Why AI Video Sometimes Gets It Backwards

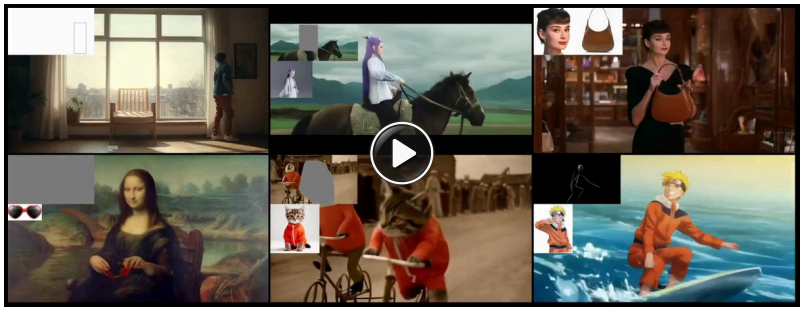
Originally published at https://www.unite.ai/why-ai-video-sometimes-gets-it-backwards/

If 2022 was the year that generative AI captured a wider public's imagination, 2025 is the year where the new breed of generative *video* frameworks coming from China seems set to do the same.

Tencent's Hunyuan Video has made a major impact on the hobbyist AI community with its open-source release of a full-world video diffusion model that users can tailor to their needs.

Close on its heels is Alibaba's more recent Wan 2.1, one of the most powerful image-to-video FOSS solutions of this period – now supporting customization through Wan LoRAs.

Besides the availability of recent human-centric foundation model SkyReels, at the time of writing we also await the release of Alibaba's comprehensive VACE video creation and editing suite:



CLICK TO PLAY VIDEO ON SITE

Click to play. *The pending release of Alibaba's multi-function AI-editing suite VACE has excited the user community.* Source: https://ali-vilab.github.io/VACE-Page/

Sudden Impact

The generative video AI research scene itself is no less explosive; it's still the first half of March, and Tuesday's submissions to Arxiv's Computer Vision section (a hub for generative AI papers) came to nearly 350 entries – a figure more associated with the height of conference season.

The two years since the launch of Stable Diffusion in summer of 2022 (and the subsequent development of Dreambooth and LoRA customization methods) have been characterized by the lack of further major developments, until the last few weeks, where new releases and innovations have proceeded at such a breakneck pace that it is almost impossible to keep apprised of it all, much less cover it all.

Video diffusion models such as Hunyuan and Wan 2.1 have solved, at long last, and after years of failed efforts from hundreds of research initiatives, the problem of temporal consistency as it relates to the generation of humans, and largely also to environments and objects.

There can be little doubt that VFX studios are currently applying staff and resources to adapting the new Chinese video models to solve immediate challenges such as face-swapping, despite the current lack of ControlNet-style ancillary mechanisms for these systems.

It must be such a relief that one such significant obstacle has potentially been overcome, albeit not through the avenues anticipated.

Of the problems that remain, this one, however, is not insignificant:

CLICK TO PLAY VIDEO ON SITE

Click to play. *Based on the prompt 'A small rock tumbles down a steep, rocky hillside, displacing soil and small stones ', Wan 2.1, which achieved the very highest scores in the new paper, makes one simple error.* Source: https://videophy2.github.io/

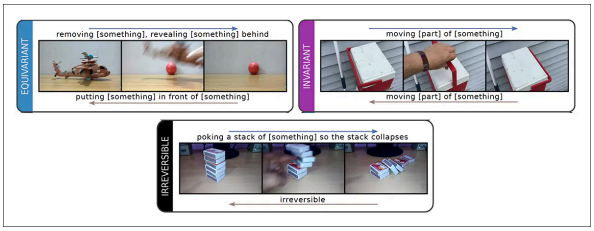
Up The Hill Backwards

All text-to-video and image-to-video systems currently available, including commercial closed-source models, have a tendency to produce physics bloopers such as the one above, where the video shows a rock rolling *uphill*, based on the prompt '*A small rock tumbles down a steep, rocky hillside, displacing soil and small stones* '.

