WorldBench: Disambiguating Physics for Diagnostic Evaluation of World Models

Anonymous CVPR submission

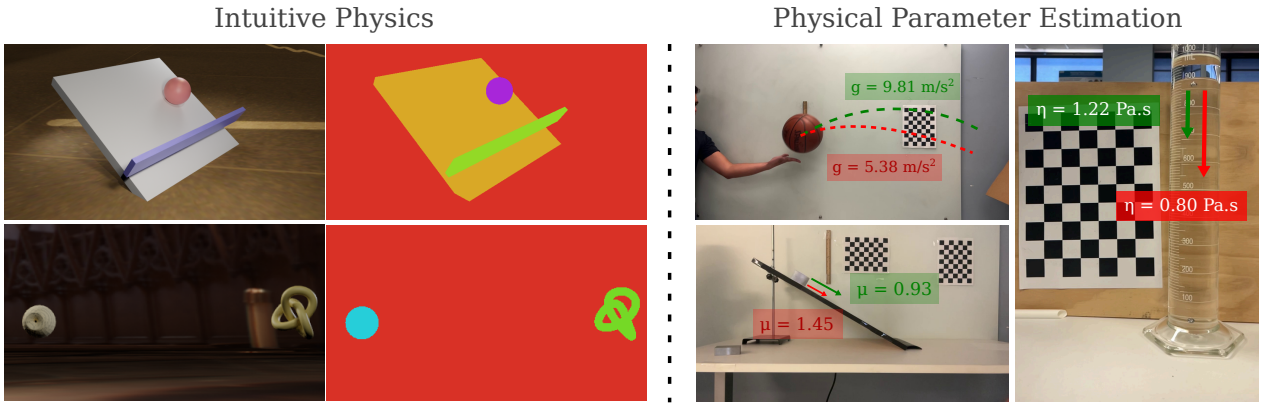

Figure 1. **We introduce WorldBench, a video-based benchmark to evaluate world foundation model performance on specific physical concepts/constants and material properties.** Prior benchmarks typically either entangle multiple concepts (making it difficult to uniquely identify model failures) or use coarse-grained metrics like binary selection (making it difficult to disambiguate between visually realistic and physically accurate trajectories). Our benchmark leverages both an intuitive physics and physical parameter estimation subset to provide greater insight into world model performance.

Abstract

Recent advances in generative foundational models, often termed "world models," have propelled interest in applying them to critical tasks like robotic planning and autonomous system training. For reliable deployment, these models must exhibit high physical fidelity, accurately simulating real-world dynamics. Existing physics-based video benchmarks, however, suffer from entanglement, where a single test simultaneously evaluates multiple physical laws and concepts, fundamentally limiting their diagnostic capability. We introduce WorldBench, a novel video-based benchmark specifically designed for concept-specific, disentangled evaluation, allowing us to rigorously isolate and assess understanding of a single physical concept or law at a time. To make WorldBench comprehensive, we design benchmarks at two different levels: 1) an evaluation of intuitive physical understanding with concepts such as object permanence or scale/perspective, and 2) an evaluation of low-level physical constants and material properties such as friction coefficients or fluid viscosity. When SOTA video-based world models are evaluated on WorldBench, we find specific patterns of failure in particular physics concepts, with all tested models lacking the physical consistency required to generate reliable real-world interactions. Through its concept-specific evaluation, WorldBench offers a more nuanced and scalable framework for rigorously evaluating the physical reasoning capabilities of video generation and world models, paving the way for more robust and generalizable world-model-driven learning.

1. Introduction

Imagine watching a tower of blocks teeter and fall, or a ball rolling its way down a staircase. As humans, we effortlessly predict its motion. However, this intuitive grasp of physical dynamics remains a core challenge for AI. Recent world foundation models, most notably NVIDIA's Cosmos [1], promise to learn such skills at scale, suggesting that these models can be used as synthetic data generators for the real world. Rigorously evaluating these claims requires benchmarks that are designed and focused on probing physical understanding at a concept-specific level.

Existing benchmarks for physical reasoning tasks often provide coarse-grained or binary metrics on scenes that entangle multiple physics concepts, limiting their diagnostic ability. For example, benchmarks such as PHYRE [4] or CLEVRER [49] contain scenes which require a core under-

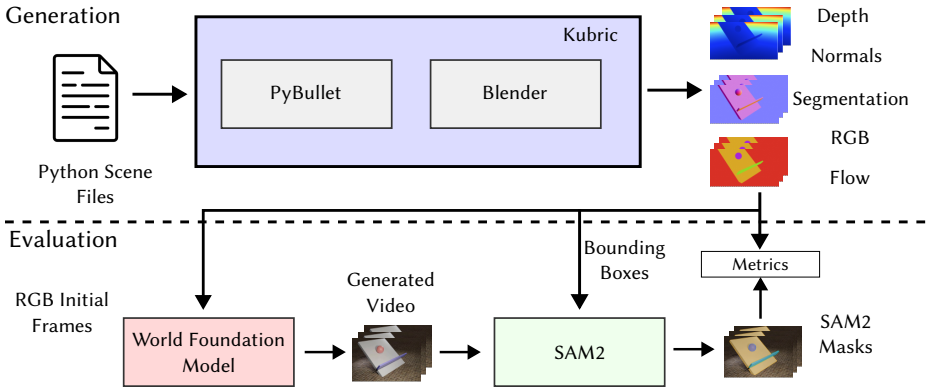

Figure 2. **Overview of our generation and evaluation process**. For generation (top), we use Kubric, which uses PyBullet and Blender under the hood. During evaluation (bottom), we first pass the initial frames of the generated video to the world foundation model which completes the video. The completed video is passed to SAM2 along with bounding boxes based on ground truth masks. The segmentations outputted by SAM2 are compared to ground truth segmentations to obtain the final metrics.

standing of multiple concepts (perspective accuracy, collision dynamics, and support relations). While some benchmarks like Physion [5] and IntPhys2 [9, 34] do provide some level of granularity in testing physical concepts, they do not contain experiment-level tests of specific physical parameters or material constants. Furthermore, a reliance on coarse-grained or binary metrics like object contact prediction limit their ability to capture nuanced physical phenomena, such as object dynamics (velocity, acceleration, rotation, etc.), deformation, or occlusion.

