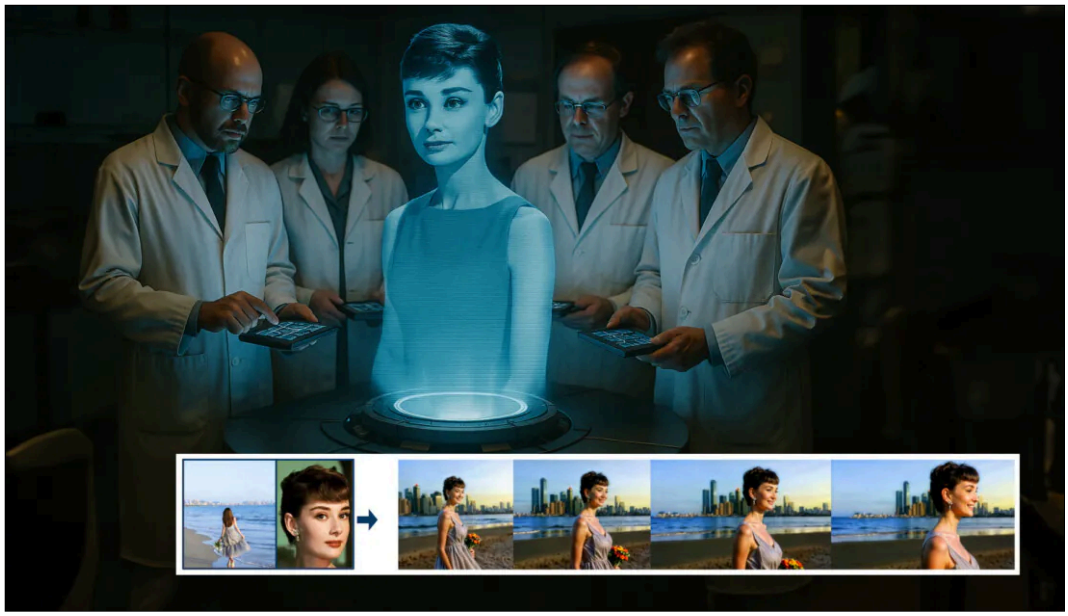
Towards Total Control in AI Video Generation

Originally published at https://www.unite.ai/towards-total-control-in-ai-video-generation/



Video foundation models such as Hunyuan and Wan 2.1, while powerful, do not offer users the kind of granular control that film and TV production (particularly VFX production) demands.

In professional visual effects studios, open-source models like these, along with earlier image-based (rather than video) models such as Stable Diffusion, Kandinsky and Flux, are typically used alongside a range of supporting tools that adapt their raw output to meet specific creative needs. When a director says, *"That looks great, but can we make it a little more [n]?"* you can't respond by saying the model isn't precise enough to handle such requests.

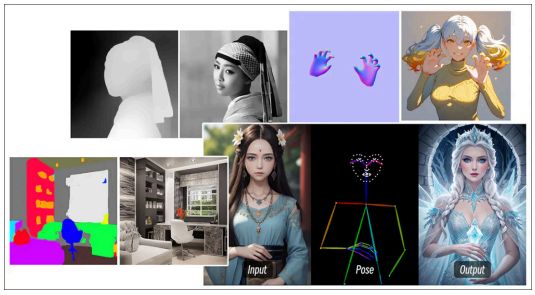
Instead an AI VFX team will use a range of traditional CGI and compositional techniques, allied with custom procedures and workflows developed over time, in order to attempt to push the limits of video synthesis a little further.

So by analogy, a foundation video model is much like a default installation of a web-browser like Chrome; it does a lot out of the box, but if you want it to adapt to your needs, rather than vice versa, you're going to need some plugins.

Control Freaks

In the world of diffusion-based image synthesis, the most important such third-party system is ControlNet.

ControlNet is a technique for adding structured control to diffusion-based generative models, allowing users to guide image or video generation with additional inputs such as edge maps, depth maps, or pose information.



ControlNet's various methods allow for depth>image (top row), semantic segmentation>image (lower left) and pose-guided image generation of humans and animals (lower left).

Instead of relying solely on text prompts, ControlNet introduces separate neural network branches, or *adapters*, that process these conditioning signals while preserving the base model's generative capabilities.

