Do generative video models understand physical principles?

Saman Motamed[†]
INSAIT, Sofia University "St. Kliment Ohridski"

Laura Culp
Google DeepMind

Kevin Swersky
Google DeepMind

Priyank Jaini[*]
Google DeepMind

Robert Geirhos[*]
Google DeepMind

Abstract

AI video generation is undergoing a revolution, with quality and realism advancing rapidly. These advances have led to a passionate scientific debate: Do video models learn "world models" that discover laws of physics—or, alternatively, are they merely sophisticated pixel predictors that achieve visual realism without understanding the physical principles of reality? We address this question by developing Physics-IQ, a comprehensive benchmark dataset that can only be solved by acquiring a deep understanding of various physical principles, like fluid dynamics, optics, solid mechanics, magnetism and thermodynamics. We find that across a range of current models (Sora, Runway, Pika, Lumiere, Stable Video Diffusion, and VideoPoet), physical understanding is severely limited, and unrelated to visual realism. At the same time, some test cases can already be successfully solved. This indicates that acquiring certain physical principles from observation alone may be possible, but significant challenges remain. While we expect rapid advances ahead, our work demonstrates that visual realism does not imply physical understanding. Our project page is at Physics-IQ website; code at Physics-IQ benchmark.

1. Introduction

Can a machine truly understand the world without interacting with it? This question lies at the heart of the ongoing debate surrounding the capabilities of AI video generation models. While the generation of realistic videos has, for a long time, been considered one of the major unsolved challenges within deep learning, this recently changed. Within a relatively short period of time, the field has seen the development of impressive video generation models [16, 45, 47], capturing the imagination of the public and researchers alike. A major milestone towards general-purpose artificial intelligence is to build machines that understand the world, and if you cannot understand what you cannot create (as Feynman would say), then the ability of those models to create visually realistic scenes is an essential step towards that capability. However, the degree to which successful *generation* signals successful *understanding* is the subject of a passionate debate. Is it possible to understand the world without ever interacting with it? Phrased differently, do generative video models learn the physical principles that underpin reality purely from "watching" videos?

Proponents argue that the way the models are trained—predicting how videos continue, a.k.a. next frame prediction—is a task that forces models to understand physical principles. According to this line of argument, it is hard to predict the next frame if the model has no understanding of how objects move (trajectories), that things fall down instead of up (gravity), and how pouring juice into a glass of water changes its color (fluid dynamics). As an analogy, large language models (LLMs) are trained in a similar fashion to predict the next tokens; a task formulation that is equally simple but has proven sufficient to enable impressive text understanding. Moreover, predicting the future is a core principle of biological perception, too: The brain constantly generates predictions about incoming sensory input, enabling energy-efficient processing of information [9] and building a mental model of the world as postulated by von Helmholtz [68] and later the predictive coding hypothesis [18]. In short, successful prediction signals successful understanding.

On the other hand, there are also important arguments contra understanding through observation. According to the causality rationale, "watching" videos (or to be more precise, training models to predict how videos continue) is a passive process, with models unable to interact with the world. This lack of interaction means that a model cannot observe the causal effects of an intervention (as, for instance, children are able to when playing with toys). Therefore, a model is faced with the nearly impossible task of distinguishing correlation from causation if it is to succeed in understanding physical principles. Furthermore, video models that are touted as "a promising path towards build-

[†]Work done while at Google DeepMind.
[*]Joint last authors.

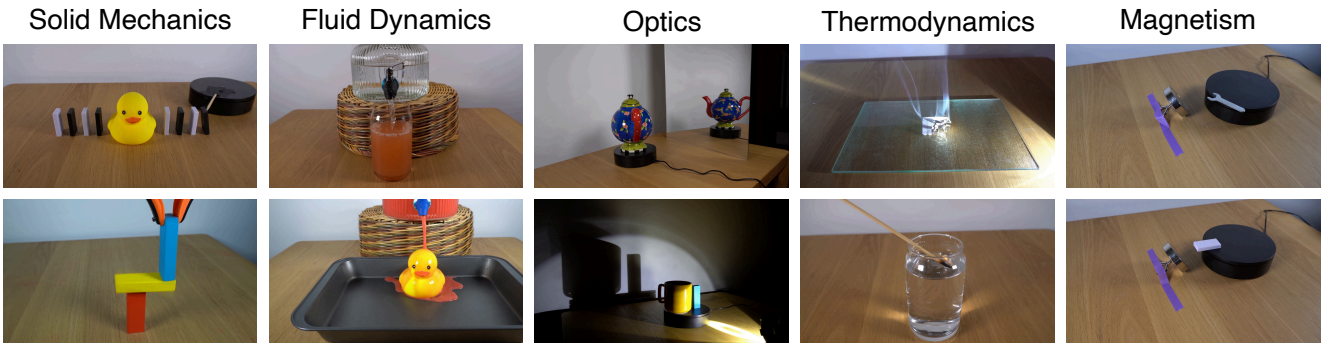

Figure 1. Sample scenarios from the Physics-IQ dataset for testing physical understanding in generative video models. Models are shown the beginning of a video (single frame for image2video models; 3 seconds for video2video models) and need to predict how the video continues over the next 5 seconds, which requires understanding different physical properties: Solid Mechanics, Fluid Dynamics, Optics, Thermodynamics, and Magnetism. See here for an animated version of this figure.

ing general purpose simulators of the physical world" [47] arguably experience a different world to begin with: the digital world as opposed to the real world that an embodied system (like a robot, or virtually all living beings) experience. As a consequence, skeptics argue that visual realism by no means signals true understanding: All it takes to produce realistic videos is to reproduce common patterns from the model's vast sea of training data—shortcuts without understanding [21, 32].

In light of these two diametrically opposed perspectives, how can we tell whether generative video models indeed learn physical principles? To address this question in a quantifiable, tractable way, we created a challenging testbed for physical understanding in video models: the "Physics-IQ" benchmark. The core idea is to prompt video models to do what they do best: predict the continuation of a video. In order to test understanding, we designed a range of diverse scenarios where predicting the continuation requires a deep understanding of physical principles, going beyond pattern reproduction and testing out-of-distribution generalization. For instance, models are asked to predict how a domino chain falls—normally, vs. when a rubber duck is placed in the middle of the chain; or how pillows react when a kettlebell vs. a piece of paper is dropped onto the pillow. The diverse set of scenarios encompass solid mechanics, fluid dynamics, optics, thermodynamics and magnetism, totalling 396 high-quality videos that we filmed from three different perspectives in a controlled environment. Samples are shown in 1. We then compare the model's prediction to the ground truth continuation using a set of simple metrics that capture different desiderata, and analyze a range of current models: Sora [47], Runway Gen 3 [63], Pika 1.0 [62], Lumiere [7], Stable Video Diffusion [12], and VideoPoet [35].

