Overview of all Physics-IQ scenarios

Figure 7 shows our recording setup. Figure 11 presents the switch frames (center view) from all 66 scenarios in the Physics-IQ dataset. These frames represent the last frame of the conditioning signal, after which a model is asked to generate a prediction for the future frames.

Figure 7. Illustration of recording setup (top) and perspectives (bottom).

Model specifications

Table 2 presents the specifications of the models evaluated in this work, including their conditioning criteria, generated FPS, and resolution.

Table 2. Specifications of evaluated video models, including input conditioning, frame rate (FPS), and resolution.

Model	Text Condition	Multi-frame Condition	Single-frame Condition	FPS	Resolution
VideoPoet (i2v)	✓	✗	✓	8	128224
VideoPoet (multiframe)	✓	✓	✗	8	128224
Lumiere (i2v)	✓	✗	✓	16	128128
Lumiere (multiframe)	✓	✓	✗	16	128128
Stable Video Diffusion (i2v)	✗	✗	✓	8	1024576
Runway Gen 3 (i2v)	✓	✗	✓	24	1280768
Pika 1.0 (i2v)	✓	✗	✓	24	1280720
Sora (i2v)	✓	✗	✓	30	854480

Adjusting video frame rate

Pseudocode 1 outlines the method for changing the frame rate (FPS) of a video using linear interpolation. It generates a smooth transition between original frames while optionally resizing the output resolution. This technique ensures temporal consistency, making it well-suited for generating videos with desired FPS to adapt Physics-IQ for models with different FPS.

Algorithm 1 Change video FPS with linear interpolation

Require: video file V, original FPS $\text{fps}_{\text{original}}$, new FPS fps_{new}, output dimensions (w, h) (optional)
Ensure: video V' with adjusted FPS and resolution

$f_{\text{original}} \leftarrow$ extract frames from V at $\text{fps}_{\text{original}}$
$\text{duration} \leftarrow$ length of of V
$n_{\text{original}} \leftarrow$ number of frames in $\text{fps}_{\text{original}}$
$n_{\text{new}} \leftarrow \text{duration} \cdot \text{fps}_{\text{new}}$
Initialize empty list n_{new}
for $j \leftarrow 0$ to $n_{\text{new}} - 1$ **do**
    $\alpha \leftarrow j \times (n_{\text{original}} - 1)/(n_{\text{new}} - 1)$
    $i \leftarrow \lfloor \alpha \rfloor$ ▷ Index of the first frame for interpolation
    $\beta \leftarrow \alpha - i$ ▷ Weight for linear interpolation
    $f_1 \leftarrow f_{\text{original}}[i]$
    $f_2 \leftarrow f_{\text{original}}[\min(i + 1, n_{\text{original}} - 1)]$
    $f_{\text{interpolated}} \leftarrow (1 - \beta) \cdot f_1 + \beta \cdot f_2$
    if (w, h) is not None **then**
        resize $f_{\text{interpolated}}$ to (w, h)
    end if
    append $f_{\text{interpolated}}$ to f_{new}
end for
$V' \leftarrow$ recreate video from f_{new} with fps_{new}
Save V'

Generating binary mask videos

Pseudocode 2 describes a method to generate binary mask videos that highlight moving objects. The algorithm combines background subtraction with adaptive updates and morphological operations to detect and cleanly segment motion in video frames. This approach is useful for creating spatial and temporal masks in Physics-IQ evaluations.

MLLM evaluation prompt

The following prompt was used in the two-alternative forced-choice paradigm: "Your task is to help me sort my videos. I mixed up real videos that I shot with my camera and similar videos that I generated with a computer. I only know that exactly one of the two videos is the real one, and exactly one of the following two videos is the generated one. Please take a look at the two videos and let me know which of them is the generated one. I'll tip you $100 if you do a great job and help me identify the generated one. First explain your reasoning, then end with the following statement: 'For this reason, the first video is the generated one' or 'For this reason, the second video is the generated one'."

Algorithm 2 Generate binary mask video for moving objects

Require: Video V, output file V', threshold τ, update rate α, averaging window size w
Ensure: Binary mask video V' highlighting moving objects
Initialize video reader for V and writer for V'
Read first w frames $\{f_1, \ldots, f_w\}$ and preprocess: grayscale and blur
Initialize background model $\mathrm{B} \leftarrow \frac{1}{w} \sum_{i=1}^{w} f_i$ ▷ Initial average reduces noise
for each frame f_t in V **do**
    Preprocess f_t: grayscale and blur
    Update background $\mathrm{B} \leftarrow (1 - \alpha) \cdot \mathrm{B} + \alpha \cdot f_t$
    Compute difference $d_t \leftarrow |f_t - \mathrm{B}|$
    Threshold $m_t \leftarrow 255$ if $d_t > \tau$, else 0
    Morphologically clean m_t (opening and closing)
    Write m_t to V'
end for
Save and close V'

Human evaluation details

We conducted two human studies to evaluate alignment between Physics-IQ scores and human judgments of physical plausibility.

