GRASP: A Novel Benchmark for Evaluating Language GRounding and Situated Physics Understanding in Multimodal Language Models

Serwan Jassim[1] , **Mario Holubar**[2] , **Annika Richter**[1] , **Cornelius Wolff**[1] , **Xenia Ohmer**[1] and **Elia Bruni**[1]
[1]Osnabrück University
[2]University of Amsterdam
serwan.jassim@uos.de, mario.holubar@student.uva.nl,
{annrichter, cowolff, xenia.ohmer, elia.bruni}@uos.de

Abstract

This paper presents GRASP, a novel benchmark to evaluate the language grounding and physical understanding capabilities of video-based multimodal large language models (LLMs). This evaluation is accomplished via a two-tier approach leveraging Unity simulations. The first level tests for language grounding by assessing a model's ability to relate simple textual descriptions with visual information. The second level evaluates the model's understanding of "Intuitive Physics" principles, such as object permanence and continuity. In addition to releasing the benchmark, we use it to evaluate several state-of-the-art multimodal LLMs. Our evaluation reveals significant shortcomings in the language grounding and intuitive physics capabilities of these models. Although they exhibit at least some grounding capabilities, particularly for colors and shapes, these capabilities depend heavily on the prompting strategy. At the same time, all models perform below or at the chance level of 50% in the Intuitive Physics tests, while human subjects are on average 80% correct. These identified limitations underline the importance of using benchmarks like GRASP to monitor the progress of future models in developing these competencies.

1 Introduction

The remarkable progress of large language models (LLMs) has sparked intense debate about their potential for achieving genuine machine cognition and human-level comprehension. A very important question is whether these models possess an understanding of the physical world, even though they learn primarily through next-word prediction, so in a fundamentally ungrounded manner. Recently, multimodal LLMs have started to make their mark, learning not just from vast text databases but also from extensive visual inputs. However, it remains unclear whether multimodality enhances language grounding and physical understanding. With **GRASP**, which stands for **GR**ounding **A**nd **S**ituated **P**hysics, we propose a novel evaluation benchmark to answer these questions for video-based LLMs.

Most research on physical understanding in LLMs has been centered around tasks that assess a model's ability to correctly link a textual question to its corresponding textual answer [Gordon *et al.*, 2012; Hendrycks *et al.*, 2021]. For instance, consider a data point in the "Choice of plausible alternatives" (COPA) dataset [Gordon *et al.*, 2012]: Given the text "I poured water into the glass.", the model has to choose one of the two effects "The water quenched my thirst." or "The glass became full." Under a conservative interpretation, this test evaluates whether the model has learned to map the input text pattern onto the correct output text pattern from examples of similar sentence pairs in its vast training dataset. It certainly does not evaluate whether the model knows what water looks and behaves like and whether the model genuinely understands that pouring liquid into a glass will fill up that glass.

GRASP advances the evaluation of language grounding and physical grounding by leveraging multimodality: Questions about objects and their behaviors are connected to an external environment. Specifically, we use the Unity[1] game engine to simulate various scenes (in the form of videos) as a proxy for the real world. The model has to answer textual questions about these scenes, which are designed to assess its ability to relate text to the physical world (*grounding*) as well as its ability to understand the basic principles of physics that govern how objects behave in the world (*Intuitive Physics*).

Based on these two aspects, the GRASP dataset is comprised of two levels. Level 1 tests for grounding: The model has to detect simple objects and recognize object features, relative positions, and (direction of) motion. This level tests whether a model can map between simple textual descriptions and visual inputs. Passing Level 1 is necessary, if the model is to generate statements about the physical properties and behaviors of objects in a visual scene. It is therefore a requirement for passing Level 2 which tests for Intuitive Physics: The model has to judge the physical plausibility of video sequences, where simple objects behave according to, or violate, Intuitive Physics concepts such as *object permanence* or *continuity*. The knowledge and abilities of LLMs are tied to their accessibility through language. Thus, the two levels evaluate a model's ability to "perceive" the environment and to "reason" about the physical events therein within the constraints of its language interface.

[1]https://unity.com/

This language interface differentiates GRASP from other image- or video-based datasets for Intuitive Physics, which were largely developed to train and evaluate dedicated Intuitive Physics models (lacking language) [Battaglia *et al.*, 2016; Watters *et al.*, 2017; Piloto *et al.*, 2022]. Most relevant to our approach is the recently developed Physical Concepts[2] dataset [Piloto *et al.*, 2022], which, similar to our Level 2, consists of simulated videos of physically plausible and implausible events. Instead of targeting dedicated Intuitive Physics models, GRASP is designed to evaluate whether language grounding and Intuitive Physics are a subset of the abilities that emerge in multimodal LLMs.

To this end, GRASP extends and improves existing datasets in several ways. First, we introduce novel stimuli to test for grounding (Level 1). While grounding is naturally absent in vision-only models it is a key prerequisite for a (question-based) evaluation of Intuitive Physics in multimodal LLMs. Second, we significantly broaden the range of Intuitive Physics concepts that can be tested (Level 2). While *Physical Concepts* comprises five concepts, our benchmark comprises eight and provides multiple experiments for some concepts as well as combinations of concepts. Lastly, by disseminating the Unity source code along with the benchmark data, we enable the community to customize and expand the benchmark as multimodal LLMs become more sophisticated.[3]

Accompanying the release of GRASP, we provide scores for five state-of-the-art multimodal LLMs. Our findings reveal that, despite their impressive capabilities, current multimodal LLMs are still lacking in both language grounding and intuitive physics understanding. While the tested models demonstrate certain grounding capabilities, specifically regarding colors and shapes, they universally fail the Intuitive Physics tests. These shortcomings emphasize the necessity for using benchmarks like GRASP to monitor the progress of future models in terms of these capabilities.