One theory as to why this happens, recently proposed in an academic collaboration between Alibaba and UAE, is that models train always on single images, in a sense, even when they're training on videos (which are written out to single-frame sequences for training purposes); and they may not necessarily learn the correct temporal order of *'before'* and *'after'* pictures.

However, the most likely solution is that the models in question have used data augmentation routines that involve exposing a source training clip to the model both forwards *and* backwards, effectively doubling the training data.

It has long been known that this shouldn't be done arbitrarily, because some movements work in reverse, but many do not. A 2019 study from the UK's University of Bristol sought to develop a method that could distinguish *equivariant*, *invariant* and *irreversible* source data video clips that co-exist in a single dataset (see image below), with the notion that unsuitable source clips might be filtered out from data augmentation routines.



Examples of three types of movement, only one of which is freely reversible while maintaining plausible physical dynamics. Source: https://arxiv.org/abs/1909.09422

The authors of that work frame the problem clearly:

> *'We find the realism of reversed videos to be betrayed by reversal artefacts, aspects of the scene that would not be possible in a natural world. Some artefacts are subtle, while others are easy to spot, like a reversed 'throw' action where the thrown object spontaneously rises from the floor.*
>
> *'We observe two types of reversal artefacts, physical, those exhibiting violations of the laws of nature, and improbable, those depicting a possible but unlikely scenario. These are not exclusive, and many reversed actions suffer both types of artefacts, like when uncrumpling a piece of paper.*
>
> *'Examples of physical artefacts include: inverted gravity (e.g. 'dropping something'), spontaneous impulses on objects (e.g. 'spinning a pen'), and irreversible state changes (e.g. 'burning a candle'). An example of an improbable artefact: taking a plate from the cupboard, drying it, and placing it on the drying rack.*
>
> *'This kind of re-use of data is very common at training time, and can be beneficial – for example, in making sure that the model does not learn only one view of an image or object which can be flipped or rotated without losing its central coherency and logic.*
>
> *'This only works for objects that are truly symmetrical, of course; and learning physics from a 'reversed' video only works if the reversed version makes as much sense as the forward version.'*

Temporary Reversals

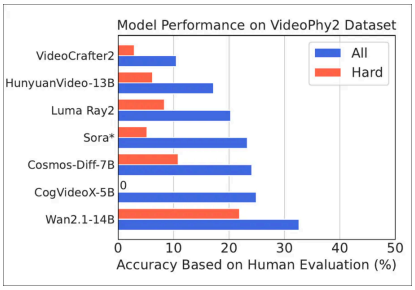
We don't have any evidence that systems such as Hunyuan Video and Wan 2.1 allowed arbitrarily 'reversed' clips to be exposed to the model during training (neither group of researchers has been specific regarding data augmentation routines).

Yet the only reasonable alternative possibility, in the face of so many reports (and my own practical experience), would seem to be that hyperscale datasets powering these model may contain clips that actually feature movements occurring in reverse.

The rock in the example video embedded above was generated using Wan 2.1, and features in a new study that examines how well video diffusion models handle physics.

In tests for this project, Wan 2.1 achieved a score of only 22% in terms of its ability to consistently adhere to physical laws.

However, that's the *best* score of any system tested for the work, indicating that we may have found our next stumbling block for video AI:



Scores obtained by leading open and closed-source systems, with the output of the frameworks evaluated by human annotators. Source: https://arxiv.org/pdf/2503.06800

The authors of the new work have developed a benchmarking system, now in its second iteration, called *VideoPhy*, with the code available at GitHub.

Though the scope of the work is beyond what we can comprehensively cover here, let's take a general look at its methodology, and its potential for establishing a metric that could help steer the course of future model-training sessions away from these bizarre instances of reversal.

The study, conducted by six researchers from UCLA and Google Research, is called *VideoPhy-2: A Challenging Action-Centric Physical Commonsense Evaluation in Video Generation*. A crowded accompanying project site is also available, along with code and datasets at GitHub, and a dataset viewer at Hugging Face.

