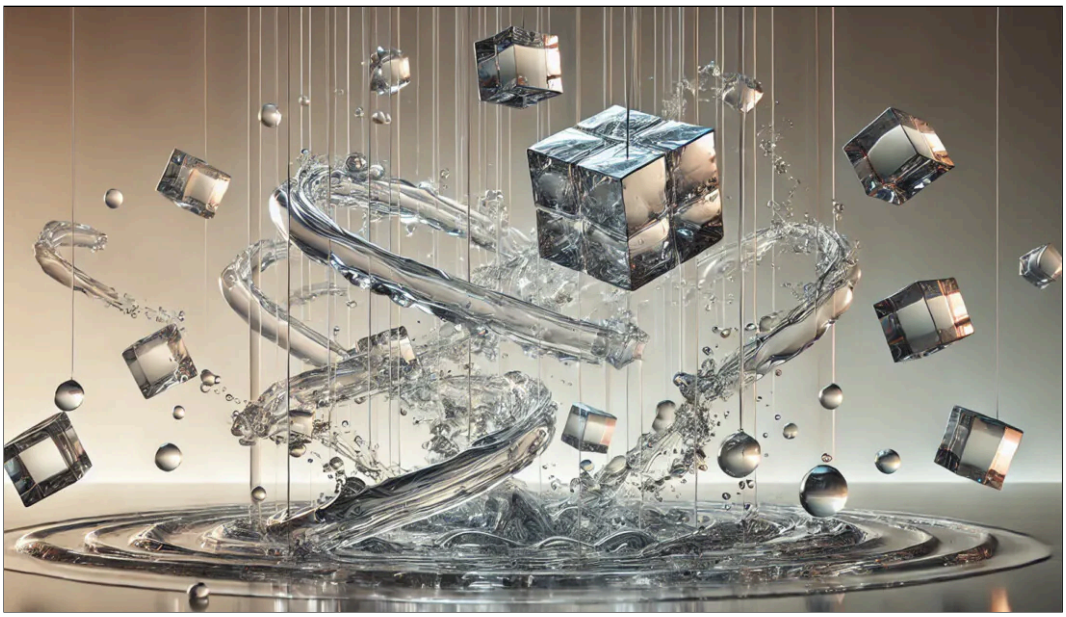
Can AI World Models Really Understand Physical Laws?

Originally published at https://www.unite.ai/can-ai-world-models-really-understand-physical-laws/



The great hope for vision-language AI models is that they will one day become capable of greater autonomy and versatility, incorporating principles of physical laws in much the same way that we develop an innate understanding of these principles through early experience.

For instance, children's ball games tend to develop an understanding of motion kinetics, and of the effect of weight and surface texture on trajectory. Likewise, interactions with common scenarios such as baths, spilled drinks, the ocean, swimming pools and other diverse liquid bodies will instill in us a versatile and scalable comprehension of the ways that liquid behaves under gravity.

Even the postulates of less common phenomena – such as combustion, explosions and architectural weight distribution under pressure – are unconsciously absorbed through exposure to TV programs and movies, or social media videos.

By the time we study the *principles* behind these systems, at an academic level, we are merely 'retrofitting' our intuitive (but uninformed) mental models of them.

Masters of One

Currently, most AI models are, by contrast, more 'specialized', and many of them are either fine-tuned or trained from scratch on image or video datasets that are quite specific to certain use cases, rather than designed to develop such a general understanding of governing laws.

Others can present the *appearance* of an understanding of physical laws; but they may actually be reproducing samples from their training data, rather than really understanding the basics of areas such as motion physics in a way that can produce truly novel (and scientifically plausible) depictions from users' prompts.

At this delicate moment in the productization and commercialization of generative AI systems, it is left to us, and to investors' scrutiny, to distinguish the crafted marketing of new AI models from the reality of their limitations.

One of November's most interesting papers, led by Bytedance Research, tackled this issue, exploring the gap between the apparent and real capabilities of 'all-purpose' generative models such as Sora.

The work concluded that at the current state of the art, generated output from models of this type are more likely to be *aping examples from their training data* than actually demonstrating full understanding of the underlying physical constraints that operate in the real world.

The paper states*:

> *'[These] models can be easily biased by "deceptive" examples from the training set, leading them to generalize in a "case-based" manner under certain conditions. This phenomenon, also observed in large language models, describes a model's tendency to reference similar training cases when solving new tasks.*
>
> *'For instance, consider a video model trained on data of a high-speed ball moving in uniform linear motion. If data augmentation is performed by horizontally flipping the videos, thereby introducing reverse-direction motion, the model may generate a scenario where a low-speed ball reverses direction after the initial frames, even though this behavior is not physically correct.'*

We'll take a closer look at the paper – titled *Evaluating World Models with LLM for Decision Making* – shortly. But first, let's look at the background for these apparent limitations.

