PhysBench: Benchmarking and Enhancing Vision-Language Models for Physical World Understanding

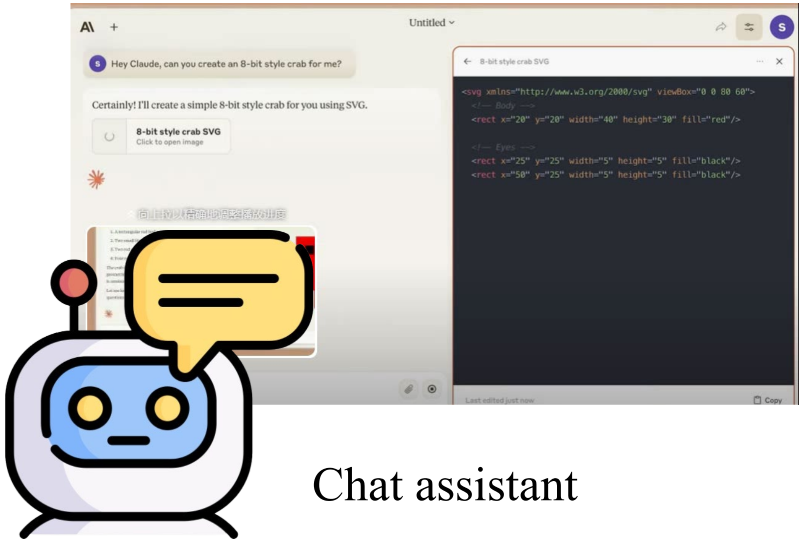
Wei Chow[1]*, Jiageng Mao[1]*, Boyi Li[2], Daniel Seita[1], Vitor Guizilini[3], Yue Wang[1]

[1]University of Southern California, [2]UC Berkeley, [3]Toyota Research Institute
*Equal Contribution

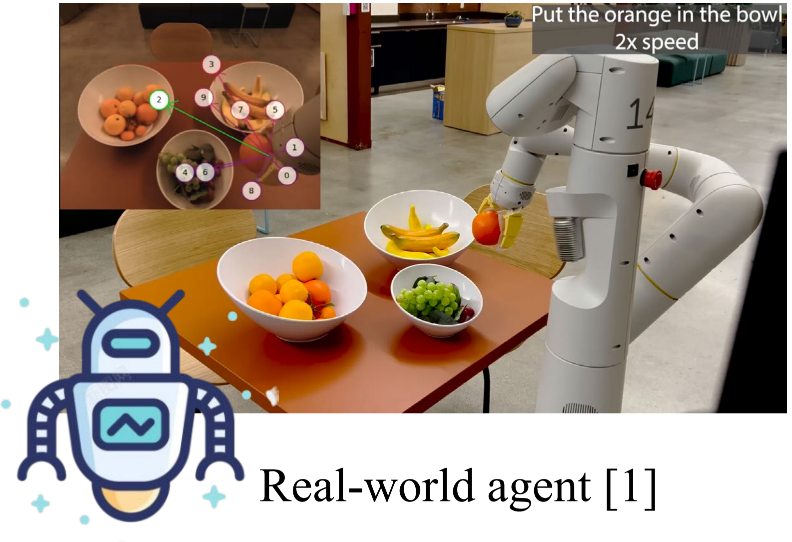
Singapore, Thu 24 Apr 04:18 PM +08

● Background

VLM not only serves as **a chat assistant**, but also has a broader application scenario as a **robotics agent** deployed in real-world environments to **solve practical problems**.



Chat assistant



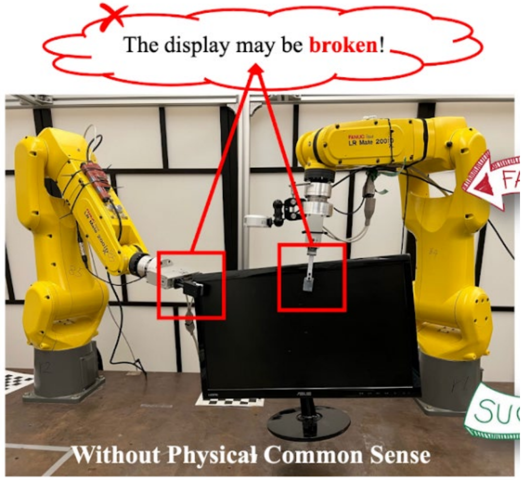
Real-world agent [1]

[1] VLM as robotics agent: PIVOT (2024), MOKA (2024), Voxposer (2023)

● Motivation

However, numerous studies have shown that VLM's **lack of basic perception of the physical world** leads to operational errors.

