arXiv:2505.21649v8 [cs.CV] 12 Jul 2026

Seeing Isn't Orienting: A Cognitively Informed Hierarchical Benchmark for Object Orientation in MLLMs

Nazia Tasnim*, Keanu Nichols*, Yuting Yan**, Nicholas Ikechukwu, Elva Zou, Deepti Ghadiyaram, and Bryan A. Plummer

Boston University
{nimzia, kmn5409, ncholas, elzou030, dghadiya, bplum}@bu.edu

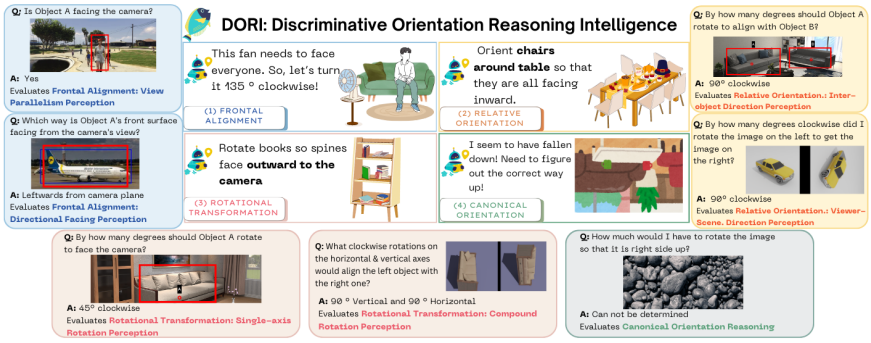


Fig. 1: DORI organizes object orientation evaluation into four diagnostic dimensions: (1) object's **directional alignment**, (2) its **orientation relative to** viewers, scenes, and other objects, (3) required **rotational transformation** for different objectives, and (4) its natural/**canonical orientation** in the world. Each dimension evaluates a targeted aspect of object orientation reasoning across diverse visual settings.

Abstract. Humans develop object orientation understanding progressively, from recognizing which way an object faces, to mentally rotating it, to reasoning about how multiple objects are oriented relative to each other. Yet current vision-language benchmarks treat orientation as an afterthought, conflating it with positional relationships and general scene understanding. We introduce **D**iscriminative **O**rientation **R**easoning **I**ntelligence (**DORI**), a cognition-informed hierarchical benchmark that establishes object orientation as the primary evaluation target. Specifically, DORI decomposes orientation into four dimensions, each assessed at both coarse (categorical) and granular (metric) levels. This results in 33,656 multiple-choice questions over diverse open-vocabulary objects within 13,652 real-world and synthetic images taken from 14 sources.

* Equal contribution

** Work done while at Boston University; currently at Johns Hopkins University - yyan79@jh.edu

DORI's coarse-to-granular design isolates orientation from confounds such as object recognition difficulty, scene clutter, and linguistic ambiguity through bounding-box isolation, standardized spatial reference frames, and structured prompts. Our evaluation of 26 state-of-the-art vision-language models reveals a consistent pattern: *models competent on general spatial benchmarks remain near-random on object-centric orientation tasks*. Even the best models achieve only 64.2% on coarse and 42.9% on granular judgments, with the largest drops on compound rotations and inter-object reference frame shifts. Large coarse-to-granular gaps further expose that models rely on categorical heuristics rather than geometric reasoning, a limitation invisible to existing benchmarks. These findings establish orientation understanding as an unsolved challenge in multimodal systems, with direct implications for robotic manipulation, 3D scene reconstruction, and human-AI interaction. Find the dataset here: `https://huggingface.co/datasets/appledora/DORI-Benchmark`

1 Introduction

Object orientation understanding demands complex, multi-stage processing of intrinsic object features, viewer perspective, angular relationships, and reference frame transformations [12,46,54,101]. Humans master this fundamental aspect of visual cognition [3, 17] through various inherent sensory-motor experiences, proprioceptive integration, and neural formation [96]. In the human cognitive system, these mechanisms develop progressively from basic frontal orientation recognition to complex rotational transformations [74, 97, 99]. This equips us with a crucial aptitude for real-world visual tasks that require precise spatial interaction with the environment, such as tool manipulation and navigation. This capability also underpins real-world tasks such as robotic grasping [14, 18], autonomous navigation [35, 50], and augmented reality object alignment [77, 81, 105]. Multimodal large language models (MLLMs) [7, 95, 103] are currently attractive for many of these applications, as they perform well on various spatial reasoning tasks [36,43,65,77,81,109]. Yet, existing benchmarks reveal consistent failures in orientation tasks [13,15,58,59,71,93] while overlooking key limitations: narrow focus on simple directions [110], synthetic-only data [108], ambiguous prompts [76], egocentric bias [53], or small scales [33]. These limitations lead to an incomplete assessment of a model's orientation reasoning abilities, potentially failing to identify critical weaknesses in real-world scenarios and mask true geometric reasoning deficits, favoring memorized shortcuts over human-like cognition.

As summarized in Fig. 1, we address these shortcomings by introducing **D**iscriminative **O**rientation **R**easoning **I**ntelligence (**DORI**), a multifaceted benchmark to evaluate object orientation understanding in multimodal language models. Tab. 1 shows how we draw on developmental literature to motivate a principled decomposition of orientation understanding into four dimensions of increasing complexity: (1) frontal alignment perception, (2) rotational transformations, (3) ego and allocentric relative orientation understanding, and (4) natural or canonical orientation of objects. Our benchmark further employs a

Table 1: Four diagnostic dimensions in DORI, motivated by prior work on human visuospatial cognition. The cited literature provides task-level motivation for evaluating orientation across perception, transformation, reference-frame reasoning, and canonical pose understanding.

Dimension	Motivating Process	Description	Relevant Human Cog. Dev. Studies
(1) Frontal Alignment	Basic Perception	Identifying a front-facing surface's orientation relative to the viewer.	Early visual cortex [5, 9] supports viewpoint-invariant frontal recognition via pathways distinct from angular estimation [46, 56].
(2) Rotational Transformation	Mental rotation and spatial transformation.	Simulating orientation changes through single- or multi-axis rotation.	Overlapping premotor and parietal circuits activate during physical and imagined rotation [25, 78, 113], with load scaling with transformation complexity [21, 30, 61, 91].
(3) Relative Orientation	Multi-object frame-of-reference integration.	Inter-object and viewer-relative orientation reasoning.	Hippocampal-parietal circuits [42, 55, 72, 80, 125] support cross-viewpoint orientation tracking essential for scene comprehension.
(4) Canonical Orientation	Semantic, physics-grounded orientation knowledge.	Detecting deviations from expected poses and inferring corrective transformations.	Ventral stream and prefrontal regions [40, 85] encode intuitive physics [6, 46, 75], constrained by gravity, feature positioning, and ecological validity [10, 31, 45].

two-tiered assessment framework to provide a more holistic understanding of MLLM performance: **coarse-grained** questions to evaluate basic categorical understanding (*e.g.*, "is the car facing toward or away from the camera?") and **fine-grained** questions to probe precise metric relationships (*e.g.*, "at what angle is the car oriented relative to the camera?"). DORI leverages **13652** images from **14** diverse sources to generate **33,656** multiple-choice questions, combining real-world images (37%) with simulated renders (63%) to ensure we have a large dataset with varying levels of visual complexity.

We generate clear, unambiguous prompt-answer pairs through a rigorous three-step process: (1) isolating objects with bounding boxes to tackle cluttered scenes, (2) employing standardized orientation terminology (*e.g.*, "frontal alignment") with explicit spatial frames of reference (*e.g.*, egocentric, allocentric, and object-centric), examples, and task descriptions, and (3) ensuring difficulty progression from simple categorical judgments to precise angular measurements across all orientation dimensions. This systematic approach isolates orientation from scene perception skills, minimizes confounding factors such as object recognition difficulty, scene clutter, linguistic ambiguity, and contextual distractions that plague existing benchmarks [13, 16, 60]. Our extensive experiments with 26 state-of-the-art MLLMs on DORI reveal several key findings:

– Models struggle most with tasks requiring spatial transformation and perspective shifts: performance drops by 30% on dynamic rotational tasks relative to static pose identification, and by 25% on allocentric tasks requiring non-egocentric viewpoint adaptation (*e.g.*, determining if two objects face each other from their own frames of reference).