This enables fine-tuned outputs that adhere more closely to user specifications, making it particularly useful in applications where precise composition, structure, or motion control is required:

With a guiding pose, a variety of accurate output types can be obtained via ControlNet. Source: https://arxiv.org/pdf/2302.05543

However, adapter-based frameworks of this kind operate externally on a set of neural processes that are very internally-focused. These approaches have several drawbacks.

First, adapters are trained independently, leading to *branch conflicts* when multiple adapters are combined, which can entail degraded generation quality.

Secondly, they introduce *parameter redundancy*, requiring extra computation and memory for each adapter, making scaling inefficient.

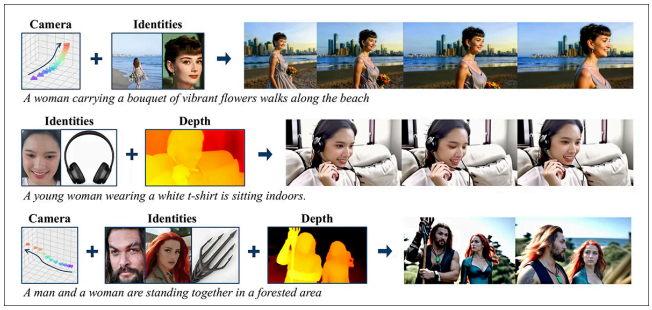
Thirdly, despite their flexibility, adapters often produce *sub-optimal* results compared to models that are fully fine-tuned for multi-condition generation. These issues make adapter-based methods less effective for tasks requiring seamless integration of multiple control signals.

Ideally, the capacities of ControlNet would be trained *natively* into the model, in a modular way that could accommodate later and much-anticipated obvious innovations such as simultaneous video/audio generation, or native lip-sync capabilities (for external audio).

As it stands, every extra piece of functionality represents either a post-production task or a non-native procedure that has to navigate the tightly-bound and sensitive weights of whichever foundation model it's operating on.

FullDiT

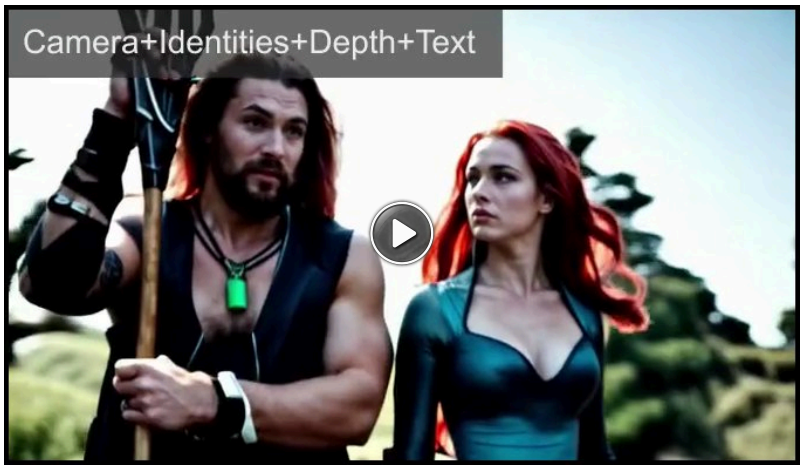
Into this standoff comes a new offering from China, that posits a system where ControlNet-style measures are baked directly into a generative video model at training time, instead of being relegated to an afterthought.



From the new paper: the FullDiT approach can incorporate identity imposition, depth and camera movement into a native generation, and can summon up any combination of these at once. Source: https://arxiv.org/pdf/2503.19907

Titled *FullDiT*, the new approach fuses multi-task conditions such as identity transfer, depth-mapping and camera movement into an integrated part of a trained generative video model, for which the authors have produced a prototype trained model, and accompanying video-clips at a project site.

In the example below, we see generations that incorporate camera movement, identity information and text information (i.e., guiding user text prompts):



CLICK TO PLAY VIDEO ON SITE

Click to play. *Examples of ControlNet-style user imposition with only a native trained foundation model.* Source: https://fulldit.github.io/

It should be noted that the authors do not propose their experimental trained model as a functional foundation model, but rather as a proof-of-concept for native text-to-video (T2V) and image-to-video (I2V) models that offer users more control than just an image prompt or a text-prompt.

Since there are no similar models of this kind yet, the researchers created a new benchmark titled *FullBench*, for the evaluation of multi-task videos, and claim state-of-the-art performance in the like-for-like tests they devised against prior approaches. However, since FullBench was designed by the authors themselves, its objectivity is untested, and its dataset of 1,400 cases may be too limited for broader conclusions.