2 Related Work

GRASP takes inspiration from psychology research on *Intuitive Physics in early development*. It is related to other datasets that have been created to develop *neural network models of Intuitive Physics* but targets language models. We study *grounding and physics understanding in LLMs* but instead of looking at text-based models, we focus on recently developed *video-based multimodal LLMs*.

Multimodal LLMs. With the recent widespread success of LLMs, researchers have also explored their use for processing multi-modal inputs. A key idea utilized by seminal works such as Flamingo [Alayrac *et al.*, 2022] and BLIP-2 [Li *et al.*, 2023a] is to align a pretrained vision model with the textual embedding space of an LLM. LLaVA [Liu *et al.*, 2023] and MiniGPT-4 [Zhu *et al.*, 2023] combine this technique with instruction tuning to deliver an end-to-end chatbot with visual reasoning abilities. Recent work has extended this approach to video data. VideoChat [Li *et al.*, 2023b] was among the first to do so, followed by Video-ChatGPT [Maaz *et al.*, 2023]. Video-LLaMA [Zhang *et al.*, 2023] adds an audio processing branch to the pipeline, and PandaGPT [Su *et al.*, 2023] supports a total of six different modalities including thermal images and IMU sensor data. Although these models differ in implementation and training data, they are all based on the premise of aligning the embeddings of pretrained foundation models through a learnable interface.

Intuitive Physics in early development. To evaluate Intuitive Physics in LLMs, we rely on extensive research on the subject from developmental psychology. In particular, we focus on fundamental Intuitive Physics concepts that have been identified in this research, such as object permanence [Baillargeon *et al.*, 1985; Baillargeon, 1987; Spelke *et al.*, 1992], gravity [Kim and Spelke, 1992; Kim and Spelke, 1999; Spelke *et al.*, 1992], and inertia [Kim and Spelke, 1999; Spelke *et al.*, 1992; Spelke *et al.*, 1994]. Studies in this field typically employ simple experimental setups with geometric shapes that are part of physically possible or physically impossible events and assess understanding with the so-called *violation-of-expectation* (VoE) paradigm. This paradigm is based on the idea that infants will show surprise—measured through their behavioral or physiological response—when an event violates their expectations. We draw inspiration from known experimental setups to develop the videos for GRASP. Since LLMs possess language, we can inquire about the plausibility or implausibility of the scenes directly, mapping the VoE paradigm onto a binary classification problem.

Neural network models of Intuitive Physics. Advances in deep learning and AI have paved the way for the development of dedicated models capable of learning Intuitive Physics from data (for an overview, see Duan et al. [2022]). One of the earliest examples is the so-called *Interaction Network* [Battaglia *et al.*, 2016]. It received explicit information about objects and their relationships for a given scene and was trained on next-state prediction. Scenes included n-body systems, bouncing balls, and springs colliding with rigid bodies. Since then, the research focus has shifted to learning Intuitive Physics from raw visual inputs. For example, Watters et al. [2017] extended the Interaction Network with a convolutional neural network to process image inputs. Other prominent examples include neural networks predicting the behavior of block towers [Lerer *et al.*, 2016] or the dynamics of robot-object interactions [Agrawal *et al.*, 2016].

As a part of model development and evaluation, several Intuitive Physics benchmarks have been developed. Among others, the Physical Concepts [Piloto *et al.*, 2022], IntPhys [Riochet *et al.*, 2022], InfLevel [Weihs *et al.*, 2022], and AVoE [Dasgupta *et al.*, 2021a; Dasgupta *et al.*, 2021b] benchmarks utilize simulated videos to assess physical concepts in computer vision models. Like GRASP, they focus on established concepts and VoE experiments from the developmental psychology literature. GRASP can be considered an expansion of these benchmarks: It adds an entirely new level to test for grounding, as well as more concepts, more scenes per concept, and combinations of concepts. Besides, unlike all the approaches above, it is designed to evaluate LLMs and therefore uses a language interface, allowing for analyses within and across

[2]https://github.com/deepmind/physical_concepts

[3]The benchmark including all our code as well as supplementary material is available at https://github.com/i-machine-think/grasp.

the modalities of language and vision.

Grounding and physics understanding in LLMs. Before the advent of multimodal LLMs, it has often been argued that LLMs—being trained on text—fail to relate language to the physical world [Bisk *et al.*, 2020]. Progress on neural networks with grounded language abilities largely happened in research fields originating from image captioning [Mitchell *et al.*, 2012], such as visual question answering (VQA) [Antol *et al.*, 2015], instruction following [Anderson *et al.*, 2018; Das *et al.*, 2018], and visual commonsense reasoning [Zellers *et al.*, 2019]. Among others, the VQA datasets CLEVR [Johnson *et al.*, 2017] and CLEVRER [Yi *et al.*, 2020] have been developed in the context of this research. The former tests for compositional language and elementary visual reasoning using images of 3D shapes. The latter tests for various visual reasoning capabilities based on videos of colliding objects (again 3D shapes). At the time CLEVRER was released, SOTA models performed well on descriptive tasks but poorly on causal ones requiring explanation, prediction, or counterfactual reasoning. Now, multimodal LLMs promise to provide extensive language capabilities while being able to relate between text and images or videos. For example, GPT-4 outperforms SOTA models on various VQA benchmarks (https://openai.com/research/gpt-4). GRASP provides a novel, and extensive benchmark to test fundamental grounding and Intuitive Physics skills in a question-answering setting.