CLICK TO PLAY VIDEO ON SITE

Click to play. *Here, the feted OpenAI Sora model fails to understand the interactions between oars and reflections, and is not able to provide a logical physical flow either for the person in the boat or the way that the boat interacts with her.*

Method

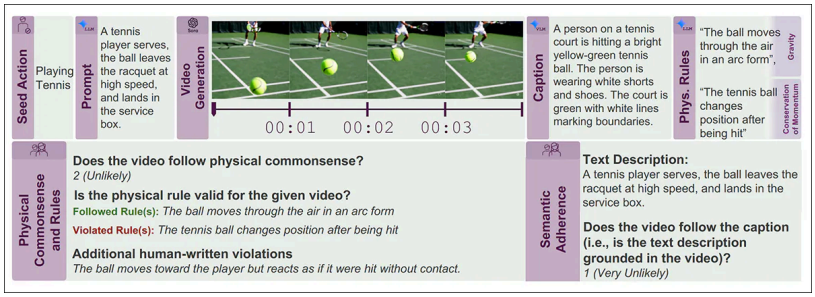
The authors describe the latest version of their work, *VideoPhy-2*, as a 'challenging commonsense evaluation dataset for real-world actions.' The collection features 197 actions across a range of diverse physical activities such as *hula-hooping*, *gymnastics* and *tennis*, as well as object interactions, such as *bending an object until it breaks*.

A large language model (LLM) is used to generate 3840 prompts from these seed actions, and the prompts are then used to synthesize videos via the various frameworks being trialed.

Throughout the process the authors have developed a list of 'candidate' physical rules and laws that AI-generated videos should satisfy, using vision-language models for evaluation.

The authors state:

> *'For example, in a video of sportsperson playing tennis, a physical rule would be that a tennis ball should follow a parabolic trajectory under gravity. For gold-standard judgments, we ask human annotators to score each video based on overall semantic adherence and physical commonsense, and to mark its compliance with various physical rules.'*



Above: A text prompt is generated from an action using an LLM and used to create a video with a text-to-video generator. A vision-language model captions the vi rules at play. Below: Human annotators evaluate the video's realism, confirm rule violations, add missing rules, and check whether the video matches the original p

Initially the researchers curated a set of actions to evaluate physical commonsense in AI-generated videos. They began with over 600 actions sourced from the Kinetics, UCF-101, and SSv2 datasets, focusing on activities involving sports, object interactions, and real-world physics.

Two independent groups of STEM-trained student annotators (with a minimum undergraduate qualification obtained) reviewed and filtered the list, selecting actions that tested principles such as *gravity*, *momentum*, and *elasticity*, while removing low-motion tasks such as *typing*, *petting a cat*, or *chewing*.

After further refinement with Gemini-2.0-Flash-Exp to eliminate duplicates, the final dataset included 197 actions, with 54 involving object interactions and 143 centered on physical and sports activities:

mopping floor, blowing out candles, throwing water balloon, passing american football, billiards, lifting a surface with something on it until it starts sliding down, using a sledge hammer, swing, hitting baseball, hammerthrow, playing tennis, longjump, sewing, roller skating, wading through water, riding mechanical bull, pole vault, blowdryhair, tying necktie, paragliding, something falling like a rock, playing ice hockey, sailing, gymnastics tumbling, pushing something so that it slightly moves, folding clothes, poking something so lightly that it doesn't or almost doesn't move, javelinthrow, surfing, snatch weight lifting, chiseling stone, spinning something that quickly stops spinning, poking a stack of something without the stack collapsing, carving ice, throwing something, twisting (wringing) something wet until water comes out, putting something onto something else that cannot support it so it falls down, wiping something off of something, playing field hockey, juggling balls, wading through mud, shooting basketball, welding, hoverboarding, javelin throw, catching or throwing softball, hammering, knitting, putting something that can't roll onto a slanted surface so it slides down, playing polo, pouring something onto something, pulling two ends of something so that it gets stretched, using circular saw, flint knapping, backflip (human), long jump, uncovering something, pouring something into something until it overflows, dribbling basketball, poking a hole into some substance, bulldozing, peeling fruit, parallelbars, playing darts, spinning something so it continues spinning, luge, pushing something so it spins, curling (sport), riding unicycle, throwing discus, folding paper, ripping paper, trapezing, playing pinball, burying something in something, throwing axe, wrapping present, yarn spinning, tying shoe laces, flying kite, tightrope walking, using a paint roller, using a wrench, sharpening pencil, pizzatossing, catching or throwing baseball, catching or throwing frisbee, playing kickball, golf, nunchucks, pouring something out of something, opening bottle (not

Samples from the distilled actions.

