Measuring Vision Language Models' Ability to Detect Physical Anomalies

Murdock Aubry**
University of Toronto
murdock@cs.toronto.edu

Adam Barroso*
University of Toronto
abarroso@cs.toronto.edu

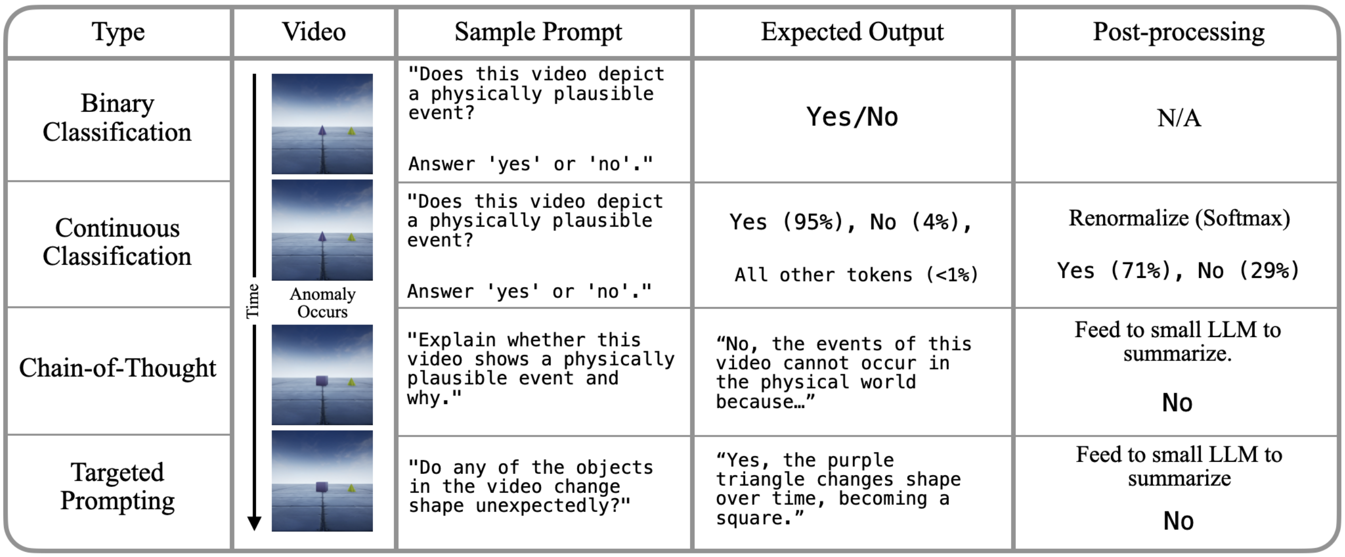

Figure 1: Schmematic diagram summarizing the experiment types used in this report.

Abstract

This study explores the intuitive physics capabilities of Vision-Language Models (VLMs) by introducing a novel set of language-native evaluation protocols. While models such as BLIP-2 and LLaVA have demonstrated efficacy in various multimodal tasks, they are not explicitly trained to detect physical anomalies. In this work, we propose a framework that re-contextualizes physical plausibility assessments as natural language tasks, allowing VLMs to reason about physical laws when appropriately prompted. Our experimental methodology encompasses binary classification, confidence-based plausibility scoring, chain-of-thought reasoning, and targeted prompting across several intuitive physics concepts. Despite the limitations imposed by computational resources and the use of smaller model variants, our results offer initial insights into the conditions under which VLMs are capable of detecting physical violations. This research contributes to the development of more accurate and equitable evaluation benchmarks for assessing the physical reasoning abilities of VLMs, emphasizing the need for future studies employing larger models and more robust training approaches.

*Both authors contributed equally to this research.

Conference'17, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

Keywords

Vision Language Models, Physics, Anomalies

ACM Reference Format:
Murdock Aubry and Adam Barroso. 2025. Measuring Vision Language Models' Ability to Detect Physical Anomalies. In . ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

Vision-Language Models (VLMs) have rapidly emerged as a dominant paradigm in multimodal AI, combining the perceptual capabilities of deep vision models with the flexible reasoning and compositionality of large language models. Systems such as BLIP-2, LLaVA, and Qwen-VL demonstrate impressive generalization across tasks like visual question answering, captioning, and multi-modal dialogue [11]. These models are often framed as a step toward more general forms of machine intelligence, capable of interpreting complex visual scenes and communicating about them in natural language.

Despite these advances, recent studies have highlighted a critical limitation: VLMs often lack an intuitive understanding of physical dynamics. While they may identify objects and describe surface-level relationships, they frequently fail at detecting violations of basic physical principles such as object permanence, shape constancy, or causality [3]. This gap limits their reliability in real-world applications that require grounded reasoning, such as robotics, embodied AI, and assistive technologies.

In contrast, recent work in video-based representation learning, particularly predictive architectures such as V-JEPA, has shown promise in capturing intuitive physics by modeling the dynamics of visual scenes directly. These models are evaluated using violation-of-expectation (VoE) metrics that quantify "surprise" when the observed outcome deviates from a model's internal prediction of the future. While powerful, these evaluation protocols are tightly coupled to the architectural assumptions of predictive models and are not directly applicable to VLMs, which reason through text rather than latent or pixel-based prediction.

In this project, we propose a set of language-native evaluation protocols to more fairly assess the intuitive physics capabilities of VLMs. By re-framing physical plausibility judgments as natural language tasks, we aim to determine whether VLMs can reason about violations of physical laws when appropriately prompted. Our experiments include binary classification, confidence-based plausibility scoring, chain-of-thought explanation, and targeted prompting across a range of intuitive physics concepts. We also compare base models with parameter-efficient fine-tuned variants (LoRA) to explore whether lightweight adaptation can improve physical reasoning.

Although our results are limited by compute constraints, such as the use of small models and short video clips, they provide initial insights into how and when VLMs can (or cannot) detect physical violations. More broadly, this work contributes toward defining fairer and more informative benchmarks for evaluating the physical reasoning abilities of general-purpose vision-language systems.

2 Related Work and Background

2.1 Physical Reasoning in Neural Networks

This work is largely inspired by the recent findings of [5]. Building on this, our work investigates whether fine-tuning a VLM on natural videos further enhances its ability to detect physical anomalies, quantifying improvements using the similar evaluation metrics.

Other recent works have expanded on similar goals using synthetic data and symbolic reasoning. Balazadeh et al. in "Synthetic Vision" [3] introduce a synthetic training pipeline that improves a VLM's understanding of physical interactions. By generating structured simulations using Physics Context Builders, they show that physics comprehension can be efficiently bootstrapped from synthetic environments, with improvements transferring to real-world settings.

Lastly, [19] develop a framework that learns to infer latent physical properties from 3D videos. Their autoregressive, prediction-based architecture is optimized for transfer learning across scenarios, offering a complementary perspective on learning generalizable physical representations from video.

To the best of our knowledge, however, very few works analyze the emergent ability to intuite physical laws, and more specifically, detect physical anomalies, directly from training data. Additionally, limited benchmarks have been widely adopted, requiring us to closely follow the methodology of [5].

2.2 Video Joint Embedding Predictive Architecture (V-JEPA)

V-JEPA [4] builds upon the Joint Embedding Predictive Architecture (JEPA) framework, which formulates representation learning as a prediction task in latent space rather than reconstructing raw input signals. In JEPA, the model learns to predict the latent representation of a masked or missing region of the input using contextual information from the visible regions, with the objective of aligning the predicted and target embeddings within a shared representation space [2]. This approach encourages the learning of abstract, high-level features while avoiding the complexity and limitations of pixel-level reconstruction or contrastive sampling.

Extending this formulation to the video domain, V-JEPA predicts latent representations of masked spatiotemporal regions using context from unmasked regions. The architecture comprises a frozen encoder, a context encoder, and a predictor network. The frozen encoder produces target latent embeddings for the masked regions, serving as the ground truth for training. The context encoder processes the visible (unmasked) portions of the video input to generate contextual embeddings, which are then passed to the predictor to estimate the embeddings of the masked regions. All learning takes place in latent space, decoupling the optimization objective from low-level signal reconstruction and instead promoting semantic abstraction and temporal coherence.