2. Physics-IQ benchmark

2.1. Dataset

Our goal is to develop a dataset that tests the physical understanding capabilities of video generative models on different physical laws like solid mechanics, fluid dynamics, optics, thermodynamics, and magnetism. We therefore created the Physics-IQ dataset which consists of 396 videos (8 seconds each) covering 66 different physical scenarios. Each scenario dataset focuses on a specific physical law and aims to test a video generative model's understanding of physical events. These events include examples like collisions, object continuity, occlusion, object permanence, fluid dynamics, chain reactions, trajectories under the influence of forces (e.g., gravity), material properties and reactions, as well as shadows, reflections, and magnetism.

Each scenario was filmed at 30 frames per second (FPS) with a resolution of 3840×2160 (16:9 aspect ratio) from three different perspectives: left, center, and right using high-quality Sony Alpha a6400 cameras equipped with 16-50mm lenses. Each scenario was shot twice (*take 1* and *take 2*) under identical conditions to capture the inherent variability of real-world physical interactions. These variations are expected in real-world due to factors like chaotic motion, subtle changes in friction, and variations in force trajectory. In this paper, we refer to the differences observed between these two recordings of the same scenario as *physical variance*. This results in a total of 396 videos (66 scenarios $\times$ 3 perspectives $\times$ 2 takes). All our videos are shot from a static camera perspective without camera motion. The setup for filming the videos is illustrated in Supp. 7.

2.2. Evaluation protocol

Physical understanding can be measured in different ways. One of the most stringent tests is whether a model can predict how a challenging, unusual video continues—such as a

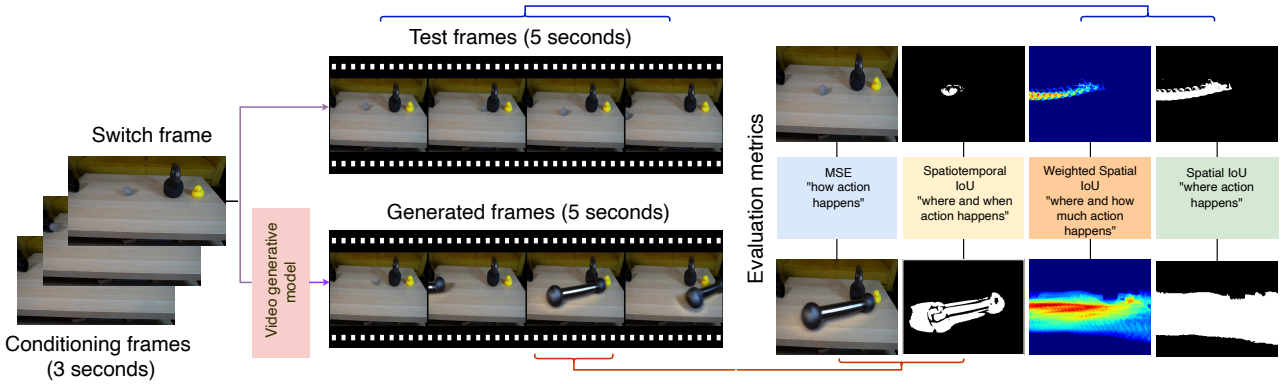

Figure 2. Overview of the Physics-IQ evaluation protocol. A video generative model produces a 5 second continuation of the conditioning frame(s), optionally including a textual description of the conditioning frames for models that accept text input. They are compared against the ground truth test frames using four metrics that quantify different properties of physical understanding. The metrics are defined and explained in the methods section.

chain of dominoes with a rubber duck in the middle interrupting the chain. Out-of-distribution scenarios like these test true understanding, since they cannot be solved by reproducing patterns seen or memorized from the training data [e.g. 20, 21, 25, 60]. We therefore test physical understanding of video generative models by taking a full video of 8 seconds in which a physically interesting event occurs, splitting the video into a 3-second conditioning video and a 5-second ground truth continuation video. The model is then given the conditioning signal: either the 3-second video for video2video models (named *multiframe models* in figures), or the last frame of this conditioning video—called the *switch frame*—in the case of image2video models (named *i2v models* in figures). Since video models are trained precisely to generate the next frames given the previous frame(s) as conditioning signal, our evaluation protocol matches the paradigm these models were trained for. The switch frame is a careful manual selection for each scenario such that enough information about the physical event and objects in the scenario is provided, while at the same time making sure that successfully predicting the continuation requires some understanding of physics (e.g., in the scenario involving the chain reaction when a domino falls, the switch frame corresponds to the moment when the first domino is tipped but has not yet contacted the second domino). We provide video models that support multiframe conditioning with as many conditioning frames (up to a maximum of 3 seconds) as they can accommodate. Some video models (e.g., Runway Gen 3, Pika 1.0, and Sora) generate subsequent frames based on a single image. For these models, we provide just the switch frame as the conditioning signal. Supp. 11 shows the switch frame for all scenarios in the Physics-IQ dataset.

Both multiframe and single-frame conditioned video models can additionally be conditioned on a human-written text description of the scene that describes the conditioning part without, however, giving away the answer of how the future unfolds. For evaluating image-to-video (i2v) and multiframe video models, we provide both these captions and the conditioning frame(s) as conditioning signals. Stable Video Diffusion is the only model in our study that does not accept text as a conditioning signal.

2.3. Why create a real-world Physics-IQ dataset

The question of whether video generative models can understand physical principles has been explored through a range of benchmarks designed to evaluate physical reasoning. Physion [10] and its successor Physion++ [65] use object collisions and stability to assess a model's ability to predict physical outcomes and infer relevant properties of objects (e.g., mass, friction) during dynamic interactions. Similarly, CRAFT [3] and IntPhys [52] assess causal reasoning and intuitive physics, testing whether models can infer forces or understand object permanence. Intuitive physics has a rich history in Cognitive Science and is concerned with understanding how humans build a commonsense intuition for physical principles [e.g. 2, 23, 34, 36, 42, 43, 48, 53, 58, 59, 64]. Recent efforts have extended physical reasoning evaluation to generative video models. VideoPhy [6] and PhyGenBench [44] focus on assessing physical commonsense through text-based descriptions rather than visual data. These works emphasize logical reasoning about physical principles but do not incorporate real-world videos or dynamic visual contexts. PhysGame [13] focuses on gameplay, while the Cosmos project [1] aims to enable better embodied AI, including robotics. LLM-Phy [15] combines a large language model with a non-differentiable physics simulator to iteratively estimate physical hyperparameters (e.g., friction, damping, layout) and predict scene dynamics. Other benchmarks, such as CoPhy

[8] and CLEVERER [71], emphasize counterfactual reasoning and causal inference in video-based scenarios. ESPRIT [50] couples physical reasoning tasks with explainability via natural language explanations, and PhyWorld [33] evaluates the ability of generative video models to encode physical laws, focusing on physical realism. A comprehensive overview of recent models and methods is provided by [40].