In this paper, we introduce **WorldBench**, a novel benchmark designed to rigorously evaluate the disentangled, concept-specific, physical reasoning capabilities of WFMs through video prediction. Our testing framework uses fine-grained categories of evaluation to fill in a critical gap in in the current research landscape. The video-based output requires models to accurately forecast the physically plausible evolution of full visual scenes over time, providing a more robust signal than previous binary metrics. To achieve this nuanced assessment while ensuring repeatable, interpretable outcomes, we design simplified, yet physically rich and visually realistic scenes.

Our benchmark is split into two subsets. The intuitive physical understanding subset targets four critical principles: motion physics, object permanence, support relations, and scale/perspective. This subset is meant to benchmark a model's ability to generate plausible dynamics governed mostly by their respective core principles (e.g. ball rolling behind pillars, object moving towards the camera, etc.).

WorldBench also includes a physical parameter estimation subset. This feature requires the video output to accurately adhere to specific, known physical parameters that govern the scene's dynamics, such as gravitational acceleration, fluid viscosity, and friction coefficients. By enforcing adherence to these measurable parameters, we provide a definitive, quantifiable metric for assessing a model's true grasp of each underlying physical law.

Our proposed task, which we term "constrained video prediction," allows for a significantly more nuanced and detailed assessment of physical laws than previous work, including a more fine-grained diagnosis of which concept a model underperforms on, and an accurate representation of object dynamics (velocity, acceleration, rotation, deformation, and occlusion). The first subset's focus on specific, real-world physics properties, such as object permanence and scale/perspective—which are absent in prior benchmarks like Physion—ensures that **WorldBench** provides a richer signal of how close these models are to truly learning and understanding real-world dynamics. The second subset's focus on real-world physical parameters also enables us to directly measure the physical accuracy of world models, which is particularly critical if these models aim to be synthetic data generators. For future work in the evaluation of these world models, our benchmark also leads to a wider array of downstream tasks, such as object tracking, anomaly detection, action planning, etc.

Using **WorldBench**, we extensively test state-of-the-art world models, including the Cosmos architecture, revealing substantial gaps in physical consistency and generalization when compared to both real-world captured and simulated, physically accurate expectations. Importantly, we find that models tend to generate visually realistic scene evolutions (e.g. ball follows parabolic trajectory), but fail to adhere to physical parameters (e.g. ball accelerates downward at $9.8\frac{m}{s}^2$). By leveraging video outputs rather than binary selection tasks, our benchmark provides a way to quantitatively disambiguate between visually plausible and physically accurate model outputs. Our results highlight the critical limitations of current world model architectures and motivate further research into physically grounded learning.

Below is a summary of our key contributions:

- We introduce a video-based benchmark comprised of real and synthetic videos to evaluate WFM performance on specific physical concepts and parameters.
- Our first subset provides scenes designed to test a model's qualitative understanding of foundational concepts such as object permanence, scale/perspective, support relations, and motion physics.
- Our second subset provides carefully designed experiments which directly measure a models ability to reproduce specific well-defined physics constants and behaviors such as gravitational acceleration, fluid viscosity, and friction coefficients.
- We perform an empirical analysis of SOTA WFMs and Image-to-Video models to identify concept-specific shortcomings and gaps in physical understanding.

2. Related Work

2.1. World Foundation Models

A significant body of work has emerged around "world models", models that can understand and predict the real world, in recent years. Initial work in this space focused on vision-language models [27, 29, 35], but recent work has been on using video generation models [3]. These models typically leverage transformer architectures and either latent diffusion models [6, 7, 11, 16, 23, 24, 28, 36, 41] or autoregressive models [15, 32, 42–47, 50] to achieve temporally consistent video synthesis. However, while these models are able to generate visually realistic and aesthetically pleasing outputs, there has also been a recent growth of research surrounding physically accurate generations. The recent Cosmos [1, 2] aims to be a "world foundation model", which can output temporally and physically accurate videos that can be leveraged for training downstream AI models that interact with the physical environment. Cosmos can generate these videos using either a transformer-based autoregressive model or a transformer-based diffusion model, training a large corpus of over 100M video clips, labeled by numerous different vision-language models [38]. Similarly, other models such as Genie [10], also attempt at creating a "world foundation model" capable of generating physically accurate interactive environments. It uses a novel video tokenizer and a causal action model, passing both the video tokens and action latents to an autoregressive dynamics model for prediction. However, note that Genie is currently closed-source and not available for evaluation under our proposed benchmark. These "world foundation models" claim to be physically accurate enough for their outputs to be used as simulated data, but little to no evaluations have been developed so far to validate this claim.

2.2. Image-to-Video Models

A closely related and highly active area of research is Image-to-Video (I2V) generation, which focuses on synthesizing a video sequence from a single static image input. The fundamental challenge in I2V models is temporal consistency: ensuring that the generated frames maintain the identity, structure, and appearance of the initial image while introducing plausible motion and scene evolution.

Modern I2V models typically "inflate" standard image diffusion models with a temporal dimension, using temporal convolution or attention layers [6, 7, 20–22, 24, 25, 39, 40, 48]. Notable models include CogVideoX [24, 48], which leverage a 3D-VAE for enhanced compression and an expert transformer for better text-video alignment, as well as WAN [39] which leveraged large-scale data training and a spatio-temporal VAE to achieve SOTA results on previous benchmarks.

While I2V models excel at animating scenes with high visual fidelity and maintaining the initial scene's identity, their primary design objective has been aesthetic realism and adherence to user-specified motion (often via text prompt), rather than physical plausibility. The evaluation of I2V models is typically focused on metrics like Fréchet Inception Distance (FID), Inception Score (IS), and temporal coherence measures, which assess visual quality and smoothness but do not inherently check for adherence to physical laws like gravity, friction, or object dynamics. This gap underscores the need for benchmarks like WorldBench, which can systematically test whether the generated motion reflects a true understanding of the physical parameters required for real-world simulation.