Table 2: Comparison with standard benchmarks for orientation-reasoning. Asterisks (*) indicate counts refer only to orientation-related tasks, not the full dataset. DORI is larger and/or more diverse than similar datasets. *C = Coarse, G = Granular; View: A = Allocentric, E = Egocentric; Ω = Open-vocabulary*

Dataset	Obj.	Granul.	View	#Sources	N-Imgs	S-Imgs	VQA
Spatial-MM [86]	342	C	AE	1	537	–	0.6K
ScanQA [4]	20	C	A	1	800	–	7.4*K
BLINK [34]	71*	C	AE	2	552*	–	0.5*K
KiVA [123]	62*	G	A	2*	–	400*	0.6*K
EgoOrientBench [53]	197	C	E	5	6K	2.3K	33.4K
EmbSpatial-Bench [27]	294	C	E	3	1.5K*	683*	2.4*K
CLEVR-3D [115]	160	C	A	2	1.0K	–	2.3*K
3DSRBench [73]	Ω	C	A	1*	2.0K*	–	2.7*K
PO3D-VQA [106]	5	C	A	1	–	∼30K	300*K
Mind the Gap [89]	197	C	E	2*	∼280*	∼520*	1.2*K
Spatial457 [107]	5	C	A	1	–	992*	4.5*K
VSI-Bench [119]	47	C	A	3	288	–	5K
OmniSpatial [51]	Ω	C	AE	9	8.4K	–	8.4K
VLM4D [128]	Ω	C	AE	4	600	400	1.8K
MMSI-Bench [120]	Ω	C	AE	8	1.9K	–	1K
DORI (Ours)	**Ω**	**CG**	**AE**	**14**	**5K**	**9K**	**33.6K**

– Architectural design and prompt tuning matter more than scale: token-based integration (*e.g.*, Mantis-Idfs-8B [52]) consistently outperforms linear projection, and smaller instruction-tuned models (*e.g.*, DeepSeek-1.3B-Chat [7]) often surpass larger base counterparts (*e.g.*, DeepSeek-7B-Base [7]).
– LoRA finetuning on DORI generalizes broadly, boosting performance by up to 27% on standard spatial reasoning benchmarks including BLINK [34], SAT [84], and 3DSRBench [89].

2 Related Work

Understanding where an object is pointing, how it has rotated, or whether it sits in its expected pose are deceptively simple questions, yet MLLMs consistently struggle with them [13, 15, 93]. Tab. 2 contextualizes DORI against prior spatial and orientation-related benchmarks. Existing datasets have made important progress in broader visual reasoning, but orientation is often treated as a secondary subset rather than the primary evaluation target. For example, BLINK [34], Spatial-MM [86], and EmbSpatial-Bench [27] include orientation-related questions, but primarily emphasize coarse directional judgments and do not systematically probe rotational transformation, allocentric reasoning, or canonical pose understanding.

Tab. 2 also highlights limitations in scale, granularity, and visual diversity. Most prior benchmarks support only coarse or categorical questions, making it difficult to distinguish failures in basic facing-direction perception from failures in precise angular reasoning. Synthetic datasets such as PO3D-VQA [106] and KiVA [123] provide precise ground truth but contain limited or no natural images, while EgoOrientBench [53] is restricted to egocentric frames and discrete

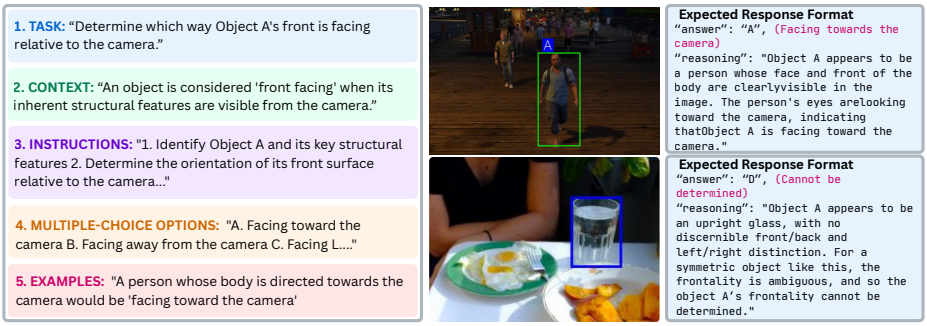


Fig. 2: Structured prompt design and example question–answer pairs from DORI. Each query follows a consistent format comprising a task description, contextual definition, step-by-step instructions, multiple-choice options, and illustrative examples, enabling systematic evaluation of orientation perception. All questions include the option "Cannot be determined," maintaining a uniform answer space while explicitly modeling cases of frontality ambiguity or insufficient visual evidence.

orientation classes. Other benchmarks such as EmbSpatial-Bench [27], 3DSR-Bench [73], OmniSpatial [51], MMSI-Bench [120], and RoboSpatial [87] evaluate spatial reasoning more broadly but do not treat object orientation as a primary, multi-dimensional target. Pose estimation approaches [23, 26, 126] offer precise 6-DoF outputs but require costly pipelines [20, 29] and resist natural-language interpretation.

DORI addresses these limitations by evaluating object orientation as a primary task across complementary dimensions and difficulty levels. It combines open-vocabulary objects, both egocentric and allocentric coverage, and coarse-plus-granular assessment, a combination not jointly provided by prior benchmarks in Tab. 2. Drawing from 14 sources across natural and simulated environments, DORI yields 13,652 images and 33,656 questions, providing the diagnostic resolution needed to identify where and why orientation reasoning fails in current MLLMs.

3 *DORI*: Discriminative Orientation Reasoning Intelligence

Object orientation understanding spans a wide spectrum of perceptual and cognitive demands - from static pose recognition and mental rotation to egocentric and allocentric spatial reasoning and physics-grounded semantic inference. Evaluating MLLMs across this full range requires a principled framework that can expose not just *whether* models fail, but *where* and *why*. For example, how closely do they match human-level orientation performance, do their reasoning patterns reflect those observed in human visuospatial cognition [8, 88, 99], and what architectural or training factors drive their shortcomings? These questions jointly motivate DORI's task structure, data selection, and evaluation design which we will discuss further in the next section.

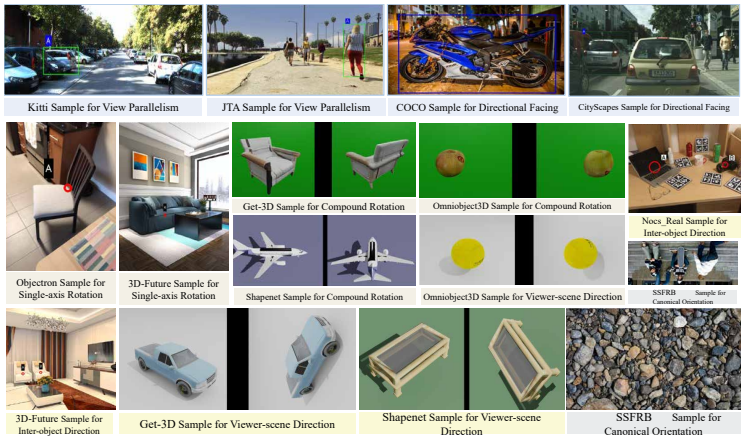


Fig. 3: Representative samples from each dataset, including natural (Kitti, Cityscapes, Coco, SSFRB, etc.) and simulated (JTA, 3D-Future, Get-3D, etc.) sources.

3.1 Core Capabilities

Drawing on established frameworks in cognitive neuroscience, DORI decomposes object orientation understanding into the four dimensions. Tab. 1 shows how each axis corresponds to a separable perceptual or reasoning mechanism identified in the visuospatial literature as diagnostically meaningful [30, 46, 85, 100]. We do not claim a strict developmental mapping. Instead, this decomposition motivates *why* orientation understanding must be evaluated along multiple targeted axes, and *what* visual and cognitive processing types each axis demands. Below we describe how each dimension manifests as concrete evaluation tasks in DORI.

1. Frontal Alignment evaluates the fundamental ability to perceive how an object's front-facing surface is oriented relative to the viewer - a prerequisite for any orientation-based reasoning. Humans rapidly identify which way an object is facing (*e.g.*, deciding if a vehicle is approaching or departing) by recognizing structural and functional features such as faces, headlights, or entrances. However, we require additional reasoning steps for orientation interpretation [56], such as assessing objects' angular relationship with the viewing plane defined as the imaginary plane onto which a scene is projected onto.

