How Close are World Models to the Physical World?

Rishi Upadhyay[1] **Howard Zhang**[1] **Zhirong Lu**[1*]

Lakshman Sundaram[1*] **Ayush Agrawal**[1*] **Yilin Wu**[1*] **Yunhao Ba**[2] **Alex Wong**[3]

Celso M de Melo[4] **Achuta Kadambi**[1]

[1]University of California, Los Angeles [2]Sony

[3]Yale University [4]DEVCOM Army Research Laboratory

Abstract

Recent advances in world modeling, such as the Cosmos foundation model, highlight the growing need for physically accurate representations of dynamic environments. Despite this ambition, existing evaluation benchmarks fall short of capturing the full complexity of physical interactions, often relying on discrete or binary proxy tasks like object contact prediction. We introduce WorldBench, a new video-based benchmark that directly evaluates a model's ability to predict the evolution of physical scenes over time. Our dataset comprises four physically rich scenarios (motion physics, object permanence, support relations, scale/perspective) with 425 total configurations that assess both visual fidelity and physical plausibility. We additionally add natural language to a subset of this dataset, allowing us to benchmark text-generation models as well. Evaluating on SOTA world foundation models, we find that all configurations lack the physical consistency required to generate reliable real-world interactions. Furthermore, evaluating SOTA vision-language models, we find that the best models perform only slightly better than chance, highlighting a need for better object tracking and temporal consistency. Combined, this benchmark offers a more nuanced and scalable framework for evaluating the physical reasoning capabilities of world models, paving the way for more robust and generalizable simulation-driven learning. Our benchmark and evaluation code can be found at: https://huggingface.co/datasets/worldbenchmark/WorldBench

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 [2], promise to learn such skills at scale, with some even 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 beyond simple outcomes, but existing benchmarks for physical reasoning tasks often provide only coarse-grained or binary metrics. For example, Physion [6] evaluates physical reasoning using the object contact prediction task, which determines whether two targeted objects in a scene touch." While its visual realism and variety is useful for early progress, such evaluation frameworks fail to capture nuanced physical phenomena, such as object dynamics (velocity, acceleration, rotation, etc.), deformation, or occlusion.

In this paper, we introduce a new benchmark designed to evaluate the physical reasoning capabilities of world foundation models through video prediction. Rather than predicting discrete or binary outcomes, our benchmark requires models to forecast the evolution of full visual scenes over time. To

Preprint. Under review.

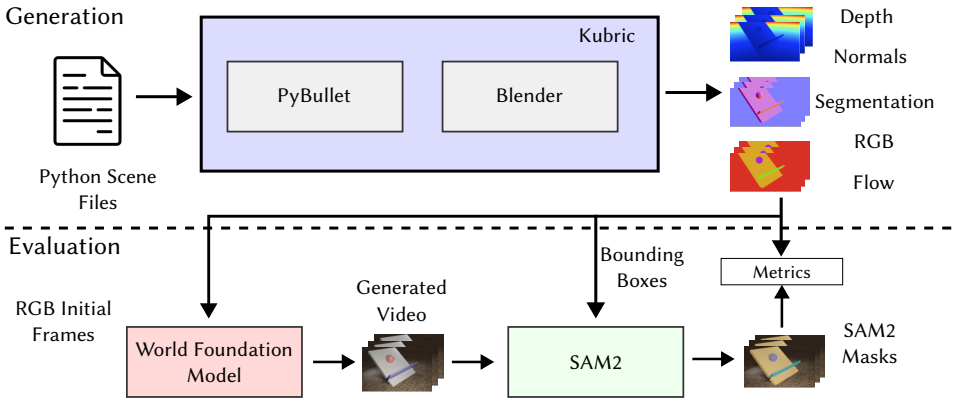

Figure 1: **Overview of our generation and evaluation process**. For generation (top), we use Kubric, which uses PyBullet and Blender under the hood. During evaluation (bottom), we first pass the initial frames of the generated video to the world foundation model which completes the video. The completed 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.

ensure the output dynamics converge toward a single interpretable outcome, we design simplified yet physically rich and visually realistic scenes across four categories: motion physics, object permanence, support relations, and scale/perspective. The dataset includes 425 diverse video sequences across all categories. In addition, in order to further validate the physical reasoning of popular vision-language models such as Gemini [36, 37] or Qwen [39] with the same set of benchmark videos, we include a set of scene-specific question-answer metrics.