2.3. Physics Datasets and Benchmarks

There has been a growing interest in the community to evaluate the physical understanding and reasoning abilities of modern vision models [14, 19, 26, 31, 34, 37]. Datasets like PHYRE [4] focus on simplistic 2D scenarios constructed from balls and rectangular bars, with dynamics like collision, gravity, and friction. CLEVRER [49] is a video reasoning benchmark designed with simple structures for tasks including description, explanation, prediction, and counterfactuals. The MOVi set of datasets [18], are multi-object video datasets, targeting object-centric models and their ability to detect and discover object boundaries in videos. More recently, the Physion dataset [5] compiles a set of visually realistic videos separated between 8 different physics scenarios: dominoes, support, collide, contain, drop, link, roll, and drape. It leverages the object contact prediction (OCP) task to evaluate the physical understanding ability of models. While considerable progress has been made in this space, all prior work are deficient in at least one key area. Datasets like PHYRE and CLEVRER lack in visual realism and are made up of overly simplistic objects and structures.

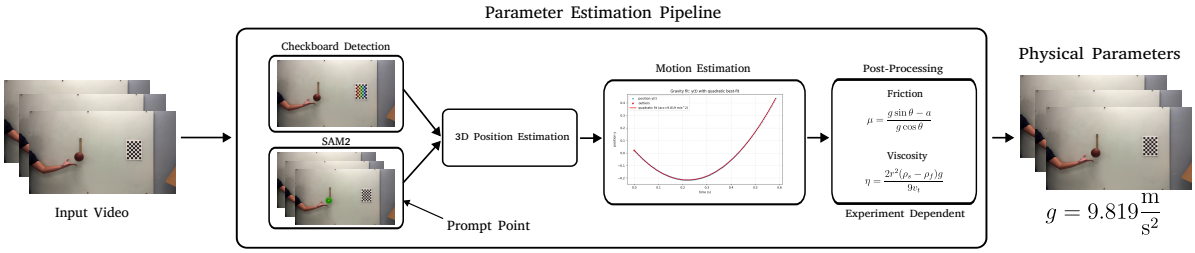

Figure 3. **Overview of our physical parameter estimation pipeline.** Given an input video, we first use checkerboard detection and SAM2 to extract 3D positions for objects in the video. We then fit curves to these parameters to estimate relevant physical properties such as acceleration or terminal velocity. These are then post-processed, if needed, to calculate the relevant physical parameters.

Table 1. Our proposed benchmark is the first to have concept-specific evaluation of both intuitive physics concepts as well as parameter-based experiments. Our proposed **WorldBench** is also enables distengled evaluation by video outputs, enabling more nuanced and fine-grained evaluation.

Benchmark	3D	Real-World Data	Video Output	Disentangled Intuitive Physics	Disentangled Parameter Estimation	Task
PHYRE [4]	✗	✗	✗	✗	✗	Scene Modification
CLEVRER [49]	✓	✗	✗	✗	✗	Visual Question Answering
MOVi [18]	✓	✓	✓	✗	✗	Object Tracking
Physion [5]	✓	✗	✗	✓	✗	Object Contact Prediction
IntPhys2 [9, 34]	✓	✗	✗	✓	✗	Binary Selection
Ours	✓	✓	✓	✓	✓	Frame Prediction

The MOVi datasets has visually diverse scenes and objects, but focuses on object discovery rather than physical reasoning tasks. As described in earlier sections, while Physion does have a wide variety of different tasks and visually realistic video inputs, the sole use of the OCP task for physical understanding evaluation limits its ability to be used to evaluate the new wave of world foundation models such as Cosmos [1] or Genie [10]. Compared to these, our benchmark is the first benchmark where the inputs and outputs are both video based. This aligns much more closely with the architectures of today's models, making it a better fit for physics evaluation. Reference Table 1 for a comparison summary of notable physics datasets and benchmarks.

3. Benchmark

In order to test concept-specific physical understanding in "world foundation models" (WFMs), we introduce a novel benchmark, **WorldBench**, designed to evaluate their physics prediction capabilities. The core methodology is to provide these models with a short input video and tasking them to generate a continuation. Each video is designed to evaluate a single physics concept or law.

WorldBench is divided into two distinct, yet complementary components, designed to probe both the intuitive and the engineering-grade understanding of the physical world. This structure ensures a comprehensive assessment of both high-level "intuitive physics" and low-level physics constants.

All simulated videos in our benchmark are rendered using Kubric, an open-source physics simulation pipeline [17]. Kubric uses PyBullet [13] as the physics simulator and Blender [8] as the renderer. This allows us to combine the physically accurate simulation of PyBullet with the high-quality rendering of Blender.

3.1. Intuitive Physics Understanding

The first subset is designed to assess the model's implicit understanding of core, foundational physics concepts, often referred to as "intuitive physics." This section takes inspiration from developmental psychology, where infants quickly acquire an understanding of the world's basic operational rules through observation. The goal is to determine if large-scale models, trained on vast quantities of video data, have successfully internalized these necessary cognitive building blocks. Our benchmark focuses on four fundamental physics concepts: Motion Physics (how objects move and interact), Support Relations (how objects are supported or balanced), Object Permanence (understanding that objects continue to exist when hidden), and Scale/Perspective (how size and spatial relationships change with motion/viewpoint). This is not an exhaustive list of physical concepts, but is designed to cover a range of common real-world scenarios.

The difficulty in the design of this subset is narrowing the object trajectories to ensure consistent scene evolution, but allowing for enough variation for a rich and robust benchmark. To that end, for each concept, we construct 3-5 scenarios, where each of these scenarios is hand-designed to capture some element of the concept it is testing. Each scenario has 25 videos, each of which is generated by randomizing various components such as object type, location and material. In addition, we collect 10-14 real videos for each high-level concept, generally from a subset of the scenarios. In total, this subset of **WorldBench** is made up of 469 videos spanning the 4 concepts: 425 simulated, and 44 real. Each video is 132 frames long and includes ground truth object segmentations. The synthetic videos additionally include ground truth depth, normals, and optical flow. All meshes and objects used in our simulations were taken

Table 2. **Validation of Physics Parameters on the PerfectPhysics subset.** We show the evaluation pipeline on our captured ground truth videos here. All estimated parameters are within an acceptably small margin of error