For MLLMs, we assess this frontal alignment capability through two complementary tasks: **view parallelism analysis**, which quantifies the degree to which an object's frontal surface deviates from being parallel to the image plane, providing angular measurements (*e.g.*, is a chair directly facing the camera or at a 45-degree angle). **Directional facing perception** asks the cardinal direction an object's front is oriented relative to the camera position (*e.g.*, whether a desk is facing toward, away, leftward, or rightward from the viewer's perspective). This dual assessment is supported by research that reports viewpoint-invariant recognition of frontal features operates through different neural mechanisms than precise angular estimation [46]. Prior work also demonstrates MLLMs' inability to perceive object frontality (*e.g.*, confusing left/right perspectives) even when

provided with bounding boxes [86, 94], often being on par with random predictions [122]. For objects whose frontality is difficult to define (*e.g.*, symmetric objects like a ball), models are expected to respond with the "Cannot be determined" option, which is provided as a choice in every question (see example in Fig. 2).

2. Rotational Transformation examines the ability of an MLLM to comprehend orientation changes through rotation, reflecting requirements such as embodied object manipulation (*e.g.*, fitting a key into a lock), and viewpoint-dependent navigation (*e.g.*, reorienting a map for wayfinding). Mental rotation capabilities allow humans to predict how objects align when rotated [78], with neuroimaging evidence showing premotor and parietal activation during both physical & imagined rotations [25, 78]. Both processes trigger similar neural activations [113], highlighting the inherent complexity of spatial transformation tasks that MLLMs must simulate. Inspired by this, we design the rotational tasks in our benchmark to progress from simple to complex levels mirroring human cognitive processing demands [61, 91, 113]. Thus, the first subtask examines **single-axis rotation** & evaluates basic angular transformations rotated along one spatial dimension (*e.g.*, determining the shortest rotation to face a chair toward the camera). This establishes baseline capabilities before progressing to more cognitively demanding compound rotation [91], involving **multiple-axes rotation** (*e.g.*, aligning objects through a sequence of horizontal/vertical rotations). These subtasks represent common scenarios in assembling products through item manipulation and real-world scene planning by embodied agents, where object orientation understanding directly impacts final task performance.

3. Relative Orientation examines the understanding of how objects are oriented in relation to each other and with respect to the viewer. Humans navigate a complex visual world by seamlessly tracking orientation changes across scenes and viewpoints, which is crucial for scene comprehension and geometric coordination. The human brain contains a specific interconnected region facilitating "mental orientation" -the ability to effectively spatially orient an object to different viewpoints and perspectives [42, 72, 80, 125]. In contrast, studies have shown that MLLMs struggle significantly with questions posed from non-egocentric perspectives [86, 105, 121] and with maintaining consistent dimensional relationships between multiple objects across time and viewpoints [119] We systematically probe this aspect of object's relative orientation via the following sub-tasks: (1) **inter-object directional relationships**, to assess the relative facing directions of objects (*e.g.*, determine if two cars are facing the same or opposite directions), and (2) **image-pair rotational relationships**, to measure the ability to track orientation changes between two images (*e.g.*, identify the degree of rotation between two views of the same object).

3.2 Benchmark Creation Process

DORI comprises seven orientation-reasoning tasks spanning the four dimensions above (Fig. 1), each evaluated at two tiers: **coarse-grained** questions for ba-

sic categorical judgments (*e.g.*, "Are objects A and B facing the same direction?" for relative orientation; "Has the object rotated between the two images?" for rotational transformation) and **fine-grained** questions for precise quantitative estimation (*e.g.*, "How many degrees clockwise did the object rotate?"; "What specific transformations would restore this image to its canonical orientation?"). This hierarchical design enables systematic evaluation from fundamental perception to advanced metric reasoning. We generated the questions via two pipelines: converting existing 3D annotations into orientation questions, drawing from JTA [28], 3D-Future [32], Get3D [37], ShapeNet [44], OmniObject3D [112], NOCS REAL [102], Objectron [1], SSFRB [41, 70, 79, 98, 111], and the OmniNOCS [57] subsets of KITTI [38] and Cityscapes [19]; and manually annotating real-world samples from COCO [63]. Source selection is also task-specific: for instance, inter-object directional relationship questions require multi-object scenes, so single-object datasets like ShapeNet [44] are excluded in favor of 3D-Future [32] and NOCS REAL [102] (further details in Sec. 3.3). To handle orientation ambiguity in symmetric objects, every question includes a "Cannot be determined" option regardless of whether it is correct for that instance (Fig. 2). We describe the full curation details in the supplementary; representative image samples are shown in Fig. 3.

We designed the prompts to isolate orientation perception from confounding factors such as object recognition difficulty, scene clutter, linguistic ambiguity, and contextual distractions [11] (see supplementary for detailed discussion). The prompts follow a structured five-component format (Fig. 2): a **task description** specifying the orientation dimension being tested; contextual grounding of key concepts (*e.g.*, "An object is considered 'front facing' when its inherent structural features are visible from the camera"); **step-by-step analysis instructions** (*e.g.*, "1. Identify Object A and its key structural features. 2. Determine the orientation..."); **multiple-choice options**; and concrete reasoning **examples** (*e.g.*, "A person whose body is directed towards the camera would be 'facing toward the camera.'"). This design draws on instruction-tuning principles showing that explicit task framing and example-driven guidance improve model comprehension [47]. We have iteratively refined the prompts through multiple cycles of human feedback from non-expert annotators to address ambiguities, clarify terminology, and improve the task specificity [64]. For instance, early rotational transformation prompts yielded inconsistent axis interpretations, addressed by introducing concrete analogies (*e.g.*, "like a ballerina spinning clockwise" to distinguish vertical-axis rotation from abstract directional descriptions). Response formats were standardized to require both an explicit answer and a chain-of-thought explanation, ensuring that performance differences reflect genuine orientation understanding rather than prompt interpretation variance. See supplementary for full prompt templates and examples across all question types.

Label reliability. DORI's labels are primarily derived from structured 3D annotations, rotation matrices, and controlled rendering pipelines, which provide direct access to object orientation parameters. The main exception is the

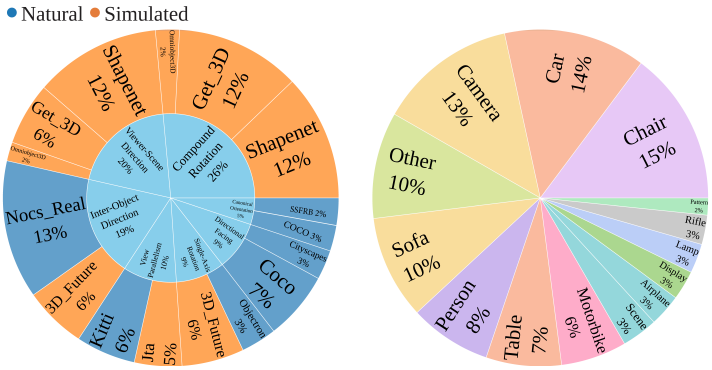


Fig. 4: Dataset composition in DORI. The left chart shows the distribution of examples across the seven orientation-reasoning tasks and natural/simulated image sources. The right chart shows the distribution of examples across broad object groups. Together, the plots summarize DORI's task, source, and object-group diversity.

directional-facing subset from COCO, where orientation labels require manual judgment from real-world images. To assess ambiguity in this manually annotated subset, we conducted an agreement analysis with three annotators and observed complete agreement in 96.7% of cases. This suggests that the manual labels are largely unambiguous, while the majority of DORI's labels are obtained from deterministic geometric annotations rather than subjective judgment.

3.3 DORI Statistics

Collectively, DORI encompasses **13,652** images spanning both natural (37%) and simulated (63%) environments, and includes **33,656** carefully constructed multiple-choice questions over diverse open-vocabulary objects across **14** computer vision datasets, including KITTI [38], Cityscapes [19], COCO [63], JTA [28], 3D-FUTURE [32], Objectron [1], ShapeNet [44], and OmniObject3D [112], among others (full list in Sec. 3.2). Fig. 4 illustrates the broad object-group and dataset distribution. This multi-source approach allows us to balance complex, real-world environments (37% of images) with controlled synthetic environments (63%) where orientation parameters are precisely known. Knowing precise orientation parameters in synthetic data is crucial as it provides ground truth angular measurements with known accuracy, eliminating visual ambiguity or occlusion that might confound assessment. Meanwhile, real-world images introduce natural complexity and diversity while maintaining clear ground truth through expert annotation. Additional statistics on fine-grained and broad categories across tasks are available in the supplementary.

