Physics Context Builders:
A Modular Framework for Physical Reasoning in VLMs

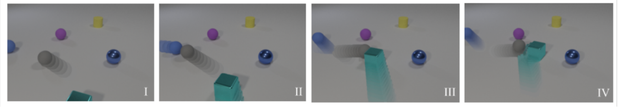
Vahid Balazadeh[1], Mohammadmehdi Ataei[2], Hyunmin Cheong[2,3], Amir Khasahmadi[2], Rahul G. Krishnan[1]

[1]University of Toronto, Vector Institute [2]Autodesk Research [3]Spectral Labs (current)

AUTODESK Research · UNIVERSITY OF TORONTO · VECTOR INSTITUTE

(0) Motivation

Vision-language models (VLMs) struggle at physical reasoning, even in simple tasks like collision detection. Their training datasets often lack physics-based annotations.



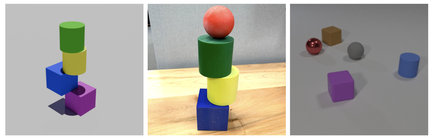
Illustrative Example: How many collisions happen after the cube enters the scene? Answer: 3

VLM	Response
VideoLLaMA 2	There are two collisions that occur between the cube and the purple sphere.
MiniGPT4-video	The cube hits one of the other eggs, causing a collision.
Video-LLaVA	After the cube enters the scene, there are two more collisions.
Gemini 1.5 Pro	There are 4 collisions that occur after the cube enters the scene.
GPT-4o	There appear to be no collisions detected after the cube enters the scene.

(1) TL;DR

- **Problem:** VLMs lack physics-grounded supervision and exhibit weak physical reasoning.
- **Core Idea:** Fine-tune *smaller* VLMs on simulator-generated physics annotations to create Physics Context Builders (**PCBs**). PCBs translate visual scenes into detailed text descriptions (e.g., object properties and spatial relations) and feed those to foundation models. The main idea is to separate perception from reasoning.
- **Results:** i) +10–20 % on CLEVRER descriptive/explanatory; ii) +20–32 % on stability detection; iii) strong Sim2Real transfer. All without modifying the foundation model.
- **Takeaway:** Training data is a major bottleneck; simulator-trained PCBs offer a modular and composable solution to enhance perception.

(2) Datasets and Setup



Falling Tower (Static, Simulated) · Falling Tower (Static, Real) · CLEVRER (Dynamic, Simulated)

CLEVRER. (Dynamic reasoning benchmark) ~ 10K training and ~ 5K test videos, with 151K questions across descriptive, explanatory, predictive, and counterfactual tasks on object interactions.

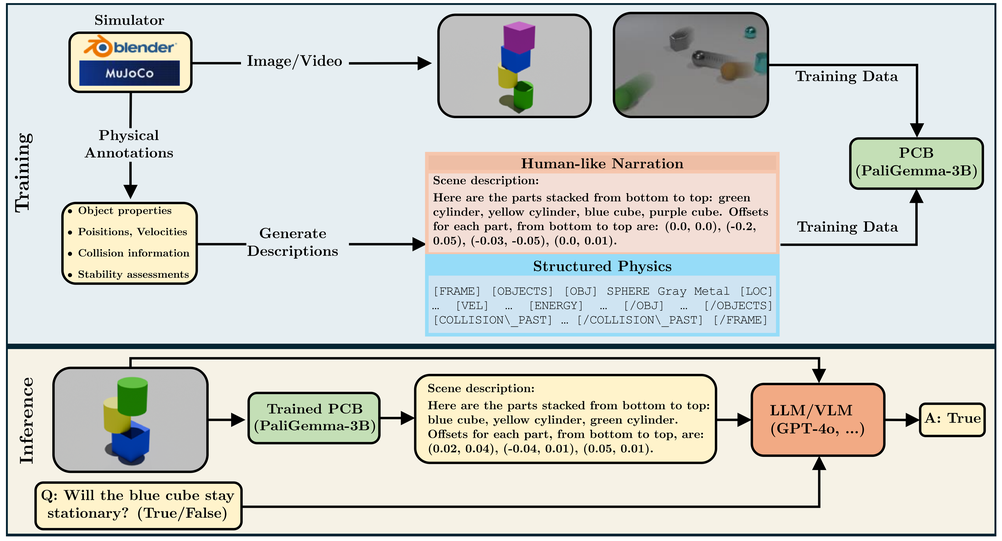
Falling Tower. (Static stability detection) ~ 5K simulated images with 15 object types and 70K question-answer pairs + Sim2Real data. Includes descriptive and stability detection tasks.

