

Physics Context Builders: A Modular Framework for Physical Reasoning in Vision-Language Models

Vahid Balazadeh*
 University of Toronto
 vahid@cs.toronto.edu

Mohammadmehdi Ataei
 Autodesk Research
 mehdi.ataei@autodesk.com

Hyunmin Cheong
 Autodesk Research
 hyunmin.cheong@autodesk.com

Amir Hosein Khasahmadi
 Autodesk Research
 amir.khasahmadi@autodesk.com

Rahul G. Krishnan
 University of Toronto
 rahulgk@cs.toronto.edu

Abstract

Physical reasoning remains a significant challenge for Vision-Language Models (VLMs). This limitation arises from an inability to translate learned knowledge into predictions about physical behavior. Although continual fine-tuning can mitigate this issue, it is expensive for large models and impractical to perform repeatedly for every task. This necessitates the creation of modular and scalable ways to teach VLMs about physical reasoning. To that end, we introduce Physics Context Builders (PCBs), a modular framework where specialized smaller VLMs are fine-tuned to generate detailed physical scene descriptions. These can be used as physical contexts to enhance the reasoning capabilities of larger VLMs. PCBs enable the separation of visual perception from reasoning, allowing us to analyze their relative contributions to physical understanding. We perform experiments on CLEVRER and on Falling Tower, a stability detection dataset with both simulated and real-world scenes, to demonstrate that PCBs provide substantial performance improvements, increasing average accuracy by up to 13.8% on complex physical reasoning tasks. Notably, PCBs also show strong Sim2Real transfer, successfully generalizing from simulated training data to real-world scenes.

1. Introduction

Physical reasoning is a fundamental component of human intelligence, enabling interpretation of complex interactions, prediction of future events, and understanding of causal relationships in real-world environments [22]. For humans, the ability to understand the physical world is developed early and operates intuitively [3, 30, 38]. However, phys-

ical reasoning remains a significant challenge for artificial intelligence (AI) systems [10, 23, 35, 45], despite advances in computer vision and Vision-Language Models (VLMs) [2, 25, 26, 32, 36, 39, 46].

VLMs are remarkably successful on many predictive problems, spanning a broad range of tasks. This makes their use central to real-world applications requiring expertise in a variety of tasks, such as robotics and embodied AI [17, 28]. However, current VLMs consistently fail at physical reasoning tasks, struggling with basic spatial relationships (e.g., object positioning, counting) [33], object attributes [40], and physical interactions (e.g., stability assessment, dynamics prediction) [10, 14]. While humans leverage causal or physically guided knowledge for physical understanding of the world [5], the mechanisms by which VLMs make predictions are not well understood. This leads to two important questions: *what factors contribute to the lack of physical understanding in VLMs, and how can we improve it?*

One potential explanation for such limitations lies in VLMs' training data. VLMs are created by fusing the representations of image and text encoders using datasets like MSCOCO [27] and Conceptual Captions [34], which focus on general scene descriptions but lack annotations of physical relationships. Our experiments with standard benchmarks like CLEVRER [45] confirm this hypothesis, demonstrating that fine-tuning on physics-focused data can enable a relatively small VLM to approach state-of-the-art results achieved by specialized architectures [11, 45]. However, as the ecosystem of open and closed source models grows, repeatedly fine-tuning each VLM for every regime of physical concepts it cannot reason about becomes impractical; there is a need for practical, performant and modular tools to augment VLM capabilities.

To this end, we introduce Physics Context Builders (PCBs): specialized VLMs that are fine-tuned on simula-

*This project was completed during an internship at Autodesk Research.

tion data to generate fine-grained physical descriptions that can be used by larger VLMs. Simulation can provide precise annotations of controllable physical interactions of interest, which can serve as scalable training data to teach models about physical phenomena. We show how to embed the knowledge of physical concepts into PCBs and how they can transfer this knowledge to larger VLMs; this enables us to effectively separate visual perception from reasoning, using PCBs as perception modules, while leaving reasoning to larger models. See Fig. 1 for a demonstration of PCBs.

PCBs offer a practical and modular approach to enhance the physical reasoning of existing large-scale VLMs without requiring expensive or infeasible modification of the larger foundation models. Our experiments on standard benchmarks demonstrate effective integration of PCBs with standard commercial VLMs. Moreover, they show that PCBs trained with simulation data successfully generalize to real-world scenarios, effectively performing simulation-to-reality (Sim2Real) transfer. In summary, our contributions are:

1. We introduce a modular framework where specialized VLMs are fine-tuned to generate detailed physical scene descriptions, enhancing the physical reasoning capabilities of foundation models without their modification.
2. We demonstrate the effectiveness of separating visual perception from reasoning through our PCB approach, providing insights into how each component contributes to physical understanding.
3. We show how simulation can be leveraged to train specialized modules that transfer successfully to real-world scenarios, avoiding expensive simulations at inference.

2. Related Work

2.1. Physical Reasoning in Vision Models

Physical reasoning presents significant challenges for vision models, including both specialized architectures and large-scale VLMs. Recent benchmarks demonstrate that current VLMs struggle with basic physical understanding, often failing at tasks such as counting, depth reasoning, and physical interaction prediction [14, 26, 39]. These limitations are not unique to VLMs; physical reasoning has also proven difficult for single-purpose models, necessitating specialized architectures such as physics-inspired predictive models [4, 13, 16], neural-symbolic executors [9, 20, 48], differentiable physics engines [12, 42], simulation-in-the-loop approaches [29, 49], and task-specific architectures [11]. However, large VLMs are not easily amenable to architectural modifications and require alternative approaches to enhance their capabilities, which we consider in this work.