4 Experiments

We evaluate 26 state-of-the-art multimodal models spanning diverse architectures, parameter scales, and pretraining methodologies across both open-source

Table 3: Performance of several open-source MLLMs on DORI. Most models perform poorly, particularly on granular questions. These experiments reveal a systemic gap in object orientation understanding across all four dimensions studied in DORI. See supplementary for results with model variance. C: coarse, G: granular.

	Frontal Alignment				Rotational Transformation				Relative Orient.				Canonical		Avg.	
	View Parallel.		Dir. Facing		Single-axis Rot.		Compound Rot.		Inter-Obj. Dir.		Viewer-scene Dir.		Orient.			
	C	G	C	G	C	G	C	G	C	G	C	G	C	G	C	G
Random	35.7	25.5	20.2	19.8	20.7	17.4	20.4	6.6	15.0	10.1	33.1	20.3	34.8	16.0	25.7	16.5
LLaVA-v1.6- [69]	55.9	26.8	22.6	19.8	23.4	16.5	32.8	**10.6**	7.2	12.4	46.6	20.1	34.7	11.3	31.9	16.7
LLaVA-v1.6-34B [69]	52.8	35.3	32.6	26.4	22.1	26.2	13.0	4.3	16.2	14.8	34.8	25.4	40.1	11.6	30.2	20.5
LLava-Next-8B [68]	33.2	25.2	21.3	21.0	20.5	17.9	40.6	6.3	10.3	12.0	61.5	25.2	39.5	9.2	32.4	14.4
Yi-VL-6B [2]	38.9	31.0	25.0	25.2	28.5	14.7	29.6	3.0	15.2	12.3	41.4	10.8	32.8	14.4	30.2	15.9
Yi-VL-34B [2]	53.1	35.1	23.3	24.0	28.1	21.2	32.5	4.4	13.4	14.5	61.1	19.7	36.7	12.5	35.4	18.7
Mantis-CLIP [52]	60.1	24.3	26.1	16.3	15.4	20.9	9.1	5.8	13.2	10.1	37.7	17.9	30.1	34.5	27.3	15.0
Mantis-Idfs-8B [52]	57.8	33.0	22.5	12.7	25.7	23.4	25.9	6.6	17.6	9.0	55.4	24.5	**51.2**	**41.0**	36.5	17.5
DS-1.3B-Base [7]	58.1	24.8	15.0	19.4	21.1	15.2	17.0	5.3	17.0	11.9	61.3	26.0	1.8	2.3	27.3	14.7
DS-1.3B-Chat [7]	58.1	24.8	22.6	22.0	21.2	15.5	33.9	5.8	15.3	10.5	47.6	18.0	20.8	17.3	31.3	14.1
DS-7B-Base [7]	26.3	31.1	14.6	16.4	18.3	14.3	35.5	2.8	3.6	4.1	35.7	6.8	28.3	10.7	23.1	9.6
DS-7B-Chat [7]	59.4	31.3	31.3	24.5	28.2	17.8	35.8	6.0	**20.2**	14.9	18.5	32.2	44.4	32.2	33.9	18.6
Qwen2.5-3B-Inst. [90]	56.3	29.3	23.4	3.1	18.2	17.7	25.4	0.21	16.3	13.8	62.8	16.9	43.2	11.4	35.0	13.2
Qwen3-8B-Inst. [116]	73.7	45.6	50.8	54.8	**32.9**	**45.9**	**49.0**	7.5	18.2	**21.9**	**79.1**	37.2	42.0	21.1	**49.3**	33.4
Qwen3-32B-Inst [116]	**75.8**	**54.4**	**61.1**	**59.2**	31.6	44.1	39.6	8.6	9.2	20.5	78.9	46.1	28.1	25.5	46.3	**36.9**
Magma-8B [118]	51.5	21.4	25.4	20.9	20.6	18.7	32.0	5.4	16.3	11.5	32.6	21.0	37.5	16.1	30.8	16.4
RP-v1-13B [124]	45.7	20.5	24.7	22.3	24.9	14.7	37.6	5.6	12.9	12.7	61.2	20.9	39.8	10.6	35.2	15.3
IntVL-14B-Inst. [104]	63.4	37.2	37.9	48.6	30.3	21.4	26.4	7.2	4.3	15.0	94.3	**44.1**	45.8	16.0	43.2	27.1
IntVL-30B-A3B-Inst. [104]	72.1	33.9	46.6	50.0	17.4	17.7	33.3	6.8	19.3	10.0	90.1	35.0	51.1	22.2	47.1	25.1

Table 4: Comparing open and closed-source MLLMs on a randomly selected 100-sample-per-task subset of DORI. Closed-source models show improved performance, but all models still perform poorly overall.

	Frontal Alignment				Rotational Transformation				Relative Orient.				Canonical		Avg.	
	View Parallel.		Dir. Facing		Single-axis Rot.		Compound Rot.		Inter-Obj. Dir.		Viewer-scene Dir.		Orient.			
	C	G	C	G	C	G	C	G	C	G	C	G	C	G	C	G
LLaVA-v1.6-34B [68]	45.5	30.0	22.5	23.5	15.0	15.5	19.3	7.6	15.0	10.5	37.3	34.3	66.0	19.0	31.5	20.0
LLava-Next-8B [68]	34.0	13.0	18.0	19.0	12.5	17.5	40.3	7.0	12.0	13.5	64.0	25.0	66.0	16.0	35.2	15.8
DS-7B-Chat [7]	52.5	28.0	24.5	24.5	21.0	15.0	35.3	5.0	24.0	14.0	23.6	32.0	67.0	2.0	35.4	19.7
Qwen3-8B-Inst. [116]	61.5	41.0	38.5	36.0	23.5	37.0	**69.6**	9.3	17.5	13	88.3	47	62.0	26.0	51.5	29.9
Qwen3-32B-Inst [116]	60.0	39.5	37.5	33.5	20.5	37.5	68.3	10.3	16.5	13.5	86.6	44.6	45.0	26.0	47.7	29.2
IntVL-14B-Inst. [104]	61.5	32.0	32.5	42.5	29.0	16.0	31.3	8.3	6.5	14.0	**93.3**	45.6	71.0	25.0	46.4	26.2
IntVL-30B-A3B-Inst. [104]	67.5	34.0	40.5	39.0	18.0	11.0	35.0	8.0	26.0	13.0	89.0	33.6	79.0	33.0	50.7	24.5
Gemini 1.5 Pro	68.5	53.0	43.5	39.0	24.0	37.5	65.3	12.3	25.0	14.5	91.0	47.3	83.0	28.0	57.1	33.0
Gemini 2.0 Flash	67.0	35.0	37.5	33.5	28.5	27.0	52.0	14.3	23.0	**17.0**	92.0	43.6	78.0	29.0	54.0	28.5
Gemini 3 Flash pre.	**68.5**	**71.0**	**63.0**	**63.0**	59.0	18.5	37.5	**63.7**	**44.0**	8.0	91.7	**57.0**	86.0	**51.0**	**64.2**	42.0
Claude-sonnet-4.6	61.5	41.0	42.5	45.5	26.0	35.5	66.0	21.0	23.5	16.0	82.7	45.0	90.1	42.9	56.0	35.3
GPT-4o	44.5	45.0	30.5	35.5	29.5	34.0	39.3	15.3	23.5	13.5	88.3	53.3	88.0	45.0	49.0	34.5
GPT-4-1	48.5	41.0	39.0	40.5	32.5	42.0	31.6	14.0	22.5	8.5	87.6	44.6	**93.0**	46.0	50.6	33.8
GPT-5-mini	61.5	55.5	55.0	60.0	**37.5**	**47.5**	46.3	22.6	19.0	10.5	88.0	57.3	86.0	41.0	56.1	**42.9**
GPT-5.5-Pro	67.0	65.5	61.0	65.0	25.5	34.5	48.7	24.7	42.5	8.5	92.7	48.7	75.0	47.0	58.9	41.9

and proprietary systems. The models are as follows: LLaVA-v1.6-13B [67], LLaVA-v1.6-34B [67], LLava-Next-8B [68], Yi-VL-6B [2], Yi-VL-34B [2], Mantis-CLIP [52], Mantis-Idfs-8B [52], DS-1.3B-Base [7], DS-1.3B-Chat [7], DS-7B-Base [7], DS-7B-Chat [7], Qwen2.5-3B-Inst. [90], Qwen3-8B-Inst. [116], Qwen3-32B-Inst. [116], Magma-8B [118], RP-v1-13B [124], IntVL-14B-Inst [104], and IntVL-30B-A3B-Inst. [104]. Proprietary models are queried via their official APIs; open-source models are run on 2–4 NVIDIA RTX A6000 GPUs, with models exceeding 34B parameters evaluated on a single NVIDIA A100-80G. All models are evaluated at temperature 0 for deterministic outputs. Answer options are shuffled randomly across runs to control for position bias, and responses are parsed via a standardized extraction pipeline that handles free-form text, letter-based, and JSON-formatted outputs. We report answer accuracy separately for coarse and granular question types across all task categories.