Participants. We recruited **five PhD students in Computer Science** (none of whom are authors), all familiar with generative video or computer vision research.

Procedure. Both studies used the same set of **25 randomly sampled scenarios** from the benchmark, covering diverse physical events. Participants were shown pairs of videos and asked:

Which video appears more physically plausible?

To reduce bias from resolution or sharpness differences across models, we applied a uniform **Gaussian blur** to all videos.

Study 1: Real vs. Model. Each participant viewed **40 video pairs**, each consisting of a real video and a generated one (randomly sampled across 8 models). We aggregated responses per model to compute how often model outputs were preferred over real videos.

Study 2: Model vs. Model. Each participant viewed **20 video pairs**, where both videos were generated by different models depicting the same scenario. Model pairings followed rank-based matchups (e.g., 1st vs. 8th).

Interface. All comparisons were displayed side-by-side with randomized left-right order and a consistent prompt. An example is shown in Figure 8.

These studies confirm that Physics-IQ rankings broadly align with human judgments, particularly when differences in model quality are substantial.

Physics-IQ Score and Mean Model Rank

Figure 10 compares the Physics-IQ score rankings of models with their mean rank across four evaluation metrics.

Visualizing different MSE values

9 illustrates the relationship between a distortion applied to a video and MSE (Mean Squared Error) in a scene. Note that none of the videos in the benchmark have a distortion applied to them; instead, this is intended as an visual intuition for how much a certain MSE value distorts an image.

Performance breakdown by physical principle

Figure 12 breaks down model performance into different categories that include solid mechanics, fluid dynamics, optics, thermodynamics, and magnetism. While there is no category that can be considered "solved", performance varies across categories, with some showing promising indications and differences across models. Interestingly, all models perform much better on Spatial-IoU, a metric that has the weakest requirement in the sense that it is only sensitive to *where* an action occurred, not whether it occurred at the right time (as Spatiotemporal-IoU would track) or whether it had just the right *amount* of action (as measured by Weighted-spatial-IoU). Furthermore, even a relatively simple metric like MSE shows a large gap between physically realistic videos and model-generated predictions.

Outlook: Understanding without interaction?

Our findings are connected to a larger, interdisciplinary debate at the heart of intelligence: Does an understanding of the world emerge from predicting what happens next (next-video-frame prediction in artificial intelligence, predictive coding in neuroscience)—or, alternatively, is it necessary to interact with the world in order to understand it (as argued by proponents of embodied cognition and robotics)? In cognitive science, being able to interact with the world is seen as an important component for developing intuitive physics [5, 14, 57, 67], in combination with predicting the outcome of a person's actions [17, 38, 56]. In contrast, deep learning's current bread-and-butter approach is scaling models and datasets without interactions. Will these models essentially solve physical understanding—or instead, hit a limit after which one can only improve one's understanding of the world by interacting with it? The jury is still out on this question, but the benchmark and analyses introduced in this article might help quantifying this either way. In addition to future models, improvements could also come from inference-time scaling [30, 41, 55] such as sampling more. If this would indeed lead to strong results, it would raise the following question: from a model's perspective, is reality but one possibility among infinitely many others?

