PAC Bench: Do Foundation Models Understand Prerequisites for Executing Manipulation Policies?

Anonymous Authors

Abstract

Vision-Language Models (VLMs) are increasingly pivotal for generalist robot manipulation, enabling tasks such as physical reasoning, policy generation, and failure detection. However, their proficiency in these high-level applications often assumes a deep understanding of low-level physical prerequisites, a capability that is largely unverified. To perform actions reliably, robots must comprehend intrinsic object properties (e.g., material, weight), action affordances (e.g., graspable, stackable), and physical constraints (e.g., stability, reachability, or an object's state like being closed). Despite their ubiquitous use in manipulation, we argue that off-the-shelf VLMs may lack this granular, physically-grounded understanding, as these specific prerequisites are often overlooked in their pre-training. Addressing this critical gap, we introduce **PAC Bench**, a comprehensive benchmark designed to systematically evaluate VLM comprehension of these core **P**roperties, **A**ffordances, and **C**onstraints (PAC) from a task executability perspective. PAC Bench features a diverse dataset with over 30,000 annotations, comprising 673 real-world images (115 object classes, 15 property types, 1–3 affordances defined per class), 100 real-world humanoid-view scenarios and 120 unique simulated constraint scenarios across four tasks. Our evaluations reveal significant gaps in the ability of VLMs to grasp fundamental physical concepts, underscoring their current limitations for reliable robot manipulation and pointing to key areas that require targeted research. PAC Bench also serves as a standardized benchmark for rigorously evaluating VLM physical reasoning and guiding the development of more robust and physically grounded models for robotic manipulation. Project Page: Anonymous Project Webpage.

1 Introduction

The quest for generalist robots capable of intelligently and safely interacting with the complexities of the physical world represents a grand challenge in artificial intelligence. Recent breakthroughs in Large Language Models (LLMs) and Vision-Language Models (VLMs) have catalyzed remarkable progress, particularly enabling the development of versatile Vision-Language-Action (VLA) [1, 2, 3]. These systems leverage the powerful representational capabilities of pre-trained models to interpret multimodal sensory input, generate language-grounded plans, and execute a diverse range of manipulation tasks, showcasing impressive generalization. However, their impressive capabilities often mask a critical, yet largely unverified, assumption: that the underlying foundation models possess a sufficiently deep and physically grounded understanding of the fundamental prerequisites for safe, effective, and truly generalizable manipulation.

This assumption demands rigorous scrutiny. Foundation models, despite their exposure to vast quantities of text and video, often lack explicit grounding in the fine-grained physical interplay of objects, actions, and their environmental context knowledge that is intuitive to humans and essential for robust robotic interaction. Consequently, high performance on standard vision-language benchmarks (e.g., VQA [4]) does not reliably translate to the nuanced physical reasoning required

Preprint. Under review.

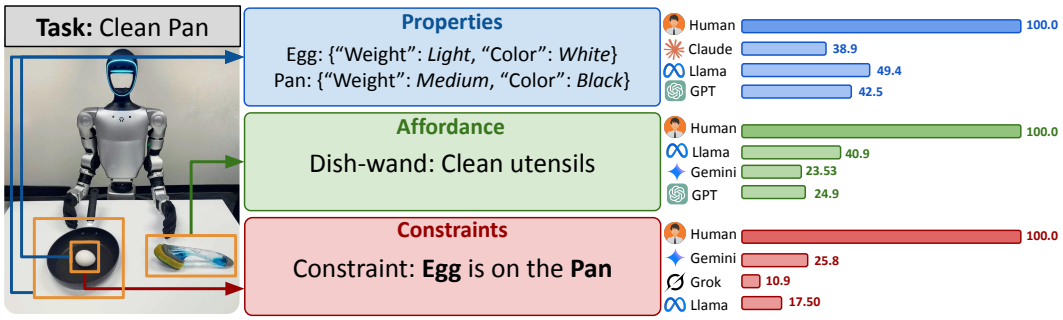

Figure 1: Evaluating foundation models' grasp of Properties, Affordances, and Constraints (PAC) for robotic manipulation. (Left) PAC Bench uses scenarios requiring nuanced physical understanding. (Right) We present example performance of leading VLMs (e.g., GPT-4o, Llama, Claude, Deepseek) on tasks related to Properties (blue), Affordances (green), and Constraints (red), indicating varied strengths and weaknesses across these fundamental reasoning skills.

to anticipate action outcomes or adapt to novel physical scenarios. Before a robot can confidently execute any manipulation, it must implicitly or explicitly reason about the world: assessing intrinsic object **P**roperties (e.g., Is it heavy? Is it fragile?), discerning valid action **A**ffordances (e.g., Can this be stacked?), and recognizing critical physical **C**onstraints (e.g., Is the target reachable without collision?). Relying on superficial correlations learned from web-scale data, without a robust grasp of these Properties, Affordances, and Constraints (PAC), can lead to unpredictable failures, unsafe operations, and a fundamental brittleness that severely limits their deployment in safety-critical or economically vital open-world applications. As these powerful models are increasingly positioned at the core of autonomous systems, rectifying these gaps in physical understanding is not merely an academic pursuit but a prerequisite for trustworthy and scalable robotic intelligence.

Despite the critical importance of this granular physical understanding, existing benchmarks predominantly focus on end-to-end task performance [5], broad physical knowledge question-answering [6, 7], or other aspects of model behavior like trustworthiness [8] or safety from a policy perspective [9]. A targeted evaluation framework to specifically dissect and measure foundation models' comprehension of the core *prerequisites* for manipulation has been notably absent. This absence hinders targeted improvements, as developers lack precise diagnostics to identify *why* end-to-end policies fail or *which specific aspects* of physical reasoning are underdeveloped in their foundation models.

To bridge this crucial diagnostic gap, we introduce **PAC Bench** (Figure 1): the first benchmark meticulously engineered to evaluate foundation models' understanding of **Properties**, **Affordances**, and **Constraints** essential for robotic manipulation. PAC Bench moves beyond holistic task success by decomposing physical reasoning into these three core, queryable components. Through a diverse suite of targeted evaluations across both simulated and real-world scenarios, our benchmark enables researchers to pinpoint specific deficiencies in models' internal representations of the physical world. We envision PAC Bench not just as an evaluation tool, but as a catalyst for a new wave of research into building more robustly and verifiably grounded foundation models. This detailed diagnostic capability is vital for accelerating the development of VLA systems that can reason causally about their actions, adapt to unforeseen circumstances, and ultimately operate with greater safety and efficacy, advancing the frontier of general-purpose robotics. Our primary contributions are as follows.

