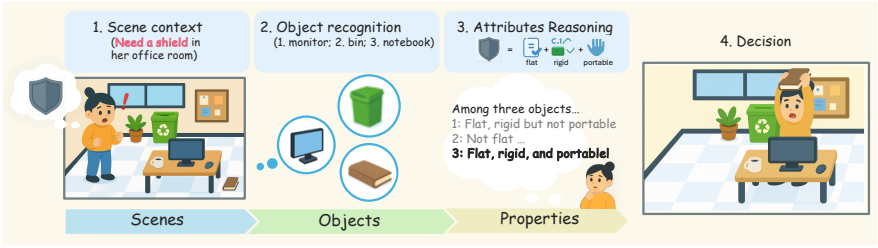
PercepTax: Benchmarking Physical Intelligence through Cross-Property Reasoning in Vision-Language Models

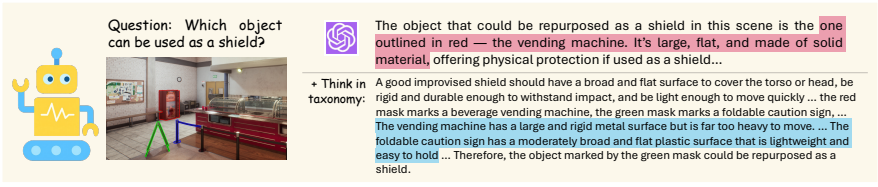
Jonathan Lee[1*] Xingrui Wang[1*✉] Jiawei Peng[1] Luoxin Ye[1]
Zehan Zheng[1] Tiezheng Zhang[1] Tao Wang[1] Wufei Ma[1] Siyi Chen[1]
Yu-Cheng Chou[1] Prakhar Kaushik[1†] Alan Yuille[1†]

[1]Johns Hopkins University, Baltimore, MD, USA
[*]Equal contribution [†]Joint senior authors ✉Corresponding author

Project Page Dataset Card Code



(a)



(b)

Fig. 1: Illustration of our proposed Perceptual Taxonomy for structured scene understanding. (a) A motivating example of human reasoning: when needing protection, a person hierarchically perceives the scene (*scene-level*), identifies available objects (*object-level*), and infers their properties *across multiple dimensions of properties* (flat, rigid), affordances (holdable, portable) to select the most suitable object (book as shield). (b) An example from **PercepTax** instantiating our key contribution: **Cross-Property Reasoning**, where a model must *jointly* reason over physical properties, affordances, and spatial context to identify to perform task, going beyond single property recognition toward human-like perceptual decision-making.

Abstract. Reasoning about objects in the physical world requires understanding both the spatial structure of a scene and the properties of

objects within it. To perform reasoning tasks, a model must identify candidate objects from the scene and then reason over their properties, such as material, physical attributes, affordance, and function to satisfy the task constraints. We formalize this reasoning process as **Perceptual Taxonomy**, a hierarchical framework that organizes visual understanding into structured scene–object–property representations. Existing benchmarks, however, address only parts of this framework: some focus on spatial understanding alone, while others evaluate individual property families—material, affordance, or physical attributes—each under a separate schema. They fail to cover a challenging yet practically important capability: **cross-property reasoning**, where a model must simultaneously integrate reasoning across all property families to answer questions about physical environments. To evaluate this capability, we introduce **PercepTax**, a benchmark for perceptual taxonomy reasoning. We annotate **3,173** object classes with **54** fine-grained attributes across four property families: material, physical attributes, affordance, and function. The benchmark contains **5,802** images from both synthetic and real domains and **28,033** questions spanning object description, single-property recognition, 3D spatial reasoning, and cross-property reasoning. Experiments show that state-of-the-art vision language models (VLMs) perform strongly on object description and single-property tasks, but accuracy drops substantially on cross-property reasoning that requires integrating multiple attributes simultaneously. Oracle experiments with ground-truth objects' properties with 3D positions and taxonomy-guided in-context learning both improve performance substantially yet remain well below human levels, confirming that both visual perception of individual properties and their cross-property integration are bottlenecks for VLMs.

1 Introduction

When interacting with the physical world, humans often perform tasks by reasoning over both the spatial structure of a scene and the physical properties of objects. As illustrated in Fig. 1(a), when seeking protection but no shield is available, a person must first understand the spatial layout of the environment and identify candidate objects (e.g., monitor, bin, notebook), then infer which object possesses the properties required for the task, such as flatness, rigidity, and portability.

We formalize this structured reasoning process as ***Perceptual Taxonomy***. In this framework, scene understanding is organized hierarchically: a model first perceives objects and their 3D spatial structure, then infers physically grounded properties such as material, affordance, and function, and finally integrates these properties to support reasoning. Prior work in visual cognition confirms that humans naturally organize scene understanding in this hierarchical, property compositional manner

[13, 15, 50]. However, it remains unclear whether current vision language models (VLMs) can perform the same reasoning, particularly the final step of

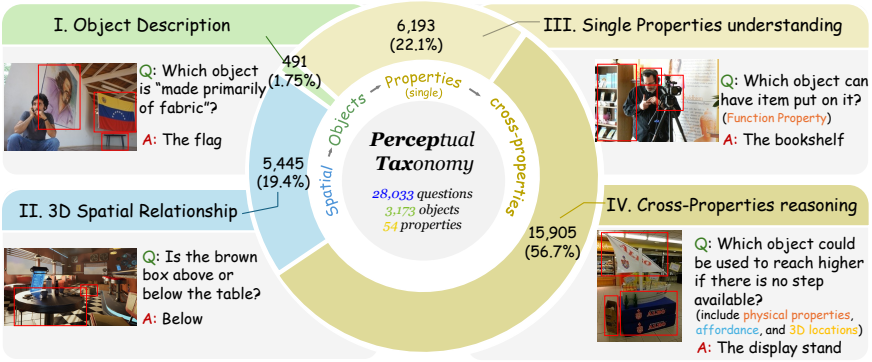


Fig. 2: Overview and examples of PercepTax. Our benchmark organizes perceptual taxonomy problem from scene to object to properties, covering 28,033 questions across 3,173 objects and 54 properties. As a new task, we introduce the **cross-property Reasoning** task (56.7%), which requires jointly reasoning over physical properties, affordances etc., going beyond existing single property tasks.

integrating multiple properties simultaneously. As VLMs have become the leading systems for multimodal reasoning [8, 9, 26, 35, 41, 45], prior benchmarks have addressed parts of perceptual taxonomy only in isolation. Some work starts a first step from 3D spatial reasoning [2, 20, 46], and individual properties such as material [19], affordance [37, 44], and physical attributes [7] separately. However, they miss the challenging yet practically important task of **cross-property reasoning**: simultaneously integrating material, affordance, function, and physical constraints to answer questions in physical scenes. As shown in Tab. 1, every prior benchmark covers at most one property dimension and none supports cross-property reasoning.