3 Benchmark Design

GRASP is a two-level benchmark, with each level containing multiple visual tests. These tests were modeled in the Unity simulator and compiled into a dataset in the form of videos.[4] All videos are ten seconds long and were generated at 50 frames per second.

3.1 Level 1 (Grounding)

The initial stage of GRASP evaluates the elementary visual understanding capabilities of LLMs. This stage comprises tests that assess basic visual comprehension and lay the groundwork for higher-order reasoning required in the subsequent level. We premise that models struggling at this foundational phase will likely encounter difficulties in the next stage, where they must discern and reason about more complex physical interactions. This approach ensures a sequential increase in task complexity, aligning with the natural progression of cognitive development. More specifically, Level 1 comprises six test categories:

- **Shape**: A cube or a sphere of random size is spawned at a random location on a table.
- **Color**: A black, blue, green, or red sphere of random size is spawned at a random location on a table.
- **Directionality**: A ball rolls forward, backward, right, or left.

[4]Although most Level 1 tests do not entail dynamics and could therefore be represented as images, we decided to capture them as videos to ensure consistency across all Level 1 and 2 tests.

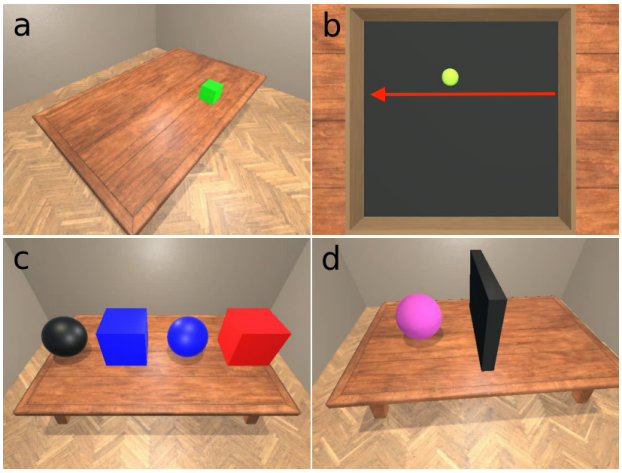

Figure 1: Examples from GRASP's Level 1: (a) Shape & Color (both have the same setup, but differ in the randomized property), (b) Directionality & Movement (ball is rolling from right to left as indicated by red arrow), (c) Object Ordering, (d) Relational Position.

- **Movement**: A ball rolls in a random direction or stands still.
- **Object Ordering**: A random sequence of balls or cubes (between two and four) of random color are spawned on a table. Each object in one video is unique.
- **Relational Position**: A ball is spawned either to the left or right of a barrier.

For each test, we generate 128 videos. Examples of these tests are displayed in Figure 1. Importantly, the elements used in this phase—such as objects in various shapes, colors, and positional relationships—form the fundamental components of the videos in the next stage.

Prompts

For this level of the benchmark, we introduce two different sets of prompts, each giving rise to a different classification task. To maintain uniformity with Level 2, one set is designed to induce a **binary classification** problem. We generate positive and negative samples by combining each input video with a prompt that proposes an observation and asks whether this observation is true. The observation is either a true (pos. sample) or a false (neg. sample) statement about the video. In Figure 1c, for example, the model is prompted with "From left to right, the following objects are on the table: black ball, blue cube, blue ball, red cube. Is this true?" (pos. sample) and "From left to right, the following objects are on the table: red cube, blue ball, black ball, blue cube. Is this true?" (neg. sample). A complete list of the prompts is provided in the supplementary material. For the Object Ordering test, we exclusively create negative samples by permuting the existing objects, to specifically assess ordering accuracy. For all other tests, we sample entirely different observations.

To evaluate the models' sensitivity to the instructions, we propose a second set of prompts that allow for open-ended answers and frame the problem as **multi-class classification**. For example, the prompt for the Directionality test is "Which