Our proposed task of "constrained video prediction" allows for the more nuanced and detailed assessment of physical laws described above (i.e. dynamics, deformation, occlusion, etc). It also permits qualitative human judgment for dynamic realism (i.e. how realistic was the object movement?). In addition, the task design sets it apart from its predecessors. It is focused on specific physics properties such as object permanence or scale/perspective, that consistently show up in the real world and are needed for real tasks. These are not present in previous benchmarks such as Physion. We believe our benchmark provides a more rich signal of how close these models are to truly learning and understanding real-world dynamics. For future work, our benchmark also leads to a wider array of downstream tasks, such as object tracking, anomaly detection, action planning, etc.

We use our benchmark to extensively test the Cosmos world models, revealing substantial gaps in physical consistency and generalization when compared to simulated, physically accurate expectations. Our results highlight the limitations of current architectures and motivate further work on physically grounded learning.

Below is a summary of our key contributions:

- We introduce a novel fully video-based benchmark, **WorldBench**, for evaluating physical reasoning in world foundation models with 425 unique, hand-designed scenarios.
- We additionally introduce a new language-based subset of benchmark for evaluating physical reasoning in vision-language models.
- We perform an empirical analysis of the performance of state of the art WFMs and VLMs to identify 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 [25, 28, 33], but recent work has been on using video generation models [3]. These models

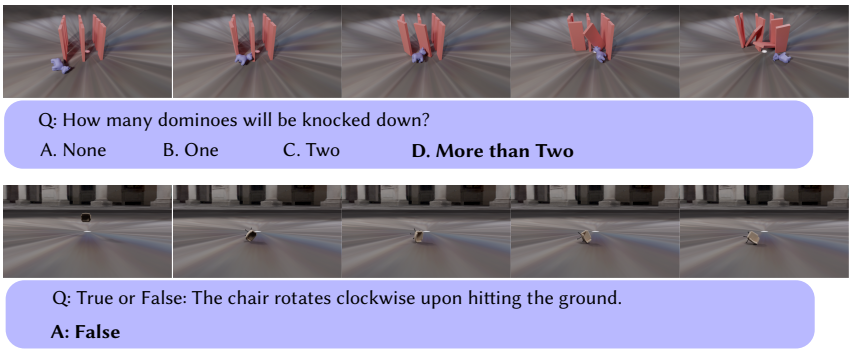

Figure 2: **Qualitative Examples of the Language-based subset of WorldBench**. VLMs are given access to a 9 frame video (same as what is inputted to COSMOS) and ask to answer a True/False or multiple choice question based on the video and future predictions.

typically leverage transformer architectures and either latent diffusion models [7, 8, 16, 21, 22, 11, 27, 34, 40] or auto-regressive models [15, 42, 45, 49, 44, 41, 46, 43, 30] 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 [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 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, 20, 29, 32, 35]. Datasets like PHYRE [5] focus on simplistic 2D scenarios constructed from balls and rectangular bars, with dynamics like collision, gravity, and friction. CLEVRER [48] 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 [6] 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. 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 [2] or Genie [10]. Compared to these, our benchmark is the first fully video-based 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.

2.3 Multi-modal Vision Language Models

Accompanying the wave of popularity of large language models is the vision-language model, a multi-modal model capable of processing both text and visual input [37, 39, 1, 17, 24, 26, 47, 50]. These models are typically capable of video understanding and reasoning. The recent Gemini model [36, 37] is an example of a multi-modal model, capable of flexibly taking in any order of visual, textual, or audio input. The more recent Qwen2.5-VL model [39, 4], from the Qwen series of vision-language models, uses dynamic resolution processing and absolute time encoding, and particularly targets a visual agent's ability to perform visual reasoning, tool usage, and task execution. To evaluate the dynamics prediction accuracy and physical reasoning abilities of these models, we extend our proposed benchmark with a language-based visual reasoning framework.

3 Benchmark

In order to test 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. To assess the accuracy of the generated physics, we segment objects in the generated videos and compare them against ground truth segmentations. All of the data used in the benchmark is obtained from a physics simulator, ensuring that it is physically accurate and that we can obtain real ground truth. Our benchmark specifically probes 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 viewpoint). This is not an exhaustive list, but is designed to cover a range of common real-world scenarios.

For each concept, we construct 3-5 scenarios. Each of these scenarios is hand-designed to capture some element of the concept it is testing. For each scenario, we have 25 videos, each of which is generated by randomizing various components such as object type, location and material. In total, **WorldBench** is made up of 425 videos spanning the 4 concepts. Each video is 132 frames long and includes depth, normals, object segmentations, and optical flow. All meshes and objects used in our simulations were taken 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.

All videos are rendered using Kubric, an open-source physics simulation pipeline [19]. Kubric uses PyBullet [13] as the physics simulator and Blender [9] as the renderer. This allows us to combine the physically accurate simulation of PyBullet with the high-quality rendering of Blender. We will now provide a brief description on each of the concepts and scenarios individually. More details are included in the supplemental material.