Granular evaluation tolerance. For granular questions, we use task-specific angular tolerances that reflect the resolution of each orientation judgment. The

Table 5: Model rankings across MMSI-Bench, OmniSpatial, and DORI. Cell color encodes rank (green = high, red = low), superscripts encode rank. DORI's coarse (C) and granular (G) split reveals patterns invisible to single-score benchmarks.

Model	MMSI-Bench	OmniSpatial	DORI (C)	DORI (G)
GPT-5-mini	41.9[1]	–	**56.1**[1]	**42.9**[1]
Gemini 2.0 Flash	–	44.0[4]	54.0[2]	28.5[4]
GPT-4.1	30.9[2]	**51.8**[1]	50.6[3]	33.8[3]
InternVL3- [104]	26.8[4]	45.9[3]	46.4[5]	27.1[5]
RoboPoint-v1-13B [124]	–	34.6[6]	35.2[6]	15.3[6]
Qwen2.5-VL-3B [90]	26.5[5]	40.3[5]	35.0[7]	13.2[7]
GPT-4o	30.3[3]	47.8[2]	49.0[4]	34.5[2]

main results use a strict matching criterion to evaluate whether models recover the intended metric orientation rather than only the correct coarse category. In the supplementary, we additionally report a more tolerant evaluation setting. Model rankings are largely stable under this relaxed criterion, indicating that the observed failures are not solely due to small angular miscalibrations, but often reflect larger errors in orientation reasoning.

4.1 Results

Tab. 3 benchmarks open-source MLLMs across all DORI question types, and reveals a fundamental limitation: many open-source MLLMs perform at or near random chance. Granular questions are especially challenging. Most models rely on coarse categorical judgments rather than precise angular reasoning. Even robotics-specialized models, such as, Magma-8B [118] and RP-V1-13B [124] - fail to exceed general-purpose architectures, indicating that domain-specific pretraining alone is insufficient. Larger model families outperform smaller ones on coarse questions, but this advantage largely disappears on fine-grained tasks.

Tab. 4 extends this comparison to closed-source systems. On coarse accuracy, the leading proprietary models outperform the best open-source alternatives by up to 12.7 percentage points, yet several closed-source systems fail to consistently exceed top open-source models. On granular accuracy, the picture is more mixed: GPT-5-mini and GPT-5.5-Pro show a more meaningful lead, while Gemini 2.0 Flash remains near the best open-source level. Across both question types, all models exhibit severe deficits on rotational transformation and relative orientation. This confirms that robust geometric understanding remains absent even in state-of-the-art proprietary systems. We attribute this to pretraining methodology. Most MLLMs employ CLIP-style contrastive objectives that optimize for semantic alignment rather than geometric structure [127]. Tong *et al.* [94] similarly noted that pretraining creates a "dimensional collapse" in the embedding space, where continuous orientation variations become compressed into discrete semantic clusters (*e.g.*, treating "left" and "right" as opposing categorical concepts rather than points along a continuous angular spectrum). While this may be slightly mitigated by using generative objectives [62], MLLMs lack the necessary equivalent neural inductive biases utilized by humans [22,39,66,82]. Instead,

Table 6: Performance of Qwen2.5-3B-Inst [90] model finetuned using LoRA using Linear Projection, Linear Projection with a bottleneck, & Token-based Fusion on DORI

	Frontal Alignment				Rotational Transformation				Relative Orient.				Canonical		Avg.	
	View Parallel.		Dir. Facing		Single-axis Rot.		Compound Rot.		Inter-Obj. Dir.		Viewer-scene Dir.		Orient.			
	C	G	C	G	C	G	C	G	C	G	C	G	C	G	C	G
Linear Proj	55.9	39.4	19.4	0.0	36.6	28.0	55.9	5.6	25.0	17.8	77.0	26.2	18.8	11.7	41.2	19.1
Linear Proj+Btlneck	55.6	32.4	26.1	0.0	19.1	23.8	30.1	5.0	19.0	10.8	41.6	18.8	33.5	14.1	32.1	14.8
Token based fusion	87.2	72.4	75.0	2.5	64.9	65.6	63.1	14.9	67.7	56.5	94.1	69.5	67.6	55.8	74.3	46.6

MLLMs approximate neural mechanisms through suboptimal attention patterns leading to hallucinations [24, 49, 114, 117].

Tab. 5 compares model rankings against MMSI-Bench [120] and OmniSpatial [51] - two benchmarks covering general scene-level and multi-image spatial reasoning. In contrast, DORI isolates orientation, enabling a direct test of whether general spatial competence transfers to geometric reasoning. The ranking shifts are substantial: GPT-4o places 2nd on OmniSpatial and 3rd on MMSI-Bench, but drops on DORI's coarse orientation split, despite remaining competitive on granular questions. This suggests that strong performance on broader spatial reasoning benchmarks does not necessarily imply robust performance across all forms of object-centric orientation reasoning, a blind spot that DORI is designed to expose.

4.2 Analysis

Architecture critically determines orientation capacity. Tab. 3 also reveals clear architectural hierarchies among open-source models. Token-based fusion (*e.g.*, Mantis-Idfs-8B [52]) consistently outperforms linear projection, preserving richer spatial information. Instruction-tuned variants universally surpass base models: DeepSeek-7B-Chat outperforms DeepSeek-7B-Base [7] by ∼47% relative on average coarse accuracy. Strikingly, smaller tuned models can surpass larger base ones, e.g: DeepSeek-1.3B-Chat [7] exceeds DeepSeek-7B-Base [7] on coarse questions. This confirms that architectural design, pretraining data, and training objectives matter more than parameter count [83]. The 41% relative coarse gain from Qwen2.5-3B [90] to Qwen3-8B [116] further shows that advances in vision-language alignment directly benefit geometric understanding.

To further isolate the contribution of fusion strategy, we finetune Qwen2.5-3B-Inst. with LoRA under three conditions for 2000 steps: standard linear projection, linear projection with a bottleneck (hidden dim 256, GELU activation), and the model's default token-based fusion. As shown in Tab. 6, token-based fusion substantially outperforms both projection variants across all tasks (74.3% vs. 41.2% average coarse accuracy), corroborating the pattern observed in Tab. 3. Adding a bottleneck to linear projection further degrades performance (32.1% average coarse), indicating that information compression at the fusion stage is particularly harmful for fine-grained orientation reasoning. We note, however, that this experiment carries a confounding variables: Qwen was pretrained with

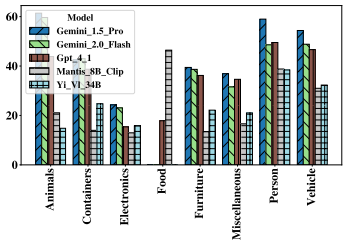


Fig. 5: Performance of MLLMs by source category (additional models in supplementary). Although for many categories the relative ranking of methods is relatively stable, in a few cases, like the food category, most models perform poorly.

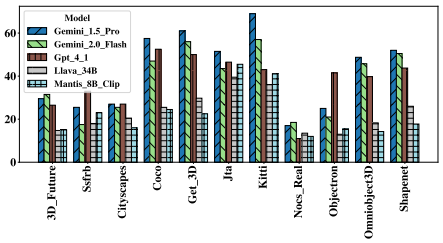


Fig. 6: Performance by source dataset. Models generalize better to simulated data and coarse questions, exposing a gap on natural images and fine-grained tasks (additional models in supplementary).

token-based fusion, so the linear projection variants are not part of their native training regime. A fully controlled comparison would require retraining comparable MLLMs from scratch under each fusion strategy, which is beyond the scope of this work. We therefore interpret these results as suggestive rather than causal: within this adaptation setting, token-based fusion performs substantially better, indicating that fusion strategy is an important design axis that should be considered alongside model scale.