2.2. Evaluating Physical Reasoning in VLMs

Several recent works have evaluated physical reasoning capabilities in vision-language models. Nagar et al. [31] bench-

mark zero-shot visual reasoning in both large language models (LLMs) and VLMs. They find that underlying LLMs, when provided with textual scene descriptions, consistently outperform VLMs that use visual embeddings. Their analysis shows that this performance gap is due to VLMs’ difficulty in translating visual information into accurate representations for reasoning. Our work builds on this insight by developing a modular approach to bridge this gap without requiring extensive retraining of VLMs.

Concurrently, Chow et al. [10] introduce PhysBench, a comprehensive benchmark for evaluating physical understanding in VLMs. Their work reveals that VLMs’ physical reasoning does not scale proportionally with model size, training data, or input frame count. They identify perceptual and knowledge gaps as the primary sources of errors and propose using vision foundation models like Depth Anything [44] and SAM [21] to enhance visual perception. While our approach shares the goal of improving physical reasoning, we take a different direction by leveraging simulation data to train specialized context builders rather than relying on generic vision foundation models.

2.3. Simulation for Physical Understanding

Simulation has long been recognized as a valuable tool for teaching machines about physical dynamics [5]. Previous approaches have incorporated simulation directly into the inference pipeline [29, 42, 43], requiring computationally expensive simulators at inference time. In contrast, our method leverages simulation only during the training phase to generate rich physical descriptions, eliminating the need for simulation during inference while still benefiting from the detailed annotations that simulations provide.

2.4. Enhancing General Capabilities of VLMs

Recent work has explored modular approaches to enhance VLM capabilities without full model retraining. Chain-of-thought prompting techniques [41] leverage large language models’ reasoning abilities by encouraging step-by-step thinking, while multimodal chain-of-thought approaches [47] extend this to vision-language tasks. Our Physics Context Builders build on these ideas by creating modular components that specialize in translating visual inputs into detailed physical descriptions.

3. Understanding the Effect of Training Data on Physical Reasoning

We begin by investigating why current VLMs struggle with physical reasoning. After outlining our experimental setup, covering datasets, models, and training procedures, we present results showing how fine-tuning with physics-focused data significantly improves performance. We then analyze the trade-off between data quantity and quality.

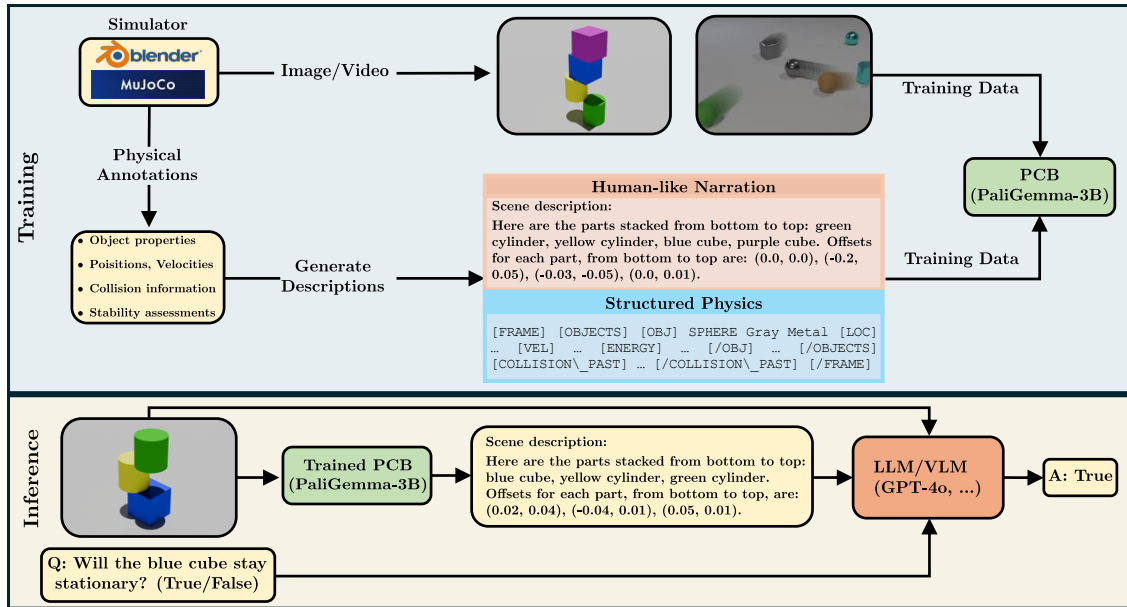


Figure 1. Physics Context Builders (PCBs) pipeline. The training phase (top) shows how physics simulators generate images/videos with corresponding annotations, which are converted into two types of physical descriptions: Human-like Narration (providing natural language scene descriptions with object properties and spatial relationships) and Structured Physics (offering frame-by-frame structured descriptions with physical properties). These descriptions serve as training data for fine-tuning a relatively small VLM into specialized PCBs. During the inference phase (bottom), a trained PCB processes a new image/video and generates detailed physical context about the scene, which is then provided to a foundation model (e.g., GPT-4o) alongside a user query to produce physically grounded responses.

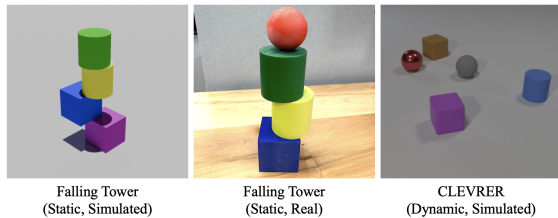


Figure 2. Datasets used for experiments: Falling Tower representing static physics in simulated and real environments; CLEVRER representing dynamic physics in a simulated environment.

3.1. Experimental Setup

Datasets. We utilize two benchmarks: CLEVRER [45] for dynamic physical reasoning and Falling Tower for static stability detection (Fig. 2).