In the second stage, the researchers used Gemini-2.0-Flash-Exp to generate 20 prompts for each action in the dataset, resulting in a total of 3,940 prompts. The generation process focused on visible physical interactions that could be clearly represented in a generated video. This excluded non-visual elements such as *emotions*, *sensory details*, and *abstract language*, but incorporated diverse characters and objects.

For example, instead of a simple prompt like '*An archer releases the arrow*', the model was guided to produce a more detailed version such as *'An archer draws the bowstring back to full tension, then releases the arrow, which flies straight and strikes a bullseye on a paper target*'.

Since modern video models can interpret longer descriptions, the researchers further refined the captions using the Mistral-NeMo-12B-Instruct prompt upsampler, to add visual details without altering the original meaning.

Action	Prompt	Physical Principle	Category
Canoeing	A person uses a kayak paddle to push their kayak up the bank of a river.	Buoyancy	Physical Activity
Riding Unicycle	A person on a unicycle stops abruptly, putting a foot down to regain balance.	Inertia	Physical Activity
Pushing something so it spins	A chef pushes a pizza spinning on a tray with their hand.	Conversation of Momentum	Object Interaction

Sample prompts from VideoPhy-2, categorized by physical activities or object interactions. Each prompt is paired with its corresponding action and the relevant physical principle it tests.

For the third stage, physical rules were not derived from text prompts but from generated videos, since generative models can struggle to adhere to conditioned text prompts.

Videos were first created using VideoPhy-2 prompts, then 'up-captioned' with Gemini-2.0-Flash-Exp to extract key details. The model proposed three expected physical rules per video, which human annotators reviewed and expanded by identifying additional potential violations.

Original Caption	Upsampled Caption
A person uses nunchucks to break a stack of wooden blocks.	In a dynamic display of martial arts prowess, a skilled practitioner wields a pair of nunchucks, their hands clad in black gloves that enhance grip and safety. The scene unfolds in a dimly lit, industrial setting, where a stack of wooden blocks, meticulously arranged in a pyramid, awaits the test of strength. The camera captures the moment with a static focus, highlighting the nunchucks' fluid motion as they swing through the air, their polished wooden surfaces glinting under the soft, ambient light.
A player throws a softball sidearm, and the ball spins as it travels through the air.	In a sun-drenched outdoor sports arena, a dynamic softball game unfolds, captured with cinematic precision. The camera, positioned at a low angle, focuses on a player in a sleek black jersey, their arm poised in a powerful sidearm throwing motion. The softball, a vibrant white against the lush green field, spins rapidly as it leaves their hand, tracing a graceful arc through the air.

Examples from the upsampled captions.

Next, to identify the most challenging actions, the researchers generated videos using CogVideoX-5B with prompts from the VideoPhy-2 dataset. They then selected 60 actions out of 197 where the model consistently failed to follow both the prompts and basic physical commonsense.

These actions involved physics-rich interactions such as momentum transfer in discus throwing, state changes such as bending an object until it breaks, balancing tasks such as tightrope walking, and complex motions that included back-flips, pole vaulting, and pizza tossing, among others. In total, 1,200 prompts were chosen to increase the difficulty of the sub-dataset.

The resulting dataset comprised 3,940 captions – 5.72 times more than the earlier version of VideoPhy. The average length of the original captions is 16 tokens, while upsampled captions reaches 138 tokens – 1.88 times and 16.2 times longer, respectively.

The dataset also features 102,000 human annotations covering semantic adherence, physical commonsense, and rule violations across multiple video generation models.

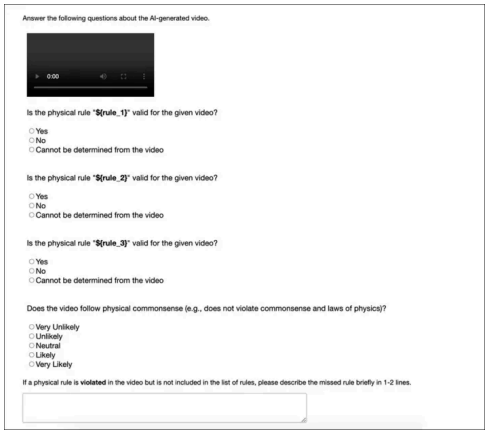
Evaluation

The researchers then defined clear criteria for evaluating the videos. The main goal was to assess how well each video matched its input prompt and followed basic physical principles.