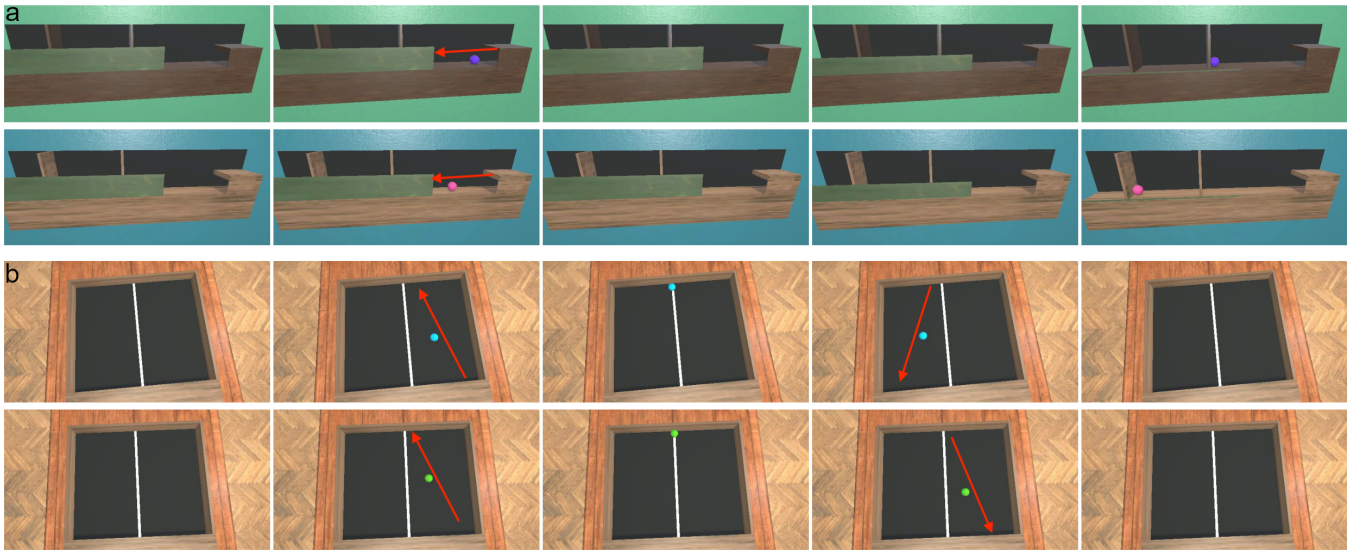

Figure 2: Examples from GRASP's Level 2. The first and second rows display plausible and implausible versions of each experiment, respectively. The examples here test the understanding of Solidity & Continuity (a), and Inertia (b). We cut off some of the background here.

direction is the ball rolling?" in this case. These prompts were not introduced for Object Ordering since it proved too difficult to parse the answers and also not for Movement which does not allow for an open-ended question. All open-ended question prompts can be found in the supplementary material.

3.2 Level 2 (Intuitive Physics)

Level 2 comprises tests of Intuitive Physics understanding, which follow the structure of VoE experiments by contrasting physically plausible and implausible scenes. Specifically, Level 2 comprises tests for the following eight concepts and events:

- **Continuity**: Objects cannot teleport in space and time, they can only move along continuous paths [Spelke *et al.*, 1992].
- **Solidity**: Objects cannot overlap in space and time, they can only move along clear paths [Spelke *et al.*, 1992].
- **Unchangeableness**: Objects cannot spontaneously change their size, shape, and color [Baillargeon *et al.*, 2012].
- **Gravity**: Objects move downward without existing support [Spelke *et al.*, 1992].
- **Support**: Objects maintain stability when on a platform, but lose stability when positioned off it [Baillargeon, 1995].
- **Collision**: Objects get displaced when hit by moving objects [Baillargeon, 1995].
- **Object Permanence**: Objects continue to exist when they are occluded [Baillargeon, 1995].
- **Inertia**: Objects do not spontaneously alter their motion [Spelke *et al.*, 1992].

We adopt 16 tests from the psychology literature and create a dataset containing 128 physically plausible and 128 physically implausible videos for each of them. Illustrations of these tests can be found in Figure 2 and the supplemental material and their distribution across the physical concepts and events is provided in Table 1. Since it is not always possible to strictly disentangle all the concepts in the tests (e.g. in the implausible scenario displayed in Figure 2a, the ball could have either teleported behind (continuity violation) or rolled through the barrier (solidity violation)), there is a discrepancy between the sum of the distribution in Table 1 and the number of tests. We also adopt the use of occluders to hide physical manipulations from experimental studies (e.g. Figure 2a). This ensures that models need to infer plausibility using the entire history of a video instead of individual frames.

Continuity	Solidity	Inertia	Gravity
4	2	4	5
Collision	**OP**	**Support**	**Unchangeableness**
1	3	1	2

Table 1: Test distribution over physical principles/events that are evaluated in Level 2 of GRASP (OP = Object Permanence).

Visual Randomization

In GRASP, models are validated against multiple videos per test to ensure that results are representative. We achieved rich visual variation for each test by randomly sampling object colors and textures, background colors, camera angles, movement speeds, start delays, orientations of experimental setups, as well as other test-specific parameters (see supplementary material) for each video, as applicable.

Prompts

The model has to evaluate the physical plausibility of the videos, that is, perform a binary classification task. In contrast

to Level 1, negative and positive samples of a test share the same prompt and are distinguished by the video contents. All prompts are based on the same template:

> The video you're seeing was generated by a simulator. Given how objects behave on earth, is ⟨*observation*⟩ plausible? Your answer should be based on the events in the video and ignore the quality of the simulation. Answer only with yes or no.

In this template, ⟨*observation*⟩ denotes a phrase that hints at what the model should pay attention to when inferring physical plausibility, e.g. "the trajectory of the ball" or "the final position of the ball". The models are instructed to ignore details related to the simulation quality to prevent their judgment from being influenced by visual inaccuracies. A complete list of prompts can be found in the supplementary material.

4 Experiments

In our experiments, we evaluate several state-of-the-art models as well as human subjects on our benchmark. Details on the model evaluation are provided in Sections 4.1–4.3 and details on the human evaluation in Section 4.4.

4.1 Models

We consider several multimodal LLMs that can perform video-based question answering.

Video-ChatGPT [Maaz *et al.*, 2023] leverages LLaVA [Liu *et al.*, 2023] as a vision-language model and adapts it to video data, fine-tuning the model on a dataset of video-instruction pairs to enable understanding of temporal dynamics. Specifically, it uses LLaVA-Lightning-7B v1.1, which is comprised of CLIP ViT-L/14 [Radford *et al.*, 2021], as a visual encoder and Vicuna-7B v1.1 [Chiang *et al.*, 2023] as a language decoder.