- **CLEVRER** [45] is an established benchmark containing 10,000 training videos, 5,000 validation videos, and 5,000 test videos, each paired with multiple questions across four categories: descriptive, explanatory, predictive, and counterfactual. The training set consists of 109,952 descriptive, 16,799 explanatory, 7,179 predictive, and 18,642 counterfactual questions. All descriptive questions are open-ended. Other question types are multiple-choice with up to four options and may have multiple correct answers. We report the accuracy of the models stratified

by the question types. For multi-choice questions, we report both per-option accuracy, which measures the model’s overall correctness on single options across all questions, and per-question accuracy that measures the model’s performance on full questions, requiring all the choices to be selected correctly. Since the test set answers are not publicly available, we evaluate on the validation set.

- **Falling Tower** is a dataset we create to complement CLEVRER by focusing on static physics. Similar to ShapeStacks [15] and block towers in [24], it features stacked objects but extends previous work by including question-answer pairs and simulation-generated annotations. The dataset contains 4,864 images of stacked objects (with 15 object types across 3 shapes and 5 colors) and 72,775 QA pairs, with 75% used for training and 25% for evaluation. Questions and corresponding answers are generated by applying transformation functions to the simulation annotations, converting it into natural language questions and calculating the answers. For instance, to assess the stability of a stacked tower of objects, we compare the initial and final positions of the objects to determine if significant movement occurred. If an object remains within a predefined threshold, it is considered stable. Questions are then categorized into descriptive (e.g., “How many objects are in the scene?”) and stability (e.g., “Will this collection of objects remain stationary?”). This

dataset also enables the evaluation of Sim2Real transfer; we include 20 real-world images captured with 3D-printed objects with 100 human-generated QAs. For more details on Falling Tower, see Appendix C.

Models and evaluation. We evaluate several zero-shot baselines, including GPT-4o, GPT-4o-mini [1, 19], Gemini-1.5-Pro [37], and PaliGemma-3B-mix (a variant fine-tuned on academic datasets) [6] with chain-of-thought prompting to encourage explicit reasoning. To study training data effects, we fine-tune PaliGemma-3B on CLEVRER and Falling Tower independently using Low-Rank Adaptation (LoRA) [18], minimizing the auto-regressive negative log-likelihood of the answers, conditioned on the questions and input video/image. For videos, we sample 8 frames and append them to the input context, except for the Gemini model, where we use the entire video as the input since they support native video processing. For all models, we take the final answer after the reasoning chain for evaluation. Appendix A.1 provides more details on the training procedure.

Framing the questions. For all questions, including the open-ended and multi-choice ones, we provide the potential options in the question statement. For CLEVRER, we reframe multi-choice questions as a set of binary questions, asking whether each option is a valid answer. This yields significant improvement in the accuracy of all the evaluated models, as reported in Appendix B.1.

3.2. Is Physics-Focused Training Data Enough?

Tabs. 1 and 2 present our results for the Falling Tower and CLEVRER benchmarks. Several key insights emerge:

Zero-shot VLMs struggle with physical reasoning despite strong descriptive capabilities. On Falling Tower, large models like GPT-4o and Gemini-1.5-Pro demonstrate near-perfect accuracy (95-100%) on descriptive questions, both in simulated and real environments. However, their performance drops substantially on stability questions, with accuracy ranging from only 55-60% (barely above random guessing). Similarly, on CLEVRER, while GPT-4o achieves 62.7% accuracy on descriptive questions, its performance falls to 30.7%, 30.3%, and 18.7% on explanatory, predictive, and counterfactual questions, respectively. This substantial performance gap highlights that while VLMs have developed strong capabilities for *high-level* scene understanding and description, they lack the abilities needed to predict outcomes or explain causal relationships accurately.

Fine-tuning with physics-focused data substantially improves physical reasoning. Unsurprisingly, the fine-tuned PaliGemma-3B model shows dramatic improvements across

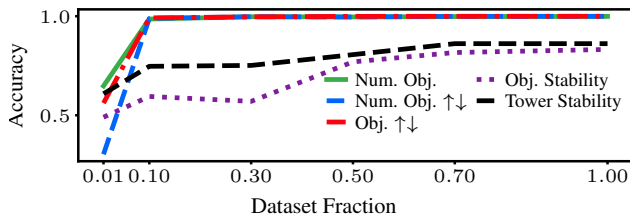


Figure 3. The amount of training data vs. performance on different QA types in Falling Tower.

all tasks. On Falling Tower, it achieves perfect descriptive accuracy (100%) and substantial gains in stability prediction (up to 28% compared to GPT-4o). On CLEVRER, the improvements are even more significant, reaching close to the state-of-the-art results by Aloe, a specialized architecture trained on CLEVRER (Descriptive: 94.0 %, Explanatory: 96.0%, Predictive: 87.5%, Counterfactual: 75.6%) [11]. These results demonstrate that targeted fine-tuning with physics-focused data can significantly enhance physical reasoning capabilities, enabling a 3B parameter model to outperform much larger state-of-the-art models. See Appendix D.1 for a comparison to other specialized baselines.

Sim2Real transfer is successful. The fine-tuned model maintains strong performance when transferring from simulated to real-world images in Falling Tower. It achieves 100% accuracy on descriptive tasks and 70.0% / 65.0% accuracy on object / tower stability questions, respectively.

3.3. Ablation Studies: Not All Data is Equal

To gain deeper insights into how training data composition affects physical reasoning capabilities, we conduct several ablation studies. We provide more ablations in Appendix B.

Specificity of QA types in fine-tuning. Tab. 3 shows the impact of using different subsets of QA pairs for fine-tuning on Falling Tower. Fine-tuning with only descriptive questions achieves perfect descriptive accuracy but fails to improve stability prediction (remaining at the baseline level). Interestingly, fine-tuning with only stability questions not only improves stability prediction substantially but also yields moderate improvements in descriptive tasks, particularly for object identification (41.2% vs. 8.5% baseline). However, the best performance comes from fine-tuning on the full QA dataset. This highlights the need to carefully select data to cover all QA types, which can be seen as a limitation of targeted fine-tuning in VLMs.

