
IntPhys 2: Benchmarking Intuitive Physics Understanding In Complex Synthetic Environments

Florian Bordes Quentin Garrido Justine T Kao Adina Williams Michael Rabbat

Emmanuel Dupoux
FAIR, Meta

Abstract

1 We present IntPhys 2, a video benchmark designed to evaluate the intuitive physics
2 understanding of deep learning models. Building on the original IntPhys bench-
3 mark [37], IntPhys 2 focuses on four core principles related to macroscopic objects:
4 Permanence, Immutability, Spatio-Temporal Continuity, and Solidity. These condi-
5 tions are inspired by research into intuitive physical understanding emerging during
6 early childhood. IntPhys 2 offers a comprehensive suite of tests, based on the
7 violation of expectation framework, that challenge models to differentiate between
8 possible and impossible events within controlled and diverse virtual environments.
9 Alongside the benchmark, we provide performance evaluations of several state-of-
10 the-art models. Our findings indicate that while these models demonstrate basic
11 visual understanding, they face significant challenges in grasping intuitive physics
12 across the four principles in complex scenes, with most models performing at
13 chance levels (50%), in stark contrast to human performance, which achieves near-
14 perfect accuracy. This underscores the gap between current models and human-like
15 intuitive physics understanding, highlighting the need for advancements in model
16 architectures and training methodologies.

17 1 Introduction

18 Understanding intuitive physics is a fundamental aspect of human cognition [34, 3, 6, 4, 44], enabling
19 individuals to effectively navigate and interact with the physical world. In recent years, there has been
20 a growing interest in replicating this intuitive understanding within artificial systems [10, 49, 35, 38].
21 However, despite advances in machine learning and computer vision, current models still fall short
22 of human capabilities in this domain [37, 50, 25, 13, 12, 8, 11]. The IntPhys benchmark [37] was
23 originally introduced to address the challenge of evaluating intuitive physics understanding in AI
24 models, providing a standardized framework for assessment. However, the benchmark had some
25 limitations, focusing on simple environments that lacked the variations and complexities found in the
26 real world. Furthermore, recent work [19] has shown that the benchmark has become saturated, with
27 predictive models such as V-JEPA[9] achieving high performance on it, highlighting the need for a
28 more challenging and diverse intuitive physics benchmark.

29 In this paper, we present IntPhys 2, an expanded and more comprehensive benchmark designed to
30 push the boundaries of intuitive physics understanding in artificial systems. IntPhys 2 evaluates four
31 key conditions inspired by human cognition: Object Permanence [3], Object Immutability [51, 52],
32 Spatio-Temporal Continuity [42], and Solidity [42]. These conditions are carefully selected to
33 encompass a broad range of physical principles, thereby providing a rigorous assessment of model
34 capabilities. The dataset contains 1416 videos that are divided in 3 different splits. The videos in
35 the Debug and Main splits are released along with their respective metadata while the last split is an
36 Held Out set, in which we release only the videos to avoid training data contamination. Unlike its

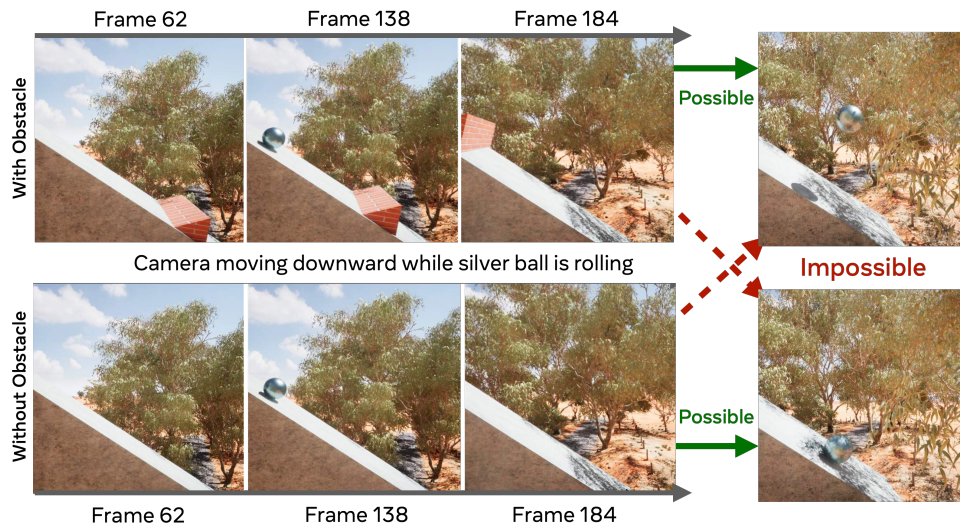


Figure 1: **Example of a scene in IntPhys2.** Each scene consists of a set of four videos. Two pairs depict possible outcomes, while the other two represent impossible outcomes. The presence of an obstacle or occluder determines the outcome: a possible outcome in the first pair becomes impossible in the second, and vice versa. In this example, a silver ball rolls down a path. If a brick obstacle is present, the ball should collide with it and change its trajectory. If the ball passes through the brick obstacle without altering its path, this outcome is deemed impossible. Conversely, when no obstacle is present, the ball’s trajectory should remain unchanged, making this the likely outcome.

predecessor that contained very basic and not fully realistic scenes, IntPhys 2 utilizes the full potential of Unreal Engine¹ using photorealistic environments made with dynamic shadows and lighting to better simulate real-world settings. IntPhys 2 improves upon the original IntPhys benchmark by introducing more realistic occlusions through the use of both fixed and moving cameras. Movement-based occlusions are more natural, capturing situations such as those that occur when an observer moves their head to look away and then back to the original point of view. By incorporating both fixed and moving cameras as well as using more complex scenes, IntPhys 2 provides a more comprehensive evaluation framework for intuitive physics understanding.

Using IntPhys 2, we performed a comprehensive performance evaluations of state-of-the-art predictive models and Multimodal Large Language Models (MLLMs)^[15]. While these models have achieved notable advancements, our findings indicate that they continue to struggle with the nuances of simple intuitive physics properties such as permanence and immutability, particularly in comparison to human performance, which remains consistently strong across all conditions. This disparity highlights the ongoing challenges in bridging the gap between artificial and human cognition, emphasizing the need for continued research and innovation in this critical area.