Video-LLaMA [Zhang *et al.*, 2023] enables simultaneous visual and auditory understanding using a multi-branch cross-modal pre-training framework. The vision-language branch uses CLIP ViT-G/14 [Radford *et al.*, 2021] and BLIP-2 Q-Former [Li *et al.*, 2023a] and is trained using video-text as well as image-text data. We test three versions of the model, which use Vicuna-7B v0, Vicuna-13B v0, and LLaMA-2-7B as their respective language decoder. They are all fine-tuned on instruction-tuning data from MiniGPT-4, LLaVA, and VideoChat.

PandaGPT [Su *et al.*, 2023] uses the joint embeddings of ImageBind [Girdhar *et al.*, 2023] to enable a Vicuna model to reason about image, video, depth, thermal, and IMU data. The multimodal encoder's feature space is aligned with the language model by training on image-language instruction-following data. The particular versions we test are the 7B version with a maximum sequence length of 1024 using Vicuna-7B v0 and the 13B version with a maximum sequence length of 400 using Vicuna-13B v0.

VTimeLLM [Huang *et al.*, 2023] adds an additional stage to the training pipeline alongside feature alignment and instruction tuning. This stage, called Boundary Perception, aims to improve the model's temporal understanding abilities by training on a dataset of time-segmented and event-annotated videos. It uses CLIP ViT-L/14 as its visual encoder and Vicuna-7B v1.5 as the language decoder.

4.2 Prompting

Apart from the prompts introduced in Section 3, we report the results for additional prompting strategies for Level 1. For Level 2 changes to the prompting strategy did not impact the results (the models still fail to perform the task). In particular, we include one-shot prompting and chain-of-thought (CoT) prompting for the Level 1 binary classification task. With one-shot prompting, we "familiarize" the models with the task by prepending an example question-answer pair to the main question. For instance, a one-shot prompt for the Color test is: "This is an example of a question about this video and the correct answer. Question: The ball on the table is red. Is this true? Answer only with yes or no. Answer: Yes. Next, I want you to answer my next question in the same way with regard to the next video." Whether a positive or negative sample is used in the initial prompt is randomized. For CoT prompting, we prepend the open-ended question (see supplementary material) to the binary classification prompt (for the Movement and Ordering tests, we use the question "What can you see in this video?"). For both, CoT and one-shot prompting, we allow the model to reply before submitting the final instruction containing the task.

4.3 Experimental Setup

Experiments are conducted on an Nvidia 3090 GPU for the 7B models and on an Nvidia A100 for the 13B models. For all models, we use their default parameters but adapt the system prompt when applicable (see supplementary material for details). Each video-prompt pair is evaluated three times with a different seed per model. For quantitative evaluation, the models' responses are classified by a simple scheme: Responses to binary yes/no questions are only counted as valid if they begin with the word "yes" or "no"; the rest of the response is considered irrelevant. We regard responses that do not adhere to this as incorrect. For open-ended questions in Level 1, we use a parsing scheme that also considers slight deviations from the ground truth as correct (e.g. ball and sphere are considered to be equivalent). The full parsing scheme is outlined in the supplementary material.

4.4 Comparison with Human Subjects

To validate and assess the difficulty of our benchmark, we submit GRASP's Level 2 tests to AWS Mechanical Turk[5] for a human trial. We focus our evaluation on Level 2, considering that human subjects can trivially solve tests in Level 1. In our experiment, participants are asked to judge the physical plausibility of each Level 2 test, resulting in 16 videos per questionnaire. For each test, we randomly sample whether to serve the plausible or implausible scene. Furthermore, we randomize the order in which the videos are displayed to each participant. Three independent submissions are collected per video from which a final answer is determined using a majority vote. We collect submissions from 120 participants, i.e. a subset of 40 videos are being classified per Level 2 test (20 per plausible and 20 per implausible scene).

[5]https://www.mturk.com

Task	Test	Video-LLaMA (7B)	Video-LLaMA (13B)	Video-LLaMA2 (7B)	PandaGPT (7B)	PandaGPT (13B)	VTimeLLM (7B)	Video-ChatGPT (7B)
Binary Classification Zero-Shot	Shape	49.1	49.1	48.3	50.0	50.0	50.0	49.9
	Color	49.6	52.1	50.8	50.0	50.0	50.0	50.1
	Movement	49.6	46.6	49.6	35.2	50.0	50.0	50.4
	Directionality	48.7	45.8	49.9	47.3	50.0	50.0	49.6
	Relational Position	50.1	47.9	49.0	50.0	50.0	50.0	49.7
	Ordering (avg.)	49.6	49.7	50.0	50.0	50.0	50.0	50.3
Binary Classification CoT	Shape	**69.4**	**69.5**	**63.8**	33.6	**80.9**	50.0	**70.1**
	Color	**79.7**	**73.3**	**84.2**	**76.2**	**70.7**	50.0	**65.2**
	Movement	48.8	48.8	48.3	44.1	35.9	50.0	43.1
	Directionality	45.3	47.5	50.1	51.2	51.2	50.0	46.1
	Relational Position	46.9	48.7	51.4	50.0	50.0	50.0	47.4
	Ordering (avg.)	48.6	49.0	50.0	49.9	50.7	50.0	50.6
Binary Classification One-Shot	Shape	41.4	23.6	45.6	50.5	46.1	41.9	48.4
	Color	43.6	23.0	49.6	29.6	44.3	32.0	49.0
	Movement	37.1	26.3	38.0	28.0	39.1	37.5	50.8
	Directionality	36.6	20.1	42.8	37.4	42.8	33.5	50.7
	Relational Position	40.2	21.0	37.4	53.4	35.4	25.7	51.4
	Ordering (avg.)	42.2	22.3	22.2	50.7	50.5	20.5	49.7
Multi-Class Classification Zero-Shot	Shape	**87.2**	**83.3**	**75.8**	49.2	14.8	14.1	14.3
	Color	**93.2**	**74.2**	**90.9**	**70.3**	**73.4**	**85.7**	**76.0**
	Directionality	14.1	15.9	11.2	26.6	25.0	20.8	17.4
	Relational Position	28.9	24.2	35.4	0.0	50.0	49.7	49.5