Benchmarking ***Perceptual Taxonomy*** reasoning is non-trivial due to two key limitations: (i) the lack of a carefully designed, large-scale annotation pipeline that defines a shared set of object properties, such as material, affordance, functionality, and physical attributes, and cluster common object categories to match with the properties; (ii) the absence of diverse and comprehensive question types and questions templates that require reasoning over these properties in varied contexts, an essential reasoning ability that current VLM benchmarks fail to evaluate.

To address this gap, we introduce **PercepTax**, a benchmark that evaluates all levels of perceptual taxonomy within a unified setting: **object description**, **spatial reasoning**, **single-property understanding**, and **cross-property reasoning** (see Fig. 2). Since no existing resource provides object annotations that jointly cover all property families, cross-property evaluation requires a new annotation effort: we curate **3,173** object classes and assign each a complete property profile across **54** fine-grained attribute clusters spanning four families (material, physical attributes, affordance, and function), using foundation model-driven clustering followed by human verification. The benchmark includes

Table 1: Comparison of existing benchmarks. Prior work addresses 3D spatial reasoning in isolation, and some focus on **Single Prop**erties (**Mat**erial, **Aff**ordance, **Funct**ion and **Phys**ical), but none supports **Cross-Prop**erty reasoning, the most challenging task under our proposed human-like Perceptual Taxonomy reasoning framework. **PercepTax** is the first to jointly cover all four property families with 3D spatial reasoning in a unified setting.

Benchmarks	**#QA**	**Synth/Real**	**3D Spatial**	**Single-Prop.**				**Cross-Prop.**
				Mat.	**Aff.**	**Func.**	**Phys.**	
CLEVR [20]	853K	Synth						
MMBench [28]	3.2K	Real						
CV-Bench [43]	2.6K	Real	✓					
GQA [19]	22M	Real		✓				
SEED-Bench-2 [24]	24K	Real		✓				
RIO [37]	130K	Real			✓			
RoboAfford [40]	1.9M	Synth+Real	✓		✓			
Hypo3D [32]	15K	Synth	✓	✓		✓		
PTR [17]	700K	Synth	✓				✓	
A4Bench [44]	1,282	Real			✓			
PhysBench [7]	10K	Real					✓	
PercepTax	28K	Synth + Real	✓	✓	✓	✓	✓	✓

5,802 images from two domains: (1) synthetic scenes rendered in Unreal Engine 5, which provide precise and complete spatial annotations including 2D instance masks and 3D bounding boxes, and (2) real images from OpenImages [23], which introduce natural visual diversity to test generalization. Building on these annotations, we design questions across four task types at graduated levels of reasoning: (1)object description questions identify objects from natural-language property cues, (2)spatial reasoning questions evaluate 3D positional relationships, (3)single-property matching questions test each property family independently; and (4)cross-property taxonomy reasoning questions are framed as goal-oriented tasks (for example, "Which object could serve as a temporary stepstool?") that require jointly satisfying constraints across multiple interacting attributes, directly reflecting how physical reasoning arises in practice rather than the isolated property recognition of prior work. In total, the benchmark contains **28,033** questions across four task types, including 50 expert-authored questions targeting the most demanding cross-property cases.

In our experiments, we evaluate both closed-source (e.g., Gemini 2.5 [9], GPT-5 [35], Claude Sonnet 4.5 [8]) and open-source (e.g., Qwen3-VL [41], InternVL [45], LLaVA-Next [26]) VLMs on **PercepTax**. Models perform strongly on object recognition and single-property tasks: Gemini 2.5 and GPT-5 achieve 84.70% and 87.06% on real-image object description. However, accuracy drops sharply on cross-property reasoning, with GPT-5 reaching only 39.84% on real images and 29.92% on simulation, a gap of more than 44 points from its object-level performance. This confirms that current VLMs struggle to integrate multiple properties jointly into coherent physical understanding.

To diagnose where failures originate, we conduct two experiments: oracle experiments that supply ground-truth property labels improve accuracy substantially by 20.11%, confirming that visual perception of individual properties is one bottleneck; and taxonomy-guided in-context learning using simulation exemplars improves real-world performance further by 9.56%, confirming that the reasoning process for integrating multiple attributes is an additional bottleneck. Both gains leave a significant gap to human performance, indicating that cross-property reasoning requires advances on both the perceptual and the reasoning paradigm

Our contributions can be summarized as follows:

1. We formalize *perceptual taxonomy* and construct a comprehensive property dictionary, annotating **54** physics-relevant attributes (14 material, 10 affordance, 13 function, 17 physical) across **3,173** object classes.
2. We introduce **PercepTax**, a benchmark containing **5,802** images (simulated and real) and **28,033** questions spanning object description, spatial reasoning, single-property recognition, and cross-property taxonomy reasoning. To our knowledge, it is the first benchmark to systematically evaluate cross-property reasoning in VLMs.
3. We benchmark state-of-the-art VLMs and reveal a systematic accuracy drop on cross-property reasoning compared to single-property tasks. Oracle and in-context learning experiments confirm that both visual perception of individual properties and their integration are bottlenecks, motivating future work on structured physical reasoning.

2 Related Work

2.1 VLM benchmark for scene understanding

Recent advances in large-scale vision and language models (VLMs) have shown strong generalization across diverse multimodal tasks [12,21,47,49]. Early benchmarks such as VQA v2.0 [14], CLEVR [20], and GQA [19] established foundations for compositional reasoning, focusing on object attributes, spatial relations, and logical consistency. Later datasets including NLVR2 [39], OK-VQA [33], VCR [51], and A-OKVQA [38] introduced language grounded and commonsense reasoning challenges. Recent works such as Winoground [42], PTR [17], and SuperCLEVR [25] target fine grained compositionality and distribution shifts. Datasets like ScanQA [2] and Spatial457 [46] expand into 3D spatial reasoning. Meanwhile, general purpose evaluation suites such as MMBench [28], SEED-Bench-2 [24], and CV-Bench [43] assess a broad range of multimodal skills. While these benchmarks advance structured multimodal reasoning, they primarily focus on semantics or geometry. A systematic framework for evaluating reasoning over attributes such as material, affordance, and function remains underexplored, yet it is essential for grounding VLMs in real world understanding.