3.1 Motion Physics

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 very 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** A sphere, initially at rest at some known height above the ground, is subject to fall freely under gravity.
- **Two Object Fall** Two distinct objects are initially at rest at different heights above the ground and at a slight horizontal offset. Both objects are released to freely fall under gravity, and then typically collide with each other and the ground.
- **Two Object Parabolic Motion** This scenario is a slight variation of the above scenario, where two distinct objects are placed on opposite sides of the scene. They are then projected towards each other at some randomized initial angle and projectile angle, not necessarily in the same vertical plane.

3.2 Object Permanence

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.

- **Block & Obj** An object is moving from left to right behind a wall. The movement is linear and predictable, starting left of the wall with a randomized initial velocity. The object disappears behind the wall and reemerges on the right side.
- **Columns** An object moving from left to right behind several thin columns. This is very similar to the Block & Obj task, except that the object periodically disappears and reappears as it passes each column.
- **Raised Block Bounce** A sphere bouncing vertically behind a raised block. A sphere appears above the block when approaching its highest vertical position, and below the block when close to the ground, As it bounces, it is periodically occluded by the block and not visible to the camera.
- **Wall Bouncing** A sphere rolling horizontally behind a block between two walls. As the sphere rolls from one side to another, it eventually collides with a wall, bounces off, and rolls back the other direction. The sphere is periodically occluded while behind the block, and reappears in a gap as it approaches a wall.
- **Two Ball Bounce** Two spheres bouncing vertically, with one larger sphere in front and one smaller sphere straight behind it. During this motion, the small sphere is periodically occluded by the large ball as their vertical positions diverge.

3.3 Support Relations

Support relations scenarios evaluate 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** This scenario features an object colliding with a series of standing dominoes on a surface. Depending on initial velocity, the object knocks over varying number of dominoes, which may subsequently topple onto one another.
- **Ramp Block** A sphere rolling down an incline until it encounters a fixed barrier at the end, causing it to stop. This tests two things, one whether the incline supports the ball as it rolls, and then if the block at the bottom supports and stops it.
- **Table Drop** An object is positioned at a table's edge with a portion extending beyond the surface. This scenario directly challenges a model's understanding of the minimum conditions required for stable support and balance. Mere contact with a support surface is insufficient–the object requires adequate support distribution relative to its mass distribution.

3.4 Perspective / Scale Relations

The perspective/scale category 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.

- **Obj/Sphere Moving Towards Camera** A single object (e.g. a sphere or miscellaneous irregular object) is launched from the background and moves towards the camera. As the object approaches, it should appear to increase in size due to perspective.
- **Obj/Sphere Moving Away From Camera** An object begins near the foreground and moves away from the camera into the distance. The object should appear to shrink as it recedes.