Data efficiency varies by task type. Fig. 3 illustrates how accuracy varies with the amount of training data for different

Table 1. Performance of zero-shot VLMs and the fine-tuned model on the Falling Tower benchmark. The second number after the first slash is the Sim2Real accuracy.

Category	Model	Descriptive [sim acc. / real acc.]			Stability [sim acc. / real acc. / real F1]	
		num. obj.	num. obj. $\uparrow\downarrow$	obj. $\uparrow\downarrow$	obj. stable	tower stable
Random	Random	33.3	25.0	14.3	50.0	50.0
Zero-shot CoT	GPT-4o	99.3 / 100.0	91.4 / 100.0	99.4 / 95.0	56.9 / 60.0 / 63.3	59.6 / 55.0 / 54.2
	GPT-4o-mini	94.9 / 89.5	61.3 / 84.2	87.9 / 73.7	49.0 / 52.6 / 55.0	53.1 / 36.8 / 19.8
	Gemini 1.5 Pro	97.2 / 95.0	89.9 / 100.0	97.8 / 95.0	54.6 / 80.0 / 80.0	60.5 / 60.0 / 60.1
	PaliGemma-3B-mix	91.4 / 90.0	44.5 / 70.0	73.0 / 65.0	51.0 / 75.0 / 68.6	39.1 / 65.0 / 51.2
Fine-tuned QA	PaliGemma-3B-Fine-Tuned	100.0 / 100.0	100.0 / 100.0	100.0 / 95.0	84.6 / 70.0 / 73.0	87.6 / 65.0 / 64.4

Table 2. Performance of zero-shot VLMs and the fine-tuned model on the CLEVRER benchmark.

Category	Model	Descriptive	Explanatory		Predictive		Counterfactual	
			per ques.	per opt.	per ques.	per opt.	per ques.	per opt.
Random	Random	28.8	11.8	50.0	25.0	50.0	7.4	50.0
Zero-shot CoT	GPT-4o	62.7	30.7	65.5	30.3	47.6	18.7	60.2
	GPT-4o-mini	49.5	9.3	51.8	44.9	45.1	15.6	51.0
	Gemini 1.5 Pro	58.6	15.7	61.2	32.0	49.6	17.6	55.6
	PaliGemma-3B-mix	38.9	6.6	33.4	44.7	48.7	7.7	49.8
Fine-tuned QA	PaliGemma-3B-Fine-Tuned	92.9	94.7	98.2	83.6	83.6	68.4	88.7

Table 3. Effect of fine-tuning PaliGemma-3B on different QA types in Falling Tower.

Fine-tuning Data	Descriptive [acc.]			Stability [acc.]	
	num. obj.	num. obj. $\uparrow\downarrow$	obj. $\uparrow\downarrow$	obj. stable	tower stable
No Fine-tuning	50.9	21.2	8.5	51.0	39.1
Stability QAs	52.4	26.0	41.2	84.4	86.5
Descriptive QAs	100.0	100.0	100.0	51.0	39.1
All QAs	100.0	100.0	100.0	84.6	87.6

QA types in Falling Tower. Descriptive questions reach near-saturation with only $\sim 10\%$ of the data, showing the model can quickly learn basic scene understanding. In contrast, stability questions benefit more from extra data, though with diminishing returns. Even with 100% of the available data, the model does not reach perfect performance on stability questions (87.6% for tower stability). This shows that increasing dataset size alone is insufficient; improvements in physical reasoning also require diverse, targeted supervision.

4. Physics Context Builders: A Modular Approach to Physical Reasoning

While fine-tuning with physics-focused data can significantly enhance physical reasoning in VLMs, it presents notable practical limitations. Fine-tuning state-of-the-art foundation models like GPT-4o or Gemini is often expensive or even impossible due to their closed-source nature and computational requirements. Furthermore, the task-specific nature of fine-tuning, as seen in Sec. 3, means separate models

may be needed for different physical reasoning tasks, which limits the generalizability of the approach. To address these challenges, we introduce *Physics Context Builders* (PCBs) – a modular and efficient approach that enhances physical reasoning capabilities of existing VLMs without modifying them directly. PCBs leverage the strong in-context learning abilities of large language models [8, 41], which have shown impressive performance with textual scene descriptions [31].

4.1. Methodology

PCBs are specialized VLMs fine-tuned on simulation data to generate detailed physical descriptions of visual scenes. Rather than directly answering questions, PCBs act as perception modules that translate visual inputs into rich textual descriptions capturing the physical properties and dynamics of a scene. These descriptions then serve as enhanced context for larger VLMs, enabling more accurate physical reasoning through in-context learning. Fig. 1 provides an overview of our approach. PCBs offer several advantages:

- **Modularity:** PCBs can be fine-tuned and deployed independently of the foundation model.
- **Efficiency:** Only smaller, specialized models need to be fine-tuned, rather than larger VLMs.
- **Flexibility:** Different PCBs can be developed for different physical phenomena, creating a toolbox of physical reasoning enhancers.
- **Compatibility:** PCBs work with any (vision) language model capable of in-context learning, including closed-source commercial models.

4.1.1. Physical Context Generation

Using annotations from the simulator, PCBs generate question-agnostic descriptions that capture the physical essence of a scene. We consider two context types:

(1) Human-like Narration (HN): Produces intuitive natural language descriptions of the scene’s physical properties and events that can be more aligned with foundation models:

Human-like Narration (HN)
Scene History: Here are the relevant observations prior to the 1st collision:
<ul style="list-style-type: none">• Object 0 (the blue rubber sphere) enters the scene and moves toward the 1st collision site.• Object 1 (the gray rubber sphere) is moving toward the 1st collision site.• Object 2 (the purple rubber sphere) remains stationary in the scene and does not participate in the collision.
Finally, Object 0 collides with Object 1...