	Gravity		Viscosity			Friction				
Model	Free-Fall (m/s^2)	Parabolic (m/s^2)	Glycerine (Pa.s)	Corn Syrup (Pa.s)	Honey (Pa.s)	Wood	Rubber	Sandpaper (80)	Sandpaper (3000)	Plastic
Estimated	9.78 $\pm$ 0.38	9.85 $\pm$ 0.36	1.22 $\pm$ 0.01	5.84 $\pm$ 0.02	13.82 $\pm$ 0.75	0.35 $\pm$ 0.05	0.93 $\pm$ 0.10	1.06 $\pm$ 0.05	0.30 $\pm$ 0.02	0.22 $\pm$ 0.03
Ground Truth	9.81	9.81	1.2	5.0 - 7.0	14.1	0.2 - 0.5	0.5 - 2.0	0.7 - 1.1	0.2 - 0.5	0.05 - 0.2

Table 3. **Foreground mIoU results on the Physical Principles Understanding subset (Simulated Videos).** Since the diffusion models generates 121 frames vs 33 for the autoregressive, we provide both comparisons. Higher is better for all columns

Model	Params	Ball Bounce	2 Obj Fall	2 Obj Para	Block/Obj	Columns	Raised Block	Walls	Two Ball
Cosmos-1 AR [1]	5B	**0.3759**	0.2675	0.2268	0.2643	0.7032	0.3798	0.4996	0.1607
Cosmos-1 (33F) [1]	7B	0.3719	0.2994	**0.2831**	**0.3476**	0.7349	**0.4555**	**0.5578**	0.2013
Cosmos-1 (121F) [1]	7B	0.1636	0.1444	0.2008	0.2047	0.5575	0.3193	0.4155	0.1403
Cosmos-2 [30]	2B	0.1903	0.1780	0.1401	0.1271	0.6958	0.4514	0.3298	**0.2055**
Cosmos-2 [30]	14B	0.1913	0.2013	0.1351	0.1672	0.4758	0.1846	0.2699	0.1751
Cosmos-2.5 [2]	2B	0.1996	**0.3035**	0.1607	0.0992	**0.7363**	0.4480	0.3929	**0.2055**
	Params	Obj Tow.	Obj Away	Sphere Tow.	Sphere Away	Dominoes	Ramp	Table	Avg.
Cosmos-1 AR [1]	5B	0.2984	0.4121	0.6349	0.4799	0.4605	**0.5292**	**0.6439**	0.4225
Cosmos-1 (33F) [1]	7B	**0.3272**	**0.4840**	**0.7123**	**0.5546**	**0.4892**	0.4861	0.4573	**0.4508**
Cosmos-1 (121F) [1]	7B	0.0996	0.2330	0.2453	0.1774	0.1568	0.3802	0.4215	0.2573
Cosmos-2	2B	0.1591	0.4341	0.4595	0.3337	0.2425	0.4082	0.4907	0.3201
Cosmos-2	14B	0.1484	0.3705	0.3721	0.3171	0.2067	0.4048	0.4061	0.2684
Cosmos-2.5 [2]	2B	0.1341	0.4321	0.4997	0.3189	0.2981	0.5049	0.4984	0.3488

from the ShapeNet dataset [12] which includes 51,000 object models across 55 different categories. We sampled across all different categories and models, allowing for diversity in object shapes, textures, sizes, and properties.

We will now provide a brief description on each of the concepts. More details on individual scenarios are included in the supplemental material.

- **Motion Physics** is focused on evaluating the kinematics and dynamics in the generated video, specifically accounting for forces such as gravity and friction. This is a common real world scenario, as it is common for these models to have to simulate moving and colliding objects. To test motion physics, we create 3 scenarios: bouncing ball, two object fall, and two object parabolic motion.
- **Object Permanence** evaluates whether video generative models understand that objects continue to exist in the scenes even when hidden from the camera. This is a fundamental physics property that significantly affects our ability to predict the world (e.g. when driving we understand cars remain even if blocked) and is generally developed in young children between the ages of only 4-7 months old. To test object permanence, we create 5 scenarios: block & object, columns, raised block bounce, wall bouncing, and two ball bounce.
- **Support Relations** evaluates how objects physically support one another, e.g. one object preventing another from falling due to gravity or external forces. This includes understanding when certain configurations of objects are stable vs. unstable: for example, a large object placed on the middle of a table would be stable while the same object placed closer to the edge would be unstable. To test support relations, we designed 3 scenarios: dominoes, ramp block, and table drop.
- **Perspective / Scale Relations** is designed to evaluate the accuracy of objects' appearance, such as size and location, with respect to the camera viewpoint. We implemented two types of scenes to evaluate whether models can reason about how object size and location change as a function of distance from the camera. To evaluate perspective / scale relations, we designed 2 scenarios: obj/sphere moving towards camera and obj/sphere moving away from camera.

3.1.1. Evaluation Methodology

Our evaluation for the intuitive physics subset is done by comparing ground truth object segmentations with segmentations obtained from the generated videos. Specifically, our pipeline (visualized in Fig. 2) works as follows: For every scenario, we use the ground truth segmentations to obtain bounding boxes of all objects in the video in the first frame. We then prompt SAM2 [33] with these bounding boxes, and use SAM2 to propogate these boxes/objects through the rest of the video. For every frame, we then compare the ground truth and predicted masks using the mIoU metric. We additionally, use the background region (calculated as all pixels not part of individual object masks) to compute the background RMSE. Since all of our backgrounds remain constant throughout the video, this metric measures how well the models maintain backgrounds.

3.2. Physical Parameter Estimation

The second subset shifts the focus from core physics principles to exact parameter estimation. One key goal of WFMs like NVIDIA's Cosmos [1] is replacing physics simulation software (where parameters are hard-coded or manually tuned) as a synthetic data generator. To that end, it is crucial that these models are able to generate videos with accurate estimations for key physical parameters, such as gravitational acceleration. For this subset, we carefully designed a total of three experimental setups testing gravitational acceleration, friction coefficients, and fluid viscosity as the respective physical parameters. These concepts have 51 videos, 103 videos, and 80 videos respectively, for a total of 234 videos. We additionally synthetically generate 30 gravity and 15 friction videos, bring the combined total to 279 videos for the physical parameter estimation set.