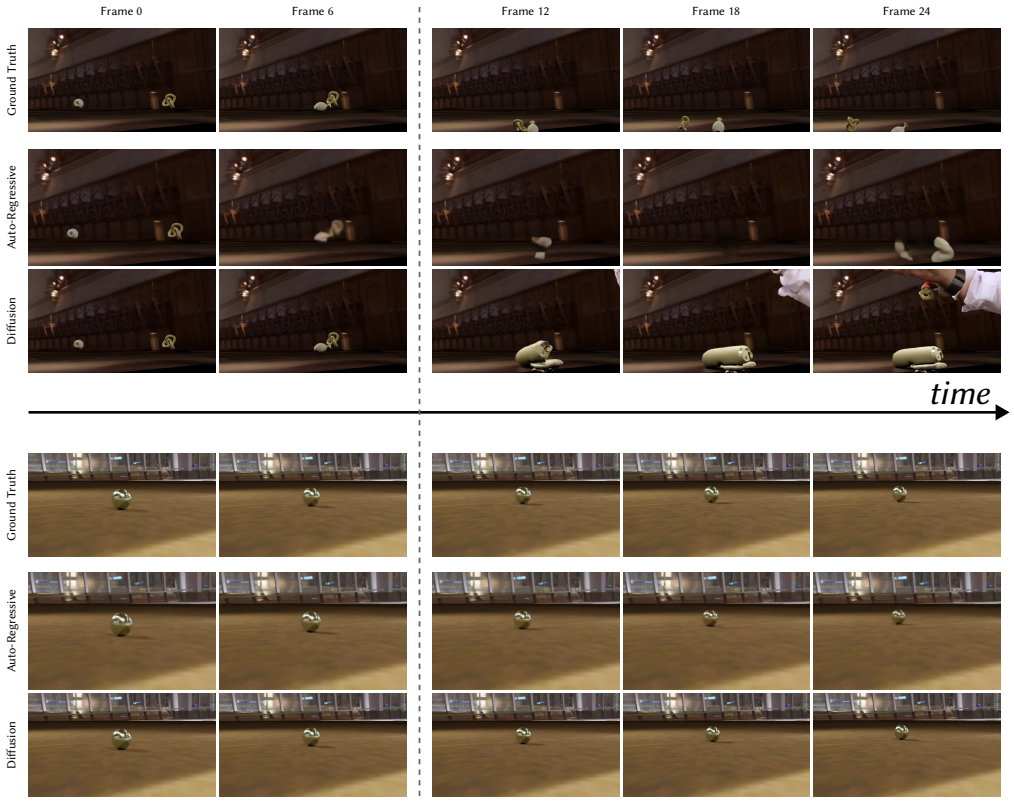

Figure 3: **Qualitative examples for the Motion Physics (top) and Scale/Perspective (bottom) scenarios.** For the motion physics example, two objects (a vase and a knot) are thrown at each other, collide, and then fall to the floor. The auto-regressive model greatly distorts the object shapes, while the diffusion model hallucinates the vase into a tank and adds a human hand. For the scale example, a metallic sphere is rolling away from the camera. Both models perform well on this sample.

3.5 Text-Enhanced Subset

In addition to the generated videos, we additionally created a language-based subset of the benchmark to evaluate todays vision-language models (VLMs) on physics understanding and prediction. This subset asks models to interpret visual details or predict physical outcomes, allowing us to assess their ability to do intuitive physical reasoning in diverse situations.

We select a subset of 181 videos and write 1 natural language question per video. Questions can be either binary True/False questions or multiple choice with up to 4 choices. Example questions and answers are shown in Fig. 2.

4 Evaluation

We evaluate both the video and language-based components of **WorldBench** in order to gain an understanding of how good today's models are, and where there might be room for improvement. For the video benchmark, we evaluate the Cosmos models. So far, these are the only models trained to be "world foundation models" that have been open sourced and therefore can be tested.

4.1 Cosmos

The Cosmos family of models includes multiple models spanning various parameter counts and architectures [2]. For this paper, we evaluate two models: a 5B-parameter auto-regressive generator, and a 7B-parameter diffusion based generator. Both of these models receive the same inputs: an input video which is 9 frames long and a prompt describing the input and continuation. The autoregressive