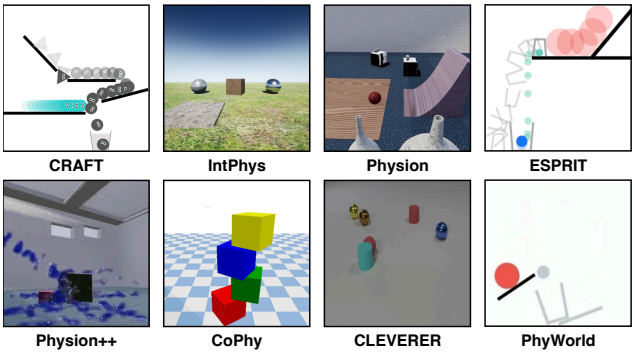

Figure 3. A qualitative overview of recent synthetic datasets related to physical understanding [3, 8, 10, 33, 50, 52, 65, 71]. These datasets are great for the purposes they were designed for, but not ideal for evaluating models trained on real-world videos due to the distribution shift.

However, a major drawback of many benchmarks is that the data they use is synthetic (see Fig 3 for samples). This introduces a real-vs-synthetic distribution shift that may confound results when testing video models trained on natural videos. The Physics-IQ dataset overcomes this limitation by providing real-world videos, capturing diverse and complex physical phenomena (see Fig 1). With three views per scenario, controlled and measured physical variance (by recording two takes for each video), and challenging out-of-distribution settings it provides a rigorous design for evaluating video generative models.

2.4. Models

We evaluate eight different video generative models on our benchmark: VideoPoet (both i2v and multiframe) [35], Lumiere (i2v and multiframe) [7], Runway Gen 3 (i2v) [63], Pika 1.0 (i2v) [62], Stable Video Diffusion (i2v) [12], and Sora (i2v) [47].

Different models have different requirements for the input conditions (single frame, multi frame, or text conditioning), frame rates (8–30 FPS), and resolution (between 256×256 and 1280×768). An overview is shown in 2. For our study, we matched the model's preferred input conditions, frame rates, and resolution exactly by performing a pre-processing step on the Physics-IQ videos (see Supp. 1 for pseudocode).

VideoPoet and Lumiere are the only two models in our study that accept multiple frames as conditioning input. These models also include a super-resolution stage, where they first generate a lower resolution video and subsequently upscale it to a higher resolution. Since we noticed that the lower resolution outputs suffice to test physical realism, we skipped the super-resolution step for these models. Our benchmark consists of physical interactions where temporal information is decidedly useful to have, thus it is generally to be expected that a perfect multiframe model should, in principle, perform better than a perfect i2v model.

2.5. Metrics

Given a model-predicted continuation, how can we tell whether it matches the ground truth continuation? Different choices of metrics are possible; we decided against using vision-language models as evaluators since, as we will see later, they are not (yet) sufficiently sensitive to violations of physics. Other widely used metrics also don't sufficiently capture physics understanding: Video generative models commonly use metrics [27, 66, 69, 72] and benchmarks [24, 28, 29] suited for evaluating the visual quality and realism of the generated videos. These metrics include Peak Signal-to-Noise Ratio (PSNR) [27], Structural Similarity Index Measure (SSIM) [69], Fréchet Video Distance (FVD) [19, 66], and Learned Perceptual Image Patch Similarity (LPIPS) [72]. While these metrics are useful for comparing the appearance, temporal smoothness, and statistics of generated videos with the ground truth; unfortunately, they do not fully capture understanding of physical laws. For instance, both PSNR and SSIM evaluate pixel-level similarities but are not sensitive to the correctness of motion and interactions in a video; FVD captures overall feature distributions but does not penalize a model for physically implausible actions and LPIPS focuses on human-like perception of similarity rather than physical plausibility. While these metrics are great for measuring what they were designed for, they are not equipped to judge real-world physics understanding.

Therefore, we need a different way of assessing whether models understand physical properties. Fortunately, we can make use of a useful property of our dataset that simplifies evaluations: our entire real-world dataset was filmed from a static perspective; additionally, we recorded ground truth continuations that we can use for comparison. Therefore, we can harness fairly simple metrics (based on well-established components) to compare generations against ground truth. We would like to highlight that the core contribution of this work is the dataset itself, and the investigation of physical understanding in generative video models that our dataset enables—while the metrics that we use are merely a means to this end, and we do not claim any novelty w.r.t. these metrics. On the contrary, precisely because of the quality of our data (static perspective, manual cutting,

not part of any model's training set) we can build on very simple metrics and established components.

Since models might fail in different ways (e.g., getting an action right but the timing wrong vs. getting the timing right and the action wrong), we use the following four metrics to track different aspects of physical understanding: *Where* does action happen? **Spatial IoU**, *Where* & *when* does action happen? **Spatiotemporal IoU**, *Where* & *how much* action happens? **Weighted spatial IoU** and *How* does action happen? **MSE**. These four metrics—explained in detail below—are then combined into a single score, the **Physics-IQ score**, by summing the individual scores (with a negative sign for MSE where lower=better). This Physics-IQ score is normalized such that physical variance (the upper limit of what we can reasonably expect a model to capture) is 100%—enabled by recording each scenario twice from each perspective to estimate this physical variance.

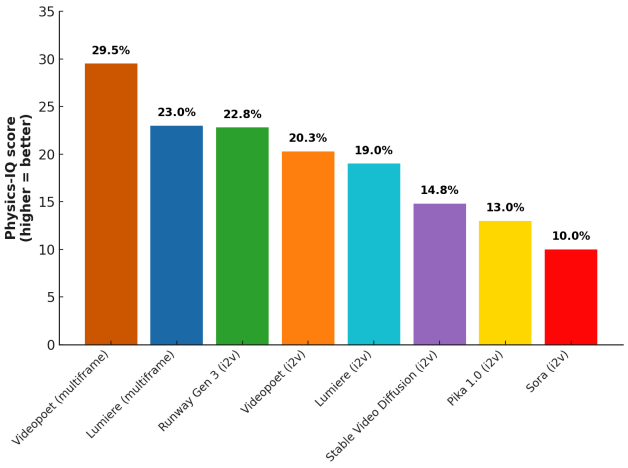

Figure 4. How well do current video generative models understand physical principles? The Physics-IQ score is an aggregated measure across four individual metrics, normalized such that pairs of real videos that differ only by physical randomness score 100%. All evaluated models show a large gap, with the best model scoring 29.5%, indicating that physical understanding is severely limited.