Our key contributions are as follows:

- A novel benchmark dataset for intuitive physics, featuring diverse scenarios with varying complexity levels. IntPhys 2 advances beyond existing benchmarks by incorporating photorealistic scenes with sophisticated visual elements (including complex lighting, shadows, occlusions, and textures), and employing both fixed and dynamic camera perspectives to simulate natural viewpoint changes.
- A comprehensive evaluation of state-of-the-art AI systems, including predictive models and Multimodal Large Language Models (MLLMs), establishing new baselines and identifying specific challenges in intuitive physics reasoning.

¹As a reminder, any use of content or technologies made available by Unreal and/or Epic Games, or any other provider, should comply with their applicable terms (such as the Content License Agreement available at <https://www.unrealengine.com/en-US/eula/content> or any other direct agreement one may have with Epic / Unreal)

2 Benchmark Design

From an early age, humans develop an innate ability to grasp basic physical principles [34][3][6][4][44], such as object permanence [3] (objects persist in space and time, even when they are out of sight), immutability [51][52] (objects maintain their shape and structure), spatio-temporal continuity [42] (objects move smoothly through space and time), and solidity [42] (objects occupy space and cannot pass through one another). These principles allow us to predict and interpret the behavior of inanimate objects in our environment, forming the foundation for more complex reasoning and decision-making processes. To systematically assess the development of these intuitive physics principles, the violation of expectations (VOE) [31][41] paradigm has been leveraged. This paradigm, which has been extensively used in studies with human infants [5][43], involves presenting them with scenarios where objects either behave in accordance with or violate these fundamental physical principles. By measuring infants' gaze time to these scenarios, researchers can infer their understanding of intuitive physics. Such a framework has been one of the main inspirations for IntPhys [37], and IntPhys 2 builds upon this foundation by adhering to the methodological framework established by its predecessor, employing a quadruplet video structure for each scene. This design comprises two possible and two impossible videos per scenario, configured such that the possible video of one scenario serves as the impossible video in another, and vice versa. This systematic arrangement is instrumental in mitigating low-level perceptual biases, thereby requiring models to engage with high-level temporal dependencies and underlying physical principles. By maintaining this rigorous structure, IntPhys 2 offers a robust and unbiased framework for assessing the depth of intuitive physics understanding in machine learning systems, preventing models from relying on shortcuts or latching onto spurious features and ensuring that model performance is more correlated with genuine cognitive capabilities rather than the exploitation of dataset-specific artifacts.

IntPhys 2 introduces several key advancements over previous benchmarks and datasets in the domain of intuitive physics understanding [25][50][37] that are illustrated in Figure 2. These enhancements are designed to provide a more rigorous and comprehensive evaluation of AI models, addressing limitations observed in earlier works. The core differences are as follows:

- **Focus on Occlusions:** Unlike previous benchmarks that may have included a variety of scenarios, IntPhys 2 exclusively considers occlusions. This focus allows for a more targeted assessment of a model's ability to maintain their understanding in the presence of visual obstructions.
- **Dynamic Camera Movements:** To create occlusions, IntPhys 2 employs static and dynamic camera movements. This approach not only increases the complexity of the scenes but also mimics real-world conditions where objects may be temporarily obscured from view due to changes in perspective.
- **Enhanced Realism:** The scenes in IntPhys 2 are crafted with improved realism, providing a more lifelike and challenging environment for models to navigate. This enhancement ensures that the benchmark more accurately reflects the complexities of the real-world.
- **Diverse Scene Variety:** IntPhys 2 significantly expands the diversity of scenes and tasks considered. Unlike traditional datasets that often feature a single scene per physical property, IntPhys 2 includes multiple tasks within each condition, offering a broader range of challenges and reducing the risk of overfitting to specific scene types.
- **Increased Short-Term Memory Demand:** The benchmark places a stronger emphasis on the need for short-term memory, requiring models to retain and utilize information over brief intervals effectively. This demand is critical for accurately predicting and understanding the dynamics of occluded objects.

We designed the benchmark with three distinct data splits to facilitate comprehensive evaluation (Table 1). The first split, known as the Debug set, includes five scenes with static camera setups and brightly colored assets. Each scene features a quadruplet of videos, supplemented by three additional videos that, while identical in configuration, display subtle variations due to environmental factors such as cloud movement or wind. These variations, though easy to miss to human observers, introduce minor pixel-level discrepancies that can influence model performance. This split is primarily intended for model calibration and to evaluate sensitivity to such noise. Ideally, a model should be robust to these negligible variations, demonstrating its ability to generalize beyond minor environmental fluctuations. We demonstrate ways to use this subset for qualitative analysis of predictive models in

Table 1: **IntPhys2 benchmark splits.** We release three separate splits. The first is intended for debugging only and provide some measurement on the model’s sensitivity to the video generation artifacts (such as mp4 compression or cloud moving the background of the scene). The second is the main evaluation set with three different sub-splits ("Easy", "Medium", "Hard"). The third is a held-out split that we release without additional metadata.

Split	Scenes	Videos	Description	Purpose
Debug Set	5	60	Static cameras, bright assets, 3 generations	Model calibration
Main Set	253	1,012	Static and moving cameras: 3 sub-splits: - Easy: Simple environments, colorful shapes - Medium: Diverse backgrounds, textured shapes - Hard: Realistic objects, complex backgrounds	Main evaluation set
Held-Out Set	86	344	Moving cameras, Mirrors hard sub-split, includes distractors	Main test set