(2) Structured Physics (SP): Provides frame-by-frame structured observations in a format similar to a physics simulation output, with standardized tags for physical properties. This structured approach helps models capture precise temporal relationships and physical properties:

Structured Physics (SP)
<pre>[FRAME] [OBJECTS] [OBJ] SPHERE BLUE RUBBER [LOC] (1.0, 2.3) [VEL] 0.5 [ENERGY] 1.2 [/OBJ] [OBJ] SPHERE GRAY RUBBER [LOC] (3.2, 1.5) [VEL] 0.8 [ENERGY] 1.7 [/OBJ] ... [/OBJECTS] [COLLISION_PAST] [COLLISION] [OBJ] SPHERE BLUE RUBBER [/OBJ] [OBJ] SPHERE GRAY RUBBER [/OBJ] [LOC] (2.1, 1.9) [/LOC] [/COLLISION] [/COLLISION_PAST] [/FRAME]</pre>

4.1.2. Training PCBs

We train PCBs by fine-tuning a pre-trained VLM to generate physical descriptions from visual inputs. Specifically:

1. We use PaliGemma-3B as the base pre-trained VLM due to its strong vision-language capabilities and reasonable computational requirements.
2. We apply LoRA-based fine-tuning [18], minimizing the auto-regressive negative log-likelihood of the context descriptions conditioned on the input video/image. See Appendix A.2 for more details.
3. For videos (e.g., in CLEVRER), we sample 8 frames and append them to the input context.
4. Fine-tuning jointly optimizes both vision and language components to ensure alignment between visual features and linguistic representations.

The training data is generated from the same simulations used to create the QA pairs, but focuses on comprehensive scene descriptions rather than specific questions and answers.

This approach enables PCBs to provide rich context independent of the specific reasoning task.

4.2. Experimental Results

We evaluate the effectiveness of PCBs by integrating them with three foundation models: GPT-4o, GPT-4o-mini, and Gemini-1.5-Pro. In each experiment, we first pass the visual input to the appropriate PCB, which generates a physical description. This description is then provided as additional context to the foundation model along with the user’s question and the video/image.

4.2.1. PCB Performance on Falling Tower

Tab. 4 shows the performance of foundation models augmented with PCBs on Falling Tower. For additional experiments, see Appendix D.2. Several key findings:

- **Substantial improvement in stability tasks:** PCBs provide remarkable gains in stability prediction, with GPT-4o-PCB showing up to 25.5% improvement in tower stability assessment. These gains are even more pronounced for smaller models like GPT-4o-mini, which sees a 31.6% improvement in tower stability prediction.
- **Modest gains in descriptive tasks:** Since foundation models already perform well in descriptive tasks, PCBs offer limited additional benefits in this area, with improvements mainly in the more challenging descriptive tasks like identifying objects above/below others (e.g., 13.0% improvement for GPT-4o-mini).
- **Effective Sim2Real transfer:** PCB-augmented models show improved generalization to real-world scenarios, with GPT-4o-PCB achieving a 15.0% gain in real-world tower stability prediction compared to zero-shot.
- **Model size effects:** Interestingly, the smaller GPT-4o-mini model shows greater relative improvements when augmented with PCBs compared to larger ones, possibly due to its more limited perception.

4.2.2. PCB Performance on CLEVRER

Tab. 5 presents results for CLEVRER, a more challenging benchmark that requires dynamic physical reasoning:

- **Human Narration vs. Structured Physics:** Human-like narration (HN) consistently outperforms structured physics (SP) descriptions across all models and question types. This can be due to foundation models’ better understanding of natural language descriptions compared to more structured, technical formats.
- **Strong improvements in descriptive and explanatory tasks:** PCBs provide substantial gains in descriptive accuracy (up to 16.2% for GPT-4o-mini) and explanatory reasoning (up to 19.9% for Gemini 1.5 Pro).
- **Limited gains in counterfactual reasoning:** While PCBs improve counterfactual reasoning (1.7–9.5% gains), the improvements are more modest, reflecting the intrinsic complexity of this task, even with enriched context.

Table 4. Performance of foundation models augmented with Physics Context Builders (PCBs) on the Falling Tower benchmark. HN refers to Human Narration style PCB. The numbers in parentheses indicate improvements over the respective zero-shot baselines. The second number after the slash is the Sim2Real accuracy and the third number after the second slash is the F1 score on Sim2Real.

Category	Model	Descriptive [sim acc. / real acc.]			Stability [sim acc. / real acc. / real F1]		
		num. obj.	num. obj. ↑↓	obj. ↑↓	obj. stable	tower stable	
Zero-shot CoT	GPT-4o	99.3 / 100.0	91.4 / 100.0	99.4 / 95.0	56.9 / 60.0 / 63.3	59.6 / 55.0 / 54.2	
	GPT-4o-mini	94.9 / 89.5	61.3 / 84.2	87.9 / 73.7	49.0 / 52.6 / 55.0	53.1 / 36.8 / 19.8	
	Gemini 1.5 Pro	97.2 / 95.0	89.9 / 100.0	97.8 / 95.0	54.6 / 80.0 / 80.0	60.5 / 60.0 / 60.1	
VLM + PCB (HN)	GPT-4o-PCB	99.5 / 100.0	97.6 / 100.0	99.5 / 95.0	76.7 / 75.0 / 73.8	85.1 / 70.0 / 65.6	
	GPT-4o-mini-PCB	(+0.2) / (0.0)	(+6.2) / (0.0)	(+0.1) / (0.0)	(+19.8) / (+15.0) / (+10.5)	(+25.5) / (+15.0) / (+11.4)	
	Gemini 1.5 Pro-PCB	99.9 / 95.0	74.3 / 90.0	97.5 / 95.0	75.0 / 70.0 / 73.0	84.7 / 40.0 / 33.8	
		(+5.0) / (+5.5)	(+13.0) / (+5.8)	(+9.6) / (+21.3)	(+26.0) / (+17.4) / (+18.0)	(+31.6) / (+3.2) / (+14.0)	
	Gemini 1.5 Pro-PCB	97.9 / 100.0	97.5 / 100.0	97.4 / 94.7	75.9 / 73.7 / 76.7	84.9 / 57.9 / 59.1	
		(+0.7) / (+5.0)	(+7.6) / (0.0)	(-0.4) / (-0.3)	(+21.3) / (-6.3) / (-3.3)	(+24.4) / (-2.1) / (-1.0)	

Table 5. Performance of foundation models augmented with Physics Context Builders (PCBs) on the CLEVRER benchmark. HN is Human Narration and SP is Structured Physics. The numbers in parentheses indicate improvements over the respective zero-shot baselines.