⇨ A gap between VLMs and real **physical world understanding**



Error affordance [1]

Excessive force [2]

Error Throwing [3]

We want to benchmark and enhance VLMs' physical understanding capability for embodied tasks.

[1] Dingkun Guo, etal. Phygrasp: Generalizing robotic grasping with physics-informed large multimodal models,
[2] Yi RuWang, et al. Newton: Are large language models capable of physical reasoning?
[3] Fangchen Liu, et al: Open-vocabulary robotic manipulation through mark-based visual prompting.

● Related Works

Most works only contain common VQA tasks that are **semantic driven**, which does not **enable physical world prior learning.**

Semantic Driven Information: Which is shown in the video?

Sure, in the video, we see a man in a red vest and a young girl in a pink outfit standing on a stage. The man appears to be a boss or a performer, while the young girl is likely his assistant or a participant in the performance. The man threw a ball in his hand, drawing a curve.

Train VLM

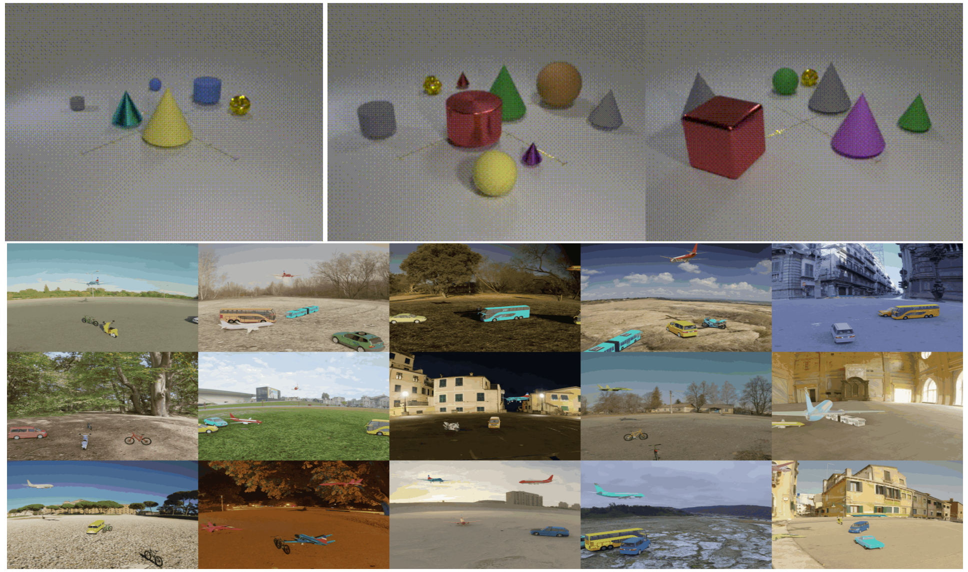
Physical World Prior : What physical prior exists in the video?

The horizontal speed of the ball remains basically unchanged, while the vertical speed first decreases to 0, then turns and increases.

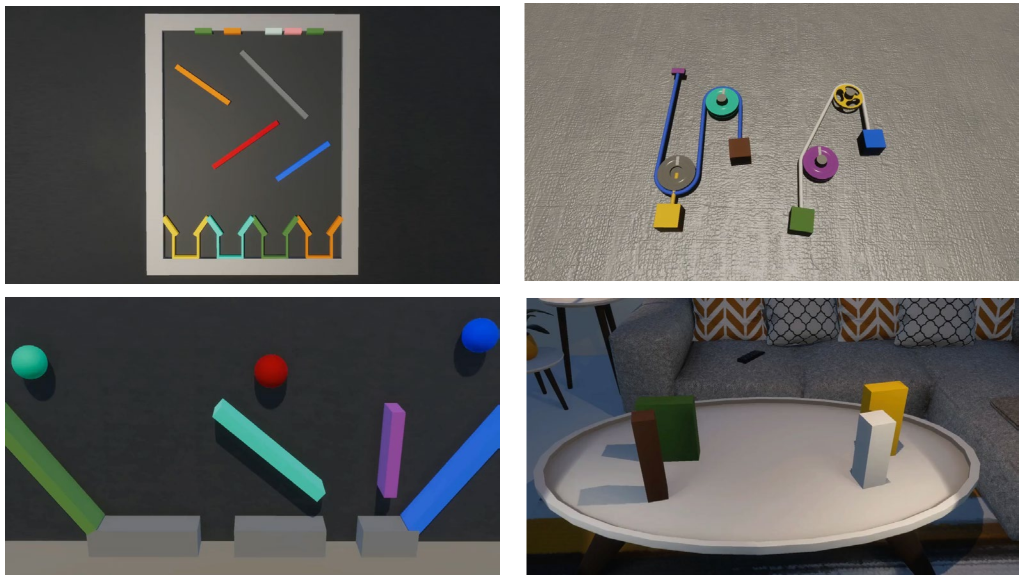
Ignored

● Related Works

Moreover, existing benchmarks for learning intuitive physics from visual inputs only includes simple visual primitives and a limited number of task types.