2.3 Vision Language Models

Vision Language Models (VLMs) integrate visual and linguistic modalities by aligning image or video encoders with language-based decoders or multimodal transformers. Pre-trained on large-scale datasets containing paired visual and textual data, these models learn to associate visual patterns with semantic concepts, enabling tasks such as image captioning, visual question answering, and video understanding [1, 7, 11]. Recent advances also highlight the importance of temporal understanding and spatial reasoning to enhance physical and causal interpretation. Fine-tuning VLMs on targeted data, whether synthetic or natural, has shown to be a promising avenue in improving their ability to recognize physical events and detect physical anomalies.

2.4 Low Rank Adaptation

Low-Rank Adapters (LoRAs) are a parameter-efficient fine-tuning technique designed to adapt large pre-trained models, such as vision-language models (VLMs), to downstream tasks with minimal computational overhead [6, 9]. Rather than updating all parameters during training, LoRA introduces trainable rank-decomposed matrices into existing model layers—typically linear projections within transformer blocks—while keeping the original weights frozen. Specifically, the weight update ΔW is parameterized as a product of two low-rank matrices $A \in^{n,m}$ and $B \in^{m,k}$ where $m << \min\{n, k\}$, significantly reducing the number of learnable parameters.

In VLMs, which couple visual encoders with language decoders or multimodal transformers, LoRA enables efficient adaptation to new tasks (e.g., image captioning, visual question answering) without full model retraining. This is particularly valuable in settings with limited computational resources (such as ours). Recent work has shown that integrating LoRAs into both visual and language components allows for effective task adaptation while preserving the generalization capacity of the base model.

3 Methods

3.1 V-JEPA

Garrido et al. [5] evaluate intuitive physics understanding in vision and vision-language models using a Violation-of-Expectation (VoE) paradigm inspired by developmental psychology. Their model, V-JEPA, is trained via self-supervised prediction of masked video representations, and evaluated on its ability to detect physical anomalies in three benchmarks: IntPhys [], GRASP [], and InfLevel-lab [17]. The primary evaluation metric is a learned "surprise" score, computed as the distance between predicted and actual latent representations of video frames. V-JEPA consistently outperforms both pixel-based predictors (e.g., VideoMAEv2 [14, 15]) and multimodal large language models (e.g., Qwen2-VL [16], Gemini 1.5 [13]), which perform near chance on each dataset. Their findings suggest that prediction in an abstract representation space—rather than pixel or text space—is key to emergent physical reasoning, even without explicit supervision or structured priors.

V-JEPA analyzes the effect of pre-training data in their Figure 3b, and finds that, while natural video does supply a significant boost to their physical intuition metrics, pre-training with all types of data still produces the strongest results. We wish to explore if this remains consistent.

[5] purports that the V-JEPA architecture has the emergent ability to detect physical anomalies that violate the intuition garnered through the training process.

We acknowledge the following limitations of the methodology presented in [5]:

- **Modality mismatch in evaluation:** VLMs are trained to produce textual outputs, but the surprise metric used in this work is based on latent representation prediction, a task that aligns with the training objective of V-JEPA but not of VLMs.
- **Prediction Objective:** Unlike V-JEPA, VLMs like Qwen2-VL and Gemini 1.5 are not trained to predict future frames or representations, making the comparison asymmetrical and biased toward models with a predictive objective.
  Forced classification: VLMs are evaluated using forced-choice classification (i.e., "which of these two videos is impossible?"), but this doesn't leverage their reasoning abilities or capacity to provide explanations or uncertainty estimates.
- **Prompt Engineering:** VLM performance can vary significantly depending on prompt phrasing, which isn't standardized in the evaluation—leading to possible underestimation of their true capabilities.

In this work, we aim to address these limitations, and provide fairer analysis and comparison of the capabilities of open-source VLMs to the V-JEPA architecture.

3.2 Vision Language Models

In this work, we evaluate physical anomalies using Vision-Language Models (VLMs), specifically employing video-based models such as LLavVa and integrating low-rank adaptation techniques. We fine-tune these models on the 100M dataset, which serves as our training data, to facilitate the model's understanding of complex visual patterns in temporal sequences, such as video frames, and to associate these patterns with corresponding linguistic concepts.

To optimize the fine-tuning process, we incorporate Low-Rank Adaptation (LoRA), a parameter-efficient technique that allows for the adaptation of large pre-trained models without the need for full retraining. LoRA introduces trainable low-rank matrices into existing model layers, typically within transformer blocks, while keeping the original weights frozen. This significantly reduces the number of learnable parameters, minimizing computational overhead while preserving the model's generalization capabilities.

For training, we specifically select a subset of trainable parameters, ensuring that the amount of training required remains computationally feasible. This selective approach, combined with the video understanding capabilities of LLavVa and the efficiency of LoRA, enables the detection and evaluation of physical anomalies in large-scale datasets, while maintaining resource efficiency and model performance.

3.3 Model Variants

Model	Language	Video	Trainable Params (%)
Video-LLaVA-7B [8, 18]	No	No	N/A
LLaVA-7B-Natural-1	Yes	No	0.15
LLaVA-7B-Natural-2	Yes	Yes	0.30

Table 1: A table summarizing the variants of the VLMs used in the experiments of this report. The language and video columns describe which modules were adapted upon.

3.4 Physical Principles

For the purposes of this study, physical principles are employed to assess whether a video adheres to the laws of physics, ensuring its physical plausibility. This encompasses various principles, such as continuity, shape consistency, gravity, and others. In this work, we primarily focus on evaluating one physical principle at a time, without addressing the complexities arising from the simultaneous violation of multiple physical laws. A more detailed description of the datasets utilized and the specific physical anomalies they are designed to evaluate can be found in §3.6.

3.5 Experimental Evalutation

To more fairly evaluate vision-language models (VLMs) on intuitive physics benchmarks, we propose a four-tiered evaluation framework that aligns with the native capabilities of language-based models. For each task, we apply the base model as well as two LoRA-tuned variants available on Hugging Face to assess the influence of task-specific fine-tuning.

(1) **Binary Decision**

- **Description:** The model is presented with a single video and prompted to answer a binary question regarding its physical plausibility, such as:
- **Sample prompt:** `"Does this video depict a physically plausible event? Answer 'yes' or 'no'."`
- **Output:** A single token ("yes" or "no").
- **Purpose:** This task assesses the model's ability to perform a crisp classification decision and is analogous to the traditional violation-of-expectation binary outcome but framed textually.

(2) **Confidence-Weighted Decision**
- **Description**: Similar to the binary task, the model is prompted for a plausibility judgment. However, the output logits or confidence scores for "yes" and "no" are extracted to compute a soft classification score.
- **Sample prompt** `"Does this video depict a physically plausible event? Answer 'yes' or 'no'."`
- **Output:** A probability distribution over "yes", "no", obtained via logit normalization (e.g., softmax).
- **Purpose:** Captures uncertainty and graded judgments, allowing for finer-grained evaluation beyond hard classification.

(3) **Chain-of-Thought Decision**
- **Description:** The model is prompted to justify its plausibility judgment using free-form language. A smaller LLM then classifies the explanation as indicating a "possible" or "impossible" event.
- **Sample Prompt to Main Model:** `"Explain whether this video shows a physically plausible event and why."`
- **Example Prompt to Classifier Model:** `"Does the following explanation describe a video that violates intuitive physics? Explain your answer."`
- **Output:** Explanation text from the main VLM, followed by binary classification from the proxy LLM.
- **Purpose:** Evaluates reasoning capabilities, causal grounding, and interpretability. The secondary classification allows scaling to large datasets by automating evaluation of free-form outputs.