Remembrance of Things Past

Without generalization, a trained AI model is little more than an expensive spreadsheet of references to sections of its training data: find the appropriate search term, and you can summon up an instance of that data.

In that scenario, the model is effectively acting as a 'neural search engine', since it cannot produce abstract or 'creative' interpretations of the desired output, but instead replicates some minor variation of data that it saw during the training process.

This is known as memorization – a controversial problem that arises because truly ductile and interpretive AI models tend to lack detail, while truly detailed models tend to lack originality and flexibility.

The capacity for models affected by memorization to reproduce training data is a potential legal hurdle, in cases where the model's creators did not have unencumbered rights to use that data; and where benefits from that data can be demonstrated through a growing number of extraction methods.

Because of memorization, traces of non-authorized data can persist, daisy-chained, through multiple training systems, like an indelible and unintended watermark – even in projects where the machine learning practitioner has taken care to ensure that 'safe' data is used.

World Models

However, the central usage issue with memorization is that it tends to convey the *illusion of intelligence*, or suggest that the AI model has generalized fundamental laws or domains, where in fact it is the high volume of memorized data that furnishes this illusion (i.e., the model has so many potential data examples to choose from that it is difficult for a human to tell whether it is regurgitating learned content or whether it has a truly abstracted understanding of the concepts involved in the generation).

This issue has ramifications for the growing interest in *world models* – the prospect of highly diverse and expensively-trained AI systems that incorporate multiple known laws, and are richly explorable.

World models are of particular interest in the generative image and video space. In 2023 RunwayML began a research initiative into the development and feasibility of such models; DeepMind recently hired one of the originators of the acclaimed Sora generative video to work on a model of this kind; and startups such as Higgsfield are investing significantly in world models for image and video synthesis.

Hard Combinations

One of the promises of new developments in generative video AI systems is the prospect that they can learn fundamental physical laws, such as motion, human kinematics (such as gait characteristics), fluid dynamics, and other known physical phenomena which are, at the very least, visually familiar to humans.

If generative AI could achieve this milestone, it could become capable of producing hyper-realistic visual effects that depict explosions, floods, and plausible collision events across multiple types of object.

If, on the other hand, the AI system has simply been trained on thousands (or hundreds of thousands) of videos depicting such events, it could be capable of reproducing the training data quite convincingly when it was trained on a *similar data point to the user's target query*; yet *fail* if the query combines too many concepts that are, in such a combination, not represented at all in the data.

Further, these limitations would not be immediately apparent, until one pushed the system with challenging combinations of this kind.

This means that a new generative system may be capable of generating viral video content that, while impressive, can create a false impression of the system's capabilities and depth of understanding, because the task it represents is not a real challenge for the system.

For instance, a relatively common and well-diffused event, such as *'a building is demolished'*, might be present *in multiple videos* in a dataset used to train a model that is supposed to have some understanding of physics. Therefore the model could presumably generalize this concept well, and even produce genuinely novel output within the parameters learned from abundant videos.

This is an *in-distribution* example, where the dataset contains many useful examples for the AI system to learn from.

However, if one was to request a more bizarre or specious example, such as *'The Eiffel Tower is blown up by alien invaders'*, the model would be required to combine diverse domains such as 'metallurgical properties', 'characteristics of explosions', 'gravity', 'wind resistance' – and 'alien spacecraft'.

This is an *out-of-distribution* (OOD) example, which combines so many entangled concepts that the system will likely either fail to generate a convincing example, or will default to the nearest semantic example that it was trained on – even if that example does not adhere to the user's prompt.

Excepting that the model's source dataset contained Hollywood-style CGI-based VFX depicting the same or a similar event, such a depiction would absolutely require that it achieve a well-generalized and ductile understanding of physical laws.

Physical Restraints

The new paper – a collaboration between Bytedance, Tsinghua University and Technion – suggests not only that models such as Sora do *not* really internalize deterministic physical laws in this way, but that scaling up the data (a common approach over the last 18 months) appears, in most cases, to produce no real improvement in this regard.

The paper explores not only the limits of extrapolation of specific physical laws – such as the behavior of objects in motion when they collide, or when their path is obstructed – but also a model's capacity for *combinatorial generalization* – instances where the representations of two different physical principles are merged into a single generative output.

CLICK TO PLAY VIDEO ON SITE

A video summary of the new paper. Source: https://x.com/bingyikang/status/1853635009611219019

The three physical laws selected for study by the researchers were *parabolic motion*; *uniform linear motion*; and *perfectly elastic collision*.