Simple visual primitives [1]
(only spheres, cubes, and collision events



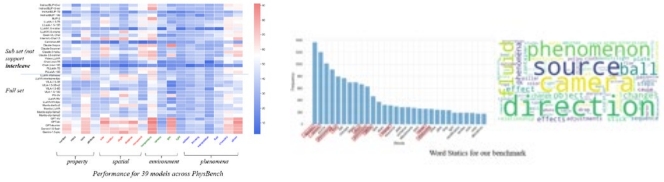
Few specific types of tasks [2]

[1] CLEVRER Yi et al. (2019) Cater Girdhar & Ramanan (2019) CRIPP-VQA Patel et al. (2022) ComPhy Chen et al. (2022) SuperCLEVR Wang et al. (2024)
[2] EmbSpatial Du et al. (2024) Physion Bear et al. (2021) Physion++ Tung et al. (2023) ContPhy Zheng et al. (2024)

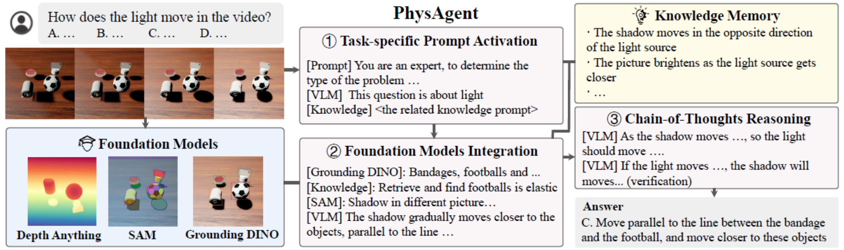
● Three Questions

VLM has the potential to serve as a **robotics agent**, **but** there is a gap in its perception of the **physical world**.

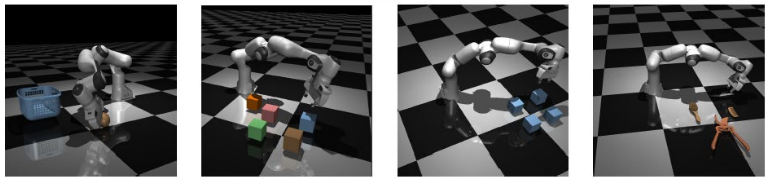
Q1 🤔 Do VLMs **perform well on** physical world understanding? If not, what **causes** them to lack this ability?



Q2 🤔 How can we improve VLM's physical world understanding ability?



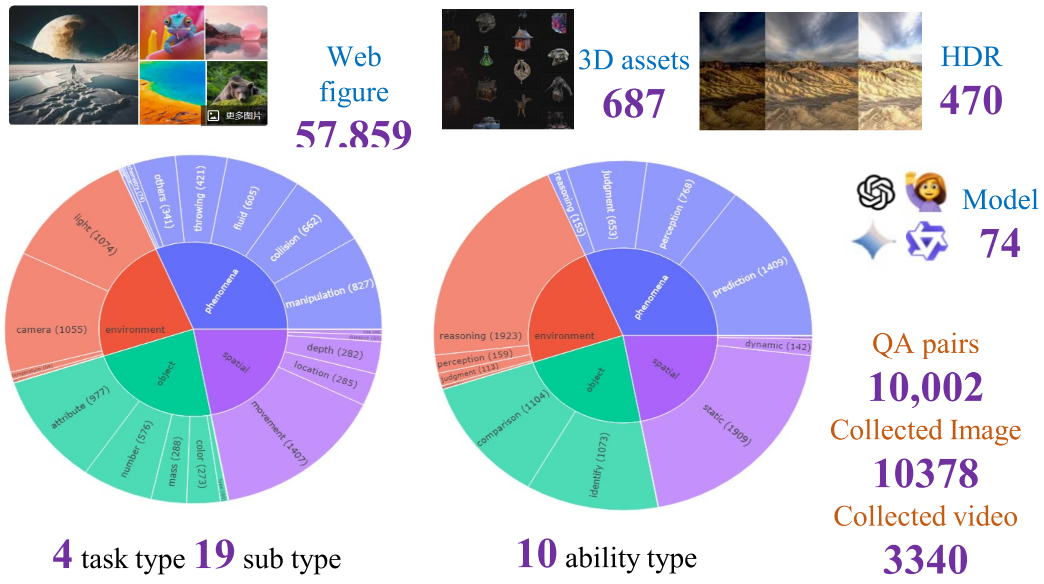
Q3 🤔 Will improving VLMs' physical world understanding facilitate the deployment of embodied agents in the real-world?