Table 2: Accuracy (%) for all models on GRASP's Level 1 using binary question (inducing binary classification) and open-ended question prompts (inducing multi-class classification). For binary classification, results are listed for zero-shot, one-shot, and chain-of-thought (CoT) prompting strategies. We highlight all accuracies that lie notably above chance performance. For binary classification, an accuracy of 50% coincides with chance performance, while for multi-class classification it is 50% for Shape and Relational Position, and 25% for Color and Directionality.

5 Results

Table 2 presents accuracy scores for the multimodal LLMs tested on **Level 1 (grounding)**. For zero-shot binary classification, the average scores across all models and tests indicate a performance close to chance (50%). Results below chance performance can be attributed to ambiguous answers that could not be parsed. Examining scores on positive and negative samples individually (see supplementary material) highlights distinctive behaviors among the models. Video-LLaMA exhibits a consistent bias toward responding "yes" to prompts, while Video-ChatGPT displays a more dynamic bias, shifting between "yes" and "no" responses across different tests. PandaGPT and VTimeLLM, in turn, consistently respond with "yes" regardless of the test category.

Furthermore, the results show that binary classification with CoT prompting leads to a consistent improvement above chance performance for the Shape and Color tasks across all models, except for PandaGPT (7B) and VTimeLLM. On the other hand, such an improvement is not observed when running binary classification with one-shot prompting. In this case, models tend to perform even worse than with zero-shot prompting due to simply repeating the answer from the provided example.

Accuracies for multi-class classification in Level 1 are also presented in Table 2. For the Directionality and the Relational Position tasks, the performance coincides with zero-shot binary classification for all models, being either chance or below chance due to ambiguous answers (in this case, accuracies of 25% and 50% equal chance performance for Directionality and Relational Position respectively). Similar to the CoT results, all models perform quite well in the Color task as well as Video-LLaMA and PandaGPT (7B) in the Shape task.

Results for **Level 2 (Intuitive Physics)** are displayed in Figure 3. We do not report individual scores since the models generally exhibit performance equivalent to, or less than, chance across all tests as indicated by the error bars. Video-LLaMA performs significantly below chance because it generates answers that cannot be parsed. The models' poor performance on Level 2 is not surprising given that they already failed on Level 1, which assesses basic grounding abilities necessary to answer questions about the videos at Level 2. With human performance at approximately 80%, the results suggest that while the task is solvable, it presents a non-trivial challenge. In particular, difficulties in the human trial were observed in one test concerning the inertia principle where participants were not able to correctly observe a discrepancy between the incoming and outgoing angle of a deflection. This test will be highlighted accordingly in the published benchmark such that researchers can decide whether to include it in their evaluations.

To understand whether the bad performance on Level 2 is due to the multimodal nature of the task, we additionally evaluate Video-LLaMA on textual descriptions of the test events (see supplementary material). Video-LLaMA and Video-LLaMA2 are the only models out of those tested that allow for text-only inputs. On the binary classification task, the 7B and 13B parameter versions of Video-LLaMA achieve an accuracy of 49.8% and 53.7%, respectively, while Video-LLaMA2 achieves 51.9%. In comparison, GPT-4 achieves an accuracy of 75.0% on this task. This analysis suggests

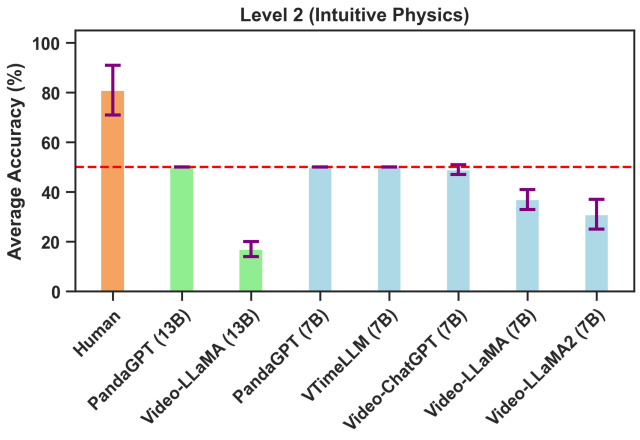

Figure 3: Average accuracies (%) over Level 2 tests for all models and the human trial. The red line indicates chance performance and error bars represent the standard deviation over all tests (positive and negative scenes combined). Models with green bars have 13B parameters and models with blue bars have 7B parameters.

that video-based LLMs might also lack an Intuitive Physics understanding when evaluated on text alone.

6 Discussion

Our results show that the tested multimodal LLMs lack a basic perceptual understanding of simulated scenes. Besides some understanding of simple shapes and colors, the models fail to answer basic questions about (relative) positions of static objects and movements of objects (Level 1). Because of that, it is not surprising that they also fail to judge the physical plausibility of simple object behaviors (Level 2).