1. A benchmark featuring over 30,000 annotations of real scenarios targeting the essential Properties, Affordances, and Constraints for robotic manipulation.
2. A comprehensive suite of tasks and metrics for fine-grained assessment of VLM physical understanding across the three PAC dimensions.
3. Extensive empirical results highlighting current VLM capabilities and critical limitations in PAC reasoning, offering a clear path for advancing physically grounded AI.

2 Related Work

The rapid evolution of LLMs and VLMs has spurred a critical need for comprehensive evaluation methodologies. General frameworks like HELM [10] and its visual counterpart VHELM [11] provide holistic assessments across a wide array of tasks and capabilities. Complementing these,

Table 1: Comparison of benchmarks evaluating physical properties (P), affordances (A), constraints (C), or related concepts. Manip: Manipulation focus. Sim/Real/Human: Data sources. (*) PhysBench evaluates rudimentary dynamics/physical relationships which imply some constraint understanding, differing in scope from PAC Bench's direct constraint prerequisites for manipulation.

	Concepts Evaluated			Focus	Data Source			Access	Size	
Benchmark	**P**	**A**	**C**	**Manip**	**Sim**	**Real**	**Human**	**Open Data**	**Size (GB)**	**(# Points)**
PAC Bench (Ours)	✓	✓	✓	✓	✓	✓	✓	✓	∼10	30,529 images-text
PhysBench [7]	✓	✓	✓*	×	✓	✓	✓	✓	∼10	10,002 video-image-text
ActAffordance [14]	×	✓	×	✓	×	✓	✓	✓	25–40	278,000 Images
EQA-phys [15]	×	×	✓	✓	✓	✓	×	✓	<1	1,300 Q&A
Physion [16]	×	×	✓	×	✓	×	×	✓	∼5	1,200 Examples
ManipVQA [17]	✓	✓	×	✓	×	✓	✓	✓	∼20	84,000 Examples
UniAff [18]	✓	✓	✓	✓	✓	✓	✓	✓	3–5	1,500 Objects
NrVLM [19]	×	✓	×	✓	✓	×	×	✓	5–10	4,500 Episodes
PHYBench [6]	✓	×	✓	×	✓	×	×	✓	<1	∼500 Problems

numerous benchmarks target specific facets of foundation models, such as trustworthiness with DecodingTrust [8], safety through regulatory lenses with Air-Bench [9], domain-specific reliability in medicine with CARES [12], and agentic capabilities in scientific discovery with MLAgentBench [13]. Public leaderboards further track ongoing performance on various safety and ethical dimensions[1]. While these efforts are crucial for understanding the broader landscape of model behavior, they do not typically delve into the nuanced, granular physical common sense specifically required as prerequisites for robust robotic manipulation.

Closer to the domain of robotics and physical interaction, several benchmarks have begun to probe foundation models' understanding of the physical world. Some focus on general physics knowledge or predictive capabilities. For instance, PHYBench [6] primarily uses text-based scenarios to assess LLMs on formal physics problems, while Physion [16] evaluates visual physical prediction, implicitly testing understanding of object properties and physical constraints governing dynamics. PhysBench [7] offers a broader multimodal evaluation of VLMs, covering aspects like explicit object properties, object relationships, scene understanding, and rudimentary physical dynamics, thus touching upon elements of properties, affordances (via relationships), and constraints (via dynamics).

Other research lines target more specific components of physical understanding relevant to manipulation. For affordances, ManipVQA [17] injects affordance knowledge into VLMs alongside property understanding, ActAffordance [14] focuses on learning bimanual affordances from human videos, and NrVLM [19] develops benchmarks for affordance-guided manipulation based on fine-grained language instructions. UniAff [18] proposes a unified representation for affordances, especially for tool use, and importantly, also incorporates the reasoning of 3D motion constraints and object properties within its framework. For constraints, EQA-phys [15] specifically targets VLM understanding of robotic physical reachability. Distinct from these, benchmarks like The Colosseum [5] are vital for assessing the generalization of end-to-end robotic manipulation policies to various environmental perturbations, rather than the underlying conceptual understanding of physical prerequisites.

Despite this valuable landscape (summarized and compared in Table 1), a critical gap remains: a dedicated, fine-grained benchmark that systematically evaluates whether foundation models comprehend the fundamental and *interconnected prerequisites* for executing manipulation actions, specifically framed through object properties, action affordances, and physical constraints. While works like PhysBench [7] and UniAff [18] evaluate aspects across P, A, and C (as indicated in Table 1), PAC Bench distinguishes itself through several key dimensions. First, it focuses on the explicit, understanding of these three components as *preconditions* for action, rather than evaluating them solely through downstream task performance. Second, PAC Bench is designed to assess these PAC dimensions with a granularity specifically tailored to common manipulation scenarios, supported by a dataset that combines diverse real-world images (for properties and affordances) with both simulated and novel real-world humanoid-view scenarios (for constraints). While existing benchmarks may test general physics knowledge, dynamic prediction, or policy generalization, PAC Bench fundamentally probes whether VLMs can reason about the specific P, A, and C conditions that make a manipulation task executable in the first place, a crucial step towards building more robust, and safe VLA systems.

[1] `https://huggingface.co/spaces/AI-Secure/llm-trustworthy-leaderboard`

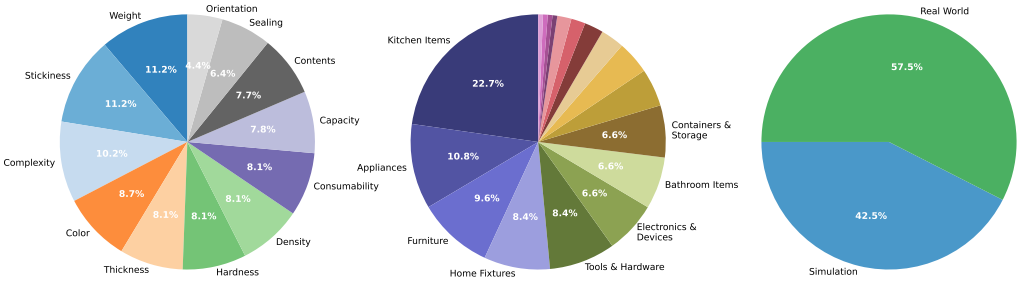

Figure 2: Distribution of annotations in PAC Bench across three dimensions: (Left) physical properties annotated in the dataset, showing the relative frequency of each property; (Center) affordance categories, with slices below 5% omitted for clarity; (Right) constraint domains, contrasting simulation (blue shades) and real-world (green shades) scenarios.