(d) (e) (f) (g)

● Dataset

To comprehensively measure how big this gap, we propose our PhysBench.



ICLR International Conference On Learning Representations

● Main Results

We tested **74 VLMs** and found that there is still a lot of room for enhancing this ability

	Size	Format	Property	Relationships	Scene	Dynamics	**Avg**
Phi-3.5V (AzureML, 2024)	4B	seq	45.72	40.15	33.02	39.40	39.75
NVILA-8B (Liu et al., 2024f)	8B	seq	55.79	40.29	33.95	43.43	43.82
NVILA-15B (Liu et al., 2024f)	15B	seq	59.16	42.34	38.78	45.72	46.91
NVILA-Lite-8B (Liu et al., 2024f)	8B	seq	53.81	39.25	34.62	41.17	42.55
NVILA-Lite-15B (Liu et al., 2024f)	15B	seq	55.44	40.15	38.11	44.38	44.93
mPLUG-Owl3-1B (Ye et al., 2024)	1B	seq	38.02	31.54	21.87	33.68	31.68
mPLUG-Owl3-2B (Ye et al., 2024)	2B	seq	40.92	35.11	26.69	35.64	34.87
mPLUG-Owl3-7B (Ye et al., 2024)	7B	seq	49.25	45.62	35.90	40.61	42.83
InternVL2-1B (Wang et al., 2024e)	1B	seq	37.05	33.06	22.84	34.92	32.35
InternVL2-2B (Wang et al., 2024e)	2B	seq	44.17	35.06	30.54	35.64	36.57
InternVL2-4B (Wang et al., 2024e)	4B	seq	47.12	39.96	30.94	39.76	39.71
InternVL2-8B (Wang et al., 2024e)	8B	seq	49.05	43.58	27.05	39.47	40.00
InternVL2-26B (Wang et al., 2024e)	26B	merge	51.92	45.20	37.94	39.34	43.50
InternVL2-40B (Wang et al., 2024e)	40B	merge	55.79	50.05	35.86	41.33	45.66
InternVL2-76B (Wang et al., 2024e)	76B	merge	57.65	52.43	38.07	40.12	46.77
InternVL2.5-1B (Gao et al., 2024b)	1B	seq	44.25	33.30	26.87	38.13	36.15
InternVL2.5-2B (Gao et al., 2024b)	2B	seq	49.63	38.15	29.44	38.39	39.22
InternVL2.5-4B (Gao et al., 2024b)	4B	seq	51.03	44.77	31.34	41.79	42.44
InternVL2.5-8B (Gao et al., 2024b)	8B	seq	55.87	48.67	29.35	41.20	43.88
InternVL2.5-26B (Gao et al., 2024b)	26B	merge	59.08	58.33	36.61	41.79	48.56
InternVL2.5-38B (Gao et al., 2024b)	38B	merge	58.77	67.51	39.04	45.00	51.94
InternVL2.5-78B (Gao et al., 2024b)	78B	merge	60.32	62.13	37.32	46.11	51.16
o1 (Jaech et al., 2024)	-	merge	59.27	73.79	40.95	49.22	55.11

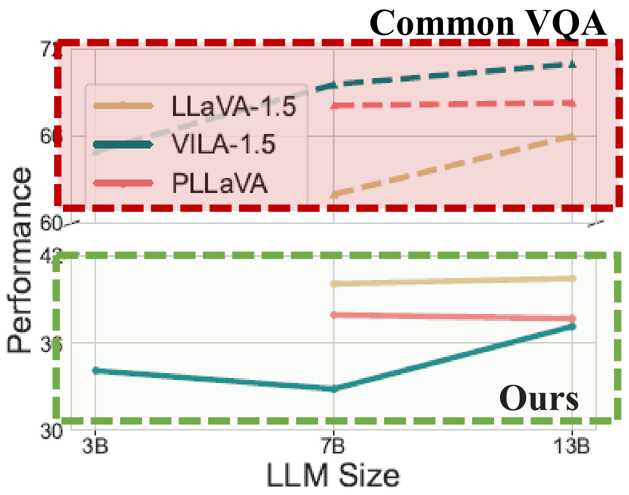
② Especially the understanding of environment and physical dynamics is poor

① The best o1 has only 55% success rate.