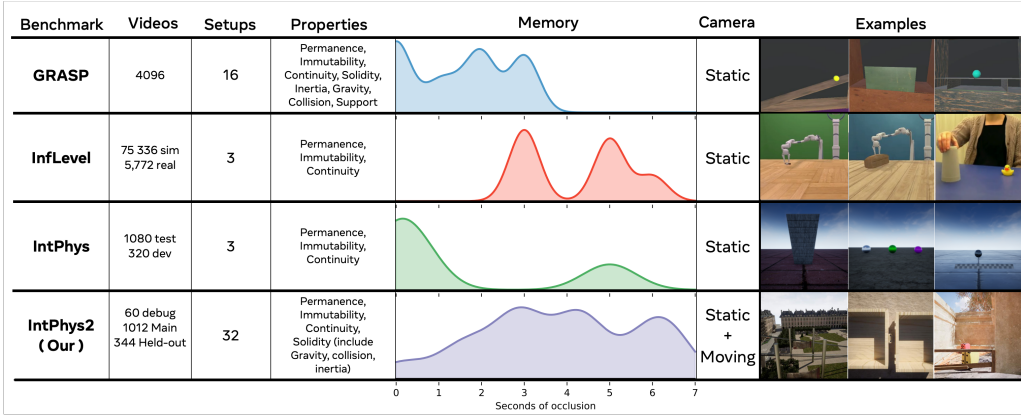


Figure 2: **Benchmark Comparison.** Our analysis compares four benchmarks: GRASP [25], InfLevel [50], IntPhys [37], and IntPhys2 (Ours). The benchmarks differ in their number of videos (simulated, real, test, development, debug, main, held-out), experimental setups, and physical properties assessed (permanence, immutability, continuity, solidity, inertia, gravity, collision, support). The density plots illustrate the distribution of occlusion durations (in seconds) for each benchmark. In contrast to other benchmarks, IntPhys2 covers a higher range of occlusion durations, allowing for a better assessment of a model’s short-term memory. Camera settings vary between static and moving configurations. Example frames from each benchmark are shown on the right.

116 appendix G The second split, termed the main set, comprises 253 scenes resulting in a total of 1,012
117 videos. This set is designed for zero-shot evaluations and serves as the primary basis for comparing
118 models. It is further divided into three sub-splits: an easy sub-split with simple environments and
119 shape-based colorful assets, a medium sub-split featuring more diverse backgrounds and simple
120 shapes with complex textures, and a hard sub-split characterized by highly diverse backgrounds
121 and assets shaped like real objects. The final split, is a held-out set for which metadata (containing
122 the ground truth about the plausibility of the events in the video) is not released. Solving this hard
123 subset requires an understanding that is robust to complex settings and that does not use metadata as
124 additional help. While the main set allows one to track progress more granularly and can be more
125 lenient on models, solving the held-out set is what will truly mean that a model understands intuitive
126 physics in complex settings.

127 3 Evaluation protocol : measuring intuitive physics understanding

128 **Human Evaluation.** To assess the gap between human and AI performance in understanding
129 intuitive physics, we conducted a human annotation task. This task was designed to evaluate the
130 ability of human participants to judge the physical plausibility of videos generated by a simulation
131 engine, similar to those used in video game development. The evaluation involved the 1,416 videos,
132 each with a duration of around 10 seconds. To ensure a comprehensive assessment, each video was

rated by three different annotators. The order of video presentation was randomized for each annotator to mitigate order effects and maintain the integrity of the evaluation process. To prevent attention fatigue and maintain high-quality ratings, each annotator was limited to evaluating a maximum of 96 videos. Prior to the main annotation task, annotators were shown only a set of 10 videos that were all physically plausible. This familiarization phase was designed to acquaint annotators with the types of videos produced by the simulation engine and to establish a baseline understanding of physical plausibility within the context of the task. Annotators were only informed that the initial set of 10 videos they saw depicted scenes where objects behaved in physically plausible ways. Following this, annotators were tasked with evaluating additional videos, some of which might contained errors that resulted in physically implausible object behavior. Annotators were instructed to watch each video carefully and in its entirety, considering the plausibility of object behavior based on real-world physics. They rated each video using a Likert scale, ranging from 1 (completely implausible) to 4 (very plausible). Then, we aggregated the results using a majority vote between the three annotators. This structured approach to human evaluation was crucial for obtaining reliable data on human performance, which serves as a benchmark for comparing the capabilities of current and future AI models in understanding intuitive physics.

Evaluating multimodal large language models. Our evaluation methodology for MLLMs diverges slightly from our human evaluation since current models 1) are not yet able to process eight videos in their input context 2) do not have long term visual memory 3) are not learning from previous context. To compensate for this lack of memory, we employed more detailed prompts for MLLMs asking only whether the video despite a plausible scenario. The prompts included explicit instructions about the video source being a simulator and that the model should base its answer solely on the events happening in the video, not on the quality of the simulation itself. To assess the models' sensitivity to prompts, we evaluated each model point-wise using the prompts presented in Table 5. The first prompt is concise and open-ended, requiring the model to respond with a simple "yes" or "no". In contrast, the second prompt is more specific, guided, and is expecting a binary digit as response. Anecdotally, MLLMs can be sensitive to the format of the required output [30], so we made a version of the second prompt in which we require a "yes/no" answer instead of the binary digit format. However, prompting is not the sole source of variance; sometimes, even with a temperature setting of 0, models can produce different answers to the same prompt and input data. Therefore, we ran each prompt at least twice to evaluate any variance in predictions. Ideally, the accuracy should remain consistent in such cases. To give MLLMs an advantage to compensate for their short-term memory limitations, we decided to show the best accuracy that can be obtained by a model across multiple different runs, instead of doing a majority vote like we did in the human evaluation. Lastly, since the number of frames that can be fed into an MLLM depends on its input context size, we had to run several experiments using a different number of input frames.

Evaluating prediction-based models. Inspired by the Violation of Expectation framework, Garrido et al. [19] introduced a model-based evaluation setup that measures how much a model is surprised when viewing an unexpected event compared to an expected event. A higher surprise indicates that the events violated the model's expectations, with impossible videos expected to elicit more surprise than possible ones. A proxy for surprise is the prediction error over a video for models that can predict the future [37][40][38][19]. We split a video into overlapping windows (typically 16-32 frames) that the model can process. For each window, the model predicts the target part based on the context, and the prediction error measures the model's surprise. Comparing surprise across videos probes the model's understanding, and we adapt the protocol for longer contexts as described in appendix F. The adaptation introduces constraints: events necessary for prediction must be in the context, and models must remember occluded objects. Models handling 16 frames at a time require a suitable framerate for prediction in order to balance memory and motion fluidity. Different experimental settings have different baseline prediction errors, making surprise comparisons challenging. Paired videos with identical content except for a physics-breaking event allow for controlled comparisons as the surprise difference can be attributed to the physics-breaking event, enabling precise probing of physics understanding, and giving rise to two evaluation protocols: pairwise and single video settings.