***Where* does action happen? Spatial IoU** The location of movement is an important indicator of physical "correctness". For instance, in the "domino with duck interrupting the chain" scenario from 1, only the part of the chain that is to the right side of the duck should tumble, while the other part should remain unchanged. Similarly, the spatial trajectory of a moving ball is indicative of whether the movement is realistic. The Spatial IoU metric compares generated videos against ground truth to determine whether the location of movement/action mirrors ground truth. Since the benchmark videos are filmed from a static perspective without camera movement, a simple threshold on pixel intensity changes across frames (see Supp. 2 for pseudocode) easily identifies where movement happens. This leads to a binary $h \times w \times t$ "motion mask video" that highlights the regions of motion in a scene at any point in time. Spatial IoU then simply generates a binary $h \times w$ spatial "motion map"—similar, in spirit, to a saliency map—by collapsing the masks across the time dimension with a max operation. A motion map thus simply has a 1 whenever action occurred, at any point in time, at a particular location; and a 0 otherwise. This motion map is compared against the motion map from the real, ground truth video, using Intersection over Union or IoU (a metric commonly used in object detection to measure overlap while penalizing areas that differ):

$$\text{Spatial-IoU} = \frac{|M_{\text{real}}^{\text{binary,spatial}} \cap M_{\text{gen}}^{\text{binary,spatial}}|}{|M_{\text{real}}^{\text{binary,spatial}} \cup M_{\text{gen}}^{\text{binary,spatial}}|}$$

where $M_{\text{real}}^{\text{spatial}}$ and $M_{\text{gen}}^{\text{spatial}}$ are the motion maps based on real and generated videos, respectively. Spatial IoU measures whether the location *where* action happens is correct.

***Where* & *when* does action happen? Spatiotemporal IoU** Spatiotemporal IoU goes a step further than Spatial IoU by also taking into account *when* an action occurs. Instead of collapsing across time as Spatial IoU does, Spatiotemporal IoU compares the two motion mask videos (based on real and generated videos) frame-by-frame, averaging across t:

$$\text{Spatiotemporal-IoU}(M_{\text{real}}, M_{\text{gen}}) = \frac{|M_{\text{real}} \cap M_{\text{gen}}|}{|M_{\text{real}} \cup M_{\text{gen}}|}$$

where M_{real} and M_{gen} are the $h \times w \times t$ binary motion masks for the real and generated videos, respectively. Spatiotemporal IoU thus tracks not only *where* an action occurs in a video, but also whether it occurs at the right time (*when*). If a model does well on Spatial IoU but poorly on Spatiotemporal IoU, this would therefore indicate that the model gets the location of the action right, but the timing wrong.

***Where* does action happen, and & *how much* action happens? Weighted spatial IoU** Weighted spatial IoU is similar to Spatial IoU in the sense that it compares two $h \times w$ "motion maps". However, instead of comparing binary motion maps (action occurred or did not occur), it also assesses *how much* action happens at any given location. This distinguishes between e.g. motion caused by a pendulum (showing repeated motion in an area) from motion by a rolling ball (which passes a location only once). Weighted spatial IoU is computed by taking the binary $h \times w \times t$ motion mask video (described above in the section on Spatial IoU) and collapsing across the time dimension t in a weighted fashion (instead of taking the maximum). The weighting simply averages per-frame action. This weighted $h \times w$

spatial "motion map" is then used to compute the metric by summing the pixel-wise minimum of two motion maps and dividing by the pixel-wise maximum:

$$\text{Weighted-spatial-IoU} = \frac{\sum_{i=1}^{n} \min\left(M_{\text{real,i}}^{\text{weighted,spatial}}, M_{\text{gen,i}}^{\text{weighted,spatial}}\right)}{\sum_{i=1}^{n} \max\left(M_{\text{real,i}}^{\text{weighted,spatial}}, M_{\text{gen,i}}^{\text{weighted,spatial}}\right)}$$

where $M_{\text{real}}^{\text{weighted,spatial}}$ and $M_{\text{gen}}^{\text{weighted,spatial}}$ are the weighted motion maps representing how much activity/action happens at any location (based on real and generated videos, respectively). Weighted spatial IoU thus measures not only *where* an action occurs, but also *how much* action is happening.

***How* does an action happen? MSE** Finally, mean squared error (MSE) calculates the average squared difference between corresponding pixel values in two frames (e.g., a real and a generated frame). Given two frames f_{real} and f_{gen}, the MSE is given by:

$$\text{MSE}(f_{\text{real}}, f_{\text{gen}}) = \frac{1}{n}\sum_{i=1}^{n}(f_{\text{real,i}} - f_{\text{gen,i}})^2$$

where n is the total number of pixels in the frame. MSE focuses on pixel-level fidelity; this is a very strict requirement that is sensitive to *how* objects look and interact. For instance, if a generative model would show a tendency to change the color of objects, this physically unrealistic event would be heavily penalized. MSE therefore tracks aspects that complement the other metrics. None of them is perfect, and none of them should be used in isolation, but collectively they provide a comprehensive assessment of different aspects that quantify physical realism. Since raw MSE values can be hard to interpret, we provide an intuition in 9.

2.6. Metric for visual realism: MLLM evaluation

In addition to measuring the physical realism, we are interested in tracking how convincingly a model can generate realistic videos, as assessed by a multimodal large language model or MLLM [here: Gemini 1.5 Pro, 22]. Instead of rating videos (which would be sensitive to model biases), we use the gold standard experimental methods from psychophysics, a 2AFC paradigm. 2AFC stands for two-alternative-forced-choice. In our case, this means that the MLLM is given pairs of real and generated videos of the same scenario in randomized order. The MLLM is asked to identify the generated video. The MLLM evaluation score is expressed as a percentage corresponding to the accuracy across all videos, with chance rate at 50%. Any accuracy that is higher indicates that the MLLM was able to correctly identify at least some of the generated videos; while accuracies close to 50% indicates that a video generative model has successfully deceived the MLLM into classifying the generated videos as real, indicating high visual realism. Details on the experiment are described in the appendix.

3. Results

3.1. Physical understanding

The goal of our Physics-IQ benchmark is to understand, and quantify, whether generative video models learn physical principles. Therefore, we test all eight models in our study on every scenario and for each camera position (left, center, right) in the benchmark dataset. These samples are visualized in 1. We first report the aggregated Physics-IQ results across all metrics related to physical understanding (Spatial IoU, Spatiotemporal IoU, Weighted-spatial IoU, MSE) in 4. The main takeaway from this figure is that all the models show a massive gap to the physical variance baseline, with the best model scoring only 29.5% out of the possible 100.0%. As we mentioned in the previous section, each scenario was recorded twice (*take 1* and *take 2*) to estimate the natural variability in real-world physical phenomena. This estimate is termed the *physical variance*; the figure is normalized such that pairs of real videos that differ only by physical randomness score 100.0%. The gap between model performance and real videos demonstrates a severe lack of physical understanding in current powerful video generative models. Across the different models, VideoPoet (multiframe) [35] ranks best; interestingly, VideoPoet is a causal model. For the two models that have both an image2video (i2v) and a version conditioned on multiple frames (multiframe), the multiframe variants outperform the i2v variants. This is expected, as predicting the future— as required by our challenging Physics-IQ benchmark—benefits from access to temporal information, which multiframe variants provide.