(3) Physics Context Builders (PCBs)

Idea. Train small VLMs on simulation annotations to output detailed physical descriptions, then provide those as **context** to a foundation model answering the question. This allows us to separate perception (using PCBs) from reasoning (using foundation VLMs).

Key Points.

- Static reasoning (Falling Tower): +20–32 %; strong Sim2Real transfer.
- Dynamic reasoning (CLEVRER): +10–20 % on descriptive/explanatory
- Sparse frames overlook short events and occlusions can mislead the model.
- Counterfactual and predictive questions remain unsolved.
- Human-like Narration (HN) > Structured Physics (SP).
- No changes to foundation models; biggest gains for smaller VLMs.



(4) Training Data is a Major Bottleneck

To see the importance of training data, we fine-tune a small VLM (3B PaliGemma) on generated question-answer pairs and evaluate it on test videos:

Key Numbers.

- CLEVRER: 92.9% descriptive, and 94.7% explanatory. Approaching state-of-the-art specialized solutions such as Aloe (94.0%, 96.0%) [1].
- Falling Tower: 100% descriptive, and 87.6% stability ≫ GPT-4o (59.6%) despite having orders of magnitude fewer parameters.
- Strong Sim2Real performance transfer.

Conclusion.

The performance gap between zero-shot foundation models and fine-tuned small models shows VLMs lack physics-grounded supervision and not necessarily model capacity. This motivates PCBs as a practical solution to bridge the data gap without expensive re-training.

(5) Results

Falling Tower — VLM + PCB

Category	Model	Descriptive [sim / real] num. obj.	Descriptive [sim / real] num. obj. ↑↓	Descriptive [sim / real] obj. ↑↓	Stability [sim / real] obj. stable	Stability [sim / real] tower stable
VLM + PCB (HN)	GPT-4o-PCB	99.5 / **100.0** (+0.2) / (0.0)	**97.6** / **100.0** (+6.2) / (0.0)	**99.5** / **95.0** (+0.1) / (0.0)	**76.7** / 75.0 (+19.8) / (+15.0)	**85.1** / **70.0** (+25.5) / (+15.0)
	GPT-4o-mini-PCB	**99.9** / 95.0 (+5.0) / (+5.5)	74.3 / 90.0 (+13.0) / (+5.8)	97.5 / **95.0** (+9.6) / (+21.3)	75.0 / 70.0 (+26.0) / (+17.4)	84.7 / 40.0 (+31.6) / (+3.2)
	Gemini 1.5 Pro-PCB	97.9 / **100.0** (+0.7) / (+5.0)	97.5 / **100.0** (+7.6) / (0.0)	97.4 / 94.7 (-0.4) / (-0.3)	75.9 / **73.7** (+21.3) / (-6.3)	84.9 / 57.9 (+24.4) / (-2.1)

CLEVRER — VLM + PCB

Category	Model	Descriptive	Explanatory per ques.	Explanatory per opt.	Counterfactual per ques.	Counterfactual 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)

(6) Limitations & Open Questions

Open Questions & Future Work.

- **Scaling to Complex Physics:** Integrate advanced simulations (fluids, deformable materials) to tackle more diverse physical reasoning tasks.
- **Learning from the Wild:** Develop methods (e.g., self-supervision) to extract physical descriptions from unannotated real-world videos.
- **Richer Representations:** Enhance PCB outputs with predictive/counterfactual information to directly improve complex reasoning.
- **Compositional Reasoning:** Explore chaining or combining specialized PCBs to analyze scenes with multiple simultaneous physical phenomena.

Current Limitations.

- **Narrow Physics Scope:** Benchmarks are constrained to rigid body dynamics, limiting evaluation on phenomena like fluid dynamics or object manipulation.
- **Reliance on Annotated Data:** The framework requires structured simulation data and struggles with unannotated real-world videos (e.g., from YouTube).
- **Complex Reasoning Ceiling:** The direct visual-to-text translation loses subtle cues, leaving performance gaps in complex counterfactual and predictive tasks.

References.

1. Ding, David, et al. "Attention over learned object embeddings enables complex visual reasoning." Advances in neural information processing systems 34 (2021): 9112-9124.

Scan for more details!