3 The PAC Bench Dataset

PAC Bench evaluates a VLM's understanding of three fundamental, interdependent components crucial for determining the executability of robotic manipulation actions:

1. **Properties:** These are the inherent physical or material characteristics of objects, as well as their states, that dictate how they behave and can be interacted with. In PAC Bench, we focus on a comprehensive suite of 12 distinct physical and material attributes, including, for instance, an object's inferred *Weight* (e.g., light, medium, heavy), its *Material* (e.g., wood, metal, plastic), its *Containment State* (e.g., lidded, open, sealed). Accurately perceiving these properties is the first step towards effective physical reasoning.
2. **Affordances:** Affordances describe the potential for action that an object offers to an agent, or that an agent can enact upon an object [20, 21], given its properties and the broader environmental context. These are specifically tailored to manipulation, covering common interactions such as *is-graspable* (by a standard gripper), *is-containable-in* (for placing objects), and *is-stackable-on* (another object). Understanding affordances bridges the gap between object perception and actionable knowledge.
3. **Constraints:** These are the physical, geometric, or environmental limitations and conditions that govern whether an intended action can be successfully executed given a task. Failure to recognize constraints often leads to task failure or unsafe robot behavior. PAC Bench evaluates understanding of constraints such as *stability limits* (e.g., predicting if stacking a specific object will cause a topple), *containment failure* (e.g., contents spilling if an open container is moved inappropriately), and *reachability issues for a robotic arm*.

A grounded understanding of these three pillars – Properties, Affordances, and Constraints – is paramount for any robotic system intended to operate robustly in the complexities of the real world. Without it, even sophisticated policies are prone to errors stemming from a superficial interpretation of the scene. For instance, attempting to lift an object perceived as light (misjudged Property) might fail if it is actually heavy. Similarly, trying to stack an object that appears stackable (misjudged Affordance) might lead to collapse if its instability (unrecognized Constraint) is not considered. PAC Bench is therefore designed to specifically probe these interconnected concepts, offering a more targeted benchmarks focusing on broader physics knowledge. By focusing on PAC, we aim to evaluate the foundational understanding that enables models to predict action feasibility *before* execution, a critical component for building more reliable and adaptable robotic agents.

3.1 Data Acquisition and Curation

The PAC Bench dataset is constructed through a multi-faceted approach, aggregating data from diverse real-world image sources and meticulously designed scenarios from both simulated and real-world robot interactions (Fig 2). This hybrid strategy ensures broad visual diversity for property and affordance, complemented by targeted and varied constraint evaluations from multiple perspectives.

Data for Properties and Affordances: Diverse Real-World and Simulated Imagery. To ground our property and affordance assessments in varied visual data, PAC Bench aggregates images from four key sources (2 real, 2 simulation): the extensive *OpenImages Dataset V7 and Extensions* [22], novel

real-world captures from multiple perspectives (agent and side views) of a *Unitree G1 humanoid robot*, multi-angle(24) capture of 45 unique objects from the *RoboCasa framework* [23], which leverages the MuJoCo physics engine for structured household environments. Across these sources, we target **115 unique object classes** (e.g., *Container, Towel, Chair, Apple, Knife*), selected for their prevalence and relevance to household manipulation. These are organized into **18 primary categories** (e.g., Appliances, Furniture, Kitchen Items; see Appendix C.2 for full taxonomy). We utilized the provided human-annotated bounding boxes for annotations. This ensures precise localization for our subsequent PAC annotations. For the VLM evaluations detailed in this paper, we curated **977** images from OpenImages and our Unitree G1 captures. The RoboCasa image data (1080 unique images), while part of the full PAC Bench dataset release to support broader research, is not included in the current VLM evaluation set due to computational costs.

Property Annotation: For each of the 977 curated images, we annotated a comprehensive suite of intrinsic and extrinsic physical properties. We defined a set of **12 distinct property types** relevant to manipulation, including: *Stickiness, Thickness, Density, Sealing, Contents, Capacity, Complexity of Parts, Consumability, Orientation, Hardness, Color*, and *Weight*. This resulted in a total of **27,674 property annotations** across the dataset. (Detailed definitions are shown in Appendix C.1.) The property annotation process was designed for high quality. Each image instance, along with a specific property query (e.g., "What is the material of the object in the bounding box?"), was presented to annotators with a set of predefined, mutually exclusive answer choices. To ensure reliability, every image instance was independently annotated for each property by **two human annotators**. The final ground-truth label for a given property was determined by consensus, requiring agreement between both annotators. Disagreements were resolved by a senior annotator or discarded if no consensus could be reached. This rigorous process yielded a high-quality set of property labels. We utilized *LabelBox* as our annotation platform, with a team of over **10 annotators** contributing to this effort (Appendix E.1).

Affordance Annotation: For each of the 115 selected object classes, we also collected affordance labels. The process involved manually identifying and listing the **top three most common action affordances** associated with each object class, ranked in order of typicality or importance. For example, for the object class *Chair*, the annotated affordances include (1) *is-sittable*, (2) *is-climbable*, and (3) *can-place-objects-on*. This initial phase of affordance annotation was conducted by assigning each object class to a primary annotator. This initial phase of affordance annotation involved a primary annotator per object class. While this provides a foundational set of common affordances, we acknowledge that future work will involve expanding this with multiple annotators to establish inter-annotator agreement and a consensus-based label set.

Data for Constraints: Simulated and Real-World Humanoid Scenarios. To evaluate the understanding of physical constraints often involving complex or dynamic interactions PAC Bench incorporates data from both simulated environments and the real-world humanoid robot perspectives. This hybrid strategy allows for scalable, controlled generation of diverse constraint types in simulation, ideal for iterative VLM testing and aligning with common policy training paradigms. These are complemented by authentic real-world humanoid scenarios that ground evaluations in genuine physical complexities and robot-centric perspectives, offering a crucial testbed for sim-to-real transfer of constraint comprehension. (Detailed specifications for all constraint domains are in Appendix C.3).

Simulated Constraint Scenarios: We leveraged the MuJoCo physics engine [24] to generate synthetic scenarios depicting various constraints (Fig 3 (left)) relevant to robotic manipulation. We designed four primary constraint domains:

1. **Impossible Placement:** Scenarios where an object cannot be stably placed on another due to factors like shape, size mismatch, or unstable support.
2. **Occlusion/Support Issues:** Challenges related to accessing an object, such as attempting to pick up a target block that is currently supporting another block.
3. **Stability Constraints:** Situations involving picking up an object that is itself part of an unstable assembly.
4. **Reachability and Access Constraints:** Scenarios where an object is present but difficult or impossible to reach due to its position or surrounding obstacles.