While we have described protocols designed for certain families of models, playing to their respective strengths, other protocols are possible. For example, in InfLevel-lab [50], models trained for action recognition are probed by measuring how out of distribution impossible videos are. This has however not yielded evidence of understanding in models. We thus chose to focus on models that have

either demonstrated previous understanding [19] or ones that can be probed akin to humans. The popularity of multimodal LLMs and predictive models also adds to the relevance of IntPhys 2 to current paradigms.

4 Experiments

In our evaluation, we investigated the capabilities of several state-of-the-art MLLMs, including both open-source and proprietary options. Our study featured the Gemini series (Gemini 1.5 Pro and Gemini 2.5 Pro Flash Preview [21]), as well as the latest versions of GPT4-o and Qwen-VL 2.5 [2]. We also evaluated three prediction-based methods: VideoMAEv2 [47] which predicts pixels directly, Cosmos-Predict1-4B [1] which predicts in the latent space of an autoencoder, and V-JEPA [9] which predicts in latent space. The main results are presented in Table 2. Notably, there is a meaningful gap between human and current model performances. The best model, Gemini 2.5 Flash, performed only slightly above random chance, except on the easy subset of our benchmark, where it achieved 64% accuracy. Table 3 provides more fine-grained results across four different conditions. The permanence condition appears to be the easiest for both models and humans, as objects are not moving by themselves. While there is no consistent trend for fixed versus moving camera scenarios among models, humans tend to perform slightly better in fixed camera settings. Overall, the gap between human and model performance remains significant across each split and data category.

Table 2: **Accuracy values showing best model performance across difficulty levels.** Most of the models were run a dozen of time with a different set of hyper-parameters. For a given model, we only report its best run in this table while the human performance is computed from a majority vote.

Model	Type	Easy	Medium	Hard	Overall	Held Out	IntPhys [37]
Human	-	96.17	97.8	95.5	96.44	92.44	-
GPT4-o [33]	MLLMs	57.69	54.75	54.17	53.75	53.19	-
Qwen-VL 2.5 [2]	MLLMs	50.96	53.25	51.49	52.27	49.12	-
Gemini-1.5 Pro [21]	MLLMs	58.65	53.0	52.67	52.27	52.10	55.81
Gemini-2.5 Flash [21]	MLLMs	64.42	56.75	54.46	55.63	56.10	56.39
VideoMAEv2-g [47]	Predictive	50.00	53.50	52.73	51.19	52.91	59.40
Cosmos-4B [1]	Predictive	46.00	52.00	48.05	49.41	48.84	85.42
V-JEPA-h + RoPE [9, 19]	Predictive	52.00	51.50	52.34	51.58	54.65	98.30

Table 3: **Accuracy across property and camera type.** For each model report the accuracy of each subset based on the best one across a set of hyperparameters. The smaller size of each subsets contributes to volatility in performance.

Model	Permanence		Immutability		Continuity		Solidity	
	Fixed	Moving	Fixed	Moving	Fixed	Moving	Fixed	Moving
GPT4-o	59.62	58.82	58.65	59.56	54.81	57.35	56.73	55.32
Qwen-VL 2.5	53.85	54.41	56.73	53.68	52.88	54.41	50.96	51.06
Gemini-1.5 Pro	55.77	55.88	56.73	56.73	54.80	54.80	56.73	56.73
Gemini-2.5 Flash	64.42	58.82	59.62	63.97	54.81	55.15	55.77	56.38
VideoMAEv2-g [47]	59.62	45.59	54.81	50.67	71.15	52.94	46.15	56.38
Cosmos-4B [1]	51.92	41.18	50.96	48.32	53.85	50.00	48.08	55.32
V-JEPA-h + RoPE [9, 19]	55.77	58.82	51.92	52.01	53.85	55.88	51.92	51.06
Human	100.0	99.26	97.11	90.44	99.04	94.44	96.15	95.21

4.1 Results: Multimodal Large Language Models

A key differentiator among these models is their language component. To assess the sensitivity of these models to the prompt, we performed an ablation study over the different prompts described in Table 5. Another key element is the method of processing video input and how many frames the model is fed with. The Gemini models are designed to accept MP4 videos directly, whereas the other models require video content to be converted into sequences of image frames. This required us to

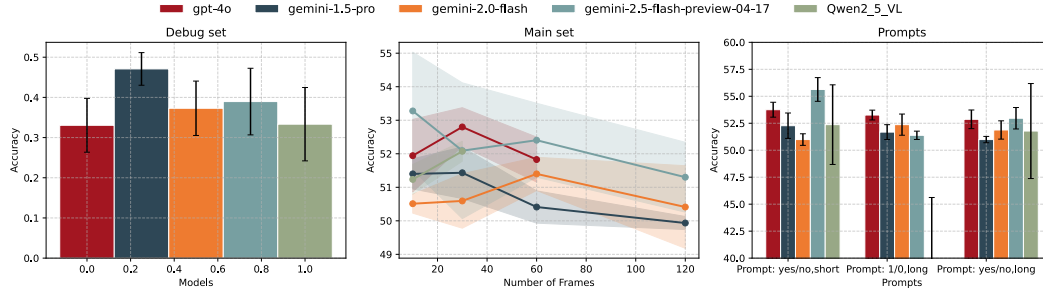


Figure 3: **Evaluation of model’s sensitivity.** (left) We conducted an ablation study examining various factors, including sensitivity to different prompts and the model’s variability in responses to identical inputs, as well as the difficulty level of the data. (right) We illustrate how a model’s performance varies with the number of frames it receives. Our findings indicate that most models struggle to effectively make use of an increased number of input frames.

adopt a customized approach for each model: for those unable to process MP4s directly, we employed uniform subsampling of frames according to the model’s input capacity. Conversely, for the Gemini models, we extended the video length to ensure the model received an adequate number of frames, as their API subsamples video at a rate of 1 frame per second. This allows us to run our ablations using either 10 frames, 30 frames, 60 frames, or 120 frames. Our evaluation included a series of ablation studies, which assessed the impact of various prompts, the number of frames inputted into each model, performance across different data difficulty levels, and sensitivity to randomness. To ensure consistency, we maintained a temperature setting of 0 across all models during testing.