The performance of each model on each individual metric is reported in 1 (See Supp. 12 for a breakdown by physical principle). VideoPoet (multiframe) performs best on a majority of the metrics (3 out of 4). Qualitatively, the generated videos from Sora are often visually and artistically superior, but they also frequently include transition cuts—despite instructed not to change the camera perspective—which is penalized by several other metrics. We expect that if a future version of this model more closely follows the prompt (static camera perspective, no camera movement), its Physics-IQ score would improve substantially. Qualitatively, success and failure cases are visualized in 6.

3.2. Human evaluation

We ran two pairwise-comparison user studies on 25 randomly sampled scenarios (covering diverse physical

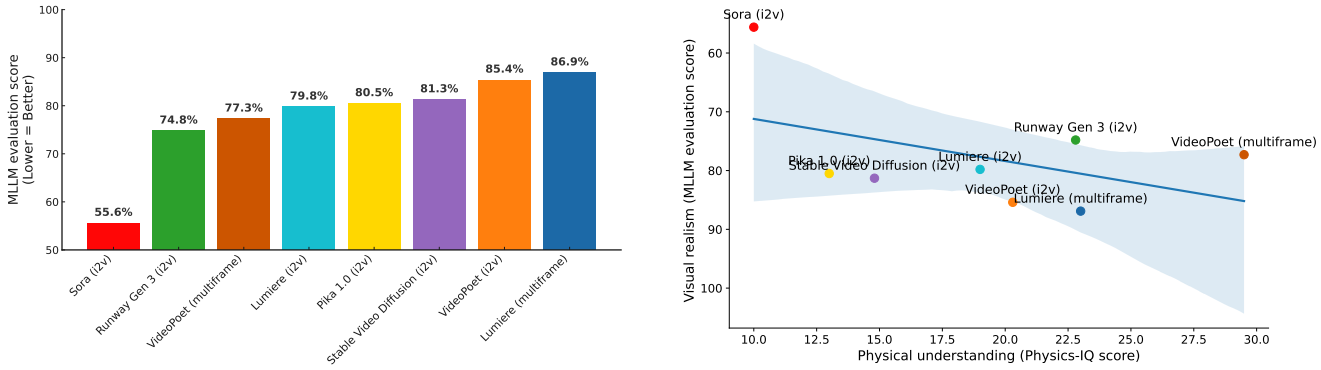

Figure 5. Relationship between visual realism and physical understanding. **Left:** A multimodal large language model (Gemini 1.5 Pro) identifies the generated video in a two-alternative forced-choice task (MLLM score). Chance is 50%; lower scores indicate greater realism. Sora-generated videos are hardest to distinguish, while Lumiere (multiframe) is easiest. **Right:** Do models that generate more realistic-looking videos also demonstrate better physical understanding (Physics-IQ score)? A scatterplot with a linear fit (95% CI shaded) shows no significant correlation (r = -0.46, p = .249). The y-axis is inverted for easier interpretation (up & right are best).

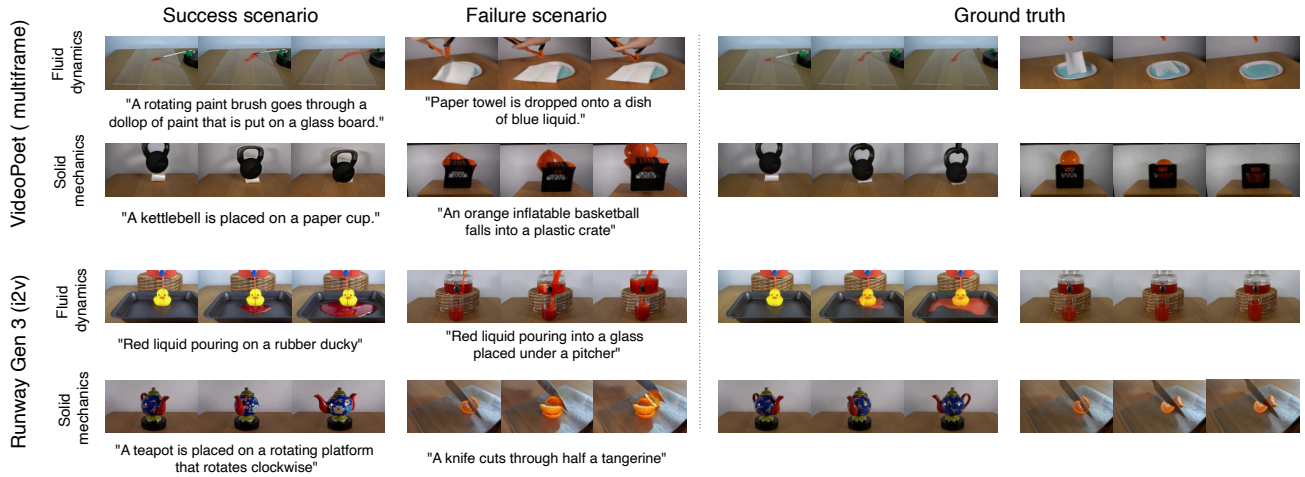

Figure 6. We visualize success and failure scenarios within the fluid dynamics and solid mechanics categories for two of the best models, VideoPoet and Runway Gen 3, according to our metrics. Both models are able to generate physics plausible frames for scenarios such as smearing paint on glass (VideoPoet) and pouring red liquid on a rubber duck (Runway Gen 3). At the same time, the models fail to simulate a ball falling into a crate or cutting a tangerine with a knife. See here for an animated version of this figure.

events), applying a light Gaussian blur to all videos to reduce resolution bias while preserving motion and boundaries.

Study 1: real vs. model. Participants chose the more physically plausible video between a real clip and a model output. Humans overwhelmingly preferred real videos. Models were selected over real only rarely: Runway Gen 3 (Rank 3) in 2/25 trials (8%), VideoPoet (i2v, Rank 4) in 1/25 (4%); all others—VideoPoet (multiframe, Rank 1), Lumiere (multiframe, Rank 2), Lumiere (i2v, Rank 5), SVD (Rank 6), Pika 1.0 (Rank 7), Sora (Rank 8)—in 0/25. This mirrors the large realism gap captured by Physics-IQ.

Study 2: model vs. model. We paired models by Physics-IQ rank (1 vs. 8, 2 vs. 7, 3 vs. 6, 4 vs. 5). The higher-ranked model was preferred in most matchups: Runway Gen 3 vs. SVD: 22–3 (88% agreement); VideoPoet (multiframe) vs. Sora: 19–6 (76%); Lumiere (multiframe) vs. Pika 1.0: 16–9 (64%). An adjacent-rank pair was the exception: VideoPoet (i2v, Rank 4) vs. Lumiere (i2v, Rank 5) favored the lower-ranked model 15–10 (40% agreement with rank; Physics-IQ 19.0% vs. 20.3%). Overall, human preferences broadly align with Physics-IQ. See Supp. 4 for protocol details.