For each simulated constraint domain, we procedurally generated **10 distinct environment instantiations** by introducing randomization in object positions, orientations, and/or distractor elements. Each instantiation was rendered from **three different camera viewpoints** (front, agent, and side view) to

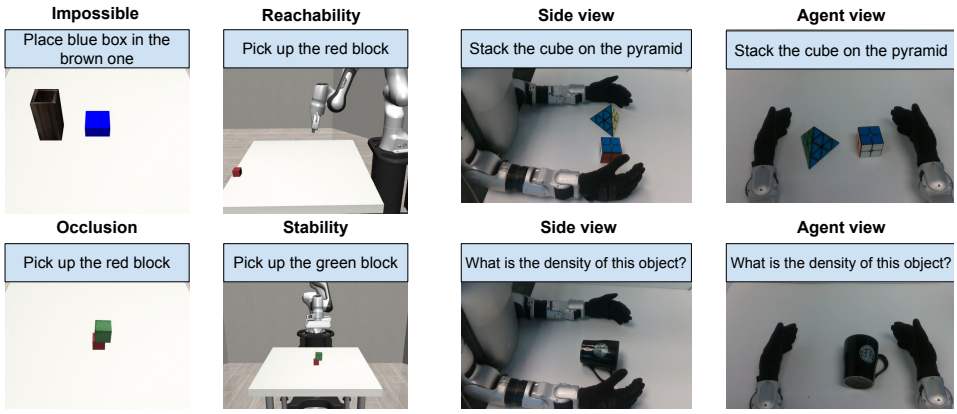

Figure 3: Examples from PAC Bench. (Left 4 Images) Scenarios designed to evaluate understanding of various physical **C**onstraints: Impossible Placement, Occlusion, Reachabilityand Stability. (Right 4 images) Example of a **P**roperty query presented with a real-world robot view from PAC Bench.

provide visual diversity and assess view-invariance. This resulted in a total of **120 unique simulated constraint scenarios**.

Real-World Humanoid Constraint Scenarios: These scenarios involve a dual-arm Unitree G1 humanoid robot attempting simple manipulation tasks in tabletop environments with everyday objects. For each scenario, we captured synchronized images from two camera views. A question was then formulated about a potential action and the physical constraints that might prevent its successful execution (see Appendix D.1 for an example prompts). The ground-truth answer provides an explanation of the relevant constraint(s). This real-world component currently comprises **2727** unique question-answer scenarios, focusing on constraints such as (Question: Can you keep the food on the plate? Expected Answer: No the plate is inverted.). (Appendix C.3 provides further examples.)

4 Experimental Results and Analysis

In this section, we present the empirical evaluation of several state-of-the-art foundation models [25, 26, 27, 28, 29, 30, 31, 32, 33] on PAC Bench. We detail our experimental setup, followed by an analysis of model performance on understanding object properties, action affordances, and physical constraints.

4.1 Experimental Setup

Models Evaluated: We evaluated a diverse suite of publicly available and proprietary VLMs to assess their PAC understanding capabilities. For some models, we also explored different prompting strategies (e.g., direct querying vs. chain-of-thought prompting in Appendix D).

Evaluation Protocol: For each task in PAC Bench, VLMs were provided with images from a scenario and a textual prompt (Appendix D.1) that queries a specific property, affordability, or constraint. Model responses were evaluated against ground-truth annotations derived from our dataset.

1. *Property* questions were multiple-choice (typically [Number, e.g., 3-5] options) targeting one of 12 predefined attributes for a specified object (e.g., "What is the density of the object in the box? A) High, B) Low...").
2. *Affordance* questions required models to provide all applicable affordances for a given object class (e.g., "What are the affordances of [object]? A) Can carry items, B) is-stackable...").
3. *Constraint* questions asked models to determine the feasibility of an action or identify the most constraining pre-condition to successfully complete a task. (e.g., "Can the robot stack X on Y? If no, why?").

4.2 Analyzing Property Awareness: Do VLMs Discern Fundamental Object Features?

This subsection presents a detailed evaluation of how well contemporary VLMs are grounded in these essential attributes. We assessed model performance across twelve distinct property categories critical for robotic manipulation: P1 (Capacity), P2 (Color), P3 (Complexity), P4 (Consumability),

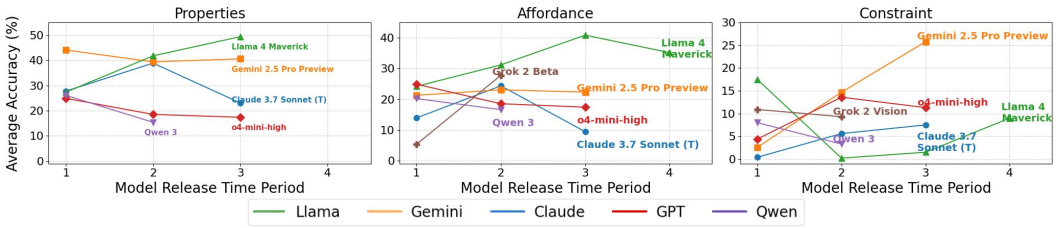

Figure 4: Comparative PAC understanding profiles of selected VLMs. The x-axis indicates nominal model release time periods (1=earlier to 4=most recent among those shown). The diverse performance signatures suggest varied developmental trajectories in acquiring physical common sense.

Table 2: Property accuracies (%) for Open Images (PAC Bench) vs. Humanoid benchmarks. Properties P1–P6 are: Color, Contents, Weight, Density, Sealing, Hardness.

	Open Images						Humanoid						Avg
Model	**P1**	**P2**	**P3**	**P4**	**P5**	**P6**	**P1**	**P2**	**P3**	**P4**	**P5**	**P6**	
Claude 3.5 Sonnet	0.0	31.9	0.0	0.0	2.7	42.3	50.2	28.9	**50.7**	52.7	19.4	55.2	**27.8**
Claude 3.7 Sonnet	20.2	23.5	32.6	36.7	66.4	37.0	47.8	30.3	48.3	55.7	13.2	55.7	**38.9**
Claude 3.7 Sonnet (T)	6.7	22.3	15.0	9.0	50.9	23.4	24.9	11.9	28.5	36.3	8.3	39.8	**23.1**
Gemini 2.0 Flash 001	19.7	**35.3**	**40.8**	**58.0**	56.1	**43.9**	**55.2**	39.8	40.3	46.8	38.2	54.7	**44.1**
Gemini 2.5 Flash P	26.9	28.8	27.1	40.1	58.9	31.1	53.2	27.9	33.8	40.3	41.8	63.2	**39.4**
Gemini 2.5 Pro Pre (T)	27.0	34.1	31.2	43.2	57.2	16.7	13.0	42.7	49.5	55.7	53.5	64.0	**40.6**
GPT-4.1 Mini	26.6	28.4	24.1	43.2	64.0	18.1	36.3	36.3	26.9	40.3	15.3	60.2	**35.0**
GPT-4.1	13.8	29.0	4.4	25.9	**91.0**	27.8	51.2	55.7	43.3	**58.2**	**43.8**	**64.2**	**42.4**
o4-mini-high (T)	17.1	0.2	4.7	26.4	72.7	26.2	20.4	36.6	31.5	52.7	43.1	63.8	**33.0**
Llama 3.2 90B Vision I	13.1	14.8	4.2	25.0	30.2	12.8	37.3	51.2	31.3	44.8	27.1	37.3	**27.4**
Llama 4 Scout	30.4	0.6	36.4	51.1	84.9	18.6	51.2	60.2	37.8	43.3	36.1	51.2	**41.8**
Llama 4 Maverick	**36.2**	34.9	37.6	47.0	90.0	14.6	43.8	**77.1**	59.2	57.7	40.3	54.2	**49.4**
Qwen 3	18.7	22.7	9.9	20.1	85.2	28.6	0.0	0.0	0.0	0.0	0.0	0.0	**15.4**
Qwen 2.5 VL	21.9	20.7	18.7	9.6	61.8	42.3	47.8	22.9	24.9	4.5	2.8	35.8	**26.1**