Category	Model	Descriptive	Explanatory		Counterfactual	
			per ques.	per opt.	per ques.	per opt.
Zero-shot CoT	GPT-4o	62.7	30.7	65.5	18.7	60.2
	GPT-4o-mini	49.5	9.3	51.8	15.6	51.0
	Gemini 1.5 Pro	58.6	15.7	61.2	17.6	55.6
VLM + PCB (HN)	GPT-4o-PCB	75.6 (+12.9)	41.6 (+10.9)	67.0 (+1.5)	28.2 (+9.5)	68.4 (+8.2)
	GPT-4o-mini-PCB	65.7 (+16.2)	26.8 (+17.5)	62.2 (+10.4)	17.3 (+1.7)	52.8 (+1.8)
	Gemini 1.5 Pro-PCB	72.8 (+14.2)	35.6 (+19.9)	70.8 (+9.6)	26.2 (+8.6)	64.9 (+9.3)
VLM + PCB (SP)	GPT-4o-PCB	70.0 (+7.3)	34.9 (+4.2)	61.1 (-4.4)	19.2 (+0.5)	63.3 (+3.1)
	GPT-4o-mini-PCB	58.6 (+9.1)	19.2 (+9.9)	58.9 (+7.1)	16.2 (+0.6)	51.8 (+0.8)
	Gemini 1.5 Pro-PCB	67.4 (+8.8)	30.5 (+14.8)	68.7 (+7.5)	21.1 (+3.5)	60.5 (+4.9)

Remark. For CLEVRER, we omit PCB evaluation on predictive questions since our current PCBs are designed to describe observed scenes rather than predict future events. Future work could address this by developing predictive PCBs that generate plausible future scene descriptions based on simulation rollouts.

4.3. Multi-Agent Framework for PCB Integration

Given the ability of foundational models to interpret the overall context of scenes effectively, we explore whether they can reliably select the appropriate PCB when provided with a question-image pair. We utilize a multi-agent triage model inspired by OpenAI’s Swarm architecture [7]. As illustrated in Fig. 4, our multi-agent framework consists of:

1. A **Triage Agent** that analyzes the user query and visual input to identify the required type of physical reasoning.
2. Multiple **PCBs**, each specialized for different physical phenomena (e.g., stability analysis, collision detection, motion tracking).
3. A **Foundation Model** that receives the PCB-generated context and the original query to produce the response.

In our evaluation, each input consists of a natural language question from a QA dataset paired with a corresponding scene image. This input is initially processed by GPT-4o or GPT-4o-mini that routes the query to one of two specialized PCBs: the PCB for the Falling Tower dataset, or PCB

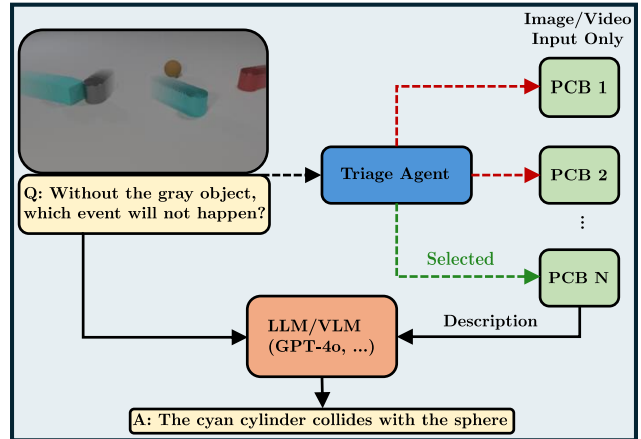


Figure 4. Multi-agent framework for PCB integration. A triage agent selects the appropriate PCB based on the user query, and the PCB generates a detailed scene description that enriches the foundation model’s context.

designed for analyzing motion and object interactions as in the CLEVRER dataset. As shown in Tab. 6, both GPT-4o and GPT-4o-mini achieve excellent F1-Scores, effectively routing queries to the appropriate PCB. These results suggest that PCBs in a multi-agent framework offer a promising approach, with foundation models capable of reliably selecting

Table 6. Accuracy and F1-scores for selecting the correct Physics Context Builder using the triage agent in a two-agent setup.

Task	Metric	GPT-4o / GPT-4o-mini
Stacked Objects	Accuracy (%)	87.67 / 97.00
	F1-Score	0.9326 / 0.9782
Dynamic Scene	Accuracy (%)	94.67 / 98.67
	F1-Score	0.9403 / 0.9785

the correct PCB based on the question-image pair.

5. Discussion

Our results demonstrate that Physics Context Builders (PCBs) offer a promising approach to enhancing physical reasoning in Vision-Language Models (VLMs). PCBs increase the average performance of GPT-4o (GPT-4o-mini) by 11.1% (11.8%) on CLEVRER and 8.2% (13.8%) on Falling Tower. These improvements are particularly notable given that no modification to the foundation models was required. Despite these gains, large VLMs still exhibit sub-optimal performance in tasks requiring deeper reasoning, such as counterfactual questions (with improvements limited to 1.7-9.5%) and explanatory questions. While enhancing visual perception through PCBs is important, especially for descriptive tasks, the subpar performance on more involved reasoning tasks calls for additional interventions to achieve a comprehensive physical understanding.