Table 1. Comparison of metric scores for different models. The best-performing model for each metric is marked in bold. Note that Physical Variance serves as a performance upper bound for each metric, representing the difference between two real videos and capturing the inherent variability in real-world scenarios.

Model	Spatial IoU ↑	Spatiotemporal IoU ↑	Weighted spatial IoU ↑	MSE ↓	Physics-IQ Score ↑
Physical Variance	**0.678**	**0.535**	**0.577**	**0.002**	**100.0**
VideoPoet (multiframe)	**0.204**	0.164	**0.137**	**0.010**	**29.5**
Lumiere (multiframe)	0.170	0.155	0.093	0.013	23.0
Runway Gen 3 (i2v)	0.201	0.115	0.116	0.015	22.8
VideoPoet (i2v)	0.141	0.126	0.087	0.012	20.3
Lumiere (i2v)	0.113	**0.173**	0.061	0.016	19.0
Stable Video Diffusion (i2v)	0.132	0.076	0.073	0.021	14.8
Pika 1.0 (i2v)	0.140	0.041	0.078	0.014	13.0
Sora (i2v)	0.138	0.047	0.063	0.030	10.0

3.3. Visual realism: Multimodal large language model evaluation

Why do many circulated samples appear visually convincing despite poor physical understanding? We quantified *visual realism* by asking Gemini 1.5 Pro [22] to pick the generated video in each real–generated pair (Physics-IQ scenarios). The *MLLM score* is this identification accuracy; values near 50% (chance) indicate harder-to-detect fakes (i.e., higher visual realism). As shown in 5 (left), the MLLM often succeeds (e.g., up to 86.9% for Lumiere multiframe). One model stands out: Sora achieves 55.6% (closest to chance), while Runway Gen 3 reaches 74.8% and VideoPoet (multiframe) 77.3%. Although the MLLM's textual justifications are frequently tangential to the visual evidence [4, 39], the accuracy pattern indicates that some models generate highly realistic videos even with limited physical understanding. Consistent with prior findings on intuitive physics and causal reasoning gaps [11, 31, 37, 46, 49, 54, 61, 70], 5 (right) shows no significant correlation between *visual realism* and *physical understanding*.

4. Discussion

We introduced Physics-IQ, a challenging and comprehensive real-world benchmark to evaluate physics understanding in video generative models. We analyzed eight models on Physics-IQ using its associated metrics to quantify physics understanding. The benchmark data and metrics cover a wide range of settings and reveal a striking discrepancy between visual realism (sometimes present in current models) and physical understanding (largely lacking in current models).

Do video models understand physical principles? Does realistic generation imply physics understanding? Our benchmark says no: all evaluated models lack deep physical competence. The best system, VideoPoet (multiframe), scores 29.5, far below the *physical variance* baseline of 100.0 (real-world variability). Progress remains possible; scaling may help, though alternative (e.g., interactive) training could be required. Prediction alone can induce structure: language models learn syntax from next-token prediction [26]. Current models sometimes succeed (e.g., VideoPoet correctly simulates paint smearing on glass) but often make basic errors (objects interpenetrate). Synthetic-data results [33] suggest that with enough data, video models can acquire physical laws. We open-source Physics-IQ to track progress.

Visual realism doesn't imply physical understanding. Visual realism and physical understanding are not significantly correlated (5). Example: when a burning match enters water, Runway Gen 3 produces a photorealistic sequence where a candle appears and lights, which is temporally impossible despite per-frame quality. Such "object-into-existence" hallucinations are common [51]. We observe them across models; stronger ones (Runway Gen 3, Sora) often hallucinate content consistent with the scene (e.g., match→candle), hinting at partial understanding.

Dataset biases are reflected in model generations. Most models respect scene/layout and object attributes; Sora and Runway Gen 3 are notably consistent across frames. Yet training biases surface: in prototyping, Lumiere changed a red pool table to green at generation start (bias toward common green tables), and Sora frequently inserted transition cuts, consistent with likely training data/style.

Metrics and their limitations. Standard quality metrics (PSNR [27], FVD [66], LPIPS [72], SSIM [69]) do not target physics. We therefore use simple, interpretable components (MSE, IoU) to score spatial, temporal, and perceptual coherence, aggregating them into the normalized *Physics-IQ* score. Each metric is a proxy: the MLLM measure captures how well videos "deceive" a multimodal model but inherits the MLLM's limits; its explanations are often wrong, and it fails more on Sora. Likewise, Stable Video Diffusion shows heavy hallucinations and implausible object motion yet attains Spatial-IoU comparable to Lumiere, Sora, Pika, and VideoPoet (i2v); conversely, Runway Gen 3 excels at Spatial-IoU but lags on Spatiotemporal-IoU. No single metric should be read in isolation.

We intentionally design Physics-IQ to be conservative, penalizing object hallucinations, camera motion (which prompts asked to avoid), and shot changes. Models like Sora exhibit these more often and thus score lower on some metrics. In a hype-prone area, we prefer benchmarks that err on the side of caution.

References

[1] Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, et al. Cosmos world foundation model platform for physical AI. *arXiv preprint arXiv:2501.03575*, 2025. 3

[2] Pulkit Agrawal, Ashvin V Nair, Pieter Abbeel, Jitendra Malik, and Sergey Levine. Learning to poke by poking: Experiential learning of intuitive physics. *Advances in neural information processing systems*, 29, 2016. 3

[3] Tayfun Ates, M Samil Atesoglu, Cagatay Yigit, Ilker Kesen, Mert Kobas, Erkut Erdem, Aykut Erdem, Tilbe Goksun, and Deniz Yuret. Craft: A benchmark for causal reasoning about forces and interactions. *arXiv preprint arXiv:2012.04293*, 2020. 3, 4

[4] Zechen Bai, Hai Ci, and Mike Zheng Shou. Impossible videos, 2025. 8

[5] Renée Baillargeon. The acquisition of physical knowledge in infancy: A summary in eight lessons. *Blackwell handbook of childhood cognitive development*, pages 47–83, 2002. Supp2

[6] Hritik Bansal, Zongyu Lin, Tianyi Xie, Zeshun Zong, Michal Yarom, Yonatan Bitton, Chenfanfu Jiang, Yizhou Sun, Kai-Wei Chang, and Aditya Grover. Videophy: Evaluating physical commonsense for video generation, 2024. 3

[7] Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Herrmann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Guanghui Liu, Amit Raj, Yuanzhen Li, Michael Rubinstein, Tomer Michaeli, Oliver Wang, Deqing Sun, Tali Dekel, and Inbar Mosseri. Lumiere: A space-time diffusion model for video generation, 2024. 2, 4