(4) **Targeted Prompting**
- **Description:** In this variant, the model is prompted with explicit, property-specific questions corresponding to the intuitive physics concept being tested in the video.
- **Sample Prompt:** For example, for a violation of shape permanence, the prompt might be: `"Do any of the objects in the video change shape unexpectedly?"` or for object permanence: `"Does any object disappear without explanation during the video?"`
- **Output:** The model can be evaluated using any of the previous schemes. See previous output samples.
- **Purpose:** Targeted prompting serves two main roles: (1) It reduces ambiguity by focusing the model's attention on the specific property being tested, improving the likelihood of a meaningful response. (2) It allows for fine-grained probing of the model's knowledge across different physical concepts, which is useful for diagnostic evaluation and detailed error analysis.
- **Applicability:** This prompting strategy is modular and can be integrated into any of the previously described evaluation formats (hard/soft/CoT). It is especially useful when the task involves subtle or less visually clear physical violations that might be missed under a general prompt.

Explicitly, we finetune the LoRAs according to the following parameters:

- **Learning rate:** `2e-4`
- **Batch size:** `32`
- **Number of training steps / epochs:** `3 epochs`
- **LoRA rank (r):** `8`
- **LoRA alpha:** `32`
- **LoRA dropout:** `0.05`
- **Optimizer:** `AdamW`
- **Weight decay:** `0.01`
- **Scheduler:** `cosine with warmup`
- **Warmup steps:** `500`
- **Gradient clipping:** `1.0`
- **Precision:** `bfloat16`

Futhermore, we ablate against the chosen frozen weights to see if certain components of the model contribute strongest to physical intuition.

3.6 Data

We make use of the following datasets:

- **Infant-Level Physical Reasoning Benchmark (InfLevel) [17].** An evaluation-only dataset designed to assess the physical reasoning capabilities of AI systems. Inspired by the violation-of-expectations (VoE) paradigm from developmental psychology, InfLevel presents AI models with *pairs* of video clips: one depicting a physically plausible event and the other an implausible. Three core physical principles are tested: Continuity, Solidity, and Gravity. The benchmark aims to probe whether AI systems, like human infants, can form and act on expectations about physical events.
- **Intuitive Physics Benchmark (Int-Phys) [12].** A synthetic video dataset designed to evaluate a model's understanding of intuitive physical principles (object individuation, kinematics, object interactions, etc). Int-Phys presents two pairs of short video clips that differ subtly in whether they conform to basic physical laws. Each model is required to assign a scalar plausibility score to individual clips, reflecting its internal estimation of physical consistency. The evaluation metric quantifies how reliably a model distinguishes possible from impossible events. This benchmark in particular probes three core physical principles: object permanence, shape constancy, and spatiotemporal continuity.
- **HowTo100M [10].** A large-scale, weakly-supervised video dataset consisting of over 100 million narrated instructional video clips collected from YouTube. Each video is paired with automatically extracted speech transcripts, resulting in temporally aligned video-text pairs without manual annotation. The dataset spans a wide variety of human activities, making it well-suited for learning generic video-language representations. Its scale and diversity enable training of models on naturalistic multimodal supervision, supporting

tasks such as cross-modal retrieval, action recognition, and temporal localization.

4 Experiments & Results

In this section, we reproduce the surprise measurement results of [5] on the pre-trained V-JEPA, and present our own results for both the base and the finetuned VLMs. the video parameters that were tuned to pass through the maximum amount of context to the VLMs is as follows in Table 2

Table 2: Frame Acquisiton Information

Dataset	Frame Skip	Total Frame Count
Inflevel	20	15
Intphys	7	15

4.1 V-Jepa

We replicated the results of Garrido et al. [5] to validate V-Jepa's ability to detect physical anomalies. Instead of using their datasets, we tested V-Jepa on a 10M subset of the 100M dataset to evaluate its performance on natural footage. We assumed all videos in the subset were physically plausible and labeled them accordingly.

Garrido et al. showed that ViT-H with ROPE (Rotary Positional Encoding) was most effective at anomaly detection, but as this model was not publicly available, we used the ViT-H model. Unlike typical pairwise input tests, our goal was to assess if the model could recognize surprise in natural videos. The model achieved around 90% accuracy, with minimal variation in performance based on context length, as expected based on it being trained on similar data.

Key insights from this experiment include:

(1) Comparative Analysis: V-Jepa performs a comparative analysis across videos to determine which is more "surprising" (e.g., normal vs. anomalous).
(2) Context Handling in Inflevel: For the Inflevel dataset, V-Jepa effectively identified the boundary between object introduction and anomaly onset, allowing for a better inclusion of context frames.

Based on these observations, we refined our methodology. Due to computational constraints, we focused on individual videos rather than pairwise comparisons, as initial tests indicated negligible performance differences between the two approaches. Furthermore, we adopted the Inflevel context management strategy in subsequent tests to handle context length more effectively.

4.2 Targeted vs General Prompting

To evaluate the Vision-Language Model (VLM), we first conducted an initial test using the Inflevel dataset, following the methodology outlined by Garrido et al. [5]. The goal was to locate and extract frames corresponding to the anomaly of interest. The evaluation prompt was as follows:

- **Prompt:**
  `"Give this video a physical feasibility score from 1 to 10, where 10 means it fully obeys the laws of physics and has no visual continuity errors (like objects disappearing or jumping positions). Respond with only a number."`

The results of this experiment, which evaluated three model types—LLaVA-7B-Natural-2 (Nat-2), LLaVA-7B-Natural-1 (Nat-1), and Video-LLaVA-7B (Base)—across the core principles of Gravity, Solidity, and Continuity, are shown in Figure 2. For these evaluations, distance was calculated by determining the mean difference from the expected value, with a score of 10 for real videos and 1 for anomalous ones.

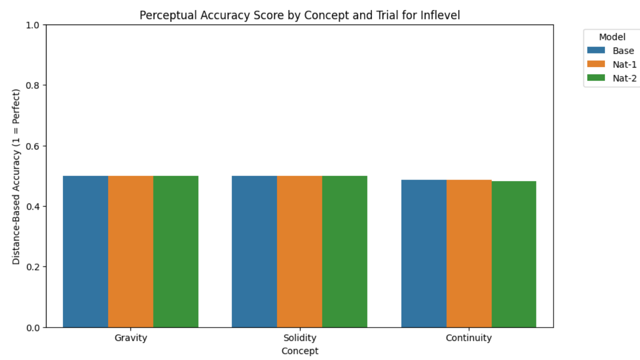

Figure 2: Mean distance-based accuracy scores across three concepts and model types for Inflevel

The results highlighted a few key insights. First, the Inflevel dataset appeared too challenging for our smaller model. When tasked with generating scene explanations, the model often focused on irrelevant subjects, such as the lady in the video, rather than the objects where anomalies were occurring. As a result, we decided to use the IntPhysics dataset for subsequent tests as it is much more focused on the anomalies. Additionally, the model consistently predicted values of '8' or occasionally '6', indicating poor sensitivity to anomalies. Furthermore, there was little to no improvement in performance across training, suggesting the model was not effectively learning. To address these issues, we concluded that a more targeted prompt would help the model focus on the anomaly, improving its ability to detect and classify physical inconsistencies, thereby that was our choice going forward for all future tests.

4.3 Binary Classification

For the binary classification task, the model was presented with a targeted prompt instructing it to respond exclusively with either "Yes" or "No." In this context, a "Yes" response indicated that the specific physical law in question was maintained, while a "No" response indicated that the law was not upheld. A summary of these prompts is provided in Table 3.

A summary of the binary classification results is presented in Figure 3

Upon examining the experimental results, it is evident that the models struggled to differentiate between the various physics concepts, despite the targeted prompts. The responses from the models showed no significant pattern, often resembling random guessing, similar to flipping a coin. Notably, the models tended to answer

Table 3: Physics Concepts and Their Yes/No Prompts

Concept	Prompt
Object Constancy	Do all objects in the video remain consistent and don't disappear or reappear unexpectedly during occlusions or movements?
Shape Consistency	Do all objects in this video maintain their shape?
Temporal Continuity	Do all objects move naturally through time without any teleporting or skipping?