As can be seen in the video above, the findings indicate that models such as Sora do not really internalize physical laws, but tend to reproduce training data.

Further, the authors found that facets such as color and shape become so entangled at inference time that a generated ball would likely turn into a square, apparently because a similar motion in a dataset example featured a square and not a ball (see example in video embedded above).

The paper, which has notably engaged the research sector on social media, concludes:

> *'Our study suggests that scaling alone is insufficient for video generation models to uncover fundamental physical laws, despite its role in Sora's broader success…*
>
> *'…[Findings] indicate that scaling alone cannot address the OOD problem, although it does enhance performance in other scenarios.*
>
> *'Our in-depth analysis suggests that video model generalization relies more on referencing similar training examples rather than learning universal rules. We observed a prioritization order of color > size > velocity > shape in this "case-based" behavior.*
>
> *'[Our] study suggests that naively scaling is insufficient for video generation models to discover fundamental physical laws.'*

Asked whether the research team had found a solution to the issue, one of the paper's authors commented:

> *'Unfortunately, we have not. Actually, this is probably the mission of the whole AI community.'*

Method and Data

The researchers used a Variational Autoencoder (VAE) and DiT architectures to generate video samples. In this setup, the compressed latent representations produced by the VAE work in tandem with DiT's modeling of the denoising process.

Videos were trained over the Stable Diffusion V1.5-VAE. The schema was left fundamentally unchanged, with only end-of-process architectural enhancements:

> *'[We retain] the majority of the original 2D convolution, group normalization, and attention mechanisms on the spatial dimensions.*
>
> *'To inflate this structure into a spatial-temporal auto-encoder, we convert the final few 2D downsample blocks of the encoder and the initial few 2D upsample blocks of the decoder into 3D ones, and employ multiple extra 1D layers to enhance temporal modeling.'*

In order to enable video modeling, the modified VAE was jointly trained with HQ image and video data, with the 2D Generative Adversarial Network (GAN) component native to the SD1.5 architecture augmented for 3D.

The image dataset used was Stable Diffusion's original source, LAION-Aesthetics, with filtering, in addition to DataComp. For video data, a subset was curated from the Vimeo-90K, Panda-70m and HDVG datasets.

The data was trained for one million steps, with random resized crop and random horizontal flip applied as data augmentation processes.

Flipping Out

As noted above, the random horizontal flip data augmentation process can be a liability in training a system designed to produce authentic motion. This is because output from the trained model may consider *both* directions of an object, and cause random reversals as it attempts to negotiate this conflicting data (see embedded video above).

On the other hand, if one turns horizontal flipping *off*, the model is then more likely to produce output that adheres to *only one direction* learned from the training data.

So there is no easy solution to the issue, except that the system truly assimilates the entirety of possibilities of movement from both the native and flipped version – a facility that children develop easily, but which is more of a challenge, apparently, for AI models.

Tests

For the first set of experiments, the researchers formulated a 2D simulator to produce videos of object movement and collisions that accord with the laws of classical mechanics, which furnished a high volume and controlled dataset that excluded the ambiguities of real-world videos, for the evaluation of the models. The Box2D physics game engine was used to create these videos.

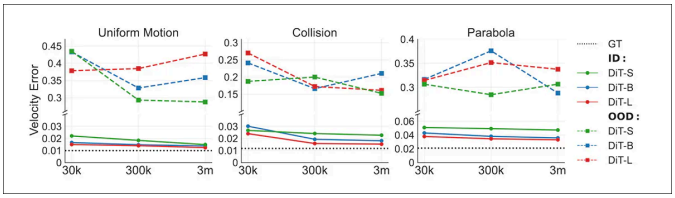
The three fundamental scenarios listed above were the focus of the tests: uniform linear motion, perfectly elastic collisions, and parabolic motion.

Datasets of increasing size (ranging from 30,000 to three million videos) were used to train models of different size and complexity (DiT-S to DiT-L), with the first three frames of each video used for conditioning.

Model	Layers	Hidden size	Heads	#Param
DiT-S	12	384	6	22.5M
DiT-B	12	768	12	89.5M
DiT-L	24	1024	16	310.0M
DiT-XL	28	1152	16	456.0M

Details of the varying models trained in the first set of experiments. Source: https://arxiv.org/pdf/2411.02385

The researchers found that the in-distribution (ID) results scaled well with increasing amounts of data, while the OOD generations did not improve, indicating shortcomings in generalization.



Results for the first round of tests.

