integrated than just face-swapping the output, you’d need to consider a broader architecture, of which a system like MDM would be just one component. But that kind of nesting, that kind of approach is beyond the scope of what we’re working on here.
“But don’t rule it out – the pure output of MDM is specifically intended to form a part of more complex work-flows, with more impressive results than the indicative stick-figures that it outputs to show you how the calculation went. It’s a potentially extensible architecture.”