2.2 3D Spatial Reasoning

Evaluating spatial relationships between objects in 3D scenes has become a key capability for scene-understanding and embodied systems. CLEVR [20] probes spatial relations in controlled synthetic settings; ScanQA [2] extends spatial QA to real indoor scenes using 3D point cloud annotations. CV-Bench [43] incorporates 3D spatial tasks within a broader multimodal evaluation suite, and Spatial457 [46] provides a systematic benchmark for 3D positional reasoning across diverse viewpoints. RoboAfford [40] couples 3D spatial understanding with affordance reasoning in robotic manipulation contexts.

While these works establish strong foundations for spatial reasoning, they evaluate object positions and relations in isolation from object properties. **PercepTax** integrates 3D spatial reasoning with property-level understanding, enabling questions that jointly require locating objects in a scene and reasoning over their physical attributes.

2.3 Object Properties Reasoning

Beyond general scene understanding, several works focus on reasoning over object properties such as material, affordance, functionality, and physical attributes. ShapeStacks [16] and Falling Tower [3] examine intuitive physics through stability prediction in stacking scenarios. PhysGame [4] challenges models to detect physically implausible scenes from gameplay videos, emphasizing material and broader physical reasoning. PhysBench [7] introduces detailed property annotations in real world images, while PhysXNet [5] provides large scale 3D assets to support VQA on physical properties.

RIO [37] provides large-scale affordance annotations for real-world images. ManipVQA [18] frames affordance and tool-use understanding as a VQA task, encouraging models to reason about object manipulation. A4Bench [44] further explores constitutive and transformative affordances, connecting static and dynamic object interaction. Most recently, Hypo3D [32] combines 3D spatial reasoning with material and function annotations in synthetic scenes, covering two property families simultaneously.

While these works provide valuable coverage of individual property families, each benchmark evaluates properties in isolation: none jointly annotates objects across multiple families, and none poses questions that require integrating several properties simultaneously. This **cross-property reasoning**, essential for practical problem-solving, remains systematically untested.

3 PercepTax: A Benchmark for Perceptual Taxonomy Reasoning

3.1 Perceptual Taxonomy and Property Definition

PercepTax organizes each scene following a three-level hierarchy: *scene* $\rightarrow$ *objects* $\rightarrow$ *properties*. (1) At scene level, we annotate the scene with a 2D instance segmentation mask and a 3D oriented bounding box, capturing its position, depth, and yaw in the camera coordinate frame to support spatial reasoning. (2)

At the object level, we annotate the object's name, and language description, to support object description questions. (3) The critical design decision is at the property level, to support both *single-property* and *cross-property* reasoning, we assign every object a complete property profile across all four families simultaneously, which is the prerequisite for cross-property reasoning:

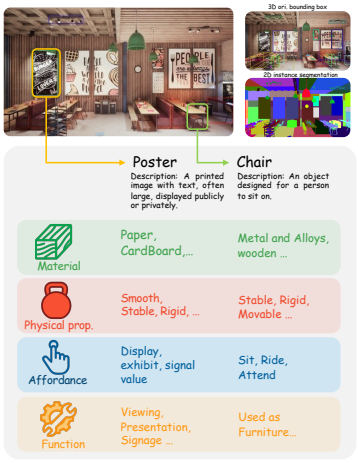
- **Material**: the object's visual and physical substance (e.g., paper, metal, wood), which determines its rigidity, weight, texture, and durability.
- **Physical properties**: intrinsic physical characteristics that govern how the object behaves under external forces, such as rigidity, fragility, elasticity, mass, or stability.
- **Affordance**: the set of actions or interactions the object enables based on its shape and components (e.g., graspable, supportable, sit-on, containable), independent of its intended purpose.
- **Functionality**: the object's intended or contextual role when invented (e.g., seating furniture, storage container, display signage), reflecting how humans typically use it in practice.

This joint annotation is the prerequisite for cross-property reasoning: a question such as "*Which object could serve as a temporary stepstool?*" requires satisfying constraints across rigidity (physical properties), load-bearing shape (affordance), and appropriate height (3D position) simultaneously. No prior benchmark supports such questions because each annotates objects under a single property schema, making cross-property constraints unresolvable. Fig. 3(a) shows an example of the annotation results in a sample scene, every object receives a complete profile: for example, a poster (*paper*, *smooth*, *display*, *signage*) and a chair (*wood and metal*, *stable*, *sit*, *furniture*). We define a taxonomy covering **14** material types, **10** affordances, **13** functionalities, and **17** physical properties, as illustrated in Fig. 3(b). A total of **3,173** unique objects are annotated with many-to-many mappings to these properties; full category-to-object mappings are provided in the Appendix.

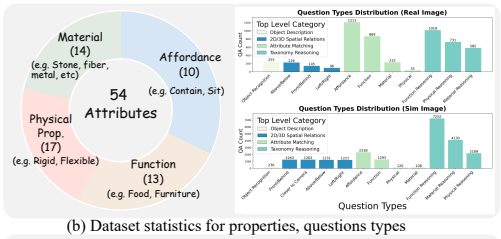
3.2 Question Generation

We begin by leveraging the comprehensive annotations described in Section 3.1, including 2D / 3D spatial cues and per-object material, affordance, function, physical properties. Building on the engine from previous work [20,46], we design 48 parameterized templates spanning four core perceptual taxonomy tasks: (1) object description, (2) spatial reasoning, (3) single-properties understanding, and (4) cross-property reasoning. We describe each task category below and we list all the templates in the Appendix.

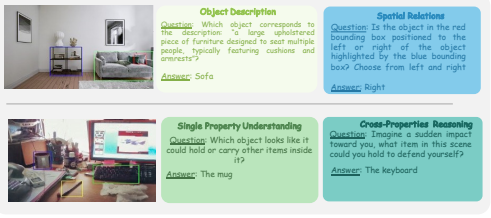
(1) Object Descriptions questions target the most basic layer of perceptual taxonomy: matching a natural-language description from an object's material, physical properties, affordances, or functional cues to the correct color coded bounding box. These questions focus on direct property object alignment rather than multi-step reasoning. For example, "Which object marked by the colored



(a) Scene annotation for Perceptual Taxonomy



(b) Dataset statistics for properties, questions types



(c) Example of the images and the corresponding questions under four question types

Fig. 3: Our **PercepTax** benchmark is structured by new scene-object-properties annotations and question-answer generation. (a) Each scene is annotated with object detections and their mapped attributes across four domains: material, physical properties, affordance, and function. (b) Summary statistics for 54 attribute categories and question distributions across synthetic and real images. (c) Example questions covering four reasoning types: object description, spatial relation, single property matching, and cross-property reasoning, together forming a unified framework for perceptual taxonomy for scene understanding.

boxes is used for writing with chalk?" requires selecting the chalkboard based on its functional property, also as the example illustrated in Fig. 3(c).