The difficulty of designing these settings lies in reducing the effect of confounding variables that can influence the physical parameter estimation ability of these models (e.g. depth ambiguity, object motion uncertainty, etc.). To address this, each experimental setup and evaluation pipeline is meticulously designed to ensure a consistent evaluation. This setup and pipeline are detailed in Sec. 3.2.1. We validate this pipeline by running it on our collected videos, and ensuring that the output physical parameter is close to the ground truth (e.g. $9.8\frac{m}{s}^2$ for gravitational acceleration). Validation results can be seen in Tab. 2.

We will now provide detail for the experimental setups, with additional information in the Supplementary material:

- **Gravity** consists of 51 videos, 17 for straight drops and 34 for parabolic motion. Dropped items vary in shape and size. For parabolic motion, launch angle and trajectory vary as well. Note that video completion models are provided enough input frames to estimate the gravitational acceleration.
- **Viscosity** consists of 80 videos, 32 for glycerine, 30 for corn syrup, and 18 for honey. In these videos, a steel ball is dropped into a test tube. The terminal velocity is estimated from the 3d position of the steel ball, which in turn gives us the viscosity. Note that viscosity varies depending on temperature and water absorption. As such, we collect all videos at $75°F$ and within the same session to as not to influence the viscosity of hygroscopic fluids. All controlled parameters are provided to the evaluated model through text conditioning. Additionally, video completion models are provided enough input frames to estimate the terminal velocity of the steel ball through the fluid.
- **Friction** consists of 103 videos, 30 for wood, 19 for rubber, 18 for sandpaper (80 grit), 18 for sandpaper (3000 grit), and 18 for plastic. In these videos, a steel block is dropped down a ramp covered with different materials. The angle of the ramp varies between runs. All controlled parameters are provided to the evaluated model through text conditioning. Additionally, video completion models are provided enough input frames to estimate the acceleration of the steel block down the ramp.

3.2.1. Evaluation Methodology

Extracting exact physical parameters from video is a difficult task because it requires estimating 3D positions of objects from only a monocular video. Estimating these 3D positions requires three key pieces of information: camera intrinsics/extrinsics, the 2D pixel location of the object and the depth of the object.

To collect camera intrinsics, we calibrate the camera using a traditional checkerboard method prior to collecting data. To estimate the extrinsics dynamically for every video/setting, we place a checkerboard in the background of all of our videos. Since we know the 3D locations of the checkerboard corners, we can use them to estimate the camera extrinsics given the intrinsics. We additionally verify the camera intrinsics and extrinsics through manual measurement (e.g. we manually measure the distance between the camera and the checkerboard and compare). To extract the 2D locations of objects, we use SAM2 [33] prompted by a manually selected prompt point to track the object through the frames. We take the centroid of the object mask as its 2D pixel location. Robustly estimating the object's depth given just a monocular video is difficult, so we instead opt to design our setup so that the depth is always constant and can be measured exactly. We do this by placing our objects at a constant depth, and ensuring that they move only in a plane parallel to the camera plane.

Once we have estimated the 3D position of the object throughout the video, we can use those positions to estimate relevant physics parameters. To do this, we first estimate the acceleration of the object. This is done by fitting a quadratic equation to the object positions over time and taking the appropriate coefficients. For our gravity scenarios, we then use this acceleration directly and compare it to g. For the friction setups, we use the following equation to compute a friction coefficient from the acceleration:

$$\mu = \frac{g \sin\theta - a}{g \cos\theta}$$

where θ is the angle of the ramp and is pre-measured.

For viscosity, instead of estimating accleration, we estimate the terminal velocity of the object in the fluid. We do this by fitting a line to the object positions and taking the slope. We then use the following equation to estimate viscosity:

$$\eta = \frac{2r^2(\rho_s - \rho_f)g}{9v_t}$$

where ρ_s and ρ_f are the densities of the sphere and fluid respectively and v_t is the terminal velocity.

Table 4. **Foreground mIoU results on the Physical Principles Understanding subset (Real Videos).** Higher is better for all columns

Model	Params	Motion Physics	Object Perm.	Scale	Support Relations	Avg.
Cosmos-1 AR [1]	5B	**0.3826**	0.3471	**0.1634**	0.7517	**0.4112**
Cosmos-1 [1]	7B	0.2156	0.3644	0.044	0.6362	0.3151
Cosmos-2 [30]	2B	0.2931	0.3641	0.0910	**0.7648**	0.3783
Cosmos-2 [30]	14B	0.2716	**0.3697**	0.0836	0.6521	0.3443
Cosmos-2.5 [2]	2B	0.2896	0.3114	0.0905	0.6691	0.3402

Table 5. **Estimated Physics Parameters on the Physical Parameter Estimation subset (Real Videos).** Both ground truth ranges as well as our real video estimations are provided as comparison (note that our video estimations are all very close to or within the range of acceptable values). For gravity and viscosity, models closest to the ground truth values are bolded. Due to the wide range of acceptable friction values, models closest to our real video estimations are bolded.