The results of our evaluation are presented in Table 2 and 3 showcasing the optimal performance outcomes for each model across the various factors we examined. In Figure 3 we present an ablation analysis focusing on several key factors: robustness to generation artifacts, number of frames, and prompt. The first plot on the left utilizes the Debug set in IntPhys2, which contains three videos of the exact same scene. Even if these videos appear identical to humans, models can be very sensitive to any compression artifact noise. The model is considered correct only if it gives the correct answer for all 3 videos generated from the same scene. Thus, this accuracy shows the model’s performance as well as its robustness to imperceptible noise. Interestingly, Gemini 1.5 Pro seems much more consistent in its answers than Gemini 2.5 Flash. The second plot in Figure 3 provides insights into how the number of frames influences model performance. Our analysis reveals that most models experience a decline in performance when additional frames are introduced, suggesting a limitation in handling extended contexts [27]. The last plot showcases a prompt ablation, in which we can clearly see that the prompt selected can have a huge impact on performance. The best prompt for one model might not be the best one for another. Interestingly, it seems that models like gemini 2.5-flash perform better when being asked to answer by a yes/no answer than a binary digit. Qwen2.5-VL is not even able to follow the instruction correctly for the 0/1 prompt. These evaluations highlight the difficulties in making fair evaluations of MLLMs, as a given choice of hyperparameters can significantly change their performance. However, even the best models are still close to random performance, highlighting that current models have not learned a good physical world model, which might result in higher variance due to the randomness of the answers.

4.2 Results: Predictive models

We evaluate three prediction-based methods — two aimed at predicting pixels, one predicting in a latent space. For the former category, we use Cosmos-Predict1-4B [1] (Cosmos-4B) and VideoMAEv2 [47]. While Cosmos is directly trained to predict the future, acting as a world model, VideoMAEv2 is trained to reconstruct masked tokens from a video. Using it to predict the future is thus different from its training objective, as is the case for the methods we discuss next. The latent prediction methods we evaluate are the V-JEPA + RoPE [9, 45] models trained in [19]. For exact hyperparameters uses, confer appendix F.3

We report the accuracy of models on the different subsets of IntPhys 2 in Table 2. We find that models exhibit performance close to chance (50%) across all subsets, as well on the held-out set.

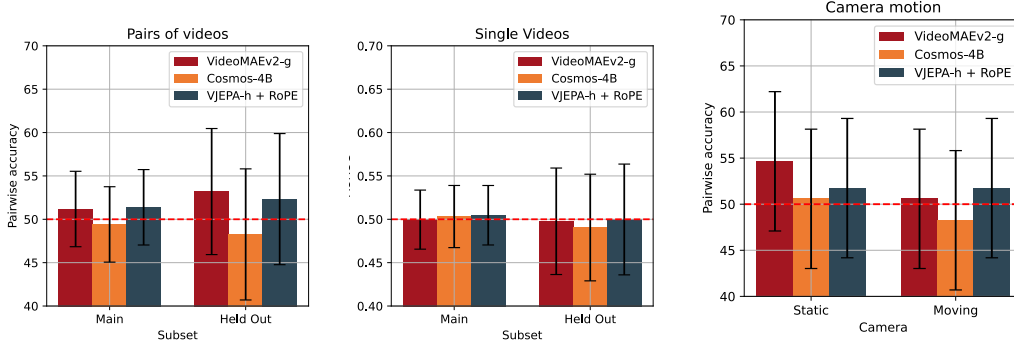


Figure 4: **Results for predictive models.** (left) When measuring whether models exhibit a higher surprise for impossible instances within a pair, we find that all tested models perform around chance level. (middle) This translates to the harder setting of single video classification, where the performance remains around chance. (right) Focusing on camera movements, one of the key chances in IntPhys 2, we find that model also struggle across camera settings. Confidence intervals obtained via bootstrapping.

This contrasts with the high accuracies that models obtain on IntPhys. Looking at per condition accuracy in Table 3 we further find that even when using more specific subsets of IntPhys 2—and thus using more specialized hyperparameters—the models struggle to surpass chance level. The exception is VideoMAEv2, which achieves a strong 71.15% accuracy on continuity for fixed camera, but these results should be contextualized against the small size of our subsets, which may introduce randomness in performance, meaning subset results should be taken with a grain of salt. In Figure 4 we further investigate the accuracy of models when classifying pairs or single videos. The latter is conceptually harder as the surprise must be as independent as possible from the general prediction difficulty of the video. It is however how humans are evaluated. We can see in the left and middle of Figure 4 that the accuracy is also close to random performance in single videos. In the right of Figure 4 we ablate further on models’ performance on the main subset of IntPhys 2. While the performance on the main set is close to random overall, it is possible that the models can perform better on certain subsets. This has been observed on InfLevel-lab, where models are able to perform well on one task, even when they perform close to random chance on others [19]. We thus isolate camera motion and find that even in the fixed camera setting—closest to existing benchmarks—the models perform close to chance level. These results demonstrate that models still struggle to understand intuitive physical concepts in complex scenarios, even if they are able to demonstrate a non-trivial understanding in simpler settings [19]. While multiple factors can explain this degradation in performance, a notable one is the stricter memory requirements. As illustrated in Figure 2 IntPhys 2 poses stricter requirements on short term memory than existing datasets, a property that video models can struggle with.

5 Related Work

Intuitive physics understanding benchmarking. Benchmarking intuitive physics understanding has been a focus of various datasets and challenges. INTPHYS [37] tests machine perception of core physical phenomena through videos with possible and impossible physical events, inspired by developmental psychology experiments with infants. GRASP [25] provides a comprehensive evaluation framework for embodied AI that includes intuitive physics as one of its core domains. INFLEVEL [50] focuses on measuring intuitive physics understanding in simulation as well as in the real world. Other notable benchmarks include PHYRE [7], which challenges models to solve physics puzzles through counterfactual reasoning, and OOPS [18], which challenges models to predict when unintended physical actions occur in in-the-wild videos.