Real versus synthetic sources. Because DORI combines real and synthetic images, we separately examine performance across source types. While performance varies across individual datasets and visual domains, the aggregate model rankings are largely consistent between real and synthetic subsets. Thus, our conclusions are not driven by a single data source or by synthetic rendering artifacts alone. Instead, the mixed-source design allows DORI to combine controlled orientation ground truth with the visual diversity of real-world scenes.

Category-specific patterns expose reliance on semantic cues. Models perform substantially better on people and animals than on furniture or containers (Fig. 5), that have clear front/back distinctions. This gap also reflects reliance on semantic shortcuts (*e.g.*, recognizing faces) rather than a fundamental geometric understanding when determining object orientation. Fig. 6 reveals a complementary pattern: performance is markedly higher on simulated datasets than natural images, suggesting that real-world clutter and variability expose geometric limitations that controlled renders conceal. In addition, strong results on datasets like COCO raise contamination concerns. To address this, we followed [92], and separated into (likely) in-pretraining and out-of-pretraining sets, and find that models generally perform similarly on both sets. Thus, confirming negligible contamination effects.

Fine-tuning on DORI enables substantial transfer gains. Tab. 7 measures zero-shot transfer to three external benchmarks (3DSRBench [73], Blink [33], and SAT [84]) after fine-tuning on DORI. Fine-tuning Qwen2.5-VL-3B [90] on 27K DORI samples (real + synthetic) with LoRA [48] yields up to 27 percentage points absolute gain. Tab. 8 further reports performance on 7K held-out

Table 7: Comparison of a base Qwen2.5-VL-3B model and a LoRA-tuned variant trained on DORI, evaluated zero-shot on 3DSRBench [73], Blink [33], and the SAT benchmark [84]. Finetuning on DORI yields up to a 27% improvement over the base model. Note: the SAT Finetuning row reports results from the official SAT model repository, which uses a Qwen2.5-VL-7B backbone - substantially larger than the Qwen2.5-VL-3B models used in all other comparisons.

	Blink			3DSRBench				SAT
	Multi-View Reasoning	Visual-Corresp.	Relative-Depth	Orient.	Multi-Obj. View To. Obj.	Multi-Obj. Parallel	Multi-Obj. Same Dir.	
Base Model	42.9	25.3	64.1	32.5	11.4	11.4	46.8	51.3
SAT Finetuning	40.6	66.2	66.9	33.7	24.4	50.0	50.4	50.8
DORI Finetuning	**45.1**	**26.5**	**65.3**	**38.6**	**14.0**	**38.1**	**50.3**	**63.3**

Table 8: Performance of Base and Fine-tuned Qwen2.5-3B-Inst. [90] models across DORI Question Types

	Frontal Alignment				Rotational Transformation				Relative Orient.				Canonical Orient.		Avg.	
	View Parallel.		Dir. Facing		Single-axis Rot.		Compound Rot.		Inter-Obj. Dir.		Viewer-scene Dir.					
	C	G	C	G	C	G	C	G	C	G	C	G	C	G	C	G
Base Model	57.7	34.8	27.2	0.0	22.9	20.4	37.7	5.6	7.3	13.7	77.7	22.0	47.0	20.2	39.6	16.6
SAT Fintuning	36.3	42.2	22.5	26.1	33.8	23.6	24.2	4.2	10.8	14.7	76.5	26.0	47.0	20.0	35.8	19.6
+ Finetuned	**87.8**	**72.8**	**81.9**	**2.6**	**76.8**	**76.1**	**64.5**	**17.3**	**79.2**	**60.6**	**99.0**	**84.9**	**67.6**	**37.1**	**79.5**	**50.2**

DORI samples, showing 40–34 percentage point absolute gains over the base model. Together, these results show that DORI finetuning generalizes across orientation and spatial reasoning tasks, capturing transferable orientation signals rather than prompt-specific patterns. However, transfer to non-orientation tasks is mixed, indicating orientation-centered rather than universal gains.

Human evaluation exposes a performance ceiling. Tab. 9 compares expert human performance against the best closed-source model across all task types. We recruited 14 experts to evaluate 50 examples per task type (700 samples total), each with three answer options: the correct answer, "Cannot be determined," and a random distractor. Humans achieved up to 98% accuracy, validating annotation quality. The best closed-source model trails by roughly 17 percentage points, revealing substantial headroom for MLLM improvement.

Prompt structure scaffolds orientation understanding. We ablated key prompt components: formatting, examples, task descriptions, and step-by-step reasoning (Fig. 2), across 8 models. Tab. 10 compares performance without structured prompting. Most models degrade sharply without structure: LLaVA-v1.6-13B [69] collapses to near-zero on Direction Facing granular, Compound Rotation granular, and Canonical Orientation granular, while LLaVA-Next-8B [69] suffers a 93% relative decline on View Parallelism granular. For these architectures, structured prompts yield up to 50–100× improvements on fine-grained tasks. Prompt sensitivity is not uniform, however. Mantis-Idfs-8B [52] shows counterintuitive *gains* under ablation: Compound Rotation coarse surges 141% relative, suggesting rigid templates interfere with its interleaved architecture's native multi-image reasoning. DS-7B-Chat [7] is similarly non-monotonic: Inter-

Table 9: Comparison of **Human performance vs. MLLMs** on 50 randomly sampled questions of each type. We find a significant performance gap exists between our expert annotators and MLLMs

	Frontal Alignment				Rotational Transformation				Relative Orient.				Canonical Orient.		Avg.	
	View Parallel.		Dir. Facing		Single-axis Rot.		Compound Rot.		Inter-Obj. Dir.		Viewer-scene Dir.					
	C	G	C	G	C	G	C	G	C	G	C	G	C	G	C	G
GPT-4o	22.0	36.0	38.0	6.0	36.0	34.0	76.0	66.6	38.0	48.0	88.0	70.6	70.0	20.0	52.5	40.1
Gemini 1.5 Pro	64.0	72.0	76.0	80.0	56.0	54.0	86.0	66.0	64.0	62.0	98.0	80.0	50.0	64.0	70.5	68.2
Human	82.0	92.0	92.0	82.0	80.0	78.0	86.0	86.0	86.0	74.0	98.0	90.0	92.0	92.0	88.0	84.8

Table 10: Models perform much worse for most of the tasks when the prompts are unstructured and have no clarification examples. The degradation is particularly prominent for Compound rotation and Canonical orientation related tasks

	Frontal Alignment				Rotational Transformation				Relative Orient.				Canonical Orient.		Avg.	
	View Parallel.		Dir. Facing		Single-axis Rot.		Compound Rot.		Inter-Obj. Dir.		Viewer-scene Dir.					
	C	G	C	G	C	G	C	G	C	G	C	G	C	G	C	G
LLaVA-v1.6-13B	57.9	33.0	25.4	0.01	20.2	16.5	27.7	0.01	16.7	12.2	60.4	19.8	43.1	0.0	35.9	11.6
Yi-VL-6B	52.4	37.8	23.9	20.3	36.8	15.9	19.7	21.6	11.0	16.5	78.6	18.6	22.3	12.9	34.9	20.5
Mantis-CLIP	40.5	9.1	6.9	6.8	6.2	2.6	0.9	1.8	0.2	0.0	57.9	1.9	41.6	46.7	22.0	6.7
Mantis-Idfs-8B	51.2	29.9	29.7	13.6	15.3	10.5	62.5	5.0	5.2	18.4	85.9	32.1	32.6	46.1	40.3	19.9
LLava-Next-8B	56.9	1.8	25.5	29.9	22.2	11.7	34.1	6.6	24.3	10.0	66.0	26.4	38.7	11.0	38.2	13.9
DS-1.3B-Chat	40.7	30.2	22.6	21.8	12.2	19.8	35.6	5.6	16.2	14.2	31.1	21.6	14.8	9.8	24.7	17.5
DS-7B-Base	45.2	35.0	21.6	21.2	25.4	18.0	2.8	5.7	15.0	11.7	33.6	25.7	3.6	16.5	21.0	19.1
DS-7B-Chat	48.1	32.4	27.9	25.7	47.0	19.5	1.0	5.1	49.6	15.8	36.6	22.7	17.3	10.7	32.5	18.8

Object Direction coarse improves 145% relative, yet Compound Rotation coarse collapses 97%. These divergences show that prompts can disambiguate some spatial transformations while disrupting others. Notably, Canonical Orientation remains difficult across all models, indicating that prompt engineering alone cannot replace geometric priors.