Perhaps the most interesting aspect of the architecture the paper puts forward is its potential to incorporate new types of control. The authors state:

> *'In this work, we only explore control conditions of the camera, identities, and depth information. We did not further investigate other conditions and modalities such as audio, speech, point cloud, object bounding boxes, optical flow, etc. Although the design of FullDiT can seamlessly integrate other modalities with minimal architecture modification, how to quickly and cost-effectively adapt existing models to new conditions and modalities is still an important question that warrants further exploration.'*

Though the researchers present FullDiT as a step forward in multi-task video generation, it should be considered that this new work builds on existing architectures rather than introducing a fundamentally new paradigm.

Nonetheless, FullDiT currently stands alone (to the best of my knowledge) as a video foundation model with 'hard coded' ControlNet-style facilities – and it's good to see that the proposed architecture can accommodate later innovations too.

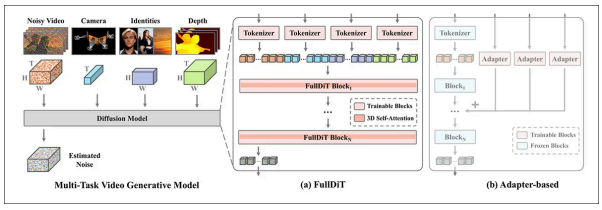
CLICK TO PLAY VIDEO ON SITE

Click to play. *Examples of user-controlled camera moves, from the project site.*

The new paper is titled *FullDiT: Multi-Task Video Generative Foundation Model with Full Attention*, and comes from nine researchers across Kuaishou Technology and The Chinese University of Hong Kong. The project page is here and the new benchmark data is at Hugging Face.

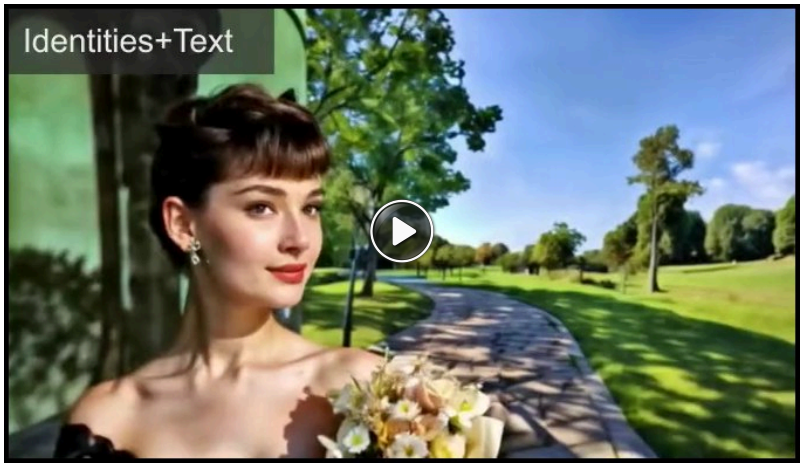
Method

The authors contend that FullDiT's unified attention mechanism enables stronger cross-modal representation learning by capturing both spatial and temporal relationships across conditions:



According to the new paper, FullDiT integrates multiple input conditions through full self-attention, converting them into a unified sequence. By contrast, adapter-based models (leftmost above) use separate modules for each input, leading to redundancy, conflicts, and weaker performance.

Unlike adapter-based setups that process each input stream separately, this shared attention structure avoids branch conflicts and reduces parameter overhead. They also claim that the architecture can scale to new input types without major redesign – and that the model schema shows signs of generalizing to condition combinations not seen during training, such as linking camera motion with character identity.



CLICK TO PLAY VIDEO ON SITE

Click to play. *Examples of identity generation from the project site*.

In FullDiT's architecture, all conditioning inputs – such as text, camera motion, identity, and depth – are first converted into a unified token format. These tokens are then concatenated into a single long sequence, which is processed through a stack of transformer layers using full self-attention. This approach follows prior works such as Open-Sora Plan and Movie Gen.

This design allows the model to learn temporal and spatial relationships jointly across all conditions. Each transformer block operates over the entire sequence, enabling dynamic interactions between modalities without relying on separate modules for each input – and, as we have noted, the architecture is designed to be extensible, making it much easier to incorporate additional control signals in the future, without major structural changes.

The Power of Three

FullDiT converts each control signal into a standardized token format so that all conditions can be processed together in a unified attention framework. For camera motion, the model encodes a sequence of extrinsic parameters – such as position and orientation – for each frame. These parameters are timestamped and projected into embedding vectors that reflect the temporal nature of the signal.