Broader physics understanding benchmarking. To evaluate broader physical understanding, benchmarks have been developed to test complex physical phenomena, including rigid bodies, fluids, soft bodies, and their interactions. PHYSION [11] presents a comprehensive evaluation suite that assesses visual physical prediction based on object properties in a scene. CATER [22] focuses

on tracking and reasoning about moving objects, while CLEVRER [54] targets causal reasoning in video through descriptive, explanatory, predictive, and counterfactual questions. The Physical Interaction QA (PIQA) benchmark [13] tests physical commonsense knowledge through everyday human interactions. More recently, PHYSION++ [46] extends the original Physion benchmark to include more complex scenarios, and PHYSICS IQ [32] proposes a real-video benchmark to assess understanding of fundamental physical principles, including fluid dynamics, optics, solid mechanics, magnetism, and thermodynamics. More recently, Wang et al. [48] introduced a new synthetic benchmark for assessing the following properties velocity, acceleration, and collisions.

Methods tackling physical understanding. Various computational approaches have been developed to tackle physical understanding challenges. World models [23, 24] learn latent dynamics of environments to predict future states and plan actions. Generative models, particularly those based on graph neural networks [39], have shown promise in modeling physical dynamics by representing objects and their interactions. Joint Embedding Predictive Architectures (JEPA) [28] represent a self-supervised approach that learns to predict representations rather than raw observations. Large Language Models (LLMs) [16, 33] have demonstrated surprising capabilities in physical reasoning despite lacking explicit physical grounding. Hybrid approaches combining simulation-based reasoning with neural networks [53, 29] have also shown promise in physical understanding tasks by leveraging both data-driven learning and explicit physical knowledge. More specialized methods have also been developed to tackle intuitive physics understanding, often relying on hardwired priors, through the use of segmentation masks [35, 38] or de-rendering [40] for example.

Datasets generated with Unreal Engine. Unreal Engine has been widely adopted in the creation of synthetic datasets and benchmarks due to its advanced rendering capabilities and flexibility in simulating complex environments. For example, the CARLA simulator leverages Unreal Engine to provide a comprehensive autonomous driving benchmark, offering diverse driving scenarios and environmental conditions that are crucial for advancing research in autonomous vehicle perception and control [17]. Similarly, the AI2-THOR framework uses Unreal Engine to generate interactive environments for training and evaluating embodied AI agents, facilitating research in robotic manipulation and navigation [26]. UnrealCV also integrates with Unreal Engine to produce photorealistic images with ground truth annotations, supporting the development and evaluation of computer vision algorithms [36]. More recently, Bordes et al. [14] have used Unreal Engine for probing robustness of vision models. These projects highlight the engine’s utility in generating high-quality datasets that enable researchers to explore new frontiers in AI and machine learning.

6 Conclusion

Our evaluations on IntPhys 2 reveal significant limitations in current models’ intuitive physics reasoning capabilities, even for those that have shown promise in other benchmarks. The increased complexity and diversity of IntPhys 2, which mirrors real-world scenarios, exposes the models’ inability to effectively process longer sequences, higher framerates, and utilize short-term memory. These limitations are evident in the almost-random performance of models on IntPhys 2, with only recent Multimodal Large Language Models like Gemini 2.5 Flash achieving non-trivial performance. Ultimately, IntPhys 2 highlights the need for novel model architectures and training methodologies that can bridge the gap between artificial and human cognition. By addressing the limitations of current models and the benchmark itself, we can pave the way for more robust and human-like AI systems that can approximate human intuitive physics understanding.

Limitations While IntPhys 2 represents a significant advancement in benchmarking intuitive physics understanding, it also has limitations. Its reliance on synthetic environments may not fully capture the complexity of real-world physics, and the scope of physical principles covered is limited. Future research should focus on integrating real-world data, expanding the benchmark to include additional dimensions, and exploring more dynamic and interactive environments. Additionally, our experimental setting has limitations, including the number of video frames that models can process, which differs from human perception. Humans can process full videos and retain long-term context, whereas current models are limited to sub-sampling a specific number of frames and cannot process multiple videos in a single context. These limitations highlight opportunities for future research and development of more advanced model architectures.

References

- [1] N. Agarwal, A. Ali, M. Bala, Y. Balaji, E. Barker, T. Cai, P. Chattopadhyay, Y. Chen, Y. Cui, Y. Ding, et al. Cosmos world foundation model platform for physical ai. *arXiv preprint arXiv:2501.03575*, 2025.
- [2] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang, H. Zhong, Y. Zhu, M. Yang, Z. Li, J. Wan, P. Wang, W. Ding, Z. Fu, Y. Xu, J. Ye, X. Zhang, T. Xie, Z. Cheng, H. Zhang, Z. Yang, H. Xu, and J. Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.
- [3] R. Baillargeon and J. DeVos. Object permanence in young infants: Further evidence. *Child Development*, 62(6):1227, Dec. 1991. ISSN 00093920. doi: 10.2307/1130803.
- [4] R. Baillargeon and S. Hanko-Summers. Is the top object adequately supported by the bottom object? young infants’ understanding of support relations. *Cognitive Development*, 5(1):29–53, Jan. 1990. ISSN 08852014. doi: 10.1016/0885-2014(90)90011-H.
- [5] R. Baillargeon, E. S. Spelke, and S. Wasserman. Object permanence in five-month-old infants. *Cognition*, 20(3):191–208, 1985.
- [6] R. Baillargeon, A. Needham, and J. Devos. The development of young infants’ intuitions about support. *Early Development and Parenting*, 1(2):69–78, Jan. 1992. ISSN 1057-3593, 1099-0917. doi: 10.1002/edp.2430010203.
- [7] A. Bakhtin, L. van der Maaten, J. Johnson, L. Gustafson, and R. Girshick. Phyre: A new benchmark for physical reasoning. In *NeurIPS*, 2019.
- [8] H. Bansal, Z. Lin, T. Xie, Z. Zong, M. Yarom, Y. Bitton, C. Jiang, Y. Sun, K.-W. Chang, and A. Grover. Videophy: Evaluating physical commonsense for video generation. *arXiv:2406.03520*, 2024.
- [9] A. Bardes, Q. Garrido, J. Ponce, X. Chen, M. Rabbat, Y. LeCun, M. Assran, and N. Ballas. Revisiting feature prediction for learning visual representations from video. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. Featured Certification.
- [10] P. W. Battaglia, J. B. Hamrick, and J. B. Tenenbaum. Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences*, 110(45):18327–18332, Nov. 2013. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1306572110.
- [11] D. Bear, C. Fan, D. Mrowca, Y. Li, S. Alter, A. Nayebi, J. Schwartz, L. Fei-Fei, J. Wu, J. Tenenbaum, and D. L. K. Yamins. Physion: Evaluating physical prediction from vision in humans and machines. In *NeurIPS*, 2021.
- [12] Y. Bencheikroun, M. Dervishi, M. Ibrahim, J.-B. Gaya, X. Martinet, G. Mialon, T. Scialom, E. Dupoux, D. Hupkes, and P. Vincent. Worldsense: A synthetic benchmark for grounded reasoning in large language models. *arXiv:2311.15930*, 2023.
- [13] Y. Bisk, R. Zellers, J. Gao, Y. Choi, et al. Piqa: Reasoning about physical commonsense in natural language. *Proceedings of the AAAI conference on artificial intelligence*, 34:7432–7439, 2020.
- [14] F. Bordes, S. Shekhar, M. Ibrahim, D. Bouchacourt, P. Vincent, and A. S. Morcos. PUG: Photorealistic and semantically controllable synthetic data for representation learning. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. URL <https://openreview.net/forum?id=KRBoWULo2w>
- [15] F. Bordes, R. Y. Pang, A. Ajay, A. C. Li, A. Bardes, S. Petryk, O. Mañas, Z. Lin, A. Mahmoud, B. Jayaraman, M. Ibrahim, M. Hall, Y. Xiong, J. Lebensold, C. Ross, S. Jayakumar, C. Guo, D. Bouchacourt, H. Al-Tahan, K. Padthe, V. Sharma, H. Xu, X. E. Tan, M. Richards, S. Lavoie, P. Astolfi, R. A. Hemmat, J. Chen, K. Tirumala, R. Assouel, M. Moayeri, A. Talattof, K. Chaudhuri, Z. Liu, X. Chen, Q. Garrido, K. Ullrich, A. Agrawal, K. Saenko, A. Celikyilmaz, and V. Chandra. An introduction to vision-language modeling, 2024. URL <https://arxiv.org/abs/2405.17247>