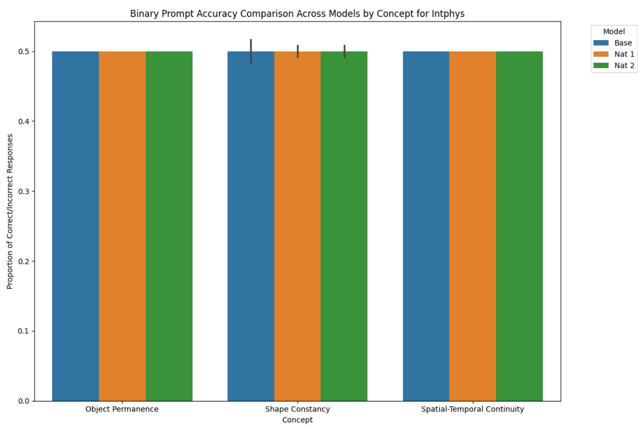

Figure 3: Accuracy for binary classification input across VLM models

"Yes" more frequently, rarely offering a "No" response. This behavior suggests that the model may have exhibited a bias toward answering affirmatively. In a subsequent attempt, the prompt was reversed, making "Yes" the anomaly with the expectation that the model would find it easier to detect a singular error. However, this adjustment did not lead to any improvement in the results.

4.4 Continuous Classification

To further assess whether any meaningful learning occurred from the natural scene training, we employed continuous classification prompting. This approach involved using the prompts outlined in Table 3 and evaluating both the "Yes" and "No" outputs. These scores were subsequently normalized using a softmax function to determine the likelihood of one outcome being selected over the other. The results were then formulated into a log loss metric, comparing the model's outputs to the ground truth. The corresponding results are presented in Figure 4.

A notable observation is the significantly higher log loss associated with the training parameters of our Nat 2 model. This suggests that the selected training parameters may not have been optimal. Additionally, the disparity between the natural video data used for fine-tuning and the simulated video game environment of the evaluation data could contribute to this performance degradation.

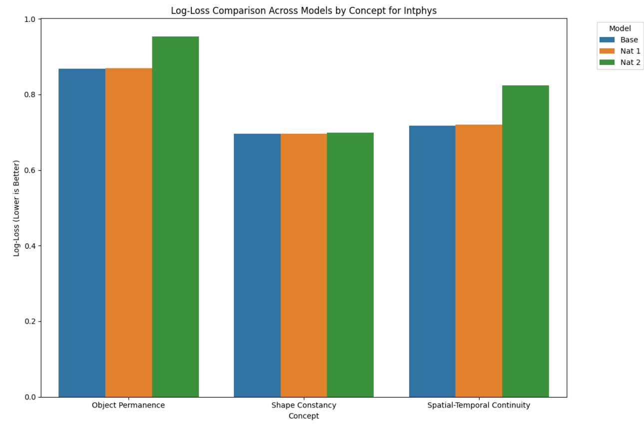

Figure 4: Continuous classification results across VLM models

This potential misalignment may be what is hindering the model's ability to effectively discern the underlying physics in the simulated scenes.

4.5 Chain of Thought

The Chain of Thought (CoT) method has gained significant popularity due to its effectiveness in enhancing large language models' (LLMs) performance by breaking down complex tasks into smaller, more manageable steps. In this study, we sought to leverage this approach to improve our results. Specifically, we adopted a variation of CoT wherein the vision-language model (VLM) was first prompted to provide reasoning for whether an anomaly was upheld or violated. This reasoning was then passed to a smaller LLM, which was tasked with summarizing the conclusion as either "Yes", the video is valid or "No", the video is invalid. The open-ended prompts, as detailed in Table 3, included the additional instruction for the VLM to explain the rationale behind its response rather than simply providing a binary answer. For details on the smaller LLM and the inputs provided, refer to Section §**??**.

The results for this method can be seen below in Figure5 and Figure6

The results of this experiment are notable and reveal several important insights. First, when the model attempts to reason through the logic of whether something is correct or incorrect, it consistently fails to provide accurate responses. Interestingly, training on the "Nat 2" dataset seems to worsen performance across most tasks, with the exception of Shape Consistency, where the "Nat 2" model outperforms others. In contrast, the "Nat 1" model fails to grasp the concept entirely, while the base model produces results that are neither "Yes" or "No" This leads to a second observation: better results could likely have been achieved with a larger LLM, but computational limitations prevented this. The sharp decline in performance for Shape Consistency, apart from model differences, suggests that further tuning of the prompt is necessary, particularly since accuracy fell below 50%. Overall, while this method did not improve results in our study, it remains possible that it could with more powerful models.

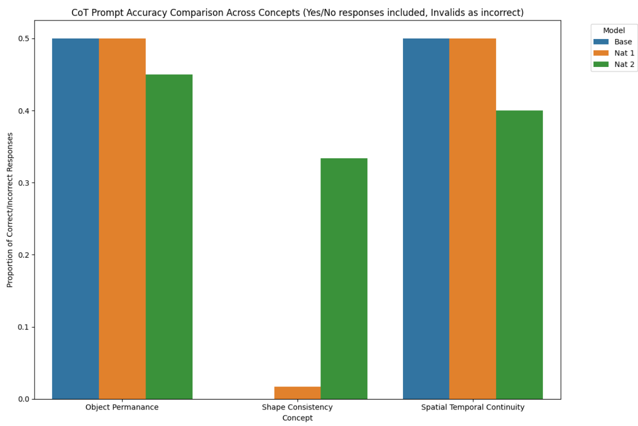

Figure 5: CoT Accuracy across Concept

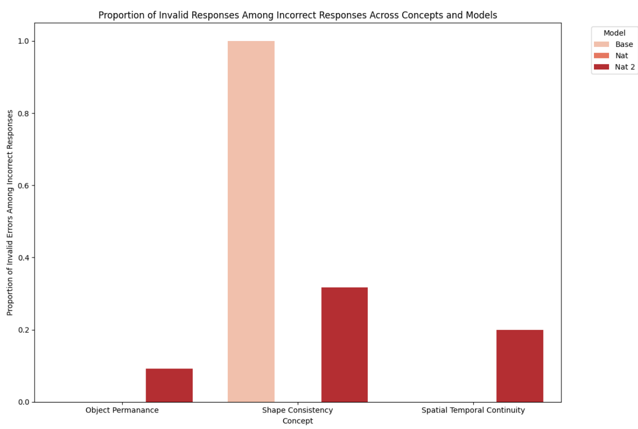

Figure 6: Amount of Invalid Results Amongst Errors

4.6 Compute Resources

The computational resources for this project included NVIDIA T4 and RTX 6000 GPUs, accessed through both the University of Toronto Computer Science department and Vector Institute GPU clusters. These GPU resources were essential for handling the memory-intensive nature of diffusion model training, even with our optimized implementation that focused on reducing VRAM requirements.

Discussion

4.7 Future Work

Our study highlights several promising directions for future research on Vision-Language Models (VLMs) in detecting physical anomalies. Our initial tests showed only minimal improvements, as detailed in §4.8, likely due to several limitations we aim to address.

A key next step is to evaluate larger language models, re-running current experiments to gain a clearer understanding of the model's true performance. Additionally, since only 0.3% of trainable parameters were used, training a larger proportion may improve generalization and anomaly detection in natural videos.

Chain-of-Thought (CoT) prompting has proven effective in many tasks and warrants further exploration in anomaly detection, especially with larger models using a step-based approach. Iterative refinement, where the model self-improves based on feedback, could also enhance performance.

These avenues offer significant potential for advancing VLMs in physical anomaly detection.

4.8 Computational Limitations

This work introduces a set of evaluation protocols aimed at more fairly assessing the intuitive physics capabilities of VLMs, in contrast to existing metrics such as latent-space surprise scores that align more closely with the training objectives of predictive video models like V-JEPA. Our focus is on designing language-native tasks that better leverage the reasoning and interpretability strengths of VLMs.

However, the experiments presented here are subject to significant limitations due to resource constraints. First, we were restricted to using relatively small open-source VLMs (e.g., Video-LLAVA-7B), and our video context was limited to a handful of frames, which was often insufficient for forming coherent visual descriptions. In many cases, the models failed to produce accurate captions or recognize key objects or interactions, making higher-level tasks such as anomaly detection or causal reasoning infeasible.