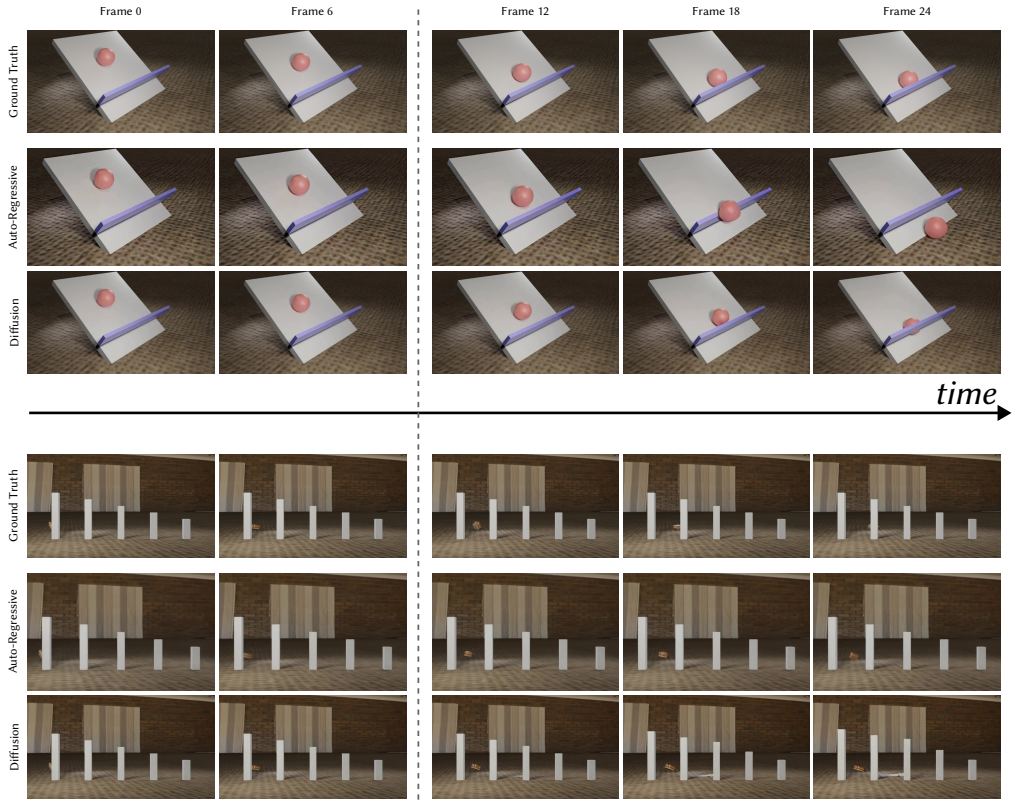

Figure 4: **Qualitative examples for the Support Relations (top) and Object Permanence (bottom) scenarios.** For the support relation example, a ball is rolled down an incline with a bar at the bottom. Both the autoregressive and diffusion models have the ball rolling through the bar violating physics principles. For the object permanence example, a box is thrown from behind columns. Both models render the box, but do not have it continue its motion after emerging from behind a column.

Table 1: **Foreground mIoU results for Cosmos models on our benchmark.** Since the diffusion models generates 121 frames vs 33 for the autoregressive, we provide both comparisons. Higher is better for all columns

Model	Ball Bounce	Two Obj Fall	Two Obj Para	Block/Obj	Columns	Raised Block	Walls	Two Ball
Autoregressive	**0.3759**	0.2675	0.2268	0.2643	0.7032	0.3798	0.4996	0.1607
Diffusion (33 frames)	0.3719	**0.2994**	**0.2831**	**0.3476**	**0.7349**	**0.4555**	**0.5578**	**0.2013**
Diffusion (121 frames)	0.1636	0.1444	0.2008	0.2047	0.5575	0.3193	0.4155	0.1403
	Obj Tow.	Obj Away	Sphere Tow.	Sphere Away	Dominoes	Ramp	Table	Avg.
Autoregressive	0.2984	0.4121	0.6349	0.4799	0.4605	**0.5292**	**0.6439**	0.4225
Diffusion (33 frames)	**0.3272**	**0.4840**	**0.7123**	**0.5546**	**0.4892**	0.4861	0.4573	**0.4508**
Diffusion (121 frames)	0.0996	0.2330	0.2453	0.1774	0.1568	0.3802	0.4215	0.2573

model then generates 33 frames, while the diffusion model generates 121 frames. 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 [31] and gives us information about how accurately the models can predict the dynamics and evolution of the scene. In order to run SAM2 [31], we extract bounding boxes from the GT segmentations and use them as prompts to match predicted masks with real masks. Background RMSE on the other hand, computes the RMSE between the background in the ground truth video and generated video. It is computed using the ground truth background segmentation mask. This metric gives us information about whether the model is able to keep the surrounding scene/environment

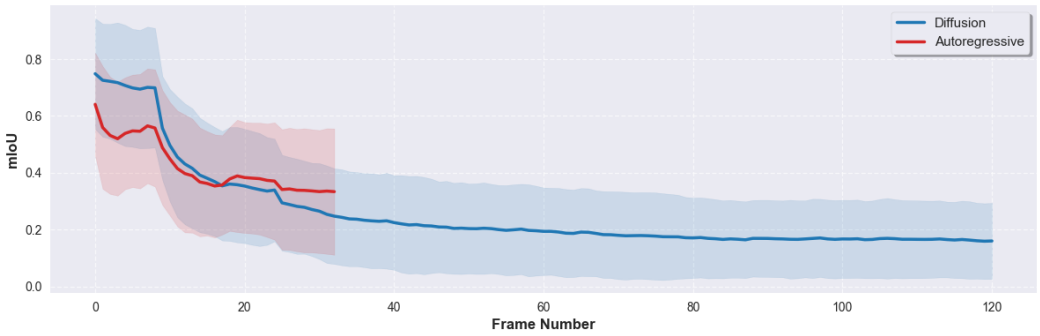

Figure 5: **mIoU results over time**. The foreground mIoU is inversely related with how far in the future the model is predicting. There is a sharp drop off after frame 9 when the model first begins predicting and this flattens after approx. 30 frames for the diffusion model and 15 frames for the auto-regressive model. The shaded region shows 1 standard deviation.

Table 2: **Background RMSE results for Cosmos models on our benchmark.** RMSE is computed over only the ground truth segmentation for "background". Lower is better for all columns.