[8] Fabien Baradel, Natalia Neverova, Julien Mille, Greg Mori, and Christian Wolf. Cophy: Counterfactual learning of physical dynamics. *arXiv preprint arXiv:1909.12000*, 2019. 4

[9] Horace B Barlow et al. Possible principles underlying the transformation of sensory messages. *Sensory communication*, 1(01):217–233, 1961. 1

[10] Daniel M. Bear, Elias Wang, Damian Mrowca, Felix J. Binder, Hsiao-Yu Fish Tung, R. T. Pramod, Cameron Holdaway, Sirui Tao, Kevin Smith, Fan-Yun Sun, Li Fei-Fei, Nancy Kanwisher, Joshua B. Tenenbaum, Daniel L. K. Yamins, and Judith E. Fan. Physion: Evaluating physical prediction from vision in humans and machines, 2021. 3, 4

[11] Marcel Binz and Eric Schulz. Using cognitive psychology to understand GPT-3. *Proceedings of the National Academy of Sciences*, 120(6):e2218523120, 2023. 8

[12] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 2, 4

[13] Meng Cao, Haoran Tang, Haoze Zhao, Hangyu Guo, Jiaheng Liu, Ge Zhang, Ruyang Liu, Qiang Sun, Ian Reid, and Xiaodan Liang. Physgame: Uncovering physical commonsense violations in gameplay videos. *arXiv preprint arXiv:2412.01800*, 2024. 3

[14] Michael B Chang, Tomer Ullman, Antonio Torralba, and Joshua B Tenenbaum. A compositional object-based approach to learning physical dynamics. *arXiv preprint arXiv:1612.00341*, 2016. Supp2

[15] Anoop Cherian, Radu Corcodel, Siddarth Jain, and Diego Romeres. LLMPhy: Complex physical reasoning using large language models and world models. *arXiv preprint arXiv:2411.08027*, 2024. 3

[16] DeepMind. Veo2: Google's state-of-the-art video generation model. https://deepmind.google/technologies/veo/veo-2/, 2024. Accessed: 2025-01-09. 1

[17] Jason Fischer and Bradford Z Mahon. What tool representation, intuitive physics, and action have in common: The brain's first-person physics engine. *Cognitive neuropsychology*, 38(7-8):455–467, 2021. Supp2

[18] Karl Friston. A theory of cortical responses. *Philosophical transactions of the Royal Society B: Biological sciences*, 360 (1456):815–836, 2005. 1

[19] Songwei Ge, Aniruddha Mahapatra, Gaurav Parmar, Jun-Yan Zhu, and Jia-Bin Huang. On the content bias in frechet video distance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7277–7288, 2024. 4

[20] Robert Geirhos, Carlos RM Temme, Jonas Rauber, Heiko H Schütt, Matthias Bethge, and Felix A Wichmann. Generalisation in humans and deep neural networks. *Advances in neural information processing systems*, 31, 2018. 3

[21] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020. 2, 3

[22] Gemini Team Google: Petko Georgiev and 1133 other authors. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024. 6, 8

[23] Alison Gopnik, Clark Glymour, David M Sobel, Laura E Schulz, Tamar Kushnir, and David Danks. A theory of causal learning in children: causal maps and bayes nets. *Psychological review*, 111(1):3, 2004. 3

[24] Xuan He, Dongfu Jiang, Ge Zhang, Max Ku, Achint Soni, Sherman Siu, Haonan Chen, Abhranil Chandra, Ziyan Jiang, Aaran Arulraj, et al. Videoscore: Building automatic metrics to simulate fine-grained human feedback for video generation. *arXiv preprint arXiv:2406.15252*, 2024. 4

[25] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2018. 3

[26] John Hewitt and Christopher D Manning. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138, 2019. 8

[27] Alain Hore and Djemel Ziou. Image quality metrics: Psnr vs. ssim. In *2010 20th international conference on pattern recognition*, pages 2366–2369. IEEE, 2010. 4, 8

[28] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin,

Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21807–21818, 2024. 4

[29] Ziqi Huang, Fan Zhang, Xiaojie Xu, Yinan He, Jiashuo Yu, Ziyue Dong, Qianli Ma, Nattapol Chanpaisit, Chenyang Si, Yuming Jiang, et al. Vbench++: Comprehensive and versatile benchmark suite for video generative models. *arXiv preprint arXiv:2411.13503*, 2024. 4

[30] Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. OpenAI o1 system card. *arXiv preprint arXiv:2412.16720*, 2024. Supp2

[31] Serwan Jassim, Mario Holubar, Annika Richter, Cornelius Wolff, Xenia Ohmer, and Elia Bruni. GRASP: A novel benchmark for evaluating language grounding and situated physics understanding in multimodal language models. *arXiv preprint arXiv:2311.09048*, 2023. 8

[32] Bingyi Kang, Yang Yue, Rui Lu, Zhijie Lin, Yang Zhao, Kaixin Wang, Gao Huang, and Jiashi Feng. How far is video generation from world model: A physical law perspective. *arXiv preprint arXiv:2411.02385*, 2024. 2

[33] Bingyi Kang, Yang Yue, Rui Lu, Zhijie Lin, Yang Zhao, Kaixin Wang, Gao Huang, and Jiashi Feng. How far is video generation from world model: A physical law perspective, 2024. 4, 8

[34] Philip J Kellman and Elizabeth S Spelke. Perception of partly occluded objects in infancy. *Cognitive psychology*, 15 (4):483–524, 1983. 3

[35] Dan Kondratyuk, Lijun Yu, Xiuye Gu, Jose Lezama, Jonathan Huang, Grant Schindler, Rachel Hornung, Vighnesh Birodkar, Jimmy Yan, Ming-Chang Chiu, Krishna Somandepalli, Hassan Akbari, Yair Alon, Yong Cheng, Joshua V. Dillon, Agrim Gupta, Meera Hahn, Anja Hauth, David Hendon, Alonso Martinez, David Minnen, Mikhail Sirotenko, Kihyuk Sohn, Xuan Yang, Hartwig Adam, Ming-Hsuan Yang, Irfan Essa, Huisheng Wang, David A Ross, Bryan Seybold, and Lu Jiang. VideoPoet: A large language model for zero-shot video generation. In *Proceedings of the 41st International Conference on Machine Learning*, pages 25105–25124. PMLR, 2024. 2, 4, 6

[36] James R Kubricht, Keith J Holyoak, and Hongjing Lu. Intuitive physics: Current research and controversies. *Trends in cognitive sciences*, 21(10):749–759, 2017. 3

[37] Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. Building machines that learn and think like people. *Behavioral and brain sciences*, 40:e253, 2017. 8

[38] Yichen Li, YingQiao Wang, Tal Boger, Kevin A Smith, Samuel J Gershman, and Tomer D Ullman. An approximate representation of objects underlies physical reasoning. *Journal of Experimental Psychology: General*, 2023. Supp2