Second, our fine-tuning was performed using parameter-efficient LoRA adapters, modifying less than 1% of the total model parameters. While this allowed us to explore two LoRA variants under limited compute, the low capacity of these adapters constrained the expressivity and adaptation potential of the models. As a result, both the absolute performance of the models and the performance differences between the base and fine-tuned variants were minimal. These narrow discrepancies limit the conclusions we can draw about the effectiveness of fine-tuning for physics reasoning in VLMs.

Taken together, the results presented should be viewed as exploratory and indicative rather than definitive. To conclude, our primary contributions are methodological: proposing fairer, VLM-aligned evaluation tasks and analyzing their feasibility under constrained conditions.

4.9 Reproducibility Statement

Our code is available in our open-sourced GitHub repository, which was directly used to produce the results presented in this paper. Additionally, each of our specialist models, as well as the cumulative merged model, have been made available on the Hugging Face Hub, and can be found here.

Conclusion

This study represents an initial exploration of the potential of Vision-Language Models (VLMs) in detecting physical anomalies, introducing a novel set of language-native evaluation protocols aimed at assessing their intuitive physics capabilities. While the experiments conducted in this work did not reveal any significant improvements or inherent ability of VLMs to detect physical anomalies, we attribute these findings primarily to computational limitations and the use of smaller models. Nevertheless, the results provide valuable insights into the current limitations and potential avenues for further advancement of VLMs in this domain.

Several key directions for future research have been identified, including the evaluation of larger language models, the expansion of model parameter training, and the further exploration of techniques such as Chain-of-Thought prompting and iterative refinement. These strategies hold the potential to significantly enhance the models' capacity to generalize and detect physical anomalies in complex visual data. Despite the existing limitations, this work contributes to the development of more effective evaluation benchmarks for VLMs, emphasizing tasks that leverage their strengths in reasoning and interpretability.

A Small LLM Information

After evaluating various small language models (LLMs), including several variations of Pythia, it was determined that an LLM specifically fine-tuned for instruction-following would yield the best results. This choice was made to mitigate instances where the model would fail to adhere to the prompt and provide responses other than "Yes" or "No." Consequently, the "gpt2-open-instruct-v1" model was selected. This model is a fine-tuned version of GPT-2 designed to enhance its ability to follow user instructions more accurately. The specific prompts given to this model are outlined in Table 4. Notably, all models received the output of the vision-language model (VLM) as the "statement."

Table 4: LLM Chain of Thought Prompts for Each Physics Concept

Concept	Prompt
Object Constancy	Answer with "yes" or "no" only. If the statement clearly indicates that **an object** does not maintain constancy or remain consistent, answer "no". If the statement suggests that **all objects** remain consistent, answer "yes".
Shape Consistency	Answer with "yes" or "no" only. If the statement clearly indicates that **an object** does not maintain its shape, answer "no". If the statement suggests that **all objects** maintain their shape, answer "yes". Statment:
Temporal Continuity	Answer with "yes" or "no" only. If the statement clearly indicates that **an object** skips or teleports, answer "no". If the statement suggests that **all objects** move without skipping or teleporting, answer "yes".

References

[1] 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. 2022. Flamingo: a Visual Language Model for Few-Shot Learning. arXiv:2204.14198 [cs.CV] https://arxiv.org/abs/2204.14198

[2] Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. 2023. Self-Supervised Learning from Images with a Joint-Embedding Predictive Architecture. arXiv:2301.08243 [cs.CV] https://arxiv.org/abs/2301.08243

[3] Vahid Balazadeh, Mohammadmehdi Ataei, Hyunmin Cheong, Amir Hosein Khasahmadi, and Rahul G. Krishnan. 2024. Synthetic Vision: Training Vision-Language Models to Understand Physics. arXiv:2412.08619 [cs.CV] https://arxiv.org/abs/2412.08619

[4] Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mido Assran, and Nicolas Ballas. 2024. V-JEPA: Latent Video Prediction for Visual Representation Learning. https://openreview.net/forum?id=WFYbBOEOtv

[5] Quentin Garrido, Nicolas Ballas, Mahmoud Assran, Adrien Bardes, Laurent Najman, Michael Rabbat, Emmanuel Dupoux, and Yann LeCun. 2025. Intuitive physics understanding emerges from self-supervised pretraining on natural videos. arXiv:2502.11831 [cs.CV] https://arxiv.org/abs/2502.11831

[6] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. LoRA: Low-Rank Adaptation of Large Language Models. arXiv:2106.09685 [cs.CL] https://arxiv.org/abs/2106.09685

[7] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. arXiv:2201.12086 [cs.CV] https://arxiv.org/abs/2201.12086

[8] Bin Lin, Bin Zhu, Yang Ye, Munan Ning, Peng Jin, and Li Yuan. 2023. Video-LLaVA: Learning United Visual Representation by Alignment Before Projection. *arXiv preprint arXiv:2311.10122* (2023).

[9] Tianyi Liu, Zuxuan Wu, Wenhan Xiong, Jingjing Chen, and Yu-Gang Jiang. 2021. Unified Multimodal Pre-training and Prompt-based Tuning for Vision-Language Understanding and Generation. arXiv:2112.05587 [cs.CV] https://arxiv.org/abs/2112.05587

[10] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. 2019. HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips. arXiv:1906.03327 [cs.CV] https://arxiv.org/abs/1906.03327

[11] 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. 2021. Learning Transferable Visual Models From Natural Language Supervision. arXiv:2103.00020 [cs.CV] https://arxiv.org/abs/2103.00020

[12] Ronan Riochet, Mario Ynocente Castro, Mathieu Bernard, Adam Lerer, Rob Fergus, Véronique Izard, and Emmanuel Dupoux. 2020. IntPhys: A Framework and Benchmark for Visual Intuitive Physics Reasoning. arXiv:1803.07616 [cs.AI] https://arxiv.org/abs/1803.07616

[13] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, Soroosh Mariooryad, Yifan Ding, Xinyang Geng, Fred Alcober, Roy Frostig, Mark Omernick, Lexi Walker, Cosmin Paduraru, Christina Sorokin, Andrea Tacchetti, Colin Gaffney, Samira Daruki, Olcan Sercinoglu, Zach Gleicher, Juliette Love, Paul Voigtlaender, Rohan Jain, Gabriela Surita, Kareem Mohamed, Rory Blevins, Junwhan Ahn, Tao Zhu, Kornraphop Kawintiranon, Orhan Firat, Yiming Gu, Yujing Zhang, Matthew Rahtz, Manaal Faruqui, Natalie Clay, Justin Gilmer, JD Co-Reyes, Ivo Penchev, Rui Zhu, Nobuyuki Morioka, Kevin Hui, Krishna Haridasan, Victor Campos, Mahdis Mahdieh, Mandy Guo, Samer Hassan, Kevin Kilgour, Arpi Vezer, Heng-Tze Cheng, Raoul de Liedekerke, Siddharth Goyal, Paul Barham, DJ Strouse, Seb Noury, Jonas Adler, Mukund Sundararajan, Sharad Vikram, Dmitry Lepikhin, Michela Paganini, Xavier Garcia, Fan Yang, Dasha Valter, Maja Trebacz, Kiran Vodrahalli, Chulayuth Asawaroengchai, Roman Ring, Norbert Kalb, Livio Baldini Soares, Siddhartha Brahma, David Steiner, Tianhe Yu, Fabian Mentzer, Antoine He, Lucas Gonzalez, Bibo Xu, Raphael Lopez Kaufman, Laurent El Shafey, Junhyuk Oh, Tom Hennigan, George van den Driessche, Seth Odoom, Mario Lucic, Becca Roelofs, Sid Lall, Amit Marathe, Betty Chan, Santiago Ontanon, Luheng He, Denis Teplyashin, Jonathan Lai, Phil Crone, Bogdan Damoc, Lewis Ho, Sebastian Riedel, Karel Lenc, Chih-Kuan Yeh, Aakanksha Chowdhery, Yang Xu, Mehran Kazemi, Ehsan Amid, Anastasia Petrushkina, Kevin Swersky, Ali Khodaei, Gowoon Chen, Chris Larkin, Mario Pinto, Geng Yan, Adria Puigdomenech Badia, Piyush Patil, Steven Hansen, Dave Orr, Sebastien M. R. Arnold, Jordan Grimstad, Andrew Dai, Sholto Douglas, Rishika Sinha, Vikas Yadav, Xi Chen, Elena Gribovskaya, Jacob Austin, Jeffrey Zhao, Kaushal Patel, Paul Komarek, Sophia Austin, Sebastian Borgeaud, Linda Friso, Abhimanyu Goyal, Ben Caine, Kris Cao, Da-Woon Chung, Matthew Lamm,