	Gravity		Viscosity			Friction				
Model	Free-Fall (m/s^2)	Parabolic (m/s^2)	Glycerine (Pa.s)	Corn Syrup (Pa.s)	Honey (Pa.s)	Wood	Rubber	Sandpaper (80)	Sandpaper (3000)	Plastic
Ground Truth	9.81	9.81	1.2	6.0	14.1	0.2-0.5	0.5-2.0	0.7-1.1	0.2-0.5	0.05-0.2
Real Videos	9.78 ± 0.38	9.85 ± 0.36	1.22 ± 0.01	5.84 ± 0.02	13.82 ± 0.75	0.35 ± 0.05	0.93 ± 0.10	1.06 ± 0.05	0.30 ± 0.02	0.22 ± 0.03
Cosmos 1 AR [1]	4.215 ± 3.713	4.297 ± 1.294	7.8 ± 1.04	**8.44** ± 2.11	> 50	**0.541** ± 0.124	1.237 ± 0.142	1.277 ± 0.185	0.528 ± 0.079	0.508 ± 0.101
Cosmos 1 DM [1]	3.506 ± 1.912	7.652 ± 2.927	0.603 ± 1.191	1.538 ± 1.478	0.17 ± 0.171	0.674 ± 0.165	1.295 ± 0.214	1.453 ± 0.257	**0.522** ± 0.139	0.481 ± 0.119
Cosmos 2 (2B) [30]	**8.927** ± 4.791	8.228 ± 3.813	0.252 ± 0.394	1.091 ± 0.309	**9.845** ± 6.198	0.696 ± 0.094	1.468 ± 0.26	1.262 ± 0.259	0.598 ± 0.111	0.614 ± 0.131
Cosmos 2 (14B) [30]	8.428 ± 3.478	**9.145** ± 4.171	0.221 ± 0.101	1.304 ± 0.844	1.805 ± 1.813	0.659 ± 0.162	**1.224** ± 0.410	**1.144** ± 0.473	0.556 ± 0.088	**0.461** ± 0.393
Cosmos 2.5 [2]	4.778 ± 4.474	5.375 ± 2.763	**0.802** ± 0.478	1.091 ± 0.309	1.678 ± 0.079	0.695 ± 0.185	1.449 ± 0.339	1.522 ± 0.278	0.611 ± 0.119	0.628 ± 0.135
Wan 2.2 [39]	0.378 ± 0.784	1.682 ± 3.065	3.418 ± 5.119	3.162 ± 3.169	3.281 ± 2.758	0.668 ± 0.101	1.321 ± 0.223	1.476 ± 0.242	0.516 ± 0.155	0.576 ± 0.105
Hunyuan Video [25]	0.369 ± 0.788	0.206 ± 0.407	7.203 ± 3.709	> 50	> 50	0.662 ± 0.086	1.432 ± 0.227	1.049 ± 0.222	0.531 ± 0.076	0.569 ± 0.104
CogVideoX [48]	-0.039 ± 0.136	0.181 ± 0.239	3.286 ± 2.854	2.247 ± 2.769	2.672 ± 1.909	0.634 ± 0.126	1.474 ± 0.3315	1.288 ± 0.498	0.516 ± 0.134	0.502 ± 0.201

4. Discussion

We evaluate the Cosmos family of models and a number of image-to-video models on both subsets of **WorldBench**. Note that as of now, Cosmos is the only video-to-video WFM that is open-source, though we expect its introduction to expand the already growing research area. We evaluate Cosmos-1 (Auto-regressive and Diffusion), Cosmos-2 (2B and 14B) and Cosmos-2.5. For image-to-video models, we evaluate Wan 2.2, Hunyuan Video, and CogVideoX. For the intuitive physics subset, we use two metrics to evaluate the accuracy of generated videos: Foreground mIoU and Background RMSE. Foreground mIoU compares ground truth object segmentations with segmentations extracted from generated videos by SAM2 [33] and gives us information about how accurately the models can predict the dynamics and evolution of the scene. Background RMSE on the other hand, computes the RMSE between the background in the ground truth video and generated video. This metric gives us information about whether the model is able to keep the surrounding scene/environment consistent while objects are in motion. Specific implementation details of our evaluation methods are provided in the supplement.

Quantitative results from our evaluations are shown in Table 3, Table 4, Table 5, and Table 6. Results for the background RMSE are in the Supp. Mat. We summarize some of our learnings about current WFMs below:

Across the board, models exhibit an extremely high variance in the parameter estimation subset. Experiments captured in the physical parameter estimation subset contain highly constrained dynamic sequences in an effort to focus solely on the parameter being tested. Despite this, all models (both Cosmos-family and image-to-video) exhibit extremely high variance between rollouts. This is especially apparent in the gravity experiments where objects experience downward acceleration at varying levels, even between rollouts with the same object and object trajectory.

Model outputs tend to follow realistic motion trajectories, while not adhering to realistic motion parameters. For example, for gravity, all models tended to exhibit realistic motion paths most of the time (both parabolic trajectories and straight drops). However, they fail to abide by the proper $9.8\frac{m}{s}^2$. Similar trends appear in the fluid viscosity and friction experiments. As remarked on before, this highlights the need for benchmarks that can target particular physical parameters, as visual realism alone is not sufficient for synthetic world data generators.

Parameter-specific performance trends. Certain models (Cosmos 2 in particular) performed well on the gravitational acceleration subtask, while others tended to underaccelerate the object. Models across the board tended to fail at fluid viscosity estimation, likely due to lack of training data involving fluid dynamics. Friction experiments, on the other hand, tended to be easier for models to predict. Importantly, models tended to abide by the correct ordering of terminal velocity from slowest to fastest (sandpaper at 80 grit, rubber, wood, sandpaper at 3000 grit, and plastic).

Image-to-video models tended to under-perform on the gravitational acceleration subtask. Image-to-video models suffer from lack of temporal information in the input, and so typically displayed a global under-performance.

Table 6. **Estimated Physics Parameters on the Physical Parameter Estimation subset (Simulated Videos).** Various different materials are tested for friction. We give the average RMSE error instead. Values closest to the ground truth value or with the lowest error are bolded.

Model	Params	Gravity		Friction
		Free-Fall	Parabolic	Varied Materials Average Error (↓)
Ground Truth		9.81	9.81	–
Cosmos-1 AR [1]	5B	10.831 ± 2.925	7.408 ± 1.15	0.298
Cosmos-1 DM [1]	7B	7.530 ± 3.672	6.274 ± 1.595	0.231
Cosmos-2 [30]	2B	4.288 ± 5.061	13.229 ± 4.312	0.232
Cosmos-2 [30]	14B	4.230 ± 6.740	2.930 ± 3.023	0.274
Cosmos-2.5 [2]	2B	10.494 ± 3.934	12.540 ± 3.296	0.217
Wan 2.2 [39]	5B	1.157 ± 2.631	0.062 ± 1.034	0.253
Hunyuan Video [25]	13B	0.939 ± 0.719	0.677 ± 0.569	0.2275
CogVideoX [48]	5B	-0.226 ± 2.392	0.309 ± 1.631	0.361

However, this trend was exacerbated in the gravitational acceleration subtask, where lack of temporal information led to severely low or even negative gravity estimates (e.g. ball slowing down).

Models perform similarly on the synthetic and real versions of both benchmarks We find that across both the intuitive physics and parameter estimation subsets, all tested models perform similar in synthetic and real scenarios. This suggests that the cause for poor performance is not the distribution gap between real world videos and synthetically generated test cases but rather due to poor physics understanding in the models.