5 Conclusion

We introduce DORI, a cognitively informed benchmark isolating orientation understanding in MLLMs. Across 26 models, a consistent pattern emerges: systems competent on general spatial benchmarks struggle severely on orientation tasks, trailing expert humans by roughly 18 percentage points. DORI's coarse-to-granular design reveals why models rely on categorical heuristics rather than continuous geometric reasoning, a limitation that broad benchmarks cannot detect. Canonical Orientation is particularly revealing: while some models improve on coarse canonical judgments, granular canonical reasoning remains difficult and highly inconsistent, suggesting that robust orientation understanding likely requires stronger architectural or pretraining-level geometric priors rather than scaling alone. Encouragingly, LoRA fine-tuning on DORI transfers broadly to held-out spatial benchmarks, showing that targeted orientation training yields

genuine geometric gains. Further discussions, error analysis, and additional experiments are provided in the supplementary.

Seeing Isn't Orienting: A Cognitively Informed Hierarchical Benchmark for Object Orientation in MLLMs . . . 1
Nazia Tasnim, Keanu Nichols, Yuting Yan, Nicholas Ikechukwu, Elva Zou, Deepti Ghadiyaram, and Bryan A. Plummer*
1 Introduction . . . 2
2 Related Work . . . 4
3 *DORI*: Discriminative Orientation Reasoning Intelligence . . . 5
3.1 Core Capabilities . . . 6
3.2 Benchmark Creation Process . . . 7
3.3 DORI Statistics . . . 9
4 Experiments . . . 9
4.1 Results . . . 11
4.2 Analysis . . . 12
5 Conclusion . . . 15
A Benchmark Construction and Statistics . . . 17
A.1 Task-wise Dataset Creation . . . 17
A.2 Compositional Diversity . . . 18
A.3 Mapping Object Categories to Broad Classes . . . 19
A.4 VQA Examples in DORI . . . 19
A.5 Enumerated List of Questions . . . 21
A.6 Confounding Factors . . . 22
B Additional Analyses of Model Performance . . . 32
B.1 Performance by Common Categories . . . 32
B.2 Performance by Data-Source . . . 33
B.3 Task Complexity Hierarchy . . . 34
C Additional Error Analysis . . . 38
C.1 Systematic Error Pattern Identification . . . 38
C.2 Component-wise Error Decomposition . . . 39
C.3 Soft Accuracy Calculation . . . 40
C.4 Performance Error Bars . . . 42
C.5 Out-Of-Pretraining Analysis . . . 46
D Human Evaluation Interface . . . 54
E Terminology Glossary . . . 57

A Benchmark Construction and Statistics

A.1 Task-wise Dataset Creation

View parallelism perception. Evaluates a model's ability to determine whether an object's front-facing surface is oriented toward, away from, or perpendicular to the camera plane. We constructed this dataset using images from the JTA [28] and KITTI [38] datasets (specifically the subset used in OmniNOCS [57]). For JTA, which contains 3D human pose annotations, we calculated orientation by analyzing shoulder positions relative to the camera and head angle to precisely determine facing direction. For KITTI, we leveraged the available rotation matrices to categorize vehicles and pedestrians based on their orientation relative to the camera. This task is critical for fundamental scene understanding, where determining which objects are facing an agent is essential for interaction and navigation decisions.

Frontal alignment perception. Extends orientation assessment to cardinal directions, requiring models to identify if objects are facing toward, away, leftward, or rightward relative to the camera. We developed this dataset using images from COCO [63] and Cityscapes [19] (from the OmniNOCS [57] subset). For COCO images, which lack orientation annotations, we conducted expert manual labeling of object orientations. For Cityscapes, we utilized rotational matrices to precisely determine directional orientation, limiting images to contain at most three objects to ensure assessment clarity. This directional understanding is vital for spatial navigation and object manipulation tasks where agents must understand not just if objects face them, but their specific directional orientation.

Single-axis rotation. Assesses understanding of rotational transformations around a vertical axis by asking models to determine the optimal rotation direction and precise angular adjustment needed for objects to face the camera. We constructed this dataset using 3D-Future [32], which provides high-resolution 3D furniture models with known 6-DoF parameters. We focused primarily on chair variants with distinctive front/back features, calculating the exact rotational adjustment needed for the object to face the camera directly. This capability forms the foundation for computational manipulation planning and scene reconfiguration understanding. Furthermore, we utilized the Objectron [1] subset of the OmniNOCS [57] dataset. This also included 6 DoF information which we used to determine the angle of the object with respect to the camera, for this dataset we utilized bikes, chairs and bottles.

Compound rotation. Evaluates comprehension of complex rotations involving sequential transformations around multiple axes, where the rotation order impacts the final orientation. We developed this dataset using 3D-rendered objects from Get3D [37], ShapeNet [44], and OmniObject3D, [112] implementing a controlled rendering pipeline in Blender. For each object, we rendered an initial third-person view, then applied precise rotations along horizontal and vertical axes in varying sequences, rendering the transformed state. This task assesses

the sophisticated mental rotation capabilities required for complex object manipulation and orientation reasoning across multiple dimensions.

Inter-object direction perception. Evaluates understanding of relative orientation between multiple objects from their own perspectives rather than the camera view. Using the 3D Future [32] and NOCS REAL [102] datasets, we leveraged 6 DoF parameters to calculate precise angular relationships between object pairs. The task requires models to determine if objects face the same direction, opposite directions, or have perpendicular orientations, progressing to granular assessment of the exact rotation needed for objects to align. This capability is essential for understanding agent-object and object-object relationships in complex scenes, particularly for collaborative robotic tasks or scene arrangement planning.

Viewer-scene direction perception. Evaluates perception of rotational changes between two images of the same object. Using Get3D [37], ShapeNet [44], and OmniObject3D [112] datasets, we rendered objects with a ground plane reference, then created corresponding images with the object rotated by specific angles around a vertical axis. Models must determine whether rotation occurred and, at the granular level, specify the exact degree of rotation. This assessment examines the ability to track orientation changes across views - a crucial capability for video understanding, temporal reasoning, and object tracking applications.

Canonical orientation reasoning. Evaluates models' understanding of normal object orientations and their ability to identify when objects appear in non-canonical positions. Using a subset of COCO [63] images with clear orientation expectations (*e.g.*, people standing upright, vehicles with wheels on the ground), we created variations with systematic flips and rotations. Models must first identify whether images appear in their canonical orientation, then determine the specific transformations (rotation, flipping, or both) needed to restore proper orientation. This capability assesses world knowledge about typical object positioning, which is critical for anomaly detection, image correction, and understanding intentional vs. unintentional orientation deviations.

A.2 Compositional Diversity

Aforementioned 14 datasets span various natural and simulated domains, capturing varied object instances, backgrounds, occlusion levels, lighting conditions, and viewpoints. Such breadth of distribution ensures that the evaluation probes both the generalization ability of MLLMs and their robustness to contextual, visual, and category shifts across domains.

Fig. 7 further complements this by presenting the distribution of the top 25 most frequent object categories across the datasets. While a few object classes, such as *car*, *person*, *camera*, and *chair* dominate in frequency, the distribution spans a broad range of object types and source domains. This heterogeneity plays a critical role in evaluating the performance of models not only on popular

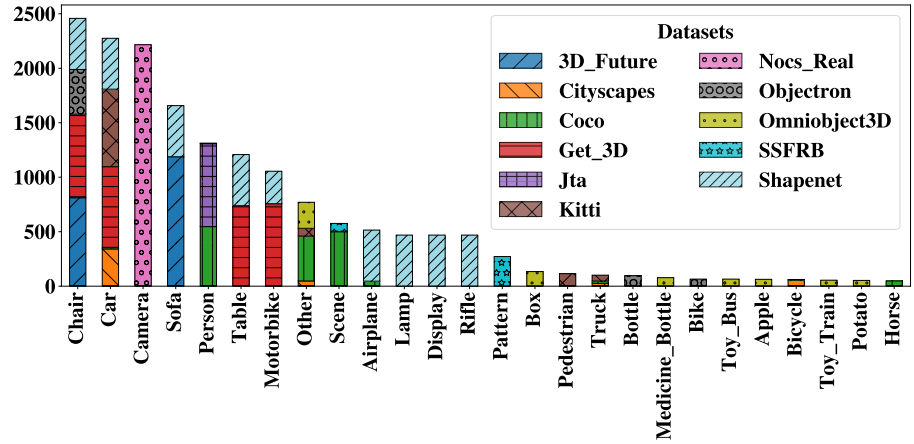


Fig. 7: Sample category distribution of top 25 object categories showing their sources

categories but also on long-tail classes, thereby encouraging more balanced and comprehensive assessment.