Our model evaluation on Level 1 revealed that using a different prompting strategy (CoT) or changing the instruction (from yes-no questions to open-ended questions) can result in a substantial performance boost. Specifically, the positive results using CoT prompting indicate that models are not able to extract necessary visual information following simple binary questions. When "guiding" models with more unspecific questions first (initial prompt in CoT), they are sometimes able to extract the necessary information from their context to solve the same subsequent binary questions. This highlights the high sensitivity to the nature of the prompts and the necessity for future models to improve upon these limitations. However, for Level 2, altering the prompting strategy did not impact the results. The models still failed to perform the task at this level, indicating that a lack of visual comprehension of more complex scenes is still at the heart of the problem.

Future work will encompass the analysis of additional multimodal LLMs. Considering the recency of video-based multimodal LLMs (all the evaluated models were released last year), their capabilities may soon improve significantly. Compared to text-based LLMs, which at this point contain hundreds of billions of parameters [Naveed *et al.*, 2023], video-based multimodal LLMs are at least one order of magnitude smaller [Zhang *et al.*, 2023; Su *et al.*, 2023; Maaz *et al.*, 2023]. Supported by the observation that GPT-4 significantly outperforms Video-LLaMA and Video-LLaMA2 in tests with scene descriptions, an increase in model size alone might lead to the emergence of relevant language grounding and physical reasoning capabilities. Image-based multimodal LLMs, such as GPT-4, have proven remarkably adept at answering complex and detailed questions about images. Therefore, we aim to create an image-based version of GRASP and to compare image- and video-based models. Throughout future developments in multimodal LLM capabilities, GRASP will prove instrumental in tracking the progression of these models, testing their grounding and physical comprehension capabilities against demanding data sets.

One potential reason for the poor performance of the models could be the discrepancy between the simulated videos in our benchmarks and the real-world training data of the models. In other words, our benchmark data is out-of-distribution (OOD) for the model. Generating a controlled dataset of real-world videos that test fundamental aspects of grounding and Intuitive Physics is difficult. Still, it would be interesting to conduct a comparison with selected examples of such stimuli in future work. Either way, GRASP is useful as a challenging benchmark. Due to its OOD nature, it tests for scene understanding that generalizes to novel, and in our case abstracted, scenarios. The limitations observed in current models' performance on GRASP's Level 1 stress the need for additional basic perceptual tests to allow for a more detailed analysis. Furthermore, a future expansion to the benchmark could involve a subsequent level that requires models to address challenges within an *interactive* simulated environment.

7 Conclusion

GRASP introduces a robust grounding and Intuitive Physics benchmark tailored for multimodal LLMs. By using simulated videos to model basic perceptual tasks and faithfully reproducing experiments from developmental psychology research within a simulation, GRASP serves as a comprehensive evaluation platform. Results across both benchmark tiers demonstrate the challenging nature of GRASP. Notably, the results indicate a lack of perceptual understanding of simulated scenes by existing models, stressing the need for further development in this domain. We plan to expand the benchmark in future work to facilitate research at the intersection of language and perception.

Acknowledgments

We would like to thank Dieuwke Hupkes for the idea that led to the creation of GRASP. We also thank Jülich Supercomputing Centre for the computing time and support to develop this project.

Contribution Statement

Xenia Ohmer and Elia Bruni share senior authorship for this work.

References

[Agrawal *et al.*, 2016] Pulkit Agrawal, Ashvin V. Nair, Pieter Abbeel, Jitendra Malik, and Sergey Levine. Learning to

poke by poking: Experiential learning of intuitive physics. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Proceedings of the 30th International Conference on Neural Information Processing Systems (NeurIPS)*, volume 29. Curran Associates, Inc., 2016.

[Alayrac *et al.*, 2022] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning. *ArXiv Preprint*, arXiv:2204.14198, 2022.

[Anderson *et al.*, 2018] Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton van den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[Antol *et al.*, 2015] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015.

[Baillargeon *et al.*, 1985] Renée Baillargeon, Elizabeth S. Spelke, and Stanley Wasserman. Object permanence in five-month-old infants. *Cognition*, 20:191–208, 1985.

[Baillargeon *et al.*, 2012] Renée Baillargeon, Maayan Stavans, Di Wu, Yael Gertner, Peipei Setoh, Audrey K. Kittredge, and Amélie Bernard. Object individuation and physical reasoning in infancy: An integrative account. *Language learning and development: The official journal of the Society for Language Development*, 8(1):4–46, 2012.

[Baillargeon, 1987] Renée Baillargeon. Object permanence in 3½- and 4½-month-old infants. *Developmental Psychology*, 23(5):655–664, 1987.

[Baillargeon, 1995] Renee Baillargeon. Physical reasoning in infancy. In Michael S. Gazzaniga, editor, *The Cognitive Neurosciences*, pages 181–204. MIT Press, 1995.

[Battaglia *et al.*, 2016] Peter Battaglia, Razvan Pascanu, Matthew Lai, Danilo Jimenez Rezende, and Koray kavukcuoglu. Interaction networks for learning about objects, relations and physics. In *Proceedings of the 30th International Conference on Neural Information Processing Systems (NeurIPS)*, pages 4509–4517, Red Hook, NY, USA, 2016. Curran Associates Inc.

[Bisk *et al.*, 2020] Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, Nicolas Pinto, and Joseph Turian. Experience grounds language. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8718–8735, Online, 2020. Association for Computational Linguistics.

[Chiang *et al.*, 2023] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing GPT-4 with 90%* ChatGPT quality. *LMSYS Org Blog*, March 2023.