(2) 3D Spatial Reasoning questions evaluate 3D positional relationships between objects in the scene. While dedicated spatial reasoning benchmarks already exist [1,6,14,19,20,29,30,46], we include spatial reasoning in **PercepTax** to provide a unified evaluation environment where all levels of perceptual taxonomy from object localization to cross-property inference are benchmarked on the same scenes and objects. Questions involve identifying 3D spatial relations such as left and right, above and below, and front and behind between objects; the closer/farther relation is additionally included on the simulation split, where exact Unreal-Engine 5 3D coordinates provide an unambiguous depth ordering (see the Supplementary Material for the real-image rationale; all relations are computed directly from ground-truth 3D object locations in the camera frame (see Sec. 4 for details). An example question is: "Is the object marked by the Green box to the left or right of the object marked by the Blue box?" Additional instances are shown in Fig. 3(c).

(3) Single Property Understanding questions evaluate whether a model can associate specific physical or functional attributes with the correct object in the scene. All queried attributes, such as material, rigidity, elasticity, thermal properties, or functional affordances are drawn directly from our properties

annotations (as in Section 3.1). Single property understanding requires models to go beyond surface visual features and recognize properties that reflect real-world physical semantics. Correctly answering these questions often requires understanding that certain attributes (e.g., "rigid," "conductive," "container-like," "made of slate") cannot be inferred purely from color or shape, but must be grounded in how objects behave or are used in the physical world. For example, in the question "Which object marked by the colored boxes is made of slate?", the correct answer is the chalkboard. Such questions therefore test the model's ability to distinguish fine-grained, attribute-level differences and ground them in both visual evidence and physical-world knowledge, rather than superficial appearance alone.

(4) Cross-Property Taxonomy Reasoning questions represent the highest level of perceptual taxonomy and are the primary focus of our benchmark. Unlike previous categories that involve direct property-object alignment, these questions require multi-step inference over multiple structured attributes, often combining affordances, material properties, physical constraints, and functional compatibility. To answer correctly, a model must perform a filter–eliminate reasoning process: identifying the attributes implied by the goal, inferring which objects possess the necessary physical or functional properties, and ruling out candidates that fail any of the constraints. This goes beyond appearance recognition and instead demands a grounded understanding of how objects behave and can be repurposed in the physical world. For example, the question *"Which object could function as a temporary stepstool?"* requires recognizing that the correct choice must be rigid, stable, and provide a flat, load-bearing surface at an appropriate height—properties only one object in the scene satisfies. Such questions probe whether models can integrate multiple dimensions of perceptual taxonomy to reach coherent and physically plausible conclusions.

Unique Candidate Constraint. To prevent trivially answerable questions, we enforce a **unique-candidate constraint**: a scene is eligible for a given question type only if exactly one bounded object in the scene satisfies all required attributes. Concretely, for each candidate question we check whether more than one bounded object passes the attribute filters; if so, the question is discarded. This constraint ensures that correct answers are unambiguous and cannot be reached by guessing among visually similar objects.

With these templates, our question engine instantiates questions at scale across scenes. We apply lightweight LLM-based rephrasing [9] for linguistic variety, followed by automatic validity checks that ensure each question is grounded, unambiguous, and has a single correct answer—filtering out objects or scenes that introduce semantic noise or multiple valid solutions. Brief human verification further ensures quality. As a **qualitative** cross-property stress-test, the benchmark additionally includes **50** expert-authored questions on real images, targeting cross-property reasoning patterns that lie beyond the reach of systematic templates. This set is intended as a focused qualitative probe rather than a statistically powered evaluation; our core claims on VLM capabilities rest on the full **28,033**-question benchmark. Full pipeline details and templates are provided in the Appendix.

4 Data Curation

Simulation image rendering. We generate our synthetic image set using 33 high-quality Unreal Engine 5 [11] assets, which provide accurate geometry, physically based materials, and photorealistic lighting. This ensures that all rendered images come with precise and complete scene annotations, including 2D instance masks, 2D/3D bounding boxes, object poses, and material and physical attributes. To obtain diverse and semantically rich images, we randomly sample camera viewpoints from fully constructed scenes to capture realistic layouts and natural spatial relationships. We further augment scene diversity by modifying environments: all placeable objects are manually curated and randomly arranged on valid surfaces (e.g., tables, floors) to produce additional compositions and layouts. These pipelines yield a 4,544 realistic and well-annotated images. The operational details of rendering, camera sampling, and quality filtering are provided in Appendix.

Real images collection and annotation. We curate real-world scenes from the OpenImages [22] training subset. Since these images do not contain scene-level annotations, we apply several foundation models to detect objects and extract their spatial information, including 2D bounding boxes, 3D bounding boxes, and pose estimates [27, 36, 48]. This allows us to obtain structured scene representations and generate 1,258 images.

Object description and properties clustering. As shown in Fig. 4, starting from the curated objects collected from both simulation scenes and real images, we use Gemini to generate concise textual descriptions covering visual characteristics and taxonomy-relevant attributes such as material, affordance, physical properties, and functionality. These descriptions are then grouped using unsupervised clustering methods (e.g., HDBSCAN [34], k-means [31]) to form initial property clusters. Human annotators review and refine these clusters to ensure semantic correctness and resolve inconsistencies, producing a reliable taxonomy that supports our QA generation pipeline. The verified clusters are then used to instantiate template-based questions grounded in the annotated scenes.

Final human quality verification. We conduct a global human quality verification stage in which annotators inspect generated QA pairs and remove problematic cases such as inaccurate bounding boxes, incorrect object identities, or mismatched attributes with image input. Approximately 31.7% of automati-

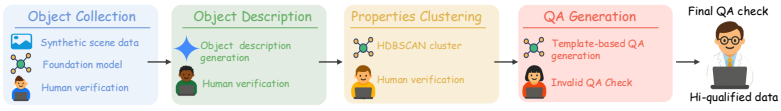


Fig. 4: Object Annotation Pipeline: Objects are first collected from synthetic and real scenes, described by foundation models, and clustered by shared material, function, and affordance attributes with human verification at each stage. Template-based QA generation then converts verified clusters into question–answer pairs, with 68.3% passing the final quality check to ensure accurate and semantically grounded annotations.

cally generated questions were discarded through this process, leaving only high-quality, visually grounded QA pairs for inclusion in the benchmark.