The authors note:

> *'These findings suggest the inability of scaling to perform reasoning in OOD scenarios.'*

Next, the researchers tested and trained systems designed to exhibit a proficiency for combinatorial generalization, wherein two contrasting movements are combined to (hopefully) produce a cohesive movement that is faithful to the physical law behind each of the separate movements.

For this phase of the tests, the authors used the PHYRE simulator, creating a 2D environment which depicts multiple and diversely-shaped objects in free-fall, colliding with each other in a variety of complex interactions.

Evaluation metrics for this second test were Fréchet Video Distance (FVD); Structural Similarity Index (SSIM); Peak Signal-to-Noise Ratio (PSNR); Learned Perceptual Similarity Metrics (LPIPS); and a human study (denoted as 'abnormal' in results).

Three scales of training datasets were created, at 100,000 videos, 0.6 million videos, and 3-6 million videos. DiT-B and DiT-XL models were used, due to the increased complexity of the videos, with the first frame used for conditioning.

The models were trained for one million steps at 256×256 resolution, with 32 frames per video.

Model	#Templates	FVD (↓)	SSIM (↑)	PSNR (↑)	LPIPS (↓)	Abnormal (↓)
DiT-XL	6	18.2 / 22.1	**0.973** / 0.943	**32.8** / 25.5	**0.028** / 0.082	3% / 67%
DiT-XL	30	19.5 / 19.7	0.973 / 0.950	32.7 / 27.1	0.028 / 0.065	3% / 18%
DiT-XL	60	**17.6** / **18.7**	0.972 / **0.951**	32.4 / **27.3**	0.030 / **0.062**	**2%** / **10%**
DiT-B	60	18.4 / 21.4	0.967 / 0.949	30.9 / 27.0	0.035 / 0.066	3% / 24%

Results for the second round of tests.

The outcome of this test suggests that merely increasing data volume is an inadequate approach:

The paper states:

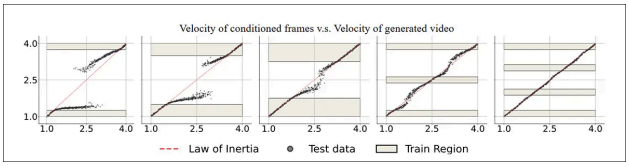
> *'These results suggest that both model capacity and coverage of the combination space are crucial for combinatorial generalization. This insight implies that scaling laws for video generation should focus on increasing combination diversity, rather than merely scaling up data volume.'*

Finally, the researchers conducted further tests to attempt to determine whether a video generation models can truly assimilate physical laws, or whether it simply memorizes and reproduces training data at inference time.

Here they examined the concept of 'case-based' generalization, where models tend to mimic specific training examples when confronting novel situations, as well as examining examples of uniform motion – specifically, how the direction of motion in training data influences the trained model's predictions.

Two sets of training data, for *uniform motion* and *collision*, were curated, each consisting of uniform motion videos depicting velocities between 2.5 to 4 units, with the first three frames used as conditioning. Latent values such as *velocity* were omitted, and, after training, testing was performed on both seen and unseen scenarios.

Below we see results for the test for uniform motion generation:

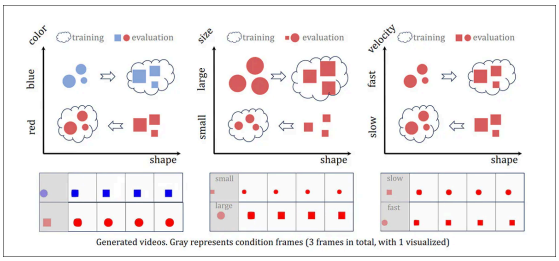


Results for tests for uniform motion generation, where the 'velocity' variable is omitted during training.

The authors state:

> *'[With] a large gap in the training set, the model tends to generate videos where the velocity is either high or low to resemble training data when initial frames show middle-range velocities.'*

For the collision tests, far more variables are involved, and the model is required to learn a two-dimensional non-linear function.



Collision: results for the third and final round of tests.

The authors observe that the presence of 'deceptive' examples, such as reversed motion (i.e., a ball that bounces off a surface and reverses its course), can mislead the model and cause it to generate physically incorrect predictions.

Conclusion

If a non-AI algorithm (i.e., a 'baked', procedural method) contains *mathematical rules* for the behavior of physical phenomena such as fluids, or objects under gravity, or under pressure, there are a set of unchanging constants available for accurate rendering.

However, the new paper's findings indicate that no such equivalent relationship or intrinsic understanding of classical physical laws is developed during the training of generative models, and that increasing amounts of data do not resolve the problem, but rather obscure it –because a greater number of training videos are available for the system to imitate at inference time.

* *My conversion of the authors' inline citations to hyperlinks.*

First published Tuesday, November 26, 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