Model	Ball Bounce	Two Obj Fall	Two Obj Para	Block/Obj	Columns	Raised Block	Walls	Two Ball
Autoregressive	**0.2112**	**0.2168**	**0.2542**	**0.1943**	0.1376	**0.0978**	0.2299	**0.1440**
Diffusion	0.2403	0.2221	0.2594	0.2103	**0.1104**	0.1223	**0.2223**	0.1709
	Obj Tow.	Obj Away	Sphere Tow.	Sphere Away	Dominoes	Ramp	Table	Avg.
Autoregressive	**0.2226**	**0.1696**	**0.1218**	**0.1848**	0.1616	**0.1027**	**0.1715**	**0.1747**
Diffusion	0.2534	0.2329	0.1325	0.2256	**0.0998**	0.2283	0.2586	0.1993

consistent while objects are in motion. All experiments were done on 1 H100 GPU. Generating a single video took approximately 70 seconds for the autoregressive model and 399 seconds for the diffusion model, resulting in 8 hours of total runtime for the autoregressive model and 47 hours for the diffusion model. Results for both the autoregressive and diffusion based models are shown in Table 1 (foreground mIoU) and Table 4. Since the diffusion model generates more frames than the autoregressive one, we provide metrics for both the entire generated video and for only the first 33 frames (to match autoregressive). Overall, neither model performs well, with the highest average mIoU being only 0.4508 for the diffusion model evaluated over 33 frames. In general, the diffusion model consistently outperforms the autoregressive model, save for the Support Relation scenarios, where the autoregressive model outperforms in 2 out of 3 scenarios. For RMSE, the autoregressive model outperforms, but both models are close in quality.

4.2 Vision-Language Models

In order to evaluate the language-based subset of **WorldBench**, we test SOTA closed- and open-source models: Qwen2.5 and Gemini. Qwen2.5 comes in 3 sizes, 7B, 32B, and 72B parameters, and is designed to handle vision inputs natively. For Gemini, we test both Gemini 2.5 Flash and Gemini 2.5 Pro. When running the evaluations, all models are provided with a system prompt which describes the tasks, defines what the output format should be, and the format of the data provided (9 frames). The Qwen models are run on 1 H100 GPU with the use of vLLM, while the Gemini models are evaluated through the provided API [23]. The total costs for evaluation were approximately $25. The outputs of the models are evaluated by directly comparing against the answers. Results for all the models are shown in Table. 5. Across all 4 scenarios, Gemini 2.5 Pro performs the best overall, achieving 49.72% accuracy. Within the open source models, Qwen2.5 32B surprisingly outperforms the larger 72B model, largely due to a very strong performance on the motion physics category. However, overall, all five models perform relatively poorly on the benchmark, achieving results only slightly better than chance. This suggests there is still much work to be done to improve the physical understanding of modern VLMs.

Table 3: **Results of SOTA closed and open models on our language-based benchmark**. Gemini 2.5 Pro, a closed model performs best overall, and Qwen2.5 32B performs best among open models.

	Model	Motion Phys	Obj. Perm.	Scale/Persp.	Support Rel.	**Avg.** ↑
Open Models	Qwen2.5-VL-7B [4]	0.5161	0.2381	0.4474	0.5357	0.3737
	Qwen2.5-VL-32B [4]	**0.8710**	0.2738	0.4737	**0.5714**	0.4641
	Qwen2.5-VL-72B [4]	0.5806	0.3333	0.4211	**0.5714**	0.4309
	GLM 4.1V 9B [22]	0.6674	0.3453	0.4473	0.6071	0.4641
	Mistral Small 3.2 24B	0.4838	0.2500	0.3684	0.3571	0.3315
	Llama-3.2-11B-Vision	0.5161	0.1548	0.3421	0.3571	0.2873
Closed Models	Gemini 2.5 Flash [37]	0.6452	0.3571	**0.6053**	0.4643	0.4751
	Gemini 2.5 Pro [37]	0.6774	0.4048	0.5000	**0.5714**	0.4972
	Claude Sonnet 4	0.7096	**0.4286**	0.5526	0.4285	**0.5027**
	GPT 4.1	0.3781	0.2619	0.5000	0.5000	0.3701

5 Discussion/Future Directions

Our empirical evaluation of current world foundation models and vision-language models shows that despite their strong performance and varied abilities, they have significant shortcomings in their ability to accurately model and predict physical interactions in dynamic environments.

For the Cosmos models — currently the only open-source WFMs — our quantitative and qualitative evaluations showed that these models often lose consistency, turning objects into different shapes, sizes, colors or even removing objects from the scene entirely. In addition, qualitative examples such as those in Fig. 3 suggest that these models have strong priors from which they generate future frames. For example, in the motion physics example, the model hallucinates a human arm that is manipulating the objects, likely because that is similar to data that it has seen during the training process. Similarly, for the support relation case, while these models can handle a ball rolling down a ramp (likely a common inclusion in synthetic datasets) it does not handle the block at the bottom of the ramp well. This suggests that is relying more on video priors in its training datasets than on real physical properties. In addition, Fig. 5 show that for both autoregressive and diffusion models, the performance is inversely correlated with the number of frames generated: as the model predicts further in the future it is worse at doing so. The evaluations of SOTA VLMs highlights similar challenges. The best performing model, Gemini 2.5 Pro, performs only slightly better than chance, and all models struggle with the object permanence questions suggesting difficulty in predictable object tracking and temporal consistency.