Instead of simply ranking videos by preference, they used rating-based feedback to capture specific successes and failures. Human annotators scored videos on a five-point scale, allowing for more detailed judgments, while the evaluation also checked whether videos followed various physical rules and laws.

For human evaluation, a group of 12 annotators were selected from trials on Amazon Mechanical Turk (AMT), and provided ratings after receiving detailed remote instructions. For fairness, *semantic adherence* and *physical commonsense* were evaluated separately (in the original VideoPhy study, they were assessed jointly).

The annotators first rated how well videos matched their input prompts, then separately evaluated physical plausibility, scoring rule violations and overall realism on a five-point scale. Only the original prompts were shown, to maintain a fair comparison across models.



The interface presented to the AMT annotators.

Though human judgment remains the gold standard, it's expensive and comes with a number of caveats. Therefore automated evaluation is essential for faster and more scalable model assessments.

The paper's authors tested several video-language models, including Gemini-2.0-Flash-Exp and VideoScore, on their ability to score videos for semantic accuracy and for 'physical commonsense'.

The models again rated each video on a five-point scale, while a separate classification task determined whether physical rules were followed, violated, or unclear.

Experiments showed that existing video-language models struggled to match human judgments, mainly due to weak physical reasoning and the complexity of the prompts. To improve automated evaluation, the researchers developed *VideoPhy-2-Autoeval*, a 7B-parameter model designed to provide more accurate predictions across three categories: *semantic adherence*; *physical commonsense*; and *rule compliance*, fine-tuned on the VideoCon-Physics model using 50,000 human annotations*.

Data and Tests

With these tools in place, the authors tested a number of generative video systems, both through local installations and, where necessary, via commercial APIs: CogVideoX-5B; VideoCrafter2; HunyuanVideo-13B; Cosmos-Diffusion; Wan2.1-14B; OpenAI Sora; and Luma Ray.

The models were prompted with upsampled captions where possible, except that Hunyuan Video and VideoCrafter2 operate under 77-token CLIP limitations, and cannot accept prompts above a certain length.

Videos generated were kept to less than 6 seconds, since shorter output is easier to evaluate.

The driving data was from the VideoPhy-2 dataset, which was split into a benchmark and training set. 590 videos were generated per model, except for Sora and Ray2; due to the cost factor (equivalent lower numbers of videos were generated for these).

(Please refer to the source paper for further evaluation details, which are exhaustively chronicled there)

The initial evaluation dealt with *physical activities/sports* (PA) and *object interactions* (OI), and tested both the general dataset and the aforementioned 'harder' subset:

Model	Class	All	Hard	PA	OI
Wan2.1-14B [55]	Open	32.6	21.9	31.5	36.2
CogVideoX-5B [58]	Open	25.0	0.0	24.6	26.1
Cosmos-Diff-7B [1]	Open	24.1	10.9	22.6	27.4
Hunyuan-13B [31]	Open	17.2	6.2	17.6	15.9
VideoCrafter-2 [14]	Open	10.5	2.9	10.1	13.1
Ray2 [40]	Closed	20.3	8.3	21.0	18.5
Sora [10]	Closed	23.3	5.3	22.2	26.7

Results from the initial round.