Gabe Barth-Maron, Thais Kagohara, Kate Olszewska, Mia Chen, Kaushik Shivakumar, Rishabh Agarwal, Harshal Godhia, Ravi Rajwar, Javier Snaider, Xerxes Dotiwalla, Yuan Liu, Aditya Barua, Victor Ungureanu, Yuan Zhang, Bat-Orgil Batsaikhan, Mateo Wirth, James Qin, Ivo Danihelka, Tulsee Doshi, Martin Chadwick, Jilin Chen, Sanil Jain, Quoc Le, Arjun Kar, Madhu Gurumurthy, Cheng Li, Ruoxin Sang, Fangyu Liu, Lampros Lamprou, Rich Munoz, Nathan Lintz, Harsh Mehta, Heidi Howard, Malcolm Reynolds, Lora Aroyo, Quan Wang, Lorenzo Blanco, Albin Cassirer, Jordan Griffith, Dipanjan Das, Stephan Lee, Jakub Sygnowski, Zach Fisher, James Besley, Richard Powell, Zafarali Ahmed, Dominik Paulus, David Reitter, Zalan Borsos, Rishabh Joshi, Aedan Pope, Steven Hand, Vittorio Selo, Vihan Jain, Nikhil Sethi, Megha Goel, Takaki Makino, Rhys May, Zhen Yang, Johan Schalkwyk, Christina Butterfield, Anja Hauth, Alex Goldin, Will Hawkins, Evan Senter, Sergey Brin, Oliver Woodman, Marvin Ritter, Eric Noland, Minh Giang, Vijay Bolina, Lisa Lee, Tim Blyth, Ian Mackinnon, Machel Reid, Obaid Sarvana, David Silver, Alexander Chen, Lily Wang, Loren Maggiore, Oscar Chang, Nithya Attaluri, Gregory Thornton, Chung-Cheng Chiu, Oskar Bunyan, Nir Levine, Timothy Chung, Evgenii Eltyshev, Xiance Si, Timothy Lillicrap, Demetra Brady, Vaibhav Aggarwal, Boxi Wu, Yuanzhong Xu, Ross McIlroy, Kartikeya Badola, Paramjit Sandhu, Erica Moreira, Wojciech Stokowiec, Ross Hemsley, Dong Li, Alex Tudor, Pranav Shyam, Elahe Rahimtoroghi, Salem Haykal, Pablo Sprechmann, Xiang Zhou, Diana Mincu, Yujia Li, Ravi Addanki, Kalpesh Krishna, Xiao Wu, Alexandre Frechette, Matan Eyal, Allan Dafoe, Dave Lacey, Jay Whang, Thi Avrahami, Ye Zhang, Emanuel Taropa, Hanzhao Lin, Daniel Toyama, Eliza Rutherford, Motoki Sano, HyunJeong Choe, Alex Tomala, Chalence Safranek-Shrader, Nora Kassner, Mantas Pajarskas, Matt Harvey, Sean Sechrist, Meire Fortunato, Christina Lyu, Gamaleldin Elsayed, Chenkai Kuang, James Lottes, Eric Chu, Chao Jia, Chih-Wei Chen, Peter Humphreys, Kate Baumli, Connie Tao, Rajkumar Samuel, Cicero Nogueira dos Santos, Anders Andreassen, Nemanja Rakićević, Dominik Grewe, Aviral Kumar, Stephanie Winkler, Jonathan Caton, Andrew Brock, Sid Dalmia, Hannah Sheahan, Iain Barr, Yingjie Miao, Paul Natsev, Jacob Devlin, Feryal Behbahani, Flavien Prost, Yanhua Sun, Artiom Myaskovsky, Thanumalayan Sankaranarayana Pillai, Dan Hurt, Angeliki Lazaridou, Xi Xiong, Ce Zheng, Fabio Pardo, Xiaowei Li, Dan Horgan, Joe Stanton, Moran Ambar, Fei Xia, Alejandro Lince, Mingqiu Wang, Basil Mustafa, Albert Webson, Hyo Lee, Rohan Anil, Martin Wicke, Timothy Dozat, Abhishek Sinha, Enrique Piqueras, Elahe Dabir, Shyam Upadhyay, Anudhyan Boral, Lisa Anne Hendricks, Corey Fry, Josip Djolonga, Yi Su, Jake Walker, Jane Labanowski, Ronny Huang, Vedant Misra, Jeremy Chen, RJ Skerry-Ryan, Avi Singh, Shruti Rijhwani, Dian Yu, Alex Castro-Ros, Beer Changpinyo, Romina Datta, Sumit Bagri, Arnar Mar Hrafnkelsson, Marcello Maggioni, Daniel Zheng, Yury Sulsky, Shaobo Hou, Tom Le Paine, Antoine Yang, Jason Riesa, Dominika Rogozinska, Dror Marcus, Dalia El Badawy, Qiao Zhang, Luyu Wang, Helen Miller, Jeremy Greer, Lars Lowe Sjos, Azade Nova, Heiga Zen, Rahma Chaabouni, Mihaela Rosca, Jiepu Jiang, Charlie Chen, Ruibo Liu, Tara Sainath, Maxim Krikun, Alex Polozov, Jean-Baptiste Lespiau, Josh Newlan, Zeyncep Cankara, Soo Kwak, Yunhan Xu, Phil Chen, Andy Coenen, Clemens Meyer, Katerina Tsihlas, Ada Ma, Juraj Gottweis, Jinwei Xing, Chenjie Gu, Jin Miao, Christian Frank, Zeynep Cankara, Sanjay Ganapathy, Ishita Dasgupta, Steph Hughes-Fitt, Heng Chen, David Reid, Keran Rong, Hongmin Fan, Joost van Amersfoort, Vincent Zhuang, Aaron Cohen, Shixiang Shane Gu, Anhad Mohananey, Anastasija Ilic, Taylor Tobin, John Wieting, Anna Bortsova, Phoebe Thacker, Emma Wang, Emily Caveness, Justin Chiu, Eren Sezener, Alex Kaskasoli, Steven Baker, Katie Millican, Mohamed Elhawaty, Kostas Aisopos, Carl Lebsack, Nathan Byrd, Hanjun Dai, Wenhao Jia, Matthew Wiethoff, Elnaz Davoodi, Albert Weston, Lakshman Yagati, Arun Ahuja, Isabel Gao, Golan Pundak, Susan Zhang, Michael Azzam, Khe Chai Sim, Sergi Caelles, James Keeling, Abhanshu Sharma, Andy Swing, YaGuang Li, Chenxi Liu, Carrie Grimes Bostock, Yamini Bansal, Zachary Nado, Ankesh Anand, Josh Lipschultz, Abhijit Karmarkar, Lev Proleev, Abe Ittycheriah, Soheil Hassas Yeganeh, George Polovets, Aleksandra Faust, Jiao Sun, Alban Rrustemi, Pen Li, Rakesh Shivanna, Jeremiah Liu, Chris Welty, Federico Lebron, Anirudh Baddepudi, Sebastian Krause, Emilio Parisotto, Radu Soricut, Zheng Xu, Dawn Bloxwich, Melvin Johnson, Behnam Neyshabur, Justin Mao-Jones, Renshen Wang, Vinay Ramasesh, Zaheer Abbas, Arthur Guez, Constant Segal, Duc Dung Nguyen, James Svensson, Le Hou, Sarah York, Kieran Milan, Sophie Bridgers, Wiktor Gworek, Marco Tagliasacchi, James Lee-Thorp, Michael Chang, Alexey Guseynov, Ale Jakse Hartman, Michael Kwong, Ruizhe Zhao, Sheleem Kashem, Elizabeth Cole, Antoine Miech, Richard Tanburn, Mary Phuong, Filip Pavetic, Sebastien Cevey, Ramona Comanescu, Richard Ives, Sherry Yang, Cosmo Du, Bo Li, Zizhao Zhang, Mariko Iinuma, Clara Huiyi Hu, Aurko Roy, Shaan Bijwadia, Zhenkai Zhu, Danilo Martins, Rachel Saputro, Anita Gergely, Steven Zheng, Dawei Jia, Ioannis Antonoglou, Adam Sadovsky, Shane Gu, Yingying Bi, Alek Andreev, Sina Samangooei, Mina Khan, Tomas Kocisky, Angelos Filos, Chintu