Models do not handle material properties on the long-tail of real-world distributions well We find that the models typically are unable to accurately simulate objects or materials which are far from average values such as Honey (very high viscosity relative) or Plastic (very low friction coefficient relative). Instead, they tend to simulate these materials closer to the average and do not differentiate well. Additionally, the variance does not meaningfully increase on these materials, showing that the model is consistent in these incorrect predictions.

Models perform better on scenarios with longer object interaction periods. In the intuitive physics subset, we find that for scenarios where objects interacts for longer, such as Ramp, Table, and Walls, perform much better than shorter interactions such as in Two Object Fall, Two Object Parabolic Motion, or Dominoes. We can also see this in the intuitive physics subset where the friction tests, in which an object is slowly sliding down a ramp, perform quite well compared to viscosity or gravity.

Models tend to perform better on scenarios where strong training priors are present. We find that across scenarios and subsets, models rely heavily on training priors for generating future frames. For example, in the intuitive physics subset, we find that the models handle balls rolling down the ramp very well, but have more trouble modeling the interaction of the ball and a block at the bottom of the ramp. This is likely because a ball rolling has been seen in training data, but obstructions on ramps are less likely. In the physical parameter subset, we also notice that most models tend to perform better on gravity prediction when the object being thrown is a basketball compared to a block or pool ball. All of this suggests that these models are relying more on priors from training data for their predictions rather than understanding of physical parameters/laws.

Limitations The goal of this benchmark is to disambiguate between different core physical concepts and parameters to identify key model failures. As such, we would like to continue expanding the number of concepts that we test. As each setup requires careful experimental design and extensive validation (for the parameter estimation subset in particular), this is an ongoing process. We plan to add additional physical tasks (collision mechanics, conservation principles, optics, etc.). Additionally, while our video-based design leads to more nuanced and diagnostic evaluation, requiring visual inputs limits the number of models we can evaluate. However, we hope that as image/video-to-video models become more popular with the introduction of Cosmos, they can benefit from our benchmark.

5. Conclusion

In this work, we introduce **WorldBench**, a new benchmark designed to evaluate concept-specific physical understanding in today's world-foundation models. Previous benchmarks assessed a model's visual realism through human-based metrics, or a model's dynamic trajectory adherence through coarse-grained metrics like object contact prediction. We leverage our benchmark's video-based and modular framework to directly measure adherence to physical concepts, constants, and material properties. This aids in our goal to 1) disambiguate between visually appealing and physically accurate generations, and 2) diagnose the core concept-specific failures of modern WFMs. We hope this benchmark can highlight the distinct challenges faced by current methods and can help guide the development and understanding of new ones as well.

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. 1, 3, 4, 5, 6, 7, 8

[2] Arslan Ali, Junjie Bai, Maciej Bala, Yogesh Balaji, Aaron Blakeman, Tiffany Cai, Jiaxin Cao, Tianshi Cao, Elizabeth Cha, Yu-Wei Chao, et al. World simulation with video foundation models for physical ai. *arXiv preprint arXiv:2511.00062*, 2025. 3, 5, 7, 8

[3] Eloi Alonso, Adam Jelley, Vincent Micheli, Anssi Kanervisto, Amos J Storkey, Tim Pearce, and François Fleuret. Diffusion for world modeling: Visual details matter in atari. *Advances in Neural Information Processing Systems*, 37: 58757–58791, 2024. 3

[4] Anton Bakhtin, Laurens van der Maaten, Justin Johnson, Laura Gustafson, and Ross Girshick. Phyre: A new benchmark for physical reasoning. *Advances in Neural Information Processing Systems*, 32, 2019. 1, 3, 4

[5] Daniel M Bear, Elias Wang, Damian Mrowca, Felix J Binder, Hsiao-Yu Fish Tung, RT Pramod, Cameron Holdaway, Sirui Tao, Kevin Smith, Fan-Yun Sun, et al. Physion: Evaluating physical prediction from vision in humans and machines. *arXiv preprint arXiv:2106.08261*, 2021. 2, 3, 4

[6] 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. 3

[7] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22563–22575, 2023. 3

[8] Blender Online Community. Blender - a 3d modelling and rendering package. 2002. 4

[9] Florian Bordes, Quentin Garrido, Justine T Kao, Adina Williams, Michael Rabbat, and Emmanuel Dupoux. Intphys 2: Benchmarking intuitive physics understanding in complex synthetic environments. *arXiv preprint arXiv:2506.09849*, 2025. 2, 4

[10] Jake Bruce, Michael D Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, et al. Genie: Generative interactive environments. In *Forty-first International Conference on Machine Learning*, 2024. 3, 4

[11] Duygu Ceylan, Chun-Hao P Huang, and Niloy J Mitra. Pix2video: Video editing using image diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23206–23217, 2023. 3

[12] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015. 5

[13] Erwin Coumans and Yunfei Bai. Pybullet, a python module for physics simulation for games, robotics and machine learning. http://pybullet.org, 2016–2021. 4

[14] Sudeep Dasari, Frederik Ebert, Stephen Tian, Suraj Nair, Bernadette Bucher, Karl Schmeckpeper, Siddharth Singh, Sergey Levine, and Chelsea Finn. Robonet: Large-scale multi-robot learning. *arXiv preprint arXiv:1910.11215*, 2019. 3

[15] Haoge Deng, Ting Pan, Haiwen Diao, Zhengxiong Luo, Yufeng Cui, Huchuan Lu, Shiguang Shan, Yonggang Qi, and Xinlong Wang. Autoregressive video generation without vector quantization. *arXiv preprint arXiv:2412.14169*, 2024. 3

[16] Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7346–7356, 2023. 3

[17] Klaus Greff, Francois Belletti, Lucas Beyer, Carl Doersch, Yilun Du, Daniel Duckworth, David J Fleet, Dan Gnanapragasam, Florian Golemo, Charles Herrmann, Thomas Kipf, Abhijit Kundu, Dmitry Lagun, Issam Laradji, Hsueh-Ti (Derek) Liu, Henning Meyer, Yishu Miao, Derek Nowrouzezahrai, Cengiz Oztireli, Etienne Pot, Noha Radwan, Daniel Rebain, Sara Sabour, Mehdi S. M. Sajjadi, Matan Sela, Vincent Sitzmann, Austin Stone, Deqing Sun, Suhani Vora, Ziyu Wang, Tianhao Wu, Kwang Moo Yi, Fangcheng Zhong, and Andrea Tagliasacchi. Kubric: a scalable dataset generator. 2022. 4