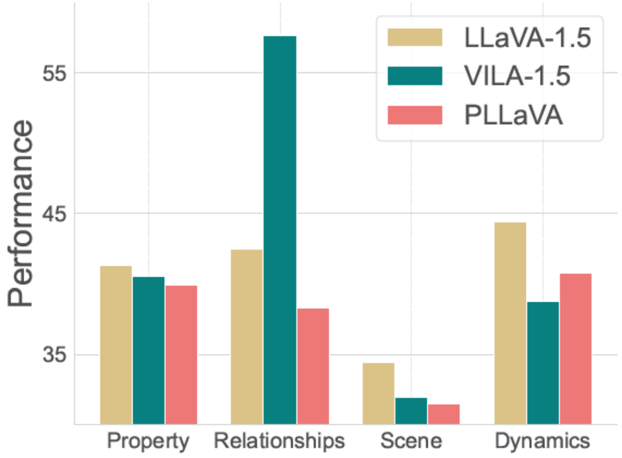
For detailed results for 74 VLMs

● Main Results

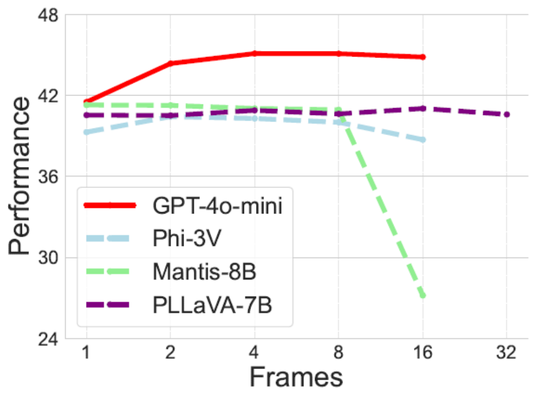
PhysBench **cannot** be improved simply through **scalability**



a) is it model size scalability ?



b) is it data scalability ?

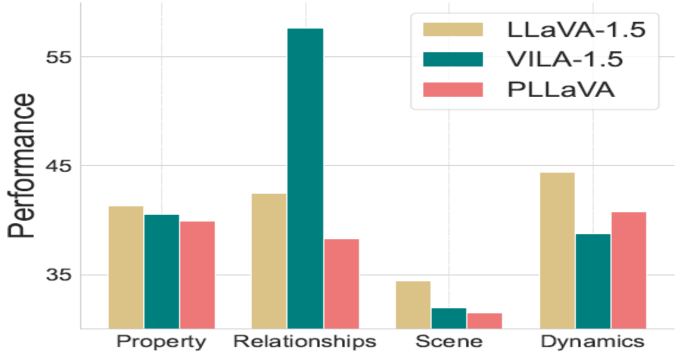


c) Is frame scalability ?

Why not?

● Main Results

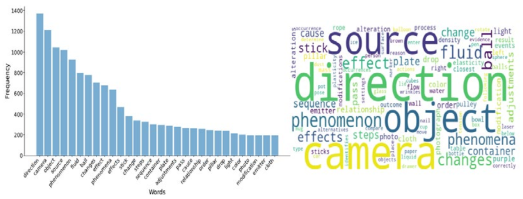
The training data for VLMs largely consists of **descriptions** of visual content, lacking physical principles and priors, which is likely a **reason** contributing to their subpar performance.



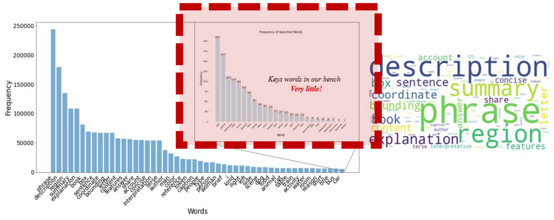
LLaVA $\xrightarrow{\text{+ more data}}$ VILA (MMC4 + SFT data)

PLLaVA (video data)