P5 (Contents), P6 (Density), P7 (Hardness), P8 (Orientation), P9 (Sealing), P10 (Stickiness), P11 (Thickness), and P12 (Weight). The comprehensive results are detailed in Table 7.

Overall Property Performance and Domain Sensitivity: Table 2 reveals considerable VLM performance disparities across models and, notably, between the Open Images and Humanoid data subsets for the six evaluated properties. No single model masters all properties across both domains, highlighting varied strengths and significant domain sensitivity. Many models, such as 'Claude 3.5 Sonnet' and 'GPT-4.1', demonstrate decent accuracy on properties like 'Color (P1)', 'Weight (P3)' when evaluated on Humanoid views compared to the more varied Open Images data. Conversely, properties like 'Sealing (P5)' frequently see higher scores on Open Images (e.g., 'Llama 4 Maverick': 90.0% vs. 40.3%). Some models show extreme domain dependence; for instance, 'Qwen 3' performs reasonably on Open Images but scores 0.0% across all six properties on the Humanoid dataset. These findings underscore that VLM property understanding is not yet consistently robust across different visual contexts, even for fundamental attributes, pointing to challenges in generalization. For detailed results across all 12 evaluated properties from our primary dataset, see Appendix D.2.

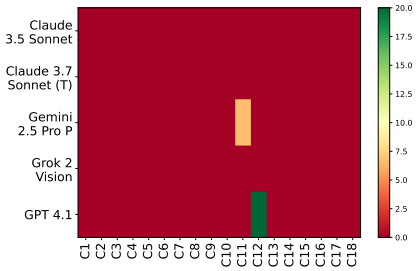

Figure 5: All affordance subset heatmap. Full heatmap in Appendix D.3

4.3 Evaluating Affordance Understanding: Can VLMs Discern Possible Interactions?

Recognizing potential actions, or affordances, that an object offers is fundamental for goal-oriented manipulation. In this subsection, we assess VLM performance on identifying common affordances for 115 object classes, primarily grouped into 14 primary categories derived from web-scale images (A1-A14). Table 3 presents results for the metric of identifying *at least one* correct affordance, and importantly, also includes overall accuracies from our distinct Humanoid dataset evaluations and

Table 3: Affordance Accuracy (%) of VLMs on recognizing at least one correct affordance for objects grouped by primary categories (Single-Category Mapping) in PAC Bench, plus overall accuracy in the humanoid dataset scores H1–H3. Categories A1–A18 are: A1 (Adhesives), A2 (Appliances), A3 (Luggage), A4 (Bathroom Items), A5 (Cleaning), A6 (Clothing), A7 (Storage), A8 (Decor), A9 (Electronics), A10 (Food & Beverage), A11 (Furniture), A12 (Home Fixtures), A13 (Kitchen Items), A14 (Instruments), H1 (Humanoid Front View), H2 (Side View), H3 (Both Views)

Model	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	A13	A14	H1	H2	H3	Avg
Claude 3.5 Sonnet	0.0	0.0	0.0	0.0	0.0	0.0	16.7	25.0	0.0	66.7	13.3	40.0	9.1	0.0	2.9	47.1	14.7	**13.9**
Claude 3.7 Sonnet	**100.0**	0.0	0.0	20.0	0.0	0.0	0.0	0.0	11.1	66.7	0.0	20.0	22.7	100.0	2.9	58.8	11.8	**24.4**
Claude 3.7 Sonnet (T)	0.0	5.6	0.0	30.0	0.0	0.0	0.0	0.0	11.1	0.0	6.7	20.0	18.2	0.0	2.9	54.4	10.3	**9.4**
Gemini 2.0 Flash 001	0.0	0.0	0.0	40.0	0.0	0.0	16.7	0.0	0.0	66.7	0.0	40.0	13.6	0.0	54.4	66.2	64.7	**21.3**
Gemini 2.5 Flash P	0.0	5.6	0.0	20.0	0.0	50.0	0.0	0.0	11.1	66.7	13.3	40.0	18.2	0.0	52.9	55.9	57.4	**23.0**
Gemini 2.5 Pro P	0.0	16.7	66.7	30.0	0.0	0.0	33.3	25.0	22.2	66.7	26.7	60.0	31.8	0.0	0.0	0.0	0.0	**22.3**
Llama 3.2 11B Vision I	**100.0**	**22.2**	0.0	30.0	0.0	50.0	33.3	0.0	22.2	66.7	0.0	0.0	13.6	0.0	20.5	27.9	25.0	**24.2**
Llama 3.2 90B Vision I	**100.0**	11.1	33.3	10.0	0.0	50.0	50.0	25.0	22.2	66.7	26.7	60.0	9.1	0.0	22.1	44.1	0.0	**31.2**
Llama 4 Scout	0.0	11.1	66.7	50.0	0.0	50.0	50.0	25.0	33.3	66.7	53.3	60.0	54.6	100.0	20.6	27.9	26.5	**40.9**
Llama 4 Maverick	0.0	**22.2**	33.3	50.0	0.0	100.0	50.0	0.0	33.3	66.7	26.7	100.0	31.8	0.0	20.6	39.7	23.5	**35.2**
GPT 4.1 Mini	0.0	5.6	0.0	30.0	0.0	0.0	50.0	25.0	0.0	100.0	13.3	60.0	36.4	0.0	20.6	57.4	25.0	**24.9**
GPT 4.1	0.0	5.6	0.0	20.0	0.0	0.0	16.7	25.0	0.0	0.0	6.7	60.0	18.2	0.0	48.5	67.6	45.6	**18.5**
o4-mini-high (T)	0.0	16.7	0.0	20.0	0.0	0.0	16.7	25.0	11.1	33.3	33.3	20.0	22.7	0.0	16.2	45.6	35.3	**17.4**
Qwen 2.5 VL	0.0	0.0	0.0	30.0	0.0	0.0	33.3	0.0	0.0	100.0	6.7	80.0	9.1	0.0	14.7	48.5	20.6	**20.2**
Qwen 3	0.0	5.5	0.0	30.0	0.0	0.0	33.3	25.0	0.0	100.0	0.0	60.0	13.6	0.0	4.4	1.4	8.8	**16.6**
Grok Vision Beta	0.0	5.6	0.0	10.0	0.0	0.0	0.0	0.0	11.1	0.0	13.3	20.0	4.6	0.0	8.8	8.8	7.4	**5.3**
Grok 2 Vision	0.0	5.6	33.3	50.0	0.0	0.0	0.0	0.0	11.1	100.0	6.7	20.0	13.6	100.0	44.1	47.1	41.2	**27.8**