Identity information is treated differently, since it is inherently spatial rather than temporal. The model uses identity maps that indicate which characters are present in which parts of each frame. These maps are divided into *patches*, with each patch projected into an embedding that captures spatial identity cues, allowing the model to associate specific regions of the frame with specific entities.

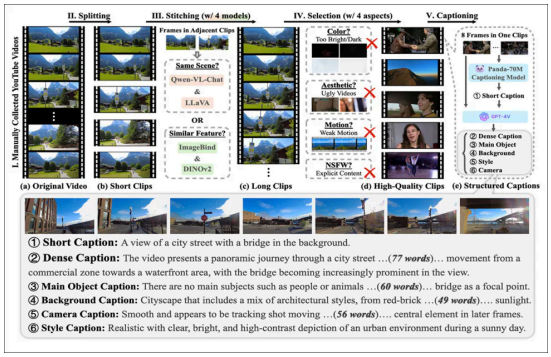
Depth is a spatiotemporal signal, and the model handles it by dividing depth videos into 3D patches that span both space and time. These patches are then embedded in a way that preserves their structure across frames.

Once embedded, all of these condition tokens (camera, identity, and depth) are concatenated into a single long sequence, allowing FullDiT to process them together using full self-attention. This shared representation makes it possible for the model to learn interactions across modalities and across time without relying on isolated processing streams.

Data and Tests

FullDiT's training approach relied on selectively annotated datasets tailored to each conditioning type, rather than requiring all conditions to be present simultaneously.

For textual conditions, the initiative follows the structured captioning approach outlined in the MiraData project.

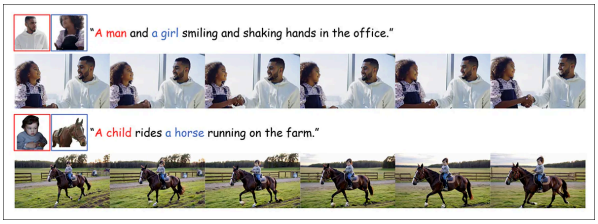


Video collection and annotation pipeline from the MiraData project. Source: https://arxiv.org/pdf/2407.06358

For camera motion, the RealEstate10K dataset was the main data source, due to its high-quality ground-truth annotations of camera parameters.

However, the authors observed that training exclusively on static-scene camera datasets such as RealEstate10K tended to reduce dynamic object and human movements in generated videos. To counteract this, they conducted additional fine-tuning using internal datasets that included more dynamic camera motions.

Identity annotations were generated using the pipeline developed for the ConceptMaster project, which allowed efficient filtering and extraction of fine-grained identity information.



The ConceptMaster framework is designed to address identity decoupling issues while preserving concept fidelity in customized videos.
Source: https://arxiv.org/pdf/2501.04698

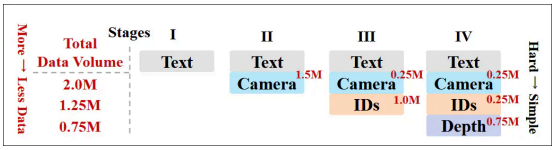
Depth annotations were obtained from the Panda-70M dataset using Depth Anything.

Optimization Through Data-Ordering

The authors also implemented a progressive training schedule, introducing more challenging conditions *earlier in training* to ensure the model acquired robust representations before simpler tasks were added. The training order proceeded from *text* to *camera* conditions, then *identities*, and finally *depth*, with easier tasks generally introduced later and with fewer examples.

The authors emphasize the value of ordering the workload in this way:

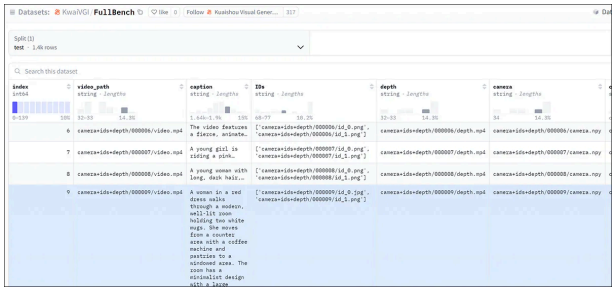
> *'During the pre-training phase, we noted that more challenging tasks demand extended training time and should be introduced earlier in the learning process. These challenging tasks involve complex data distributions that differ significantly from the output video, requiring the model to possess sufficient capacity to accurately capture and represent them.*
>
> *'Conversely, introducing easier tasks too early may lead the model to prioritize learning them first, since they provide more immediate optimization feedback, which hinder the convergence of more challenging tasks.'*



An illustration of the data training order adopted by the researchers, with red indicating greater data volume.