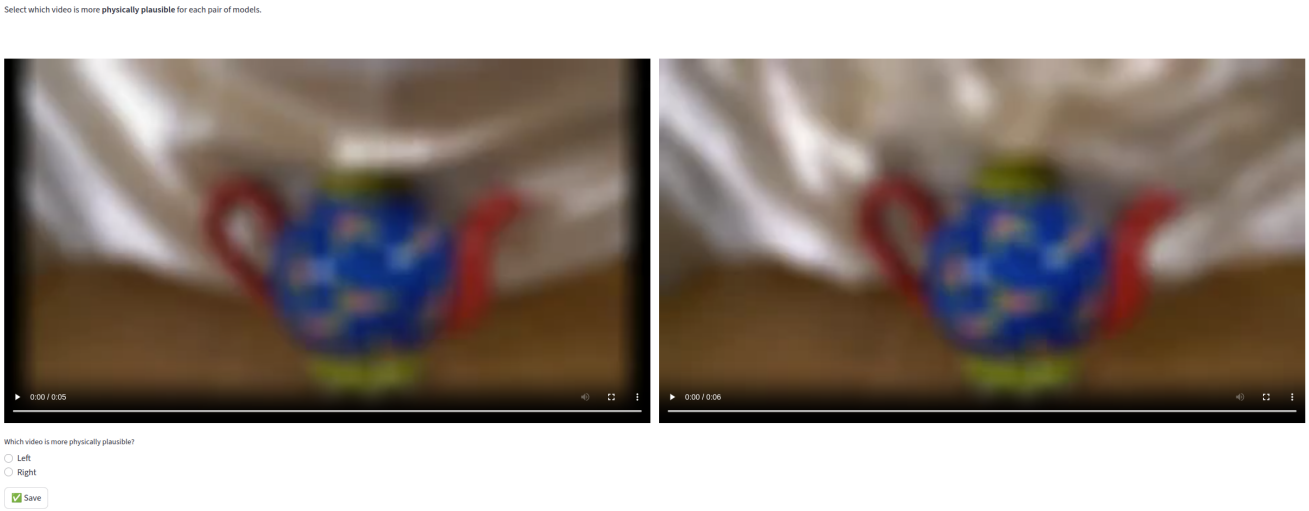

Figure 8. Human study interface used for pairwise plausibility comparisons.

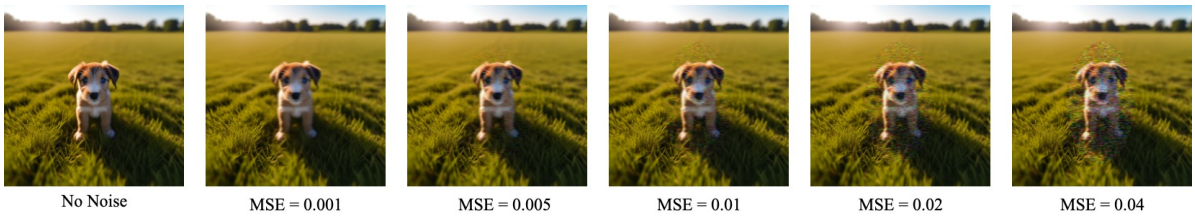

Figure 9. Since mean squared error (MSE) values can be hard to interpret, this figure shows the effect of a distortion applied to the scene, serving as a rough intuition for the effect of a MSE at different noise levels.