an aggregated average. The stricter metric, requiring identification of *all* ground-truth affordances, is detailed in Table 5 (further visualized in Figure 5). Additional results for multi-category and per-object evaluations are in Appendix D.3.

Partial Affordance Recognition and Humanoid Insights: As shown in Table 3, VLMs exhibit highly varied success in recognizing at least one correct affordance. Based on the overall average scores (Avg), which combine web-image category and humanoid task performance, models like 'Llama 4 Scout' (40.9%) and 'Llama 4 Maverick' (35.2%) demonstrate broader, albeit still moderate, capabilities. Performance peaks on specific web-image categories are notable, for instance, 'GPT 4.1 Mini' and 'Qwen 2.5 VL' achieve 100% for A10, and several Llama models along with 'Claude 3.7 Sonnet' show perfect scores for A1. However, many categories, such as A2.The Humanoid dataset scores (H1-H3) reveal further nuances; for example, 'Gemini 2.0 Flash 001' performs decent across H1-H3 (avg. 60%), while 'Gemini 2.5 Pro P' scores 0% on all Humanoid tasks despite reasonable performance on A1-A14. This suggests that affordance understanding from diverse web images may not readily transfer to specific robot-centric views or tasks without further adaptation, with models like 'Qwen VP' (0.0% Avg) struggling broadly.

Comprehensive Affordance Recognition Remains Elusive: The capacity of VLMs to identify the *full set* of an object's affordances is far more limited. As starkly illustrated in Table 5 (and the heatmap in Figure 5), when requiring models to recognize all ground-truth affordances, performance plummets to near-zero across almost all models and categories. The rare non-zero scores (e.g., 'GPT 4.1' at 20.0% for Home Fixtures. This significant drop from the "at least one" metric highlights that while VLMs might identify a primary or common affordance, they generally lack the comprehensive functional understanding critical for versatile and truly intelligent robotic interaction.

4.4 Assessing Constraint Comprehension: Can VLMs Understand Physical Limits?

PAC Bench evaluates constraints by presenting VLMs with scenarios where proposed actions might be infeasible due to underlying physical limitations. Our evaluation spans four distinct constraint domains. Furthermore, we introduce a novel set of real-world constraint scenarios captured from a humanoid robot's perspective, which will be analyzed subsequently. The performance of VLMs on the simulated constraint tasks is detailed in Table 4 (More in Appendix D.4).

Constraint Understanding: A Profound Challenge Across Simulated and Real-World Scenarios. The results presented in Table 4 underscore that reasoning about physical constraints remains a profound challenge for current VLMs, with overall average (Avg) accuracies being exceptionally low for most models. Many prominent VLMs, including 'Claude 3.5 Sonnet' (0.4% Avg), 'Llama 3.2 90B Vision I' (0.2% Avg), and 'Llama 4 Scout' (1.5% Avg), register near-zero performance across

Table 4: Constraint Accuracy (%) of VLMs on understanding physical constraints in PAC Bench across four simulated domains, three views (**F**: front-view, **A**: agent-view, **S**: side-view), and a real-world **Humonoid** split (**Both**=A+S).

Model	Simulation												Real World			Avg
	Impossible Place (↑)			Occlusion (↑)			Stability (↑)			Reachability (↑)			Humonoid (↑)			
	F	A	S	F	A	S	F	A	S	F	A	S	A	S	Both	
Claude 3.5 Sonnet	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.8	0.0	3.7	**0.4**
Claude 3.7 Sonnet	0.0	0.0	0.0	40.0	10.0	30.0	0.0	0.0	0.0	0.0	0.0	0.0	1.8	0.0	1.8	**5.6**
Claude 3.7 Sonnet (T)	0.0	0.0	0.0	20.0	20.0	30.0	0.0	0.0	10.0	10.0	0.0	0.0	0.0	0.0	3.7	**7.5**
Gemini 2.0 Flash 001	0.0	0.0	0.0	10.0	10.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	5.6	3.7	9.4	**2.6**
Gemini 2.5 Flash P	0.0	0.0	0.0	50.0	20.0	40.0	10.0	**40.0**	20.0	0.0	20.0	0.0	9.4	9.4	1.8	**14.7**
Gemini 2.5 Pro P	10.0	**20.0**	**10.0**	**90.0**	30.0	60.0	0.0	**40.0**	0.0	**30.0**	0.0	**20.0**	11.3	18.8	9.4	**25.8**
Llama 3.2 11B Vision I	**20.0**	10.0	0.0	30.0	30.0	20.0	20.0	20.0	20.0	10.0	**30.0**	0.0	0.0	1.8	0.0	**17.5**
Llama 3.2 90B Vision I	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	3.7	0.0	0.0	**0.2**
Llama 4 Scout	0.0	0.0	0.0	0.0	0.0	20.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.8	**1.5**
Llama 4 Maverick	0.0	0.0	0.0	10.0	0.0	50.0	**30.0**	10.0	10.0	0.0	0.0	0.0	9.4	7.5	7.5	**9.0**
GPT-4.1	0.0	0.0	0.0	50.0	**70.0**	50.0	0.0	0.0	0.0	0.0	0.0	0.0	11.3	13.2	9.4	**13.6**
GPT-4.1 Mini	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	**18.8**	**24.5**	**22.6**	**4.4**
o4-mini-high (T)	0.0	0.0	0.0	60.0	40.0	50.0	0.0	0.0	20.0	0.0	0.0	0.0	11.3	13.2	11.3	**11.3**
Qwen 2.5 VL	0.0	0.0	0.0	20.0	20.0	10.0	10.0	20.0	20.0	20.0	0.0	0.0	0.0	0.0	0.0	**8.0**
Qwen 3	10.0	0.0	0.0	60.0	20.0	**70.0**	30.0	80.0	80.0	10.0	10.0	0.0	3.7	0.0	0.0	**3.3**
Grok Vision Beta	0.0	0.0	0.0	33.3	50.0	0.0	25.0	0.0	**22.2**	11.1	0.0	11.1	0.0	0.0	0.0	**10.9**
Grok 2 Vision	0.0	0.0	0.0	20.0	50.0	40.0	10.0	0.0	10.0	10.0	0.0	0.0	0.0	0.0	0.0	**9.3**

the majority of both simulated and real-world tasks. This pervasive failure highlights a fundamental difficulty in inferring basic stability, support, occlusion, and reachability limits from visual input.