Kumar, Colton Bishop, Adams Yu, Sarah Hodkinson, Sid Mittal, Premal Shah, Alexandre Moufarek, Yong Cheng, Adam Bloniarz, Jaehoon Lee, Pedram Pejman, Paul Michel, Stephen Spencer, Vladimir Feinberg, Xuehan Xiong, Nikolay Savinov, Charlotte Smith, Siamak Shakeri, Dustin Tran, Mary Chesus, Bernd Bohnet, George Tucker, Tamara von Glehn, Carrie Muir, Yiran Mao, Hideto Kazawa, Ambrose Slone, Kedar Soparkar, Disha Shrivastava, James Cobon-Kerr, Michael Sharman, Jay Pavagadhi, Carlos Araya, Karolis Misiunas, Nimesh Ghelani, Michael Laskin, David Barker, Qiujia Li, Anton Briukhov, Neil Houlsby, Mia Glaese, Balaji Lakshminarayanan, Nathan Schucher, Yunhao Tang, Eli Collins, Hyeontaek Lim, Fangxiaoyu Feng, Adria Recasens, Guangda Lai, Alberto Magni, Nicola De Cao, Aditya Siddhant, Zoe Ashwood, Jordi Orbay, Mostafa Dehghani, Jenny Brennan, Yifan He, Kelvin Xu, Yang Gao, Carl Saroufim, James Molloy, Xinyi Wu, Seb Arnold, Solomon Chang, Julian Schrittwieser, Elena Buchatskaya, Soroush Radpour, Martin Polacek, Skye Giordano, Ankur Bapna, Simon Tokumine, Vincent Hellendoorn, Thibault Sottiaux, Sarah Cogan, Aliaksei Severyn, Mohammad Saleh, Shantanu Thakoor, Laurent Shefey, Siyuan Qiao, Meenu Gaba, Shuo yiin Chang, Craig Swanson, Biao Zhang, Benjamin Lee, Paul Kishan Rubenstein, Gan Song, Tom Kwiatkowski, Anna Koop, Ajay Kannan, David Kao, Parker Schuh, Axel Stjerngren, Golnaz Ghiasi, Gena Gibson, Luke Vilnis, Ye Yuan, Felipe Tiengo Ferreira, Aishwarya Kamath, Ted Klimenko, Ken Franko, Kefan Xiao, Indro Bhattacharya, Miteyan Patel, Rui Wang, Alex Morris, Robin Strudel, Vivek Sharma, Peter Choy, Sayed Hadi Hashemi, Jessica Landon, Mara Finkelstein, Priya Jhakra, Justin Frye, Megan Barnes, Matthew Mauger, Dennis Daun, Khuslen Baatarsukh, Matthew Tung, Wael Farhan, Henryk Michalewski, Fabio Viola, Felix de Chaumont Quitry, Charline Le Lan, Tom Hudson, Qingze Wang, Felix Fischer, Ivy Zheng, Elspeth White, Anca Dragan, Jean baptiste Alayrac, Eric Ni, Alexander Pritzel, Adam Iwanicki, Michael Isard, Anna Bulanova, Lukas Zilka, Ethan Dyer, Devendra Sachan, Srivatsan Srinivasan, Hannah Muckenhirn, Honglong Cai, Amol Mandhane, Mukarram Tariq, Jack W. Rae, Gary Wang, Kareem Ayoub, Nicholas FitzGerald, Yao Zhao, Woohyun Han, Chris Alberti, Dan Garrette, Kashyap Krishnakumar, Mai Gimenez, Anselm Levskaya, Daniel Sohn, Josip Matak, Inaki Iturrate, Michael B. Chang, Jackie Xiang, Yuan Cao, Nishant Ranka, Geoff Brown, Adrian Hutter, Vahab Mirrokni, Nanxin Chen, Kaisheng Yao, Zoltan Egyed, Francois Galilee, Tyler Liechty, Praveen Kallakuri, Evan Palmer, Sanjay Ghemawat, Jasmine Liu, David Tao, Chloe Thornton, Tim Green, Mimi Jasarevic, Sharon Lin, Victor Cotruta, Yi-Xuan Tan, Noah Fiedel, Hongkun Yu, Ed Chi, Alexander Neitz, Jens Heitkaemper, Anu Sinha, Denny Zhou, Yi Sun, Charbel Kaed, Brice Hulse, Swaroop Mishra, Maria Georgaki, Sneha Kudugunta, Clement Farabet, Izhak Shafran, Daniel Vlasic, Anton Tsitsulin, Rajagopal Ananthanarayanan, Alen Carin, Guolong Su, Pei Sun, Shashank V, Gabriel Carvajal, Josef Broder, Iulia Comsa, Alena Repina, William Wong, Warren Weilun Chen, Peter Hawkins, Egor Filonov, Lucia Loher, Christoph Hirnschall, Weiyi Wang, Jingchen Ye, Andrea Burns, Hardie Cate, Diana Gage Wright, Federico Piccinini, Lei Zhang, Chu-Cheng Lin, Ionel Gog, Yana Kulizhskaya, Ashwin Sreevatsa, Shuang Song, Luis C. Cobo, Anand Iyer, Chetan Tekur, Guillermo Garrido, Zhuyun Xiao, Rupert Kemp, Huaixiu Steven Zheng, Hui Li, Ananth Agarwal, Christel Ngani, Kati Goshvadi, Rebeca Santamaria-Fernandez, Wojciech Fica, Xinyun Chen, Chris Gorgolewski, Sean Sun, Roopal Garg, Xinyu Ye, S. M. Ali Eslami, Nan Hua, Jon Simon, Pratik Joshi, Yelin Kim, Ian Tenney, Sahitya Potluri, Lam Nguyen Thiet, Quan Yuan, Florian Luisier, Alexandra Chronopoulou, Salvatore Scellato, Praveen Srinivasan, Minmin Chen, Vinod Koverkathu, Valentin Dalibard, Yaming Xu, Brennan Saeta, Keith Anderson, Thibault Sellam, Nick Fernando, Fantine Huot, Junehyuk Jung, Mani Varadarajan, Michael Quinn, Amit Raul, Maigo Le, Ruslan Habalov, Jon Clark, Komal Jalan, Kalesha Bullard, Achintya Singhal, Thang Luong, Boyu Wang, Sujeevan Rajayogam, Julian Eisenschlos, Johnson Jia, Daniel Finchelstein, Alex Yakubovich, Daniel Balle, Michael Fink, Sameer Agarwal, Jing Li, Dj Dvijotham, Shalini Pal, Kai Kang, Jaclyn Konzelmann, Jennifer Beattie, Olivier Dousse, Diane Wu, Remi Crocker, Chen Elkind, Siddhartha Reddy Jonnalagadda, Jong Lee, Dan Holtmann-Rice, Krystal Kallarackal, Rosanne Liu, Denis Vnukov, Neera Vats, Luca Invernizzi, Mohsen Jafari, Huanjie Zhou, Lilly Taylor, Jennifer Prendki, Marcus Wu, Tom Eccles, Tianqi Liu, Kavya Kopparapu, Francoise Beaufays, Christof Angermueller, Andreea Marzoca, Shourya Sarcar, Hilal Dib, Jeff Stanway, Frank Perbet, Nejc Trdin, Rachel Sterneck, Andrey Khorlin, Dinghua Li, Xihui Wu, Sonam Goenka, David Madras, Sasha Goldshtein, Willi Gierke, Tong Zhou, Yaxin Liu, Yannie Liang, Anais White, Yunjie Li, Shreya Singh, Sanaz Bahargam, Mark Epstein, Sujoy Basu, Li Lao, Adnan Ozturel, Carl Crous, Alex Zhai, Han Lu, Zora Tung, Neeraj Gaur, Alanna Walton, Lucas Dixon, Ming Zhang, Amir Globerson, Grant Uy, Andrew Bolt, Olivia Wiles, Milad Nasr, Ilia Shumailov, Marco Selvi, Francesco Piccinno, Ricardo Aguilar, Sara McCarthy, Misha Khalman, Mrinal Shukla, Vlado Galic, John Carpenter, Kevin Villela, Haibin Zhang, Harry Richardson, James Martens, Matko Bosnjak, Shreyas Rammohan