[18] Klaus Greff, Francois Belletti, Lucas Beyer, Carl Doersch, Yilun Du, Daniel Duckworth, David J Fleet, Dan Gnanapragasam, Florian Golemo, Charles Herrmann, et al. Kubric: A scalable dataset generator. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3749–3761, 2022. 3, 4

[19] Oliver Groth, Fabian B Fuchs, Ingmar Posner, and Andrea Vedaldi. Shapestacks: Learning vision-based physical intuition for generalised object stacking. In *Proceedings of the european conference on computer vision (eccv)*, pages 702–717, 2018. 3

[20] Jiaxi Gu, Shicong Wang, Haoyu Zhao, Tianyi Lu, Xing Zhang, Zuxuan Wu, Songcen Xu, Wei Zhang, Yu-Gang Jiang, and Hang Xu. Reuse and diffuse: Iterative denoising for text-to-video generation. *arXiv preprint arXiv:2309.03549*, 2023. 3

[21] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023.

[22] Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. Latent video diffusion models for high-fidelity long video generation. *arXiv preprint arXiv:2211.13221*, 2022. 3

[23] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in Neural Information Processing Systems*, 35:8633–8646, 2022. 3

[24] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*, 2022. 3

[25] Team HunyuanWorld. Hunyuanworld 1.0: Generating immersive, explorable, and interactive 3d worlds from words or pixels. *arXiv preprint*, 2025. 3, 7, 8

[26] Achuta Kadambi, Celso de Melo, Cho-Jui Hsieh, Mani Srivastava, and Stefano Soatto. Incorporating physics into data-driven computer vision. *Nat. Mach. Intell.*, 5(6):572–580, 2023. 3

[27] Hao Liu, Wilson Yan, Matei Zaharia, and Pieter Abbeel. World model on million-length video and language with blockwise ringattention. *arXiv preprint arXiv:2402.08268*, 2024. 3

[28] Zhengxiong Luo, Dayou Chen, Yingya Zhang, Yan Huang, Liang Wang, Yujun Shen, Deli Zhao, Jingren Zhou, and Tieniu Tan. Videofusion: Decomposed diffusion models for high-quality video generation. *arXiv preprint arXiv:2303.08320*, 2023. 3

[29] Vincent Micheli, Eloi Alonso, and François Fleuret. Transformers are sample-efficient world models. *arXiv preprint arXiv:2209.00588*, 2022. 3

[30] NVIDIA. Cosmos-predict2: World simulation model for physical ai. https://research.nvidia.com/labs/dir/cosmos-predict2/, 2025. 5, 7, 8

[31] Luis Piloto, Ari Weinstein, Dhruva TB, Arun Ahuja, Mehdi Mirza, Greg Wayne, David Amos, Chia-chun Hung, and Matt Botvinick. Probing physics knowledge using tools from developmental psychology. *arXiv preprint arXiv:1804.01128*, 2018. 3

[32] Ruslan Rakhimov, Denis Volkhonskiy, Alexey Artemov, Denis Zorin, and Evgeny Burnaev. Latent video transformer. *arXiv preprint arXiv:2006.10704*, 2020. 3

[33] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 5, 6, 7

[34] 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. 2, 3, 4

[35] Jan Robine, Marc Höftmann, Tobias Uelwer, and Stefan Harmeling. Transformer-based world models are happy with 100k interactions. *arXiv preprint arXiv:2303.07109*, 2023. 3

[36] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022. 3

[37] Kevin Smith, Lingjie Mei, Shunyu Yao, Jiajun Wu, Elizabeth Spelke, Josh Tenenbaum, and Tomer Ullman. Modeling expectation violation in intuitive physics with coarse probabilistic object representations. *Advances in neural information processing systems*, 32, 2019. 3

[38] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 3

[39] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 3, 7, 8

[40] Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. Modelscope text-to-video technical report. *arXiv preprint arXiv:2308.06571*, 2023. 3

[41] Wenjing Wang, Huan Yang, Zixi Tuo, Huiguo He, Junchen Zhu, Jianlong Fu, and Jiaying Liu. Videofactory: Swap attention in spatiotemporal diffusions for text-to-video generation. 2023. 3

[42] Dirk Weissenborn, Oscar Täckström, and Jakob Uszkoreit. Scaling autoregressive video models. *arXiv preprint arXiv:1906.02634*, 2019. 3

[43] Wenming Weng, Ruoyu Feng, Yanhui Wang, Qi Dai, Chunyu Wang, Dacheng Yin, Zhiyuan Zhao, Kai Qiu, Jianmin Bao, Yuhui Yuan, et al. Art-v: Auto-regressive text-to-video generation with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7395–7405, 2024.

[44] Chenfei Wu, Lun Huang, Qianxi Zhang, Binyang Li, Lei Ji, Fan Yang, Guillermo Sapiro, and Nan Duan. Godiva: Generating open-domain videos from natural descriptions. *arXiv preprint arXiv:2104.14806*, 2021.

[45] Chenfei Wu, Jian Liang, Lei Ji, Fan Yang, Yuejian Fang, Daxin Jiang, and Nan Duan. Nüwa: Visual synthesis pre-training for neural visual world creation. In *European conference on computer vision*, pages 720–736. Springer, 2022.

[46] Desai Xie, Zhan Xu, Yicong Hong, Hao Tan, Difan Liu, Feng Liu, Arie Kaufman, and Yang Zhou. Progressive autoregressive video diffusion models. *arXiv preprint arXiv:2410.08151*, 2024.

[47] Wilson Yan, Yunzhi Zhang, Pieter Abbeel, and Aravind Srinivas. Videogpt: Video generation using vq-vae and transformers. *arXiv preprint arXiv:2104.10157*, 2021. 3

[48] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024. 3, 7, 8

[49] 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. 1, 3, 4

[50] Tianwei Yin, Qiang Zhang, Richard Zhang, William T Freeman, Fredo Durand, Eli Shechtman, and Xun Huang. From slow bidirectional to fast causal video generators. *arXiv preprint arXiv:2412.07772*, 2024. 3