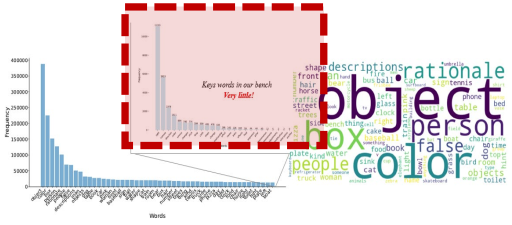
Key words of PhysBench in LLaVA



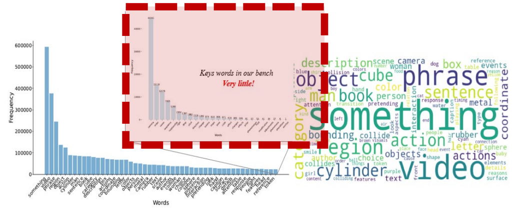
PhysBench words



LLaVA words



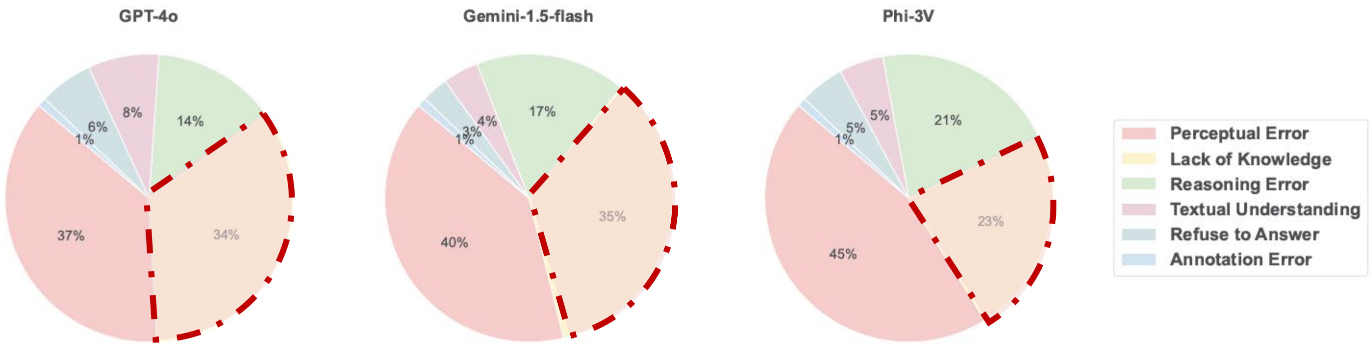
VILA words



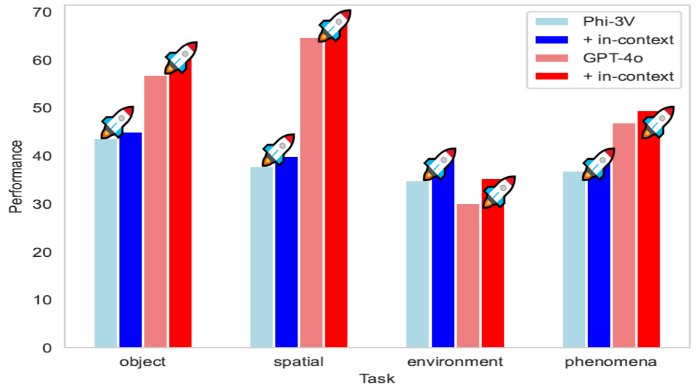
PLLaVA words

● Error Analysis

We found that the major sources of errors are the **lack of physical knowledge** and the **wrong perception** from visual inputs.



a) Error distribution



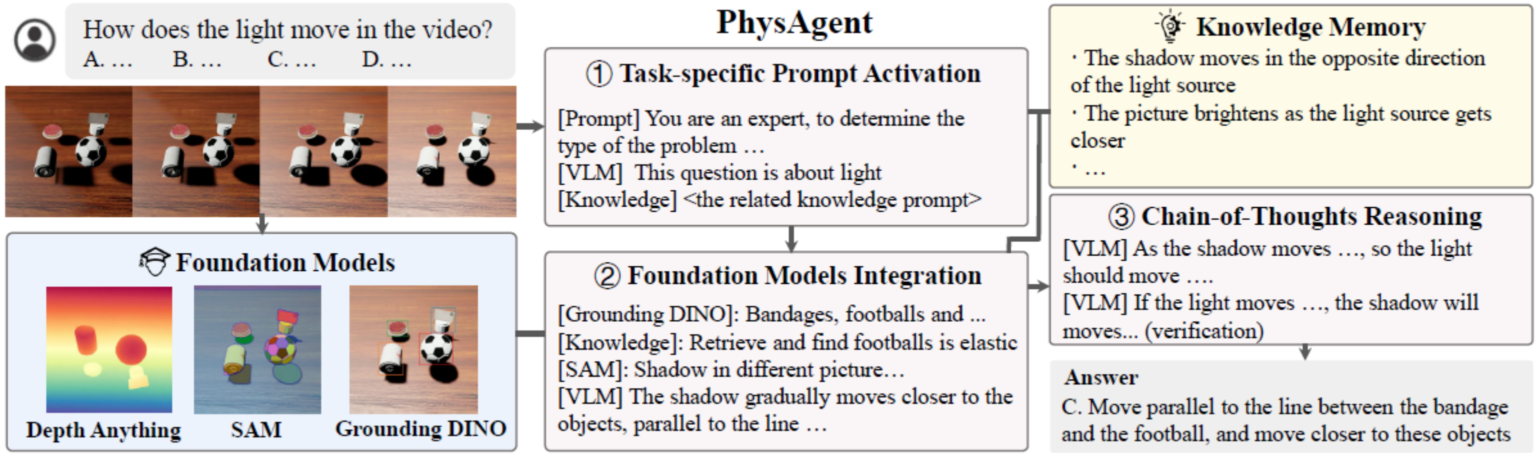
b) Physics knowledge transfer study.