Besides the performance gain, our findings confirm the value of simulation data in addressing the limitations of VLMs. Unlike simulation-in-the-loop approaches [12, 29, 42, 43], which require computationally expensive simulators during inference, our method uses simulation only to generate synthetic data (context descriptions) for fine-tuning smaller VLMs. This approach is more efficient at inference while still leveraging the rich annotations that simulations can provide. The effectiveness of PCBs in generalizing from simulated to real-world data further supports this approach, as demonstrated by the successful Sim2Real transfer in the Falling Tower benchmark.

PCBs provide a modular and efficient framework for enhancing the physical reasoning capabilities of foundation models without requiring direct fine-tuning. Generating rich physical context from visual inputs enhances the perceptual ability of foundation models, resulting in more accurate reasoning across diverse physical tasks, from static stability to dynamic collision detection. The multi-agent framework further enhances this approach by enabling adaptive selection of specialized PCBs based on the specific reasoning task.

6. Limitations and Future Work

While our work demonstrates the effectiveness of PCBs across benchmarks, several limitations remain. First, our

current benchmarks focus on a relatively constrained set of physical phenomena—primarily rigid body dynamics and stability. This limits our ability to evaluate how well these approaches generalize to the full spectrum of physical reasoning that humans perform intuitively, such as fluid dynamics or object manipulation.

Second, our framework requires annotated data, which naturally comes from simulation. For unannotated videos, such as those available on YouTube, the lack of structured annotations presents a challenge. An important open problem is how PCBs could extract detailed physical descriptions from real-world videos without explicit annotations, potentially through self-supervised learning or by leveraging other foundation models to generate pseudo-annotations.

Third, the performance improvements, while substantial, still leave room for further enhancement, particularly for complex reasoning tasks like counterfactual/predictive questions. Since PCBs inherently perform direct translation from visual signals to text, they lead to strong descriptive capabilities in the foundation models. However, this translation process does not capture all the visual cues needed for more complex reasoning. Hence, we hypothesize that more comprehensive textual descriptions, possibly containing counterfactual/predictive information, could significantly improve the performance of PCBs on these challenging tasks.

A promising future direction is to tackle more complex physical reasoning tasks that require larger-scale simulations. This might involve integrating advanced physics engines that simulate phenomena like fluid dynamics, deformable materials, and articulated mechanisms. Notably, while these simulations could become computationally expensive for simulation-in-the-loop approaches during inference, our PCB framework uses simulation data only for fine-tuning and therefore does not add significant computational burden at inference time.

Another avenue for future work is to explore how multiple PCBs could be composed or chained together to handle scenarios requiring reasoning about multiple physical phenomena simultaneously. For instance, reasoning about a scene involving both rigid body dynamics and fluid interactions might benefit from specialized PCBs for each phenomenon, with their outputs combined to provide comprehensive context to the foundation model.

Acknowledgements

We thank Brian Jeong for 3D printing the parts used in the Falling Tower dataset. Resources for this research were primarily provided by Autodesk Research, with partial support from the Province of Ontario, the Government of Canada through CIFAR, and sponsors of the Vector Institute.

References

- [1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 4
- [2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022. 1
- [3] Renée Baillargeon. Infants’ physical world. *Current directions in psychological science*, 13:89–94, 2004. 1
- [4] Peter Battaglia, Razvan Pascanu, Matthew Lai, Danilo Jimenez Rezende, et al. Interaction networks for learning about objects, relations and physics. *Advances in neural information processing systems*, 29, 2016. 2
- [5] Peter W Battaglia, Jessica B Hamrick, and Joshua B Tenenbaum. Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences*, 110(45):18327–18332, 2013. 1, 2
- [6] Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarelli, et al. Paligemma: A versatile 3B VLM for transfer. *arXiv preprint arXiv:2407.07726*, 2024. 4
- [7] Ilan Bigio, James Hills, Shyamal Anadkat, Charu Jaiswal, Colin Jarvis, and Katia Gil Guzman. GitHub - openai/swarm: Educational framework exploring ergonomic, lightweight multi-agent orchestration. Managed by OpenAI Solution team., 2024. [Accessed 12-11-2024]. 7
- [8] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901, 2020. 5
- [9] Z Chen, J Mao, J Wu, KKY Wong, JB Tenenbaum, and C Gan. Grounding physical concepts of objects and events through dynamic visual reasoning. In *International Conference on Learning Representations (ICLR) 2021*. Vienna, Austria, 2021. 2, 15
- [10] Wei Chow, Jiageng Mao, Boyi Li, Daniel Seita, Victor Campagnolo Guizilini, and Yue Wang. Physbench: Benchmarking and enhancing vision-language models for physical world understanding. In *The Thirteenth International Conference on Learning Representations*, 2025. 1, 2
- [11] David Ding, Felix Hill, Adam Santoro, Malcolm Reynolds, and Matt Botvinick. Attention over learned object embeddings enables complex visual reasoning. *Advances in neural information processing systems*, 34: 9112–9124, 2021. 1, 2, 4, 15
- [12] Mingyu Ding, Zhenfang Chen, Tao Du, Ping Luo, Josh Tenenbaum, and Chuang Gan. Dynamic visual reasoning by learning differentiable physics models from video and language. *Advances In Neural Information Processing Systems*, 34:887–899, 2021. 2, 8, 15
- [13] Jiafei Duan, Samson Yu, Soujanya Poria, Bihan Wen, and Cheston Tan. Pip: Physical interaction prediction via mental simulation with span selection. In *European Conference on Computer Vision*, pages 405–421. Springer, 2022. 2
- [14] Sadaf Ghaffari and Nikhil Krishnaswamy. Exploring failure cases in multimodal reasoning about physical dynamics. In *Proceedings of the AAAI Symposium Series*, pages 105–114, 2024. 1, 2
- [15] Oliver Groth, Fabian B Fuchs, Ingmar Posner, and Andrea Vedaldi. ShapeStacks: Learning vision-based physical intuition for generalised object stacking. In *Proceedings of the european conference on computer vision (eccv)*, pages 702–717, 2018. 3, 14
- [16] Vincent Le Guen and Nicolas Thome. Disentangling physical dynamics from unknown factors for unsupervised video prediction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11474–11484, 2020. 2
- [17] Dingkun Guo, Yuqi Xiang, Shuqi Zhao, Xinghao Zhu, Masayoshi Tomizuka, Mingyu Ding, and Wei Zhan. Phygrasp: generalizing robotic grasping with physics-informed large multimodal models. *arXiv preprint arXiv:2402.16836*, 2024. 1
- [18] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022. 4, 6
- [19] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 4
- [20] Adam Ishay, Zhun Yang, Joohyung Lee, Ilgu Kang, and Dongjae Lim. Think before you simulate: Symbolic reasoning to orchestrate neural computation for counterfactual question answering. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6698–6707, 2024. 2, 15
- [21] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023. 2