These limitations suggest that there is still much work to be done to improve the physical understanding and consistency of both video and text generation models. Although **WorldBench** is as of now only an evaluation set, we hope that future work will allow us to utilize our synthetic data generation techniques to generate useful, high quality, and abundant synthetic data that can be used to train and improve these models. In addition, physics-inspired losses and training strategies may help bridge the current gaps and help improve the real-world performance of WFMs and VLMs.

In addition to limitations in the WFMs and VLMs, there are limitations in the current form of the benchmark: because the data is synthetic, there is likely a distribution gap between the benchmark and real-world scenarios. Additionally, the use of SAM2 to obtain masks from generated video is not guaranteed to be accurate, as SAM2 can miss objects or include additional regions which would affect the resulting mIoU. Future work could address these by collecting real data and relying on human annotators for both ground truth and evaluating model predictions.

6 Conclusion

In this work, we introduce **WorldBench**, a new benchmark designed to evaluate the physics understanding and consistency of today's world-foundation models" and vision-language models. Moving

beyond binary or proxy tasks, our benchmark directly assess a model's ability to predict a scene's evolution over time, offering new insights into these model's physics understanding. Evaluating Cosmos, Gemini and Qwen on **WorldBench** shows that today's SOTA models still have significant shortcomings that limit their ability to understand and process real-world physics. These challenges highlight the need for new techniques that can instill physical understanding into models, and we hope this benchmark can help guide that development.

References

[1] Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, et al. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. *arXiv preprint arXiv:2503.01743*, 2025.

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

[3] 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.

[4] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] Blender Online Community. Blender - a 3d modelling and rendering package.

[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.

[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.

[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.

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

[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.

[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.

[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.

[17] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023.

[24] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

[25] 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.

[26] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

[27] 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.

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

[29] 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.

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

[31] 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.

[32] 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.

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

[34] 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.

[35] 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.

[36] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

[37] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

[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.

[39] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

[40] 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.

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

[42] 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.

[43] 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.

[44] 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.

[45] 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.

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

[47] Jihan Yang, Shusheng Yang, Anjali W Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. *arXiv preprint arXiv:2412.14171*, 2024.

[48] 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.

[49] 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.

[50] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

A Detailed Scene Descriptions

Here we provide a more detailed description of the scenes used in our benchmark:

Bouncing Ball A sphere, initially at rest at some known height above the ground, is subject to fall freely under gravity. The provided video includes frames after the bounce, ensuring the model has the necessary information for prediction. In this scenario, the initial height of the ball and material (including bounciness) of the ball are randomized.

Two Object Fall Two distinct objects are initially at rest at different heights above the ground and at a slight horizontal offset. Both objects are released to freely fall under gravity, and then typically collide with each other and the ground. In this scenario, we randomize the shape and initial positions of both objects.

Two Object Parabolic Motion This scenario is a slight variation of the above scenario, where two distinct objects are placed on opposite sides of the scene. They are then projected towards each other at some randomized initial angle and projectile angle, not necessarily in the same vertical plane. In this scenario, the shapes, locations, and initial velocities of the objects are randomized to generate diverse scenes.

Block & Obj An object is moving from left to right behind a wall. The movement is linear and predictable, starting left of the wall with a randomized initial velocity. The object disappears behind the wall and reemerges on the right side. In this scenario, the type of object along with its initial velocity are randomized.

Columns An object moving from left to right behind several thin columns. This is very similar to the Block & Obj task, except that the object periodically disappears and reappears as it passes each column. The provided frames generally In this scenario, the type of object and initial velocity are randomized.

Raised Block Bounce A sphere bouncing vertically behind a raised block. A sphere appears above the block when approaching its highest vertical position, and below the block when close to the ground, As it bounces, it is periodically occluded by the block and not visible to the camera. In this scenario, the mass, restitution (bounciness), and material of the sphere were randomized.

Wall Bouncing A sphere rolling horizontally behind a block between two walls. As the sphere rolls from one side to another, it eventually collides with a wall, bounces off, and rolls back the other direction. The sphere is periodically occluded while behind the block, and reappears in a gap as it approaches a wall. In this scenario, the mass, initial velocity, and friction coefficient of the sphere were randomized.