- [16] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, et al. Language models are few-shot learners. *NeurIPS*, 2020.
- [17] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun. Carla: An open urban driving simulator. In *Proceedings of the 1st Annual Conference on Robot Learning*, pages 1–16, 2017.
- [18] D. Epstein, B. Chen, and C. Vondrick. Oops! predicting unintentional action in video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 919–929, 2020.
- [19] Q. Garrido, N. Ballas, M. Assran, A. Bardes, L. Najman, M. Rabbat, E. Dupoux, and Y. LeCun. Intuitive physics understanding emerges from self-supervised pretraining on natural videos. *arXiv preprint arXiv:2502.11831*, 2025.
- [20] T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. D. Iii, and K. Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021.
- [21] Gemini Team. Gemini: A family of highly capable multimodal models, 2024. URL <https://arxiv.org/abs/2312.11805>
- [22] R. Girdhar and D. Ramanan. Cater: A diagnostic dataset for compositional actions and temporal reasoning. In *ECCV*, 2020.
- [23] D. Ha and J. Schmidhuber. World models, 2018.
- [24] D. Hafner, T. Lillicrap, J. Ba, and M. Norouzi. Learning latent dynamics for planning from pixels. In *ICML*, 2019.
- [25] S. Jassim, M. Holubar, A. Richter, C. Wolff, X. Ohmer, and E. Bruni. Grasp: A novel benchmark for evaluating language grounding and situated physics understanding in multimodal language models. In K. Larson, editor, *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pages 6297–6305, 8 2024. doi: 10.24963/ijcai.2024/696. Main Track.
- [26] E. Kolve, R. Mottaghi, W. Han, E. VanderBilt, L. Weihs, A. Herrasti, D. Gordon, Y. Zhu, A. Gupta, and A. Farhadi. Ai2-thor: An interactive 3d environment for visual ai. *arXiv preprint arXiv:1712.05474*, 2017.
- [27] Y. Kuratov, A. Bulatov, P. Anokhin, I. Rodkin, D. Sorokin, A. Sorokin, and M. Burtsev. Babilong: Testing the limits of llms with long context reasoning-in-a-haystack, 2024. URL <https://arxiv.org/abs/2406.10149>.
- [28] Y. LeCun. A path towards autonomous machine intelligence. Technical Report, Meta AI, 2022.
- [29] Y. Li, A. Torralba, A. Garg, N. Snavely, and J. Wu. Visual grounding of learned physical models. In *ICML*, 2020.
- [30] D. X. Long, H. N. Ngoc, T. Sim, H. Dao, S. Joty, K. Kawaguchi, N. F. Chen, and M.-Y. Kan. Llms are biased towards output formats! systematically evaluating and mitigating output format bias of llms, 2025. URL <https://arxiv.org/abs/2408.08656>.
- [31] F. Margoni, L. Surian, and R. Baillargeon. The violation-of-expectation paradigm: A conceptual overview. *Psychological Review*, 131(3):716–748, Apr. 2024. ISSN 1939-1471, 0033-295X. doi: 10.1037/rev0000450.
- [32] S. Motamed, L. Culp, K. Swersky, P. Jaini, and R. Geirhos. Do generative video models understand physical principles? *arXiv preprint arXiv:2501.09038*, 2025.
- [33] OpenAI. Gpt-4 technical report, 2023.
- [34] J. Piaget. *The Construction of Reality in the Child*. Basic Books, 1954.
- [35] L. S. Piloto, A. Weinstein, P. Battaglia, and M. Botvinick. Intuitive physics learning in a deep-learning model inspired by developmental psychology. *Nature Human Behaviour*, 6(9): 1257–1267, July 2022. ISSN 2397-3374. doi: 10.1038/s41562-022-01394-8.