4.1 Benchmark Statistics

Our benchmark consists of **5,802** images, including **4,544** synthetic images and **1,258** real images. Across these scenes, we annotated **3,173** distinct object classes, each matched many-to-many to **54** attribute labels spanning four property types: material (14), affordances (10), functions (13), and physical properties (17). In total, we generate **28,033** template-based open-ended questions covering the four core reasoning tasks introduced in Section 3.2. The distribution of question types is shown in the histogram of Fig. 3(b), and detailed statistics across categories and subtypes are provided in Appendix.

5 Experiments

5.1 Baselines

We evaluate our benchmark using a diverse suite of state-of-the-art Vision Language Models (VLMs) using the open-source toolkit [10]. We benchmark both proprietary and open-source VLMs spanning different architectural families. The proprietary models include *Gemini 2.5 Pro* [9], *GPT-5* [35], and *Claude Sonnet-4.5* [8]. For open-source systems, we cover two capacity tiers. At the larger model size, we evaluate *Qwen3-VL-32B* [41] and *InternVL3.5-30B-A3B* [45]. At the 7B to 8B model size, we include *Qwen3-VL-8B* [41], *InternVL3.5-8B* [45], *LLaVA-Next-Llama3* [26], *LLaVA-Next-Vicuna-7B* [26], *LLaVA-Next-Interleave-7B* [26], and *LLaVA-OneVision-7B* [26]. All models are evaluated using explicit, task-specific instructions. We adopt an open-ended question format where each model generates free-form answers rather than selecting from predefined options. All responses are evaluated by a LLM verifier [9] that judges whether the model's answer is semantically correct with respect to the ground-truth annotation, enabling more naturalistic assessment of reasoning quality.

5.2 Main Results

Tab. 2 reports model accuracy on each reasoning task. We report accuracy separately on the *simulation* and *real-image* subsets to analyze cross-domain generalization. (1) **Cross-Property reasoning is the primary focus and the hardest task.** Almost across all models, cross-property reasoning yields the lowest accuracy on the real subset, with real-image scores ranging from **9.20%** to **53.00%**—well below the human upper bound of **92.96%**. Even the strongest model achieves only **53.00%**, despite substantially higher object description scores. The accuracy is even worse on the simulation subset, where the cross-property result is **55.87%** lower than the human upper bound. The results reveal that perceptual grounding does not transfer to property integration: models can locate and describe objects but fail to simultaneously satisfy material, affordance, and functional constraints. (2) **Closed-source models consistently outperform open-source models.** Gemini 2.5 Pro leads on real images (**57.38%** overall) and simulation images (**41.87%** overall) and excels

Table 2: Performance of representative **close-source** and **open-source** VLMs on the **Simulation Set** and **Real Image Set** under the open-ended evaluation protocol (responses judged by a Gemini LLM verifier). Higher is better. **S.R.** = Spatial Relation; **Obj. Desc.** = Object Description; **Single-Prop.** = Single-Property Understanding; **Cross-Prop.** = Cross-Property Reasoning. **Human**: answers provided by human annotators, serving as an upper bound. The highest score among all models for each question type is marked in red, and the second-highest scores are underlined. The best performance among open-source models is marked in blue.

Model	Simulation Set					Real Image Set				
	S.R.	Obj. Desc.	Single-Prop.	Cross-Prop.	All	S.R.	Obj. Desc.	Single-Prop.	Cross-Prop.	All
Baselines										
Human	94.81	99.26	90.37	94.07	94.63	89.70	91.11	88.15	92.96	90.48
Close-source models										
Gemini 2.5 Pro [9]	60.63	53.81	30.77	38.20	41.87	74.30	84.70	55.29	53.00	57.38
Claude Sonnet 4.5 [8]	46.97	29.24	16.00	21.64	26.21	69.38	66.67	43.59	32.95	42.48
GPT-5 [35]	54.56	54.24	25.90	29.92	34.82	76.23	87.06	47.77	39.84	48.78
Open-source models										
Qwen3-VL-32B [41]	52.89	38.56	25.99	30.30	34.53	72.16	81.18	50.37	41.66	50.08
InternVL3.5-30B-A3B [45]	45.20	27.54	16.02	17.61	23.47	76.23	72.16	37.06	25.57	37.32
Qwen3-VL-8B [41]	52.97	43.22	24.35	25.83	31.69	73.23	77.65	45.04	39.08	46.54
InternVL3.5-8B [45]	45.56	25.42	15.35	15.33	22.09	71.95	71.76	38.21	26.77	37.95
LLaVA-OneVision-7B [26]	38.93	19.92	14.13	16.35	20.93	61.67	67.45	35.28	26.50	35.42
LLaVA-Next-Llama3 [26]	41.12	5.51	11.01	17.32	21.22	56.96	31.37	35.73	28.81	34.48
LLaVA-Next-Vicuna-7B [26]	39.39	1.69	7.25	9.96	15.83	55.89	7.45	19.77	9.20	17.91

in single-property understanding and cross-property reasoning. Among open-source models, Qwen3-VL-32B achieves the strongest results across almost all tasks (**50.08%** overall on real images and **34.53%** overall on simulation images). (3) **Object description and spatial reasoning serve as auxiliary comparisons** bounding the perceptual ceiling. High accuracy on these tasks confirms that models can localize and describe objects in scenes, yet this competence does not carry over to cross-property reasoning. The persistent gap between perceptual task performance and cross-property reasoning performance demonstrates that the bottleneck lies in property integration rather than object recognition.

5.3 Cross-Property Reasoning with In-Context Learning

To validate the importance of perceptual taxonomy, we design a **hierarchical reasoning paradigm** that guides the model from scene layout to object-level physical properties before answering. We provide one simulation exemplar with structured object–property cues for in-context learning (ICL). This taxonomy-aware prompting yields clear gains on real-image taxonomy reasoning tasks, demonstrating sim-to-real transfer of structured physical knowledge. We evaluate Gemini 2.5 Pro, GPT-5, Qwen3-VL-8B, and InternVL3.5-8B using visually disjoint real images to isolate the effect of reasoning.