After initial pre-training, a final fine-tuning stage further refined the model to improve visual quality and motion dynamics. Thereafter the training followed that of a standard diffusion framework*: noise added to video latents, and the model learning to predict and remove it, using the embedded condition tokens as guidance.

To effectively evaluate FullDiT and provide a fair comparison against existing methods, and in the absence of the availability of any other apposite benchmark, the authors introduced *FullBench*, a curated benchmark suite consisting of 1,400 distinct test cases.



A data explorer instance for the new FullBench benchmark. Source: https://huggingface.co/datasets/KwaiVGI/FullBench

Each data point provided ground truth annotations for various conditioning signals, including *camera motion*, *identity*, and *depth*.

Metrics

The authors evaluated FullDiT using ten metrics covering five main aspects of performance: text alignment, camera control, identity similarity, depth accuracy, and general video quality.

Text alignment was measured using CLIP similarity, while camera control was assessed through *rotation error* (*RotErr*), *translation error* (*TransErr*), and *camera motion consistency* (CamMC), following the approach of CamI2V (in the *CameraCtrl* project).

Identity similarity was evaluated using DINO-I and CLIP-I, and depth control accuracy was quantified using Mean Absolute Error (MAE).

Video quality was judged with three metrics from MiraData: frame-level CLIP similarity for smoothness; optical flow-based motion distance for dynamics; and LAION-Aesthetic scores for visual appeal.

Training

The authors trained FullDiT using an internal (undisclosed) text-to-video diffusion model containing roughly one billion parameters. They intentionally chose a modest parameter size to maintain fairness in comparisons with prior methods and ensure reproducibility.

Since training videos differed in length and resolution, the authors standardized each batch by resizing and padding videos to a common resolution, sampling 77 frames per sequence, and using applied attention and loss masks to optimize training effectiveness.

The Adam optimizer was used at a learning rate of 1×10^{-5} across a cluster of 64 NVIDIA H800 GPUs, for a combined total of 5,120GB of VRAM (consider that in the enthusiast synthesis communities, *24GB* on an RTX 3090 is still considered a luxurious standard).

The model was trained for around 32,000 steps, incorporating up to three identities per video, along with 20 frames of camera conditions and 21 frames of depth conditions, both evenly sampled from the total 77 frames.

For inference, the model generated videos at a resolution of 384×672 pixels (roughly five seconds at 15 frames per second) with 50 diffusion inference steps and a classifier-free guidance scale of five.

Prior Methods

For camera-to-video evaluation, the authors compared FullDiT against MotionCtrl, CameraCtrl, and CamI2V, with all models trained using the RealEstate10k dataset to ensure consistency and fairness.

In identity-conditioned generation, since no comparable open-source multi-identity models were available, the model was benchmarked against the 1B-parameter ConceptMaster model, using the same training data and architecture.

For depth-to-video tasks, comparisons were made with Ctrl-Adapter and ControlVideo.

Metrics	Text	Camera			Identities		Depth	Overall Quality		
Model	**Clip Score↑**	**RotErr↓**	**TransErr↓**	**CamMC↓**	**DINO-I↑**	**CLIP-I↑**	**MAE↓**	**Smoothness↑**	**Dynamic↑**	**Aesthetic↑**
Camera to Video										
MotionCtrl [54]	22.27	1.49	4.41	4.84	-	-	-	96.16	11.43	4.71
CameraCtrl [18]	21.36	1.57	3.88	4.77	-	-	-	95.16	13.72	4.66
CamI2V‡ [68]	-	1.43	3.81	4.62	-	-	-	94.50	19.40	-
FullDiT	**22.97**	**1.20**	**3.31**	**3.98**	-	-	-	**96.40**	**30.53**	**4.95**
Identities to Video										
ConceptMaster [23]	18.54	-	-	-	39.97	65.63	-	**95.05**	10.14	5.21
FullDiT	**18.64**	-	-	-	**46.22**	**68.59**	-	94.95	**16.68**	**5.46**
Depth to Video										
Ctrl-Adapter‡ [33]	-	-	-	-	-	-	25.63	94.23	15.47	-
ControlVideo [66]	23.38	-	-	-	-	-	30.10	94.44	18.62	**5.91**
FullDiT	**23.40**	-	-	-	-	-	**14.71**	**95.42**	**23.12**	5.26

Quantitative results for single-task video generation. FullDiT was compared to MotionCtrl, CameraCtrl, and CamI2V for camera-to-video generation; ConceptMaste identity-to-video; and Ctrl-Adapter and ControlVideo for depth-to-video. All models were evaluated using their default settings. For consistency, 16 frames were un method, matching the output length of prior models.