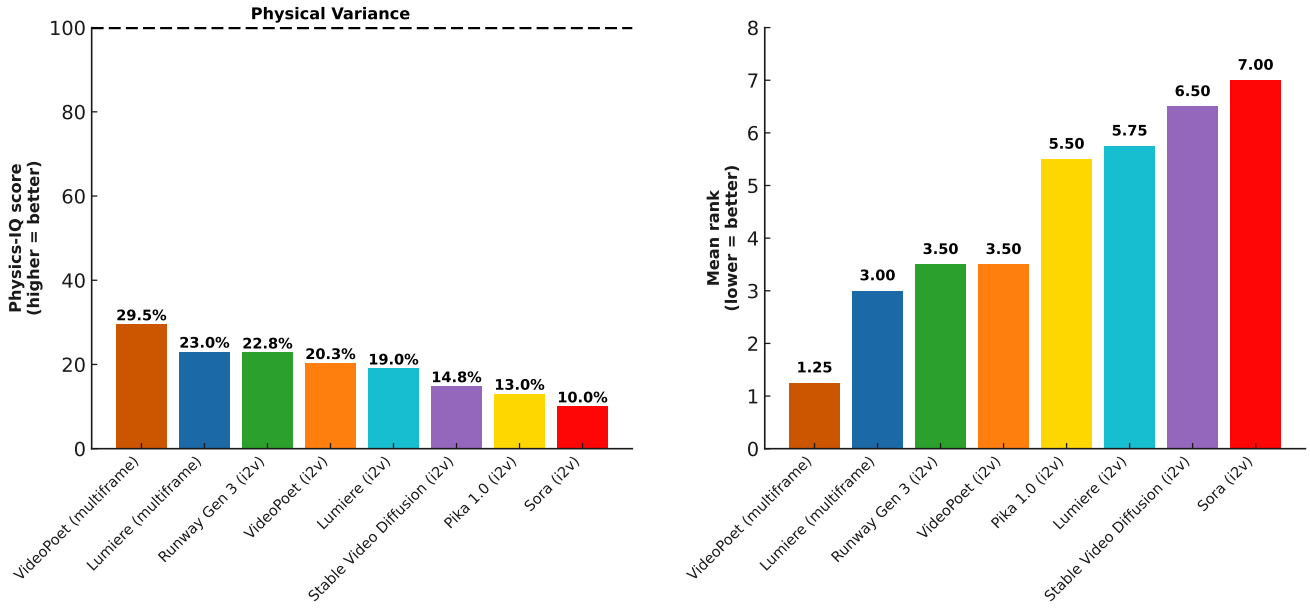

Figure 10. How well do current video generative models understand physical principles? **Left.** The Physics-IQ score is an aggregated measure across four individual metrics, normalized such that pairs of real videos that differ only by physical randomness score 100%. All evaluated models show a large gap, with the best model scoring 29.5%, indicating that physical understanding is severely limited. **Right.** In addition, the mean rank of models across all four metrics is shown here; the Spearman correlation between aggregated results on the left and mean rank on the right is high ($-.92, p < .005$), thus aggregating to a single Physics-IQ score largely preserves model rankings.

Figure 11. The switch frames (here: center view only) of all scenarios in the Physics-IQ benchmark. A switch frame is the last conditioning frame before a model is asked to predict 5 seconds of future frames.

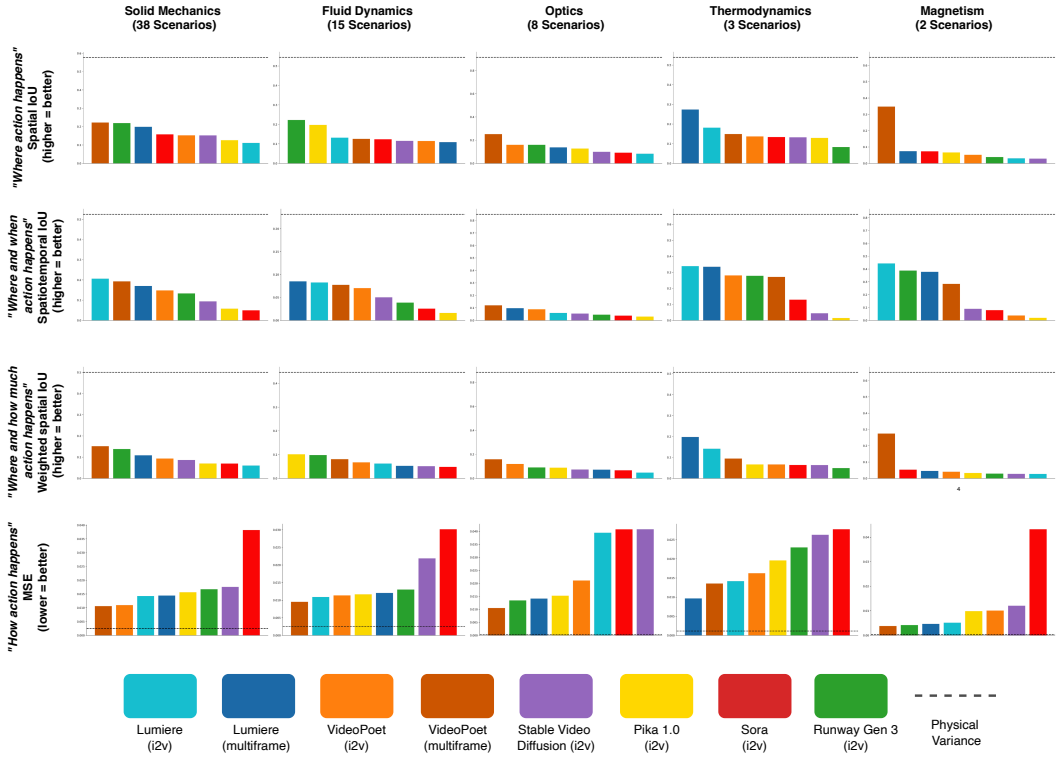

Figure 12. Performance comparison of video generative models across different physical categories (columns) and metrics (rows). For the top three metrics, higher is better; for the last metric lower values are best. Throughout, physical variance (i.e., the performance that is achievable by real videos differing only by physical randomness) is indicated by a dashed line. Across metrics and categories, models show a considerable lack in physical understanding. More lenient metrics like Spatial-IoU (top row) that only assess *where* an action occurred lead to higher scores than more strict metrics that also take into account e.g. *when* or *how much* action should be taking place.