The results in Tab. 3 report per-category cross-property reasoning accuracy on the real image set, showing consistent performance gains across both proprietary and open-source models when applying **PercepTax** in an ICL setting. Models demonstrate improved recognition of object properties and reasoning about physical functions in real images, confirming that hierarchical reasoning learned from simulation can transfer effectively to real-world scenarios.

Table 3: Sim-to-real improvement via taxonomic in-context learning. Providing one simulation exemplar with structured object–property cues improves cross-property reasoning on real images, especially for instruction-tuned models like Gemini 2.5 Pro and GPT-5.

Model	(Real Image Set) Cross-Property Reasoning			
	Physical	Material	Function	Avg.
Close-source Models				
Gemini 2.5 Pro	54.87	37.72	56.61	53.00
Gemini 2.5 Pro (w/ **PercepTax** ICL)	**55.40**	**37.98**	**59.77**	**54.86**
GPT-5	41.23	35.94	36.49	39.84
GPT-5 (w/ **PercepTax** ICL)	**45.10**	**36.81**	**41.44**	**43.60**
Open-source Models				
InternVL3.5-8B	25.86	18.50	29.89	26.77
InternVL3.5-8B (w/ **PercepTax** ICL)	**33.85**	**19.38**	**30.71**	**32.89**
Qwen3-VL-8B	44.13	26.69	42.24	39.08
Qwen3-VL-8B (w/ **PercepTax** ICL)	**46.92**	**25.66**	**39.60**	**40.19**

Notably, **Gemini 2.5 Pro** and **GPT-5** show the largest overall gains, improving from **53.00%** to **54.86%** and from **39.84%** to **43.60%** in average cross-property reasoning, respectively. These results indicate that large, instruction-tuned VLMs can effectively follow hierarchical exemplars and apply taxonomic constraints during inference. Still, Qwen3-VL-8B shows only marginal improvements, suggesting that beyond following the reasoning diagram, accurate visual perception of 3D structure and object properties remains another bottleneck for this task.

5.4 Oracle Property Analysis

To disentangle whether model failures arise from *property recognition* (failing to identify relevant object attributes) or *property integration* (failing to reason across multiple properties once identified), we conduct an oracle experiment. In the oracle condition (blue highlights in Table 4), ground-truth property labels—covering material, affordance, physical properties, and functionality—for all box-bounded objects are appended to the model prompt as structured context, providing perfect property perception. We evaluate the two top-performing models, Gemini 2.5 Pro and Qwen3-VL-32B, under this oracle setting.

As shown in Tab. 4, oracle injection substantially improves *single-property understanding* across both models, confirming that imperfect property recognition is a partial source of error on that task. However, *cross-property reasoning* remains well below the human upper bound even under oracle conditions—the gap narrows but does not close. This finding has a clear interpretation: while models sometimes fail to recognize individual properties, the deeper bottleneck is **compositional integration**, the ability to simultaneously satisfy constraints across material, affordance, and functional dimensions when reasoning toward a physically grounded conclusion. Together with the modest ICL gains in Tab. 3, these results motivate future work on property-aware architectures and reasoning paradigms that go beyond surface recognition.

Table 4: Oracle property analysis. Ground-truth property labels are appended to the model prompt, isolating **property integration** capacity from **property recognition** failures. To identify which reasoning step fails even under oracle conditions, we decompose residual oracle-wrong cross-property cases into three modes: **SB** = Category Semantic Bias (model anchors to canonical name/use); **CN** = Constraint Neglect (ignores an explicit negative or set constraint); **PCG** = Physical Commonsense Gap (lacks real-world material/functional knowledge).

Model	Split	SP		CP		Error Composition (%)†		
		Base	Oracle	Base	Oracle	SB	CN	PCG
Gemini 2.5 Pro	Sim.	30.77	**50.88**	38.20	**43.33**	37.08	7.69	55.24
Gemini 2.5 Pro	Real	55.29	**82.05**	53.00	**62.56**	21.55	27.99	49.06
Qwen3-VL-32B	Sim.	25.99	**29.41**	30.30	**33.93**	43.10	12.72	44.18
Qwen3-VL-32B	Real	50.37	**74.44**	41.66	**47.89**	21.74	28.56	48.44

†Error composition is over residual oracle-wrong *cross-property* cases only.

To further identify which kind of reasoning breaks down even when property perception is perfect, we manually categorize the residual oracle-wrong cases on the cross-property subset into three root causes (Table 4): **Physical Commonsense Gap** (PCG), where the model lacks real-world material or functional knowledge (e.g. judging water or heat resistance from a material label alone); **Category Semantic Bias** (SB), where the model anchors to an object's canonical name or use even after the oracle redefines its affordance; and **Constraint Neglect** (CN), where the model ignores an explicit negative or set constraint (e.g. selecting a container for a "non-container" task). PCG is the dominant mode under both annotation domains (44–55% of residual failures), while SB and CN shift with domain: on real images CN (≈28%) is comparable to SB (≈22%), whereas on simulation SB rises to 37–43% and CN collapses to 8–13%, likely because synthetic textures weaken the material cues that drive constraint-based reasoning. These patterns are consistent across both Gemini 2.5 Pro and Qwen3-VL-32B, indicating that the bottleneck is generic physical reasoning rather than a model-specific weakness. To further probe whether targeted prompting can overcome these failure modes, we conduct a 13-variant intervention study on representative oracle-wrong cases for each root cause; results show that SB and CN interventions fail universally across both models while PCG is recoverable with causal or material hints (see the Supplementary Material).

5.5 Qualitative Results

As shown in Fig. 5, models frequently produce fluent yet shallow explanations on taxonomy-dependent questions. In (a), GPT-5 and Qwen3 base their reasoning on spatial heuristics (e.g. proximity to the sink) while neglecting material and structural cues that determine vulnerability to water, leading to plausible but incorrect answers. In (b), Gemini 2.5 Pro exhibits a similar error, prioritizing rigidity over material suitability when reasoning about sound insulation. These

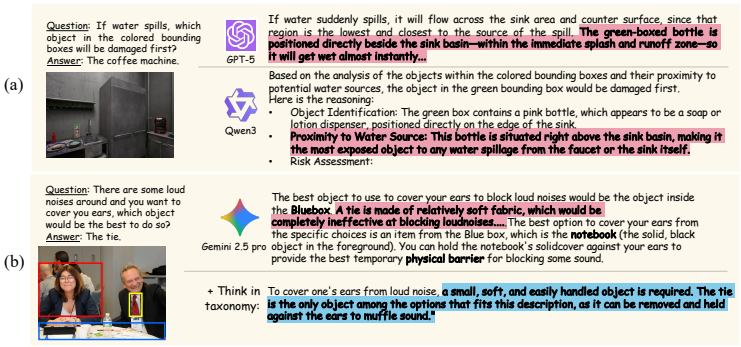


Fig. 5: Qualitative examples illustrating model errors in hierarchical reasoning. (a) Instances of hallucinated or inconsistent reasoning on simulated-image benchmark questions produced by Qwen3 and GPT-5. (b) Examples of Gemini's reasoning on real-image benchmark questions before and after in-context learning.