- [36] W. Qiu, Q. Zhou, C. Chen, and A. Yuille. Unrealcv: Virtual worlds for computer vision. *ACM Multimedia Open Source Software Competition*, 2017.
- [37] R. Riochet, M. Y. Castro, M. Bernard, A. Lerer, R. Fergus, V. Izard, and E. Dupoux. Intphys 2019: A benchmark for physical commonsense understanding. In *ECCV Workshop*, 2018.
- [38] R. Riochet, J. Sivic, I. Laptev, and E. Dupoux. Occlusion resistant learning of intuitive physics from videos. *arXiv:2005.00069*, 2020.
- [39] A. Sanchez-Gonzalez, J. Godwin, T. Pfaff, R. Ying, J. Leskovec, and P. Battaglia. Learning to simulate complex physics with graph networks. In *ICLR*, 2020.
- [40] K. Smith, L. Mei, S. Yao, J. Wu, E. Spelke, J. Tenenbaum, and T. Ullman. Modeling expectation violation in intuitive physics with coarse probabilistic object representations. *Advances in neural information processing systems*, 32, 2019.
- [41] E. S. Spelke. Preferential-looking methods as tools for the study of cognition in infancy. In G. Gottlieb and N. A. Krasnegor, editors, *Measurement of audition and vision in the first year of postnatal life: A methodological overview*, pages 323–363. Ablex Publishing, 1985.
- [42] E. S. Spelke, K. Breinlinger, J. Macomber, and K. Jacobson. Origins of knowledge. *Psychological Review*, 99(4):605–632, 1992. ISSN 1939-1471, 0033-295X. doi: 10.1037/0033-295X.99.4.605.
- [43] E. S. Spelke, R. Kestenbaum, D. J. Simons, and D. Wein. Spatiotemporal continuity, smoothness of motion and object identity in infancy. *British Journal of Developmental Psychology*, 13(2):113–142, 1995. doi: <https://doi.org/10.1111/j.2044-835X.1995.tb00669.x>. URL <https://bpspsychub.onlinelibrary.wiley.com/doi/abs/10.1111/j.2044-835X.1995.tb00669.x>
- [44] E. S. Spelke, R. Kestenbaum, D. J. Simons, and D. Wein. Spatiotemporal continuity, smoothness of motion and object identity in infancy. *British Journal of Developmental Psychology*, 13(2): 113–142, June 1995. ISSN 0261-510X, 2044-835X. doi: 10.1111/j.2044-835X.1995.tb00669.x.
- [45] J. Su, Y. Lu, S. Pan, B. Wen, and Y. Liu. Roformer: enhanced transformer with rotary position embedding. corr abs/2104.09864 (2021). *arXiv:2104.09864*, 2021.
- [46] F. Tung, M. Ding, Z. Chen, D. M. Bear, C. Gan, J. B. Tenenbaum, D. L. K. Yamins, J. Fan, and K. A. Smith. Physion++: Evaluating physical scene understanding that requires online inference of different physical properties. *arXiv*, 2023.
- [47] L. Wang, B. Huang, Z. Zhao, Z. Tong, Y. He, Y. Wang, Y. Wang, and Y. Qiao. Videomae v2: Scaling video masked autoencoders with dual masking. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14549–14560, June 2023.
- [48] X. Wang, W. Ma, A. Wang, S. Chen, A. Kortylewski, and A. Yuille. Compositional 4d dynamic scenes understanding with physics priors for video question answering. In *International Conference on Learning Representations (ICLR)*, 2025. URL <https://openreview.net/pdf?id=6Vx28LSR7f>
- [49] N. Watters, A. Tacchetti, T. Weber, R. Pascanu, P. Battaglia, and D. Zoran. Visual Interaction Networks. *arXiv:1706.01433*, June 2017.
- [50] L. Weihs, A. Yuille, R. Baillargeon, C. Fisher, G. Marcus, R. Mottaghi, and A. Kembhavi. Benchmarking progress to infant-level physical reasoning in AI. *Transactions on Machine Learning Research*, 2022.
- [51] T. Wilcox. Object individuation: Infants’ use of shape, size, pattern, and color. *Cognition*, 72(2):125–166, 1999.
- [52] T. Wilcox and C. Chapa. Priming infants to attend to color and pattern information in an individuation task. *Cognition*, 90(3):265–302, 2004.

- 480 [53] J. Wu, I. Yildirim, J. J. Lim, W. T. Freeman, and J. B. Tenenbaum. Learning to see physics in
481 everyday life. In *NeurIPS*, 2017.
- 482 [54] K. Yi, C. Gan, Y. Li, P. Kohli, J. Wu, A. Torralba, and J. B. Tenenbaum. Clevrer: Collision
483 events for video representation and reasoning. In *ICLR*, 2020.

NeurIPS Paper Checklist

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [Yes]

Justification: We propose and release a new benchmark to evaluate intuitive physic in MLLMs and predictions models.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: We do it in the conclusion.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: The paper does not include theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [\[Yes\]](#)

Justification: Yes, all the experimental details are in the appendix and the code repository that will be released upon public release.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
 - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The benchmark and code are available.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: Yes, we specify all the data splits that we made. Our benchmark is only for evaluation only which reduce the number of hyper-parameters since we are not optimizing anything.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: Yes, we report errors bar for each models accross different seed, prompts or other factors of variations to ensure that we have a good view of a model’s performance. If not possible, we use bootstrapping to compute confidence intervals.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: Compute resources are described in the appendix.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification:

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [NA]

Justification: no direct societal impact of the work performed.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: We not train or introduce new models. There aren't any safety risks in the data we are releasing

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The list of assets used will be available in our github repository.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset’s creators.

13. New assets

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: Yes, we are also releasing a Datasheet for IntPhys2 that can be found in the appendix.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. Crowdsourcing and research with human subjects

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer:[Yes]

Justification: We explain how we perform the human evaluation in section **D**

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. Institutional review board (IRB) approvals or equivalent for research with human subjects

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: [NA]

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

798 **16. Declaration of LLM usage**

799 Question: Does the paper describe the usage of LLMs if it is an important, original, or
800 non-standard component of the core methods in this research? Note that if the LLM is used
801 only for writing, editing, or formatting purposes and does not impact the core methodology,
802 scientific rigorousness, or originality of the research, declaration is not required.

803 Answer: [NA]

804 Justification: the core method development in this research does not involve LLMs as any
805 important, original, or non-standard components.

806 Guidelines:

- 807 • The answer NA means that the core method development in this research does not
808 involve LLMs as any important, original, or non-standard components.
- 809 • Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>)
810 for what should or should not be described.