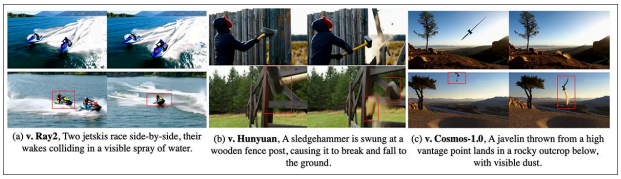
Here the authors comment:

> *'Even the best-performing model, Wan2.1-14B, achieves only 32.6% and 21.9% on the full and hard splits of our dataset, respectively. Its relatively strong performance compared to other models can be attributed to the diversity of its multimodal training data, along with robust motion filtering that preserves high-quality videos across a wide range of actions.*
>
> *'Furthermore, we observe that closed models, such as Ray2, perform worse than open models like Wan2.1-14B and CogVideoX-5B. This suggests that closed models are not necessarily superior to open models in capturing physical commonsense.*
>
> *'Notably, Cosmos-Diffusion-7B achieves the second-best score on the hard split, even outperforming the much larger HunyuanVideo-13B model. This may be due to the high representation of human actions in its training data, along with synthetically rendered simulations.'*

The results showed that video models struggled more with physical activities like sports than with simpler object interactions. This suggests that improving AI-generated videos in this area will require better datasets – particularly high-quality footage of sports such as tennis, discus, baseball, and cricket.

The study also examined whether a model's physical plausibility correlated with other video quality metrics, such as aesthetics and motion smoothness. The findings revealed no strong correlation, meaning a model cannot improve its performance on VideoPhy-2 just by generating visually appealing or fluid motion – it needs a deeper understanding of physical commonsense.

Though the paper provides abundant qualitative examples, few of the static examples provided in the PDF seem to relate to the extensive video-based examples that the authors furnish at the project site. Therefore we will look at a small selection of the static examples and then some more of the actual project videos.

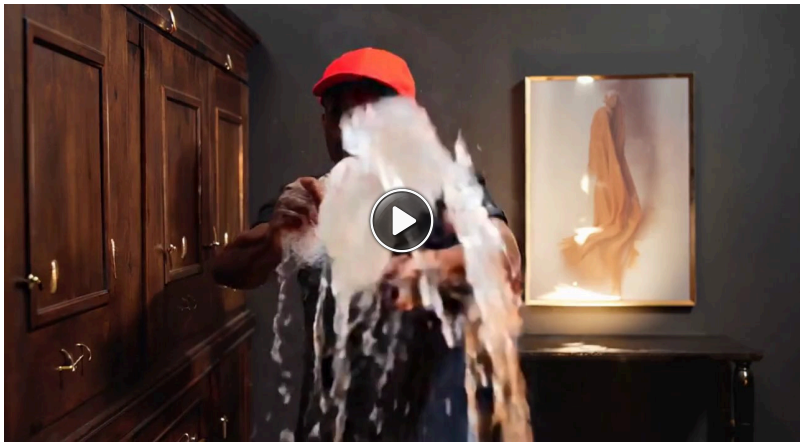


The top row shows videos generated by Wan2.1. (a) In Ray2, the jet-ski on the left lags behind before moving backward. (b) In Hunyuan-13B, the sledgehammer deforms mid-swing, and a broken wooden board appears unexpectedly. (c) In Cosmos-7B, the javelin expels sand before making contact with the ground.

Regarding the above qualitative test, the authors comment:

> *'[We] observe violations of physical commonsense, such as jetskis moving unnaturally in reverse and the deformation of a solid sledgehammer, defying the principles of elasticity. However, even Wan suffers from the lack of physical commonsense, as shown in [the clip embedded at the start of this article].*
>
> *'In this case, we highlight that a rock starts rolling and accelerating uphill, defying the physical law of gravity.'*

Further examples from the project site:



CLICK TO PLAY VIDEO ON SITE

Click to play. *Here the caption was 'A person vigorously twists a wet towel, water spraying outwards in a visible arc' – but the resulting source of water is far more like a water-hose than a towel.*

CLICK TO PLAY VIDEO ON SITE

Click to play. *Here the caption was 'A chemist pours a clear liquid from a beaker into a test tube, carefully avoiding spills', but we can see that the volume of water being added to the beaker is not consistent with the amount exiting the jug.*

As I mentioned at the outset, the volume of material associated with this project far exceeds what can be covered here. Therefore please refer to the source paper, project site and related sites mentioned earlier, for a truly exhaustive outline of the authors' procedures, and considerably more testing examples and procedural details.

** As for the provenance of the annotations, the paper only specifies 'acquired for these tasks' – it seems a lot to have been generated by 12 AMT workers.*

First published Thursday, March 13, 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