● PhysAgent

To address that, we propose PhysAgent:

1 **Lack Knowledge**: Apply predefined abstract physical world knowledge

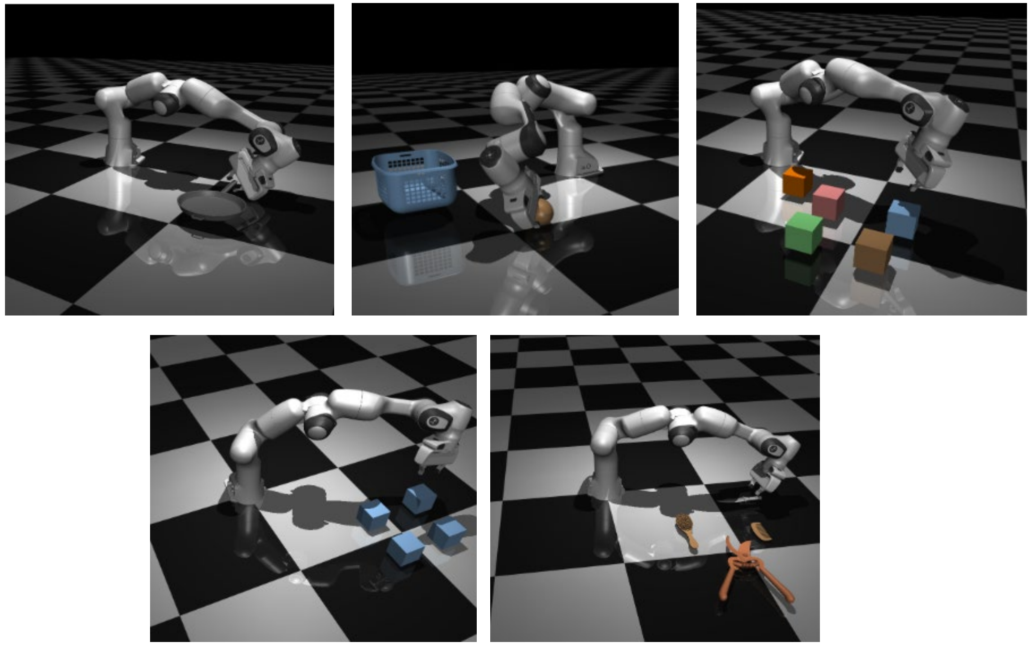
2 **Perception Error**: Advanced expert models