In the **Simulated Domains**, 'Impossible Placement' scenarios almost universally failed. The 'Occlusion' domain saw slightly more success, particularly from 'Gemini 2.5 Pro Preview' (up to 90.0%) and 'GPT-4.1' (up to 70.0%). 'Stability' and 'Reachability' tasks in simulation also proved very difficult, with only sporadic, low scores from most models, though 'Gemini 2.5 Pro P' and 'Llama 3.2 11B Vision Instruct' showed some capability in specific views for Reachability. Viewpoint (F, A, S) within simulation influenced scores inconsistently (e.g., 'Gemini 2.5 Pro P' on Sim-Occlusion: F:90.0%, A:30.0%, S:60.0%), indicating a lack of robust view-invariance.

In **Real-World Humanoid** scenarios performance is generally low, though some models show interesting divergences. 'GPT-4.1 Mini', despite near-zero performance in simulation, achieves comparatively better scores on the Humanoid tasks (18.8% Agent, 24.5% Side, 22.6% Both), although its overall average remains low (4.4%). Conversely, 'Gemini 2.5 Pro', the strongest performer in simulation (25.8% Avg), shows more modest results on the Humanoid tasks (11.3% Agent, 18.8% Side, 9.4% Both). This suggests that performance in simulated constraint scenarios does not directly translate to real-world robot-centric views, pointing to a significant sim-to-real gap in constraint understanding. Reasoning models, as seen with '(T)', provided only marginal and inconsistent benefits in these highly challenging constraint tasks. The overall poor performance across constraint evaluations clearly marks constraint comprehension as a critical area requiring substantial advancement for reliable VLM-driven robotics.

5 Limitations and Conclusion

While PAC Bench offers a significant step forward with its diverse hybrid dataset for evaluating VLM understanding of **P**roperties, **A**ffordances, and **C**onstraints (PAC), we acknowledge current limitations. These include the initial single-annotator pass for affordances, the exclusion of the RoboCasa subset from current VLM evaluations due to cost all of which suggest avenues for future expansion and refinement. Despite these, we introduced **PAC Bench** to address the critical, often unverified, assumption of deep physical grounding in VLMs for robotic manipulation. Our extensive evaluations of state-of-the-art models starkly reveal widespread deficiencies: while partial success is observed in property and basic affordance recognition, VLMs profoundly struggle with comprehensive affordance understanding and nearly all aspects of constraint reasoning in both simulated and real-world tests. These findings underscore that current VLM sophistication does not yet equate to robust physical grounding. PAC Bench thus provides the community with a crucial diagnostic tool and a structured methodology to systematically measure these foundational skills, pinpoint key weaknesses (such as poor constraint generalization or difficulty with compositional affordances), and catalyze the development of more physically intelligent, reliable, and ultimately, safer VLMs for real-world robotic interaction.interaction.

References

[1] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, Quan Vuong, Thomas Kollar, Benjamin Burchfiel, Russ Tedrake, Dorsa Sadigh, Sergey Levine, Percy Liang, and Chelsea Finn. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024.

[2] Octo Model Team, Dibya Ghosh, Homer Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Tobias Kreiman, Charles Xu, et al. Octo: An open-source generalist robot policy. *arXiv preprint arXiv:2405.12213*, 2024.

[3] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818*, 2023.

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

[5] Wilbert Pumacay, Ishika Singh, Jiafei Duan, Ranjay Krishna, Jesse Thomason, and Dieter Fox. The Colosseum: A benchmark for evaluating generalization for robotic manipulation, 2024.

[6] Shi Qiu, Chenrui Zhang, Xin Zhao, Kai Wang, Jiazheng Zhang, Ruijie Yang, Junjie Hu, Hongyang Zhang, Yi Zhou, Wenhao Wang, et al. PHYBench: Holistic evaluation of physical perception and reasoning in large language models, 2025.

[7] Kang Zhu, Yifan Mai, Qian Huang, Ze Chen, Zhi-Yong He, Yue Xu, Wei Liu, Jure Leskovec, Anima Anandkumar, Cihang Xie, and Huaxiu Yao. PhysBench: Benchmarking and enhancing vision-language models for physical world understanding, 2024.

[8] Boxin Wang, Weixin Chen, Hengzhi Pei, Chulin Xie, Mintong Kang, Chenhui Zhang, Chejian Xu, Zidi Xiong, Ritik Dutta, Rylan Schaeffer, et al. Decodingtrust: A comprehensive assessment of trustworthiness in gpt models. In *NeurIPS*, 2023.

[9] Yi Zeng, Yu Yang, Andy Zhou, Jeffrey Ziwei Tan, Yuheng Tu, Yifan Mai, Kevin Klyman, Minzhou Pan, Ruoxi Jia, Dawn Song, et al. Air-bench 2024: A safety benchmark based on risk categories from regulations and policies. *arXiv preprint arXiv:2407.17436*, 2024.

[10] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*, 2022.

[11] Tony Lee, Haoqin Tu, Chi Heem Wong, Wenhao Zheng, Yiyang Zhou, Yifan Mai, Josselin Somerville Roberts, Michihiro Yasunaga, Huaxiu Yao, Cihang Xie, et al. Vhelm: A holistic evaluation of vision language models. *arXiv preprint arXiv:2410.07112*, 2024.

[12] Peng Xia, Ze Chen, Juanxi Tian, Yangrui Gong, Ruibo Hou, Yue Xu, Zhenbang Wu, Zhiyuan Fan, Yiyang Zhou, Kangyu Zhu, et al. Cares: A comprehensive benchmark of trustworthiness in medical vision language models. *arXiv preprint arXiv:2406.06007*, 2024.

[13] Qian Huang, Jian Vora, Percy Liang, and Jure Leskovec. Mlagentbench: Evaluating language agents on machine learning experimentation. In *Forty-first International Conference on Machine Learning*, 2024.