[39] Zongxia Li, Xiyang Wu, Guangyao Shi, Yubin Qin, Hongyang Du, Tianyi Zhou, Dinesh Manocha, and Jordan Lee Boyd-Graber. Videohallu: Evaluating and mitigating multi-modal hallucinations on synthetic video understanding. *arXiv preprint arXiv:2505.01481*, 2025. 8

[40] Daochang Liu, Junyu Zhang, Anh-Dung Dinh, Eunbyung Park, Shichao Zhang, and Chang Xu. Generative physical AI in vision: A survey. *arXiv preprint arXiv:2501.10928*, 2025. 4

[41] Nanye Ma, Shangyuan Tong, Haolin Jia, Hexiang Hu, Yu-Chuan Su, Mingda Zhang, Xuan Yang, Yandong Li, Tommi Jaakkola, Xuhui Jia, et al. Inference-time scaling for diffusion models beyond scaling denoising steps. *arXiv preprint arXiv:2501.09732*, 2025. Supp2

[42] Michael McCloskey. Intuitive physics. *Scientific american*, 248(4):122–131, 1983. 3

[43] Michael McCloskey, Alfonso Caramazza, and Bert Green. Curvilinear motion in the absence of external forces: Naive beliefs about the motion of objects. *Science*, 210(4474): 1139–1141, 1980. 3

[44] Fanqing Meng, Jiaqi Liao, Xinyu Tan, Wenqi Shao, Quanfeng Lu, Kaipeng Zhang, Yu Cheng, Dianqi Li, Yu Qiao, and Ping Luo. Towards world simulator: Crafting physical commonsense-based benchmark for video generation, 2024. 3

[45] Meta AI. Meta Movie Gen: AI-powered movie generation. https://ai.meta.com/research/movie-gen/, 2024. Accessed: 2024-11-24. 1

[46] Saman Motamed, Minghao Chen, Luc Van Gool, and Iro Laina. Travl: A recipe for making video-language models better judges of physics implausibility. *arXiv preprint arXiv:2510.07550*, 2025. 8

[47] OpenAI. Sora: OpenAI's Multimodal Agent. https://openai.com/index/sora/, 2024. Accessed: 2024-11-24. 1, 2, 4

[48] Luis S Piloto, Ari Weinstein, Peter Battaglia, and Matthew Botvinick. Intuitive physics learning in a deep-learning model inspired by developmental psychology. *Nature human behaviour*, 6(9):1257–1267, 2022. 3

[49] Pooyan Rahmanzadehgervi, Logan Bolton, Mohammad Reza Taesiri, and Anh Totti Nguyen. Vision language models are blind. In *Proceedings of the Asian Conference on Computer Vision*, pages 18–34, 2024. 8

[50] Nazneen Fatema Rajani, Rui Zhang, Yi Chern Tan, Stephan Zheng, Jeremy Weiss, Aadit Vyas, Abhijit Gupta, Caiming Xiong, Richard Socher, and Dragomir Radev. Esprit: Explaining solutions to physical reasoning tasks. *arXiv preprint arXiv:2005.00730*, 2020. 4

[51] Vipula Rawte, Amit Sheth, and Amitava Das. A survey of hallucination in large foundation models. *arXiv preprint arXiv:2309.05922*, 2023. 8

[52] Ronan Riochet, Mario Ynocente Castro, Mathieu Bernard, Adam Lerer, Rob Fergus, Véronique Izard, and Emmanuel Dupoux. IntPhys: A framework and benchmark for visual intuitive physics reasoning. *arXiv preprint arXiv:1803.07616*, 2018. 3, 4

[53] Rebecca Saxe and Susan Carey. The perception of causality in infancy. *Acta psychologica*, 123(1-2):144–165, 2006. 3

[54] Luca M Schulze Buschoff, Elif Akata, Matthias Bethge, and Eric Schulz. Visual cognition in multimodal large language models. *Nature Machine Intelligence*, pages 1–11, 2025. 8

[55] Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*, 2024. Supp2

[56] Felix A Sosa, Samuel J Gershman, and Tomer D Ullman. Blending simulation and abstraction for physical reasoning. *Cognition*, 254:105995, 2025. Supp2

[57] Elizabeth S Spelke. The origins of physical knowledge. *Clarendon Press/Oxford University Press*, 1988. Supp2

[58] Elizabeth S Spelke, Karen Breinlinger, Janet Macomber, and Kristen Jacobson. Origins of knowledge. *Psychological review*, 99(4):605, 1992. 3

[59] Elizabeth S Spelke, Roberta Kestenbaum, Daniel J Simons, and Debra Wein. Spatiotemporal continuity, smoothness of motion and object identity in infancy. *British journal of developmental psychology*, 13(2):113–142, 1995. 3

[60] Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*, 2022. 3

[61] Shane Storks, Qiaozi Gao, Yichi Zhang, and Joyce Chai. Tiered reasoning for intuitive physics: Toward verifiable commonsense language understanding. In *Findings of Conference on Empirical Methods in Natural Language Processing (EMNLP) 2021*, 2021. 8

[62] Pika Labs Team. Pika labs. https://pikalabs.com, 2024. Generative AI platform for creating video and visual content. 2, 4

[63] Runway Team. Runway. https://runwayml.com, 2024. Platform for AI-powered video editing and generative media creation. 2, 4

[64] Joshua B Tenenbaum, Charles Kemp, Thomas L Griffiths, and Noah D Goodman. How to grow a mind: Statistics, structure, and abstraction. *science*, 331(6022):1279–1285, 2011. 3

[65] Hsiao-Yu Tung, Mingyu Ding, Zhenfang Chen, Daniel Bear, Chuang Gan, Josh Tenenbaum, Dan Yamins, Judith Fan, and Kevin Smith. Physion++: Evaluating physical scene understanding that requires online inference of different physical properties. *Advances in Neural Information Processing Systems*, 36, 2024. 3, 4

[66] Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018. 4, 8

[67] Michele Vicovaro. Grounding intuitive physics in perceptual experience. *Journal of Intelligence*, 11(10):187, 2023. Supp2

[68] Hermann von Helmholtz. *Handbuch der physiologischen Optik: mit 213 in den Text eingedruckten Holzschnitten und 11 Tafeln*. Voss, 1867. 1

[69] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 4, 8

[70] Luca Weihs, Amanda Yuile, Renée Baillargeon, Cynthia Fisher, Gary Marcus, Roozbeh Mottaghi, and Aniruddha Kembhavi. Benchmarking progress to infant-level physical reasoning in ai. *Transactions on Machine Learning Research*, 2022. 8

[71] Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B Tenenbaum. Clevrer: Collision events for video representation and reasoning. *arXiv preprint arXiv:1910.01442*, 2019. 4

[72] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 4, 8