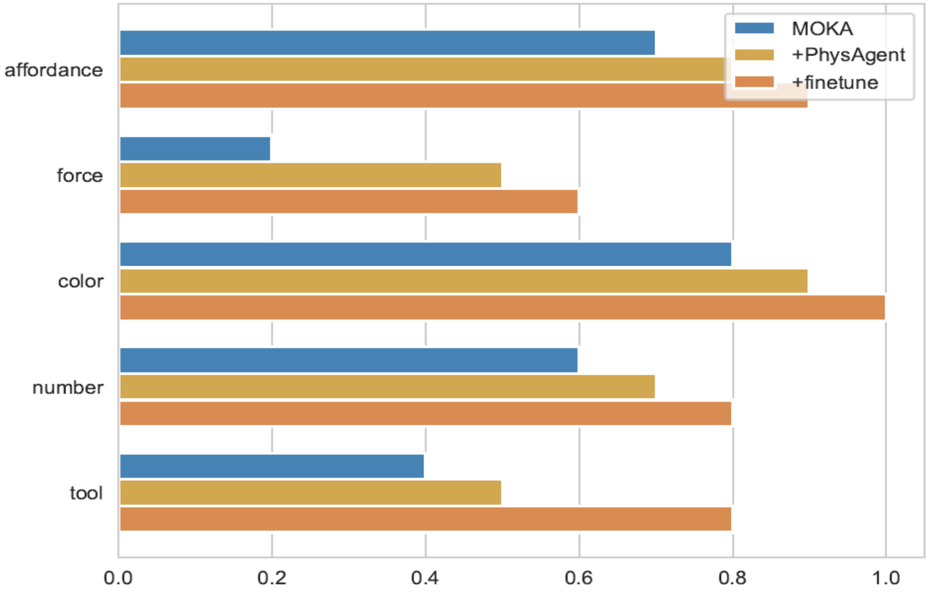


	Property	Spatial	Envir.	Phe.
Phi-3V	38.5	34.4	31.6	32.5
+ CoT	38.8	34.7	31.1	31.9
+ Desp-CoT	25.1	24.1	18.9	21.2
+ PLR	23.1	23.9	19.3	17.1
+ PhysAgent	**44.5**	**47.0**	**38.6**	**37.1**
ViperGPT	52.1	52.9	37.2	42.8
GPT-4o	53.7	61.7	27.0	34.3
+ CoT	54.5	63.2	26.4	35.1
+ Desp-CoT	51.1	58.8	27.2	32.1
+ PLR	37.8	46.2	15.4	22.1
+ PhysAgent	**58.4**	**84.2**	**45.0**	**51.3**

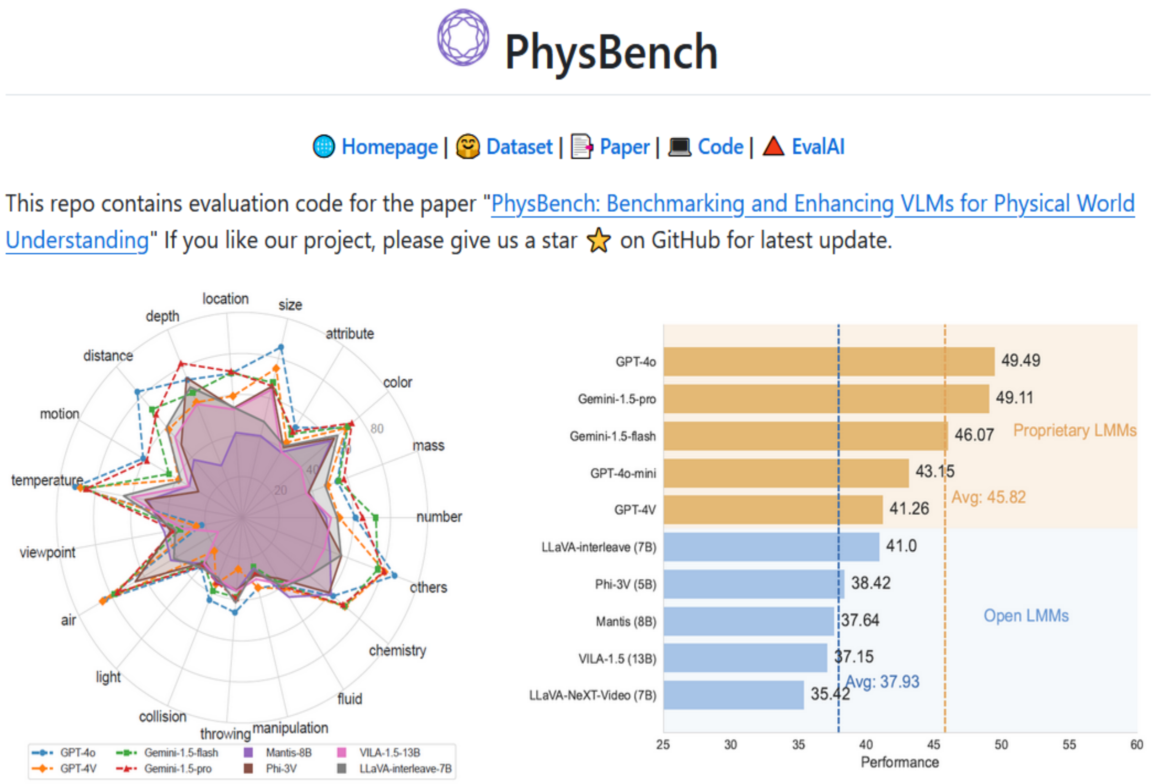
● Embodied Tasks

To further verify the effectiveness of our method and data, we also conducted experiments on **5 robotics tasks**.





● Open sources


PhysBench

Homepage | Dataset | Paper | Code | EvalAI

This repo contains evaluation code for the paper "PhysBench: Benchmarking and Enhancing VLMs for Physical World Understanding" If you like our project, please give us a star ⭐ on GitHub for latest update.

- **Code**
https://github.com/USC-GVL/PhysBench
- **Data**
https://huggingface.co/datasets/USC-GVL/PhysBench
- **Eval Planform**
https://eval.ai/web/challenges/challenge-page/2461/overview


Thanks