Belle, Jeff Seibert, Mahmoud Alnahlawi, Brian McWilliams, Sankalp Singh, Annie Louis, Wen Ding, Dan Popovici, Lenin Simicich, Laura Knight, Pulkit Mehta, Nishesh Gupta, Chongyang Shi, Saaber Fatehi, Jovana Mitrovic, Alex Grills, Joseph Pagadora, Tsendsuren Munkhdalai, Dessie Petrova, Danielle Eisenbud, Zhishuai Zhang, Damion Yates, Bhavishya Mittal, Nilesh Tripuraneni, Yannis Assael, Thomas Brovelli, Prateek Jain, Mihajlo Velimirovic, Canfer Akbulut, Jiaqi Mu, Wolfgang Macherey, Ravin Kumar, Jun Xu, Haroon Qureshi, Gheorghe Comanici, Jeremy Wiesner, Zhitao Gong, Anton Ruddock, Matthias Bauer, Nick Felt, Anirudh GP, Anurag Arnab, Dustin Zelle, Jonas Rothfuss, Bill Rosgen, Ashish Shenoy, Bryan Seybold, Xinjian Li, Jayaram Mudigonda, Goker Erdogan, Jiawei Xia, Jiri Simsa, Andrea Michi, Yi Yao, Christopher Yew, Steven Kan, Isaac Caswell, Carey Radebaugh, Andre Elisseeff, Pedro Valenzuela, Kay McKinney, Kim Paterson, Albert Cui, Eri Latorre-Chimoto, Solomon Kim, William Zeng, Ken Durden, Priya Ponnapalli, Tiberiu Sosea, Christopher A. Choquette-Choo, James Manyika, Brona Robenek, Harsha Vashisht, Sebastien Pereira, Hoi Lam, Marko Velic, Denese Owusu-Afriyie, Katherine Lee, Tolga Bolukbasi, Alicia Parrish, Shawn Lu, Jane Park, Balaji Venkatraman, Alice Talbert, Lambert Rosique, Yuchung Cheng, Andrei Sozanschi, Adam Paszke, Praveen Kumar, Jessica Austin, Lu Li, Khalid Salama, Bartek Perz, Wooyeol Kim, Nandita Dukkipati, Anthony Baryshnikov, Christos Kaplanis, XiangHai Sheng, Yuri Chervonyi, Caglar Unlu, Diego de Las Casas, Harry Askham, Kathryn Tunyasuvunakool, Felix Gimeno, Siim Poder, Chester Kwak, Matt Miecnikowski, Vahab Mirrokni, Alek Dimitriev, Aaron Parisi, Dangyi Liu, Tomy Tsai, Toby Shevlane, Christina Kouridi, Drew Garmon, Adrian Goedeckemeyer, Adam R. Brown, Anitha Vijayakumar, Ali Elqursh, Sadegh Jazayeri, Jin Huang, Sara Mc Carthy, Jay Hoover, Lucy Kim, Sandeep Kumar, Wei Chen, Courtney Biles, Garrett Bingham, Evan Rosen, Lisa Wang, Qijun Tan, David Engel, Francesco Pongetti, Dario de Cesare, Dongseong Hwang, Lily Yu, Jennifer Pullman, Srini Narayanan, Kyle Levin, Siddharth Gopal, Megan Li, Asaf Aharoni, Trieu Trinh, Jessica Lo, Norman Casagrande, Roopali Vij, Loic Matthey, Bramandia Ramadhana, Austin Matthews, CJ Carey, Matthew Johnson, Kremena Goranova, Rohin Shah, Shereen Ashraf, Kingshuk Dasgupta, Rasmus Larsen, Yicheng Wang, Manish Reddy Vuyyuru, Chong Jiang, Joana Ijazi, Kazuki Osawa, Celine Smith, Ramya Sree Boppana, Taylan Bilal, Yuma Koizumi, Ying Xu, Yasemin Altun, Nir Shabat, Ben Bariach, Alex Korchemniy, Kiam Choo, Olaf Ronneberger, Chimezie Iwuanyanwu, Shubin Zhao, David Soergel, Cho-Jui Hsieh, Irene Cai, Shariq Iqbal, Martin Sundermeyer, Zhe Chen, Elie Bursztein, Chaitanya Malaviya, Fadi Biadsy, Prakash Shroff, Inderjit Dhillon, Tejasi Latkar, Chris Dyer, Hannah Forbes, Massimo Nicosia, Vitaly Nikolaev, Somer Greene, Marin Georgiev, Pidong Wang, Nina Martin, Hanie Sedghi, John Zhang, Praseem Banzal, Doug Fritz, Vikram Rao, Xuezhi Wang, Jiageng Zhang, Viorica Patraucean, Dayou Du, Igor Mordatch, Ivan Jurin, Lewis Liu, Ayush Dubey, Abhi Mohan, Janek Nowakowski, Vlad-Doru Ion, Nan Wei, Reiko Tojo, Maria Abi Raad, Drew A. Hudson, Vaishakh Keshava, Shubham Agrawal, Kevin Ramirez, Zhichun Wu, Hoang Nguyen, Ji Liu, Madhavi Sewak, Bryce Petrini, DongHyun Choi, Ivan Philips, Ziyue Wang, Ioana Bica, Ankush Garg, Jarek Wilkiewicz, Priyanka Agrawal, Xiaowei Li, Danhao Guo, Emily Xue, Naseer Shaik, Andrew Leach, Sadh MNM Khan, Julia Wiesinger, Sammy Jerome, Abhishek Chakladar, Alek Wenjiao Wang, Tina Ornduff, Folake Abu, Alireza Ghaffarkhah, Marcus Wainwright, Mario Cortes, Frederick Liu, Joshua Maynez, Andreas Terzis, Pouya Samangouei, Riham Mansour, Tomasz Kępa, François-Xavier Aubet, Anton Algymr, Dan Banica, Agoston Weisz, Andras Orban, Alexandre Senges, Ewa Andrejczuk, Mark Geller, Niccolo Dal Santo, Valentin Anklin, Majd Al Merey, Martin Baeuml, Trevor Strohman, Junwen Bai, Slav Petrov, Yonghui Wu, Demis Hassabis, Koray Kavukcuoglu, Jeff Dean, and Oriol Vinyals. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. arXiv:2403.05530 [cs.CL] https://arxiv.org/abs/2403.05530

[14] Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. 2023. VideoMAE V2: Scaling Video Masked Autoencoders with Dual Masking. arXiv:2303.16727 [cs.CV]

[15] Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. 2023. VideoMAE V2: Scaling Video Masked Autoencoders With Dual Masking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 14549–14560.

[16] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024. Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution. arXiv:2409.12191 [cs.CV] https://arxiv.org/abs/2409.12191

[17] Luca Weihs, Amanda Rose Yuile, Renée Baillargeon, Cynthia Fisher, Gary Marcus, Roozbeh Mottaghi, and Aniruddha Kembhavi. 2022. Benchmarking Progress to Infant-Level Physical Reasoning in AI. *TMLR* (2022).

[18] Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, HongFa Wang, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, et al. 2023. LanguageBind: Extending Video-Language Pretraining to N-modality by Language-based Semantic Alignment. *arXiv preprint arXiv:2310.01852* (2023).

[19] Xiangming Zhu, Huayu Deng, Haochen Yuan, Yunbo Wang, and Xiaokang Yang. 2024. Latent Intuitive Physics: Learning to Transfer Hidden Physics from A 3D Video. arXiv:2406.12769 [cs.AI] https://arxiv.org/abs/2406.12769