The results indicate that FullDiT, despite handling multiple conditioning signals simultaneously, achieved state-of-the-art performance in metrics related to text, camera motion, identity, and depth controls.

In overall quality metrics, the system generally outperformed other methods, although its smoothness was slightly lower than ConceptMaster's. Here the authors comment:

> *'The smoothness of FullDiT is slightly lower than that of ConceptMaster since the calculation of smoothness is based on CLIP similarity between adjacent frames. As FullDiT exhibits significantly greater dynamics compared to ConceptMaster, the smoothness metric is impacted by the large variations between adjacent frames.*
>
> *'For the aesthetic score, since the rating model favors images in painting style and ControlVideo typically generates videos in this style, it achieves a high score in aesthetics.'*

Regarding the qualitative comparison, it might be preferable to refer to the sample videos at the FullDiT project site, since the PDF examples are inevitably static (and also too large to entirely reproduce here).

The first section of the qualitative results in the PDF. Please refer to the source paper for the additional examples, which are too extensive to reproduce here.

The authors comment:

> *'FullDiT demonstrates superior identity preservation and generates videos with better dynamics and visual quality compared to [ConceptMaster]. Since ConceptMaster and FullDiT are trained on the same backbone, this highlights the effectiveness of condition injection with full attention.*

'…The [other] results demonstrate the superior controllability and generation quality of FullDiT compared to existing depth-to-video and camera-to-video methods.'

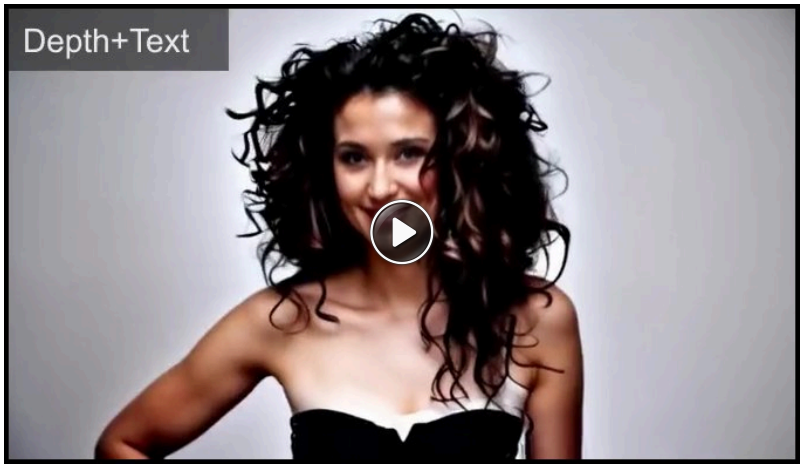
A section of the PDF's examples of FullDiT's output with multiple signals. Please refer to the source paper and the project site for additional examples.

Conclusion

Though FullDiT is an exciting foray into a more full-featured type of video foundation model, one has to wonder if demand for ControlNet-style instrumentalities will ever justify implementing such features at scale, at least for FOSS projects, which would struggle to obtain the enormous amount of GPU processing power necessary, without commercial backing.

The primary challenge is that using systems such as Depth and Pose generally requires non-trivial familiarity with relatively complex user interfaces such as ComfyUI. Therefore it seems that a functional FOSS model of this kind is most likely to be developed by a cadre of smaller VFX companies that lack the money (or the will, given that such systems are quickly made obsolete by model upgrades) to curate and train such a model behind closed doors.

On the other hand, API-driven 'rent-an-AI' systems may be well-motivated to develop simpler and more user-friendly interpretive methods for models into which ancillary control systems have been directly trained.



CLICK TO PLAY VIDEO ON SITE

Click to play. *Depth+Text controls imposed on a video generation using FullDiT.*

** The authors do not specify any known base model (i.e., SDXL, etc.)*

First published Thursday, March 27, 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

Discover More