cases reveal a broader trend: models excel at perceptual grounding and spatial localization but fail to integrate hierarchical object knowledge, often conflating surface appearance with functional taxonomy. When guided by **PercepTax** in-context exemplars, reasoning becomes more structured, explicitly referencing material softness, size, and affordance, indicating emerging hierarchical understanding. Overall, the qualitative patterns mirror the quantitative results: current VLMs exhibit strong perceptual grounding but weak integration of taxonomic and material knowledge, limiting their ability to reason about object properties and usage.

6 Conclusion

We introduce **PercepTax**, a benchmark for perceptual taxonomy reasoning, and propose **Cross-Properties Reasoning**, a new task requiring models to jointly reason over material, affordance, function, and physical properties to answer grounded visual questions. Our experiments reveal that current VLMs fall far short of human performance on this task, and oracle experiments confirm the bottleneck is structured reasoning rather than visual detection. *PercepTax ICL* and 3D spatial knowledge each yield measurable gains, yet a substantial gap remains. We hope **PercepTax** catalyzes the development of VLMs that reason over structured property hierarchies, rather than relying on superficial visual alignment.

Acknowledgements

This work was supported by ONR grant N00014-23-1-2641 (Vision-Language) and in part by the Whiting School of Engineering at Johns Hopkins University.

References

1. Armeni, I., He, Z.Y., Gwak, J., Zamir, A.R., Fischer, M., Malik, J., Savarese, S.: 3d scene graph: A structure for unified semantics, 3d space, and camera (2019). https://doi.org/10.1109/ICCV.2019.00574
2. Azuma, D., Miyanishi, T., Kurita, S., Kawanabe, M.: Scanqa: 3d question answering for spatial scene understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
3. Balazadeh, V., Ataei, M., Cheong, H., Khasahmadi, A.H., Krishnan, R.G.: Physics context builders: A modular framework for physical reasoning in vision-language models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 7318–7328 (October 2025)
4. Cao, M., Tang, H., Zhao, H., Guo, H., Liu, J., Zhang, G., Liu, R., Sun, Q., Reid, I., Liang, X.: Physgame: Uncovering physical commonsense violations in gameplay videos. arXiv preprint arXiv:2412.01800 (2024)
5. Cao, Z., Chen, Z., Pan, L., Liu, Z.: Physx-3d: Physical-grounded 3d asset generation. arXiv preprint arXiv:2507.12465 (2025)
6. Cheng, A.C., Yin, H., Fu, Y., Guo, Q., Yang, R., Kautz, J., Wang, X., Liu, S.: Spatialrgpt: Grounded spatial reasoning in vision-language models. Advances in Neural Information Processing Systems **37**, 135062–135093 (2024)
7. Chow, W., Mao, J., Li, B., Seita, D., Guizilini, V., Wang, Y.: Physbench: Benchmarking and enhancing vision-language models for physical world understanding. arXiv preprint arXiv:2501.16411 (2025)
8. Claude: Claude sonnet 4.5. https://www.anthropic.com/claude/sonnet (2025), accessed: 2026-06-30
9. Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., et al.: Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. arXiv preprint arXiv:2507.06261 (2025)
10. Duan, H., Yang, J., Qiao, Y., Fang, X., Chen, L., Liu, Y., Dong, X., Zang, Y., Zhang, P., Wang, J., et al.: Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In: Proceedings of the 32nd ACM international conference on multimedia. pp. 11198–11201 (2024)
11. Epic Games: Unreal engine 5. https://www.unrealengine.com/ (2021), version 5.x; Accessed: 2026-06-30
12. Fu, X., Hu, Y., Li, B., Feng, Y., Wang, H., Lin, X., Roth, D., Smith, N.A., Ma, W.C., Krishna, R.: Blink: Multimodal large language models can see but not perceive. In: European Conference on Computer Vision. pp. 148–166. Springer (2024)
13. Gibson, J.J.: The Ecological Approach to Visual Perception. Psychology Press, Hillsdale, NJ (1986)
14. Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., Parikh, D.: Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
15. Greene, M.R., Baldassano, C., Esteva, A., Beck, D.M., Fei-Fei, L.: Visual scenes are categorized by function. Journal of Experimental Psychology: General **145**(1), 82 (2016)
16. Groth, O., Fuchs, F.B., Posner, I., Vedaldi, A.: Shapestacks: Learning vision-based physical intuition for generalised object stacking. In: Proceedings of the European Conference on Computer Vision (ECCV) (September 2018)