Two Ball Bounce Two spheres bouncing vertically, with one larger sphere in front and one smaller sphere straight behind it. During this motion, the small sphere is periodically occluded by the large ball as their vertical positions diverge. In this scenario, the mass, restitution, and materials of both spheres were randomized.

Dominoes This scenario features an object colliding with a series of standing dominoes on a surface. Depending on initial velocity, the object knocks over varying number of dominoes, which may subsequently topple onto one another. In this scenario, the type/shape of the object thrown and initial velocity are randomized.

Ramp Block A sphere rolling down an incline until it encounters a fixed barrier at the end, causing it to stop. This tests two things, one whether the incline supports the ball as it rolls, and then if the block at the bottom supports and stops it. In this situation, the angle and length of the incline and the initial position of the sphere are randomized.

Table Drop An object is positioned at a table's edge with a portion extending beyond the surface. This scenario directly challenges a model's understanding of the minimum conditions required for

stable support and balance. Mere contact with a support surface is insufficient–the object requires adequate support distribution relative to its mass distribution. In this scenario, both the type/shape of the object and its location relative to the table were randomized. This also meant that physical parameters such as mass and restitution were in a sense randomized as they depended on the object chosen.

Obj/Sphere Moving Towards Camera A single object (e.g. a sphere or miscellaneous irregular object) is launched from the background and moves towards the camera. As the object approaches, it should appear to increase in size due to perspective. In these scenarios, the type/shape of object and the initial velocity are randomized.

Obj/Sphere Moving Away From Camera A single object (e.g. a sphere or miscellaneous irregular object) is launched from the background and moves towards the camera. As the object moves away, it should appear to decrease in size due to perspective. Similarly, in these scenarios, the type/shape of object and the initial velocity are randomized.

B Additional Quantitative VLM Metrics

In this section, we expand upon the results in Tab. 3 of the main paper and show the results for each model by scene category. Although all 5 models tested perform similarly in most categories, we notice striking differences in the Walls category where the Qwen models are all near 0.0 while both Gemini models achieve accuracys >0.6. All models have the most trouble with the Object Permanence scenes and perform the best on Motion Physics scenes overall. This is somewhat expected as it is likely that their training data included more examples of motion physics than of object permanence.

Table 4: **Results on the VLM benchmark split up by scene type.** Lower is better for all columns.

Model	Ball Bounce	Two Obj Fall	Two Obj Para	Block/Obj	Columns	Raised Block	Walls	Two Ball
Qwen2.5-VL-7B	0.8571	0.3333	0.4667	0.3846	0.5556	0.2857	0.000	0.8000
Qwen2.5-VL-32B	1.000	1.000	0.7333	0.5384	0.4444	0.7142	0.000	0.7000
Qwen2.5-VL-72B	0.8571	0.3333	0.6000	0.5384	0.6667	0.7142	0.1428	0.9000
Gemini 2.5 Flash	0.7142	0.5556	0.6667	0.4615	0.6667	0.7142	0.5714	0.9000
Gemini 2.5 Pro	0.7142	0.5556	0.7333	0.6153	0.6667	0.8571	0.8571	0.8000
	Obj Tow.	Obj Away	Sphere Tow.	Sphere Away	Dominoes	Ramp	Table	Avg.
Qwen2.5-VL-7B	0.5000	0.4444	0.3333	0.5000	0.7000	0.5000	0.4167	0.3737
Qwen2.5-VL-32B	0.6000	0.4444	0.4444	0.4000	0.8000	0.5000	0.4167	0.5714
Qwen2.5-VL-72B	0.6000	0.4444	0.3333	0.3000	0.8000	0.6667	0.3333	0.4309
Gemini 2.5 Flash	0.5000	0.5556	0.8889	0.5000	0.6000	0.5000	0.3333	0.4751
Gemini 2.5 Pro	0.8000	0.3333	0.5556	0.3000	0.7000	0.6667	0.4167	0.4972

We also summarized the performance of each model on multiple choice and True/False questions. Among the models, GLM 4.1V 9B performs the best in Multiple choice questions and Claude Sonnet 4 performs best on True/False questions.

Table 5: **Results of SOTA models on our multiple-choice and truth/false benchmark**.

Model	Multiple Choice	True/False
Qwen2.5-VL-7B	0.35	0.4262
Qwen2.5-VL-32B	0.45	0.4918
Qwen2.5-VL-72B	0.4	0.4918
GLM 4.1V 9B	0.4333	0.5345
Mistral Small 3.2 24B	0.2833	0.42622
Gemini 2.5 Flash	0.4166	0.5902
Gemini 2.5 Pro	0.4166	0.6557
Claude Sonnet 4	0.4083	0.6885
GPT 4.1	0.3333	0.4426