- [22] James R Kubricht, Keith J Holyoak, and Hongjing Lu. Intuitive physics: Current research and controversies. *Trends in cognitive sciences*, 21(10):749–759, 2017. 1
- [23] Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. Building machines that learn and think like people. *Behavioral and brain sciences*, 40:e253, 2017. 1
- [24] Adam Lerer, Sam Gross, and Rob Fergus. Learning physical intuition of block towers by example. In *International conference on machine learning*, pages 430–438. PMLR, 2016. 3
- [25] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 1
- [26] Zhiyuan Li, Heng Wang, Dongnan Liu, Chaoyi Zhang, Ao Ma, Jieting Long, and Weidong Cai. Multimodal causal reasoning benchmark: Challenging vision large language models to infer causal links between siamese images. *arXiv preprint arXiv:2408.08105*, 2024. 1, 2
- [27] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 1
- [28] Fangchen Liu, Kuan Fang, Pieter Abbeel, and Sergey Levine. Moka: Open-vocabulary robotic manipulation through mark-based visual prompting. In *First Workshop on Vision-Language Models for Navigation and Manipulation at ICRA 2024*, 2024. 1
- [29] Ruibo Liu, Jason Wei, Shixiang Shane Gu, Te-Yen Wu, Soroush Vosoughi, Claire Cui, Denny Zhou, and Andrew M Dai. Mind’s eye: Grounded language model reasoning through simulation. In *The Eleventh International Conference on Learning Representations*, 2022. 2, 8
- [30] Michael McCloskey, Allyson Washburn, and Linda Felch. Intuitive physics: the straight-down belief and its origin. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 9(4):636, 1983. 1
- [31] Aishik Nagar, Shantanu Jaiswal, and Cheston Tan. Zero-shot visual reasoning by vision-language models: Benchmarking and analysis. In *2024 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2024. 2, 5
- [32] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1
- [33] Pooyan Rahmazadehgergi, Logan Bolton, Mohammad Reza Taesiri, and Anh Totti Nguyen. Vision language models are blind. In *Proceedings of the Asian Conference on Computer Vision*, pages 18–34, 2024. 1
- [34] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual Captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, 2018. 1
- [35] Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adri Garriga-Alonso, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on machine learning research*, 2023. 1
- [36] Haoran Sun, Yijiang Li, Qingying Gao, Haiyun Lyu, Dezhi Luo, and Hokin Deng. Probing mechanical reasoning in large vision language models. In *ICLR 2025 Workshop on Bidirectional Human-AI Alignment*, 2025. 1
- [37] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024. 4
- [38] Ernő Téglás, Edward Vul, Vittorio Girotto, Michel Gonzalez, Joshua B Tenenbaum, and Luca L Bonatti. Pure reasoning in 12-month-old infants as probabilistic inference. *science*, 332(6033):1054–1059, 2011. 1
- [39] Mor Ventura, Michael Toker, Nitay Calderon, Zorik Gekhman, Yonatan Bitton, and Roi Reichart. NL-eye: Abductive NLI for images. In *The Thirteenth International Conference on Learning Representations*, 2025. 1, 2
- [40] Yi Ru Wang, Jiafei Duan, Dieter Fox, and Siddhartha Srinivasa. Newton: Are large language models capable of physical reasoning? In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023. 1
- [41] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022. 2, 5
- [42] Jiajun Wu, Ilker Yildirim, Joseph J Lim, Bill Freeman, and Josh Tenenbaum. Galileo: Perceiving physical object properties by integrating a physics engine with

- deep learning. *Advances in neural information processing systems*, 28, 2015. [2](#), [8](#)
- [43] Jiajun Wu, Erika Lu, Pushmeet Kohli, Bill Freeman, and Josh Tenenbaum. Learning to see physics via visual de-animation. *Advances in neural information processing systems*, 30, 2017. [2](#), [8](#)
- [44] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10371–10381, 2024. [2](#)
- [45] 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*, 2020. [1](#), [3](#)
- [46] Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. [1](#)
- [47] Zhuosheng Zhang, Aston Zhang, Mu Li, George Karypis, Alex Smola, et al. Multimodal chain-of-thought reasoning in language models. *Transactions on Machine Learning Research*, 2024. [2](#)
- [48] Zhicheng Zheng, Xin Yan, Zhenfang Chen, Jingzhou Wang, Qin Zhi Eddie Lim, Joshua B Tenenbaum, and Chuang Gan. Contphy: Continuum physical concept learning and reasoning from videos. In *International Conference on Machine Learning*, pages 61526–61558. PMLR, 2024. [2](#)
- [49] Erle Zhu, Yadi Liu, Zhe Zhang, Xujun Li, Xinjie Yu, Minlie Huang, Hongning Wang, et al. Maps: Advancing multi-modal reasoning in expert-level physical science. In *The Thirteenth International Conference on Learning Representations*, 2025. [2](#)