[14] Ananye Mukherjee, Priyanshu Agarwal, Ashish Kumar, Jason Corso, and Ashwin Balakrishna. 2handedafforder: Learning precise actionable bimanual affordances from human videos, 2025.

[15] Yitao Liu, Chenfei Yuan, Shiqi Liu, Renjie Li, Zixuan Wang, Li Yi, and Yang Lv. Physvlm: Enabling visual language models to understand robotic physical reachability, 2025.

[16] Daniel M Bear, Judy Fan, Elias Dyer, T K Marks, Richard Futrell, Joshua B Tenenbaum, Elizabeth S Spelke, and Daniel L K Yamins. Physion: Evaluating physical prediction from vision in humans and machines. In *Advances in Neural Information Processing Systems 34 (NeurIPS 2021)*, 2021.

[17] Haoyu Xu, Zhipeng Cao, Zhijun Zhang, Yifei Chen, Yixin Chen, and Zhiyong Wu. Manipvqa: Injecting robotic affordance and physically grounded information into multi-modal large language models, 2024.

[18] Yilong Zhou, Zeyi Liu, Zihan Ding, Wei-Chih Hung, Pieter Abbeel, and Huazhe Xu. Uniaff: A unified representation of affordances for tool usage and articulation with vision-language models, 2024.

[19] Guangzheng Chen, Hongtao Wu, Jiafeng Gu, and Qing Li. Naturalvlm: Leveraging fine-grained natural language for affordance-guided visual manipulation, 2024.

[20] Andy Zeng, Shuran Song, Johnny Lee, Stefan Lee, Alberto Rodriguez, and Silvio Savarese. Human affordances for robotic pre-training. In *Proceedings of Robotics: Science and Systems (RSS)*, 2020.

[21] Corey Lynch and Pierre Sermanet. Visual affordance-guided policy optimization. In *Conference on Robot Learning (CoRL)*, 2021.

[22] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International Journal of Computer Vision*, 2020.

[23] Soroush Nasiriany, Abhiram Maddukuri, Lance Zhang, Adeet Parikh, Aaron Lo, Abhishek Joshi, Ajay Mandlekar, and Yuke Zhu. Robocasa: Large-scale simulation of everyday tasks for generalist robots. *arXiv preprint arXiv:2406.02523*, 2024.

[24] Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ international conference on intelligent robots and systems*, pages 5026–5033. IEEE, 2012.

[25] Anthropic. Introducing Claude 3.5 Sonnet. `https://www.anthropic.com/news/claude-3-5-sonnet`, June 2024.

[26] Anthropic. Claude 3.7 Sonnet (System Prompt Release Notes). `https://docs.anthropic.com/en/release-notes/system-prompts`, February 2025. System prompt release notes.

[27] OpenAI. Introducing GPT-4.1 (mini) in the API. `https://openai.com/index/gpt-4-1/`, April 2025.

[28] xAI. Grok-2 Beta Release. `https://x.ai/blog/grok-2`, August 2024.

[29] xAI. xAI's Grok Chatbot Can Now 'See' the World Around It. *TechCrunch*, April 2025.

[30] Meta. Llama 3.2 Vision Instruct Model Card. `https://huggingface.co/meta-llama/Llama-3.2-11B-Vision-Instruct`, September 2024.

[31] TechCrunch. Meta Releases Llama 4: A New Crop of Flagship AI Models. *TechCrunch*, April 2025.

[32] QwenLM. Qwen-VL-Plus Model. `https://github.com/QwenLM/Qwen-VL`, 2025.

[33] Shuai Bai, Keqin Chen, Xuejing Liu, and *et al.* Qwen2.5-VL Technical Report. `https://arxiv.org/abs/2502.13923`, February 2025.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Yes, the abstract and introduction (Section 1) accurately reflect the paper's contributions and scope. We claim to introduce PAC Bench, a novel, hybrid benchmark for evaluating VLM understanding of physical Properties, Affordances, and Constraints (PAC) as prerequisites for manipulation. We also claim to provide a comprehensive evaluation suite and empirical insights from testing state-of-the-art VLMs. These claims are substantiated by the detailed description of the PAC Bench dataset (Section 3), its multi-source composition and annotation methodology (Section 3.1), and the presentation and analysis of experimental results on various VLMs (Section 4). The scope is clearly defined as assessing foundational physical reasoning for task executability.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.

- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 5 (Limitations and Conclusion) discusses current limitations and opportunities for future expansion.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: This paper introduces PAC Bench, a new benchmark dataset and evaluation framework for VLMs. It presents empirical findings from model evaluations rather than new theoretical results, mathematical derivations, or formal proofs.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Yes, We provide the dataset and github link and also a experiment setup which can be seen in Appendix B.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Yes, We have given open access to code.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.

- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Yes all details are mentioned in Appendix B

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Yes we show these results in Appendix D.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

Justification: Compute resource required can be seen in Appendix B.1

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research was conducted in accordance with the NeurIPS Code of Ethics. The development of PAC Bench involved using publicly available datasets (OpenImages, RoboCasa components), newly generated simulated data, and new real-world robotic data collection focused on common objects and non-sensitive scenarios. Human annotation efforts (detailed in Section 3.1 and Appendix E.1) were designed with ethical considerations, including fair practices for annotators. The benchmark aims to promote robust and grounded VLM development for safer robotic systems.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Yes we provide this in Appendix A

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.

- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The PAC Bench dataset primarily comprises images of common objects from public datasets (OpenImages), controlled simulated environments (MuJoCo, RoboCasa components), and new robotic captures of everyday tabletop scenarios which do not involve sensitive personal data. We are not releasing new pre-trained generative models or other assets typically associated with a high risk for direct misuse that would necessitate specific safeguards beyond responsible dataset curation and intended use for research.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The PAC Bench dataset, composed of public data, simulations, and new non-sensitive robotic captures, will be released with clear documentation on its intended research use, mitigating misuse risks associated with this type of evaluation benchmark.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: Yes new data collected are well documented and is provided.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [Yes]

Justification: Yes we show this in Appendix E

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: This research does not involve direct experiments with human subjects in a way that would typically require IRB approval. The human involvement was limited to data annotation of common objects and scenarios by trained annotators who were fairly compensated (details in Appendix E.1), and data collection with robotic platforms observing these objects, not interacting with human participants in an experimental context. No sensitive personal data was collected or used.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: The core focus of this paper is the introduction of PAC Bench, a benchmark specifically designed to evaluate Vision-Language Models (VLMs), which inherently involve Large Language Model components. Furthermore, our evaluation methodology for assessing constraint understanding (detailed in Section 4.1 and Appendix D) utilizes an LLM-as-a-judge approach.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.