[Das *et al.*, 2018] Abhishek Das, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. Embodied question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[Dasgupta *et al.*, 2021a] Arijit Dasgupta, Jiafei Duan, Marcelo H Ang Jr, Yi Lin, Su-hua Wang, Renée Baillargeon, and Cheston Tan. A benchmark for modeling violation-of-expectation in physical reasoning across event categories. *ArXiv Preprint*, arXiv:2111.0882, 2021.

[Dasgupta *et al.*, 2021b] Arijit Dasgupta, Jiafei Duan, Marcelo H Ang Jr, and Cheston Tan. Avoe: A synthetic 3d dataset on understanding violation of expectation for artificial cognition. *ArXiv Preprint*, arXiv:2110.05836, 2021.

[Duan *et al.*, 2022] Jiafei Duan, Arijit Dasgupta, Jason Fischer, and Cheston Tan. A survey on machine learning approaches for modelling intuitive physics. In Lud De Raedt, editor, *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 5444–5452. International Joint Conferences on Artificial Intelligence Organization, 2022. Survey Track.

[Girdhar *et al.*, 2023] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. *ArXiv Preprint*, arXiv:2305.05665, 2023.

[Gordon *et al.*, 2012] Andrew Gordon, Zornitsa Kozareva, and Melissa Roemmele. SemEval-2012 task 7: Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In Eneko Agirre, Johan Bos, Mona Diab, Suresh Manandhar, Yuval Marton, and Deniz Yuret, editors, **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 394–398, Montréal, Canada, 2012. Association for Computational Linguistics.

[Hendrycks *et al.*, 2021] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations (ICLR)*, 2021.

[Huang *et al.*, 2023] Bin Huang, Xin Wang, Hong Chen, Zihan Song, and Wenwu Zhu. Vtimellm: Empower llm to grasp video moments, 2023.

[Johnson *et al.*, 2017] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[Kim and Spelke, 1992] In K. Kim and Elizabeth S. Spelke. Infants' sensitivity to effects of gravity on visible object motion. *Journal of Experimental Psychology: Human Perception and Performance*, 18(2):385–393, 1992.

[Kim and Spelke, 1999] In K. Kim and Elizabeth S. Spelke. Perception and understanding of effects of gravity and inertia on object motion. *Developmental Science*, 2(3):339–362, 1999.

[Lerer *et al.*, 2016] Adam Lerer, Sam Gross, and Rob Fergus. Learning physical intuition of block towers by example. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning (ICML)*, volume 48, pages 430–438. PMLR, 2016.

[Li *et al.*, 2023a] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *ArXiv Preprint*, arXiv:2301.12597, 2023.

[Li *et al.*, 2023b] KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. *ArXiv Preprint*, arXiv:2305.06355, 2023.

[Liu *et al.*, 2023] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *ArXiv Preprint*, arXiv:2304.08485, 2023.

[Maaz *et al.*, 2023] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-ChatGPT: Towards detailed video understanding via large vision and language models. *ArXiv Preprint*, arXiv:2306.05424, 2023.

[Mitchell *et al.*, 2012] Margaret Mitchell, Jesse Dodge, Amit Goyal, Kota Yamaguchi, Karl Stratos, Xufeng Han, Alyssa Mensch, Alex Berg, Tamara Berg, and Hal Daumé III. Midge: Generating image descriptions from computer vision detections. In Walter Daelemans, editor, *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 747–756, Avignon, France, 2012. Association for Computational Linguistics.

[Naveed *et al.*, 2023] Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. A comprehensive overview of large language models, 2023.

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

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *ArXiv Preprint*, arXiv:2103.00020, 2021.

[Riochet *et al.*, 2022] Ronan Riochet, Mario Ynocente Castro, Mathieu Bernard, Adam Lerer, Rob Fergus, Véronique Izard, and Emmanuel Dupoux. Intphys 2019: A benchmark for visual intuitive physics understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):5016–5025, 2022.

[Spelke *et al.*, 1992] Elizabeth S. Spelke, Karen Breinlinger, Janet Macomber, and Kristen Jacobson. Origins of knowledge. *Psychological review*, 99(4):605–632, 1992.

[Spelke *et al.*, 1994] Elizabeth S. Spelke, Gary Katz, Susan E. Purcell, Sheryl M. Ehrlich, and Karen Breinlinger. Early knowledge of object motion: continuity and inertia. *Cognition*, 51(1):131–176, 1994.

[Su *et al.*, 2023] Yixuan Su, Tian Lan, Huayang Li, Jialu Xu, Yan Wang, and Deng Cai. PandaGPT: One model to instruction-follow them all. *ArXiv Preprint*, arXiv:2305.16355, 2023.

[Watters *et al.*, 2017] Nicholas Watters, Andrea Tacchetti, Théophane Weber, Razvan Pascanu, Peter Battaglia, and Daniel Zoran. Visual interaction networks: Learning a physics simulator from video. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS)*, pages 4542–4550, Red Hook, NY, USA, 2017. Curran Associates Inc.

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

[Yi *et al.*, 2020] Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B. Tenenbaum. CLEVRER: Collision events for video representation and reasoning. In *International Conference on Learning Representations (ICLR)*, 2020.

[Zellers *et al.*, 2019] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[Zhang *et al.*, 2023] Hang Zhang, Xin Li, and Lidong Bing. Video-LLaMa: An instruction-tuned audio-vidual language model for video understanding. *ArXiv Preprint*, arXiv:2306.02858, 2023.

[Zhu *et al.*, 2023] 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.