17. Hong, Y., Yi, L., Tenenbaum, J., Torralba, A., Gan, C.: Ptr: A benchmark for part-based conceptual, relational, and physical reasoning. Advances in Neural Information Processing Systems **34**, 17427–17440 (2021)
18. Huang, S., Ponomarenko, I., Jiang, Z., Li, X., Hu, X., Gao, P., Li, H., Dong, H.: Manipvqa: Injecting robotic affordance and physically grounded information into multi-modal large language models. In: 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 7580–7587 (2024). https://doi.org/10.1109/IROS58592.2024.10801993
19. Hudson, D.A., Manning, C.D.: Gqa: A new dataset for real-world visual reasoning and compositional question answering. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
20. Johnson, J., Hariharan, B., van der Maaten, L., Li, F.F., Zitnick, C.L., Girshick, R.: Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
21. Kil, J., Mai, Z., Lee, J., Chowdhury, A., Wang, Z., Cheng, K., Wang, L., Liu, Y., Chao, W.L.H.: Mllm-compbench: A comparative reasoning benchmark for multimodal llms. Advances in Neural Information Processing Systems **37**, 28798–28827 (2024)
22. Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Malloci, M., Kolesnikov, A., Duerig, T., Ferrari, V.: The open images dataset v6: A large-scale dataset for detection, segmentation and classification. International Journal of Computer Vision (IJCV) **128**(7), 1956–1981 (2020)
23. Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Malloci, M., Kolesnikov, A., et al.: The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. International journal of computer vision **128**(7), 1956–1981 (2020)
24. Li, B., Ge, Y., Ge, Y., Wang, G., Wang, R., Zhang, R., Shan, Y.: Seed-bench-2: Benchmarking multimodal large language models (2023), https://arxiv.org/abs/2311.17092
25. Li, Z., Wang, X., Stengel-Eskin, E., Kortylewski, A., Ma, W., Van Durme, B., Yuille, A.L.: Super-clevr: A virtual benchmark to diagnose domain robustness in visual reasoning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14963–14973 (2023)
26. Liu, H., Li, C., Li, Y., Li, B., Zhang, Y., Shen, S., Lee, Y.J.: Llava-next: Improved reasoning, ocr, and world knowledge (January 2024), https://llava-vl.github.io/blog/2024-01-30-llava-next/, accessed: 2026-06-30
27. Liu, S., Zhang, Z., Huang, Z., Wang, F., Zhang, H., Qiao, Y., Li, H., Wang, X., Dai, J.: Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10882–10892 (2023)
28. Liu, Y., Duan, H., Zhang, Y., Li, B., Zhang, S., Zhao, W., Yuan, Y., Wang, J., He, C., Liu, Z., et al.: Mmbench: Is your multi-modal model an all-around player? In: European conference on computer vision. pp. 216–233. Springer (2024)
29. Ma, W., Chen, H., Zhang, G., Chou, Y.C., Chen, J., de Melo, C., Yuille, A.: 3dsr-bench: A comprehensive 3d spatial reasoning benchmark. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6924–6934 (2025)
30. Ma, W., Chou, Y.C., Liu, Q., Wang, X., de Melo, C., Xie, J., Yuille, A.: Spatialreasoner: Towards explicit and generalizable 3d spatial reasoning. arXiv preprint arXiv:2504.20024 (2025), https://arxiv.org/abs/2504.20024

31. MacQueen, J.: Some methods for classification and analysis of multivariate observations. In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability. vol. 1, pp. 281–297. University of California Press (1967)
32. Mao, Y., Luo, W., Jing, J., Qiu, A., Mikolajczyk, K.: Hypo3d: Exploring hypothetical reasoning in 3d. arXiv preprint arXiv:2502.00954 (2025)
33. Marino, K., Rastegari, M., Farhadi, A., Mottaghi, R.: Ok-vqa: A visual question answering benchmark requiring external knowledge. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
34. McInnes, L., Healy, J., Astels, S.: Hdbscan: Hierarchical density based clustering. In: Journal of Open Source Software. vol. 2, p. 205 (2017). https://doi.org/10.21105/joss.00205
35. OpenAI: Gpt-5 system card. https://openai.com/index/gpt-5-system-card (2025), accessed: 2026-06-30
36. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Mazare, P., Lenc, K., Haziza, D., Massa, F., Assran, M., Ballas, N., Galuba, W., Howes, R., Smith, T., Arbel, M., Misra, I., Gelly, S., Bojanowski, P., Joulin, A., Caron, M.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
37. Qu, M., Wu, Y., Liu, W., Liang, X., Song, J., Zhao, Y., Wei, Y.: Rio: A benchmark for reasoning intention-oriented objects in open environments. In: Advances in Neural Information Processing Systems. pp. 43041–43056 (2023)
38. Schwenk, D., Khandelwal, A., Clark, C., Marino, K., Mottaghi, R.: A-okvqa: Augmented knowledge visual question answering. In: European Conference on Computer Vision (ECCV) Workshops (2022)
39. Suhr, A., Zhou, S., Zhang, A., Zhang, I., Bai, H., Artzi, Y.: A corpus for reasoning about natural language grounded in photographs. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL). Florence, Italy (2019)
40. Tang, Y., Zhang, L., Zhang, S., Zhao, Y., Hao, X.: Roboafford: A dataset and benchmark for enhancing object and spatial affordance learning in robot manipulation. In: Proceedings of the 33rd ACM International Conference on Multimedia. p. 12706–12713. MM '25, Association for Computing Machinery (2025). https://doi.org/10.1145/3746027.3758209, https://doi.org/10.1145/3746027.3758209
41. Team, Q.: Qwen3 technical report (2025), https://arxiv.org/abs/2505.09388
42. Thrush, T., Jiang, R., Bartolo, M., Singh, A., Williams, A., Kiela, D., Ross, C.: Winoground: Probing vision & language models for visio-linguistic compositional reasoning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5238–5248 (2022)
43. Tong, P., Brown, E., Wu, P., Woo, S., IYER, A.J.V., Akula, S.C., Yang, S., Yang, J., Middepogu, M., Wang, Z., et al.: Cambrian-1: A fully open, vision-centric exploration of multimodal llms. Advances in Neural Information Processing Systems **37**, 87310–87356 (2024)
44. Wang, J., Li, W., Wu, Y., Liang, Y., Guo, Y., Li, C., Duan, H., Zhang, Z., Zhai, G.: Affordance benchmark for mllms. arXiv preprint arXiv:2506.00893 (2025)
45. Wang, W., Gao, Z., Gu, L., Pu, H., Cui, L., Wei, X., Liu, Z., Jing, L., Ye, S., Shao, J., et al.: Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. arXiv preprint arXiv:2508.18265 (2025)
46. Wang, X., Ma, W., Zhang, T., de Melo, C.M., Chen, J., Yuille, A.: Spatial457: A diagnostic benchmark for 6d spatial reasoning of large mutimodal models. In: Pro-

ceedings of the Computer Vision and Pattern Recognition Conference. pp. 24669–24679 (2025)
47. Yang, J., Yang, S., Gupta, A.W., Han, R., Fei-Fei, L., Xie, S.: Thinking in space: How multimodal large language models see, remember, and recall spaces. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 10632–10643 (2025)
48. Yang, S., Zhang, C., Xu, Y., Gao, X., Zhang, L., Zhang, D.: Ram++: Enhanced regional aggregation modeling for vision-language understanding. arXiv preprint arXiv:2403.05612 (2024)
49. Yue, X., Ni, Y., Zhang, K., Zheng, T., Liu, R., Zhang, G., Stevens, S., Jiang, D., Ren, W., Sun, Y., et al.: Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9556–9567 (2024)
50. Yuille, A., Kersten, D.: Vision as bayesian inference: analysis by synthesis? Trends in cognitive sciences **10**(7), 301–308 (2006)
51. Zellers, R., Bisk, Y., Farhadi, A., Choi, Y.: From recognition to cognition: Visual commonsense reasoning. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)