Fig. 8 presents the distribution of the top 25 most frequent object categories in DORI, annotated with their corresponding orientation task types. This visualization highlights not only which categories are most prevalent, but also how they are utilized across the various tasks in DORI. While commonly occurring objects such as *chair*, *car*, and *sofa* span several task types, others are more narrowly concentrated. This overlap and separation across tasks reflect the intentional diversity of DORI, designed to evaluate model capabilities across both general-purpose and category-specific orientation challenges. The distribution also underscores the need for models to generalize effectively across unevenly represented categories and task combinations.

A.3 Mapping Object Categories to Broad Classes

To enable category-level analysis, we group the fine-grained object labels in DORI into a smaller set of semantically meaningful broad classes. This abstraction facilitates more interpretable evaluation across heterogeneous datasets and reduces sparsity in underrepresented categories. The mapping spans common categories such as `Vehicle`, `Person`, `Animals`, `Food`, `Containers`, `Furniture`, `Electronics`, and `Miscellaneous`, and is illustrated in Tab. 11 and Fig. 9.

A.4 VQA Examples in DORI

To further contextualize model behavior, we present a curated selection of Visual Question Answering (VQA) examples drawn from the DORI benchmark,

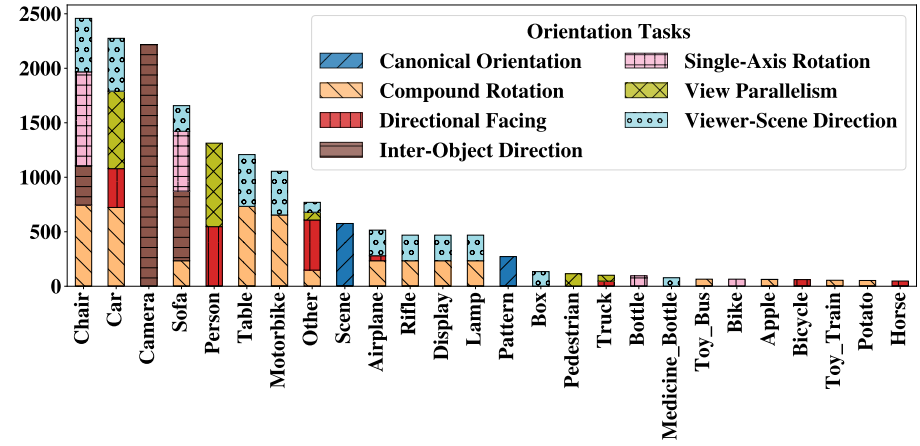


Fig. 8: Sample category distribution of top 25 object categories showing their associated orientation tasks

covering each of the seven distinct orientation reasoning tasks. These qualitative illustrations shed light on the nuanced challenges faced by state-of-the-art Multimodal Large Language Models (MLLMs), beyond aggregate metrics.

Each example is carefully chosen to represent either a canonical failure or, in rarer cases, a surprising success. The samples span both coarse-level and granular-level questions, reflecting the dual axes of abstraction and visual complexity within DORI. Despite confident language in many model predictions, we observe frequent dissonance between answer correctness and the underlying rationale, especially for tasks requiring precise spatial alignment or inter-object reasoning. For instance, in the **View Parallelism** and **Directional Facing** tasks, models often misjudge subtle orientation cues, such as limb articulation or gaze direction as we show in Fig. 10,Fig. 11, Fig. 12, and Fig. 13, leading to confidently incorrect predictions. Likewise, for **Single-Axis** and **Compound Rotation** scenarios, we show in Fig. 14,Fig. 15, Fig. 16, and Fig. 17, that even top-performing models struggle with mentally simulating object motion, resulting in angular miscalculations or overly generic justifications.

Notably, **Inter-object Direction** and **Viewer-Scene Direction** tasks as we show in Fig. 18,Fig. 19, Fig. 20, and Fig. 21, expose limitations in relational orientation reasoning, with models frequently underestimating angular disparities or misrepresenting the directional frame of reference. The final task, **Canonical Orientation** as seen in Fig. 22, underscores a broader epistemic gap: models often assert certainty in inherently ambiguous scenarios, revealing an overconfidence not grounded in the visual evidence.

Together, these examples highlight persistent limitations in visual-spatial grounding, even among the most capable contemporary MLLMs. They underscore the need for further architectural innovations and training strategies to

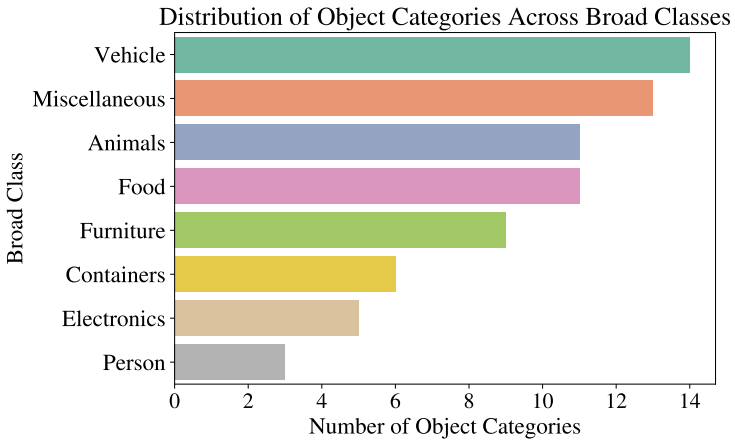


Fig. 9: Distribution of fine-grained object categories mapped to broader semantic classes. Most categories fall under *Vehicle*, *Miscellaneous*, and *Animals*, reflecting common object types in the evaluated datasets.

imbue models with a deeper, more structured understanding of object orientation dynamics.

A.5 Enumerated List of Questions

Coarse Questions

- **Q1 - View Parallelism:** Determine which way Object A's front is facing relative to the camera.
- **Q2 - Directional Facing:** Determine which direction Object A's front-facing surface is oriented from the camera's viewpoint.
- **Q3 - Single-axis Rotation:** Determine the shortest direction of rotation for Object A to face the camera.
- **Q4 - Compound Rotation:** Determine what type of rotation the object has undergone between the two images.
- **Q5 - Inter-object Direction:** Determine if objects A and B are facing each other from their own perspectives.
- **Q6 - Viewer-Scene Direction:** Determine if the object has rotated between the two images.
- **Q7 - Canonical Orientation:** Determine if the image is in its canonical orientation.

Fine-grained/Granular Questions

- **Q1 - View Parallelism:** Determine how much Object A's front surface deviates from being parallel to the camera plane.

Table 11: Mapping from object categories to broad semantic classes.

Object Category	Broad Class
car, truck, tram, van, bus, train, airplane, boat,	Vehicle
motorcycle, bicycle, coach, bike, motorbike, trailer	Vehicle
pedestrian, cyclist, person	Person
zebra, dog, elephant, giraffe, cat, horse, sheep,	Animals
bird, cow, bear, starfish	Animals
pizza, broccoli, donut, orange, apple, tomato, potato	Food
cake, egg, wine glass, bowl, bottle, onion	Food
bottle, cup, wine glass, bowl, medicine_bottle, box	Containers
chair, sofa, table, lamp, bench, sofa-chair, soft-sofa	Furniture
laptop, tv, display, camera, laptop-camera	Electronics
clock, vase, toilet, umbrella, teddy bear, rifle	Miscellaneous
stop sign, toy_train, toy_bus, house, ball	Miscellaneous
scene, pattern	Miscellaneous

- **Q2 - Directional Facing:** Identify the precise orientation of Object A's front-facing surface from the camera's viewpoint
- **Q3 - Single-axis Rotation:** Determine the closest clockwise rotation needed for Object A to face the camera.
- **Q4 - Compound Rotation:** Determine the exact rotation angles the object has undergone between the two images.
- **Q5 - Inter-object Direction:** Determine how much Object B would need to rotate to face Object A.
- **Q6 - Viewer-Scene Direction:** Determine how many degrees clockwise the object has rotated between the two images.
- **Q7 - Canonical Orientation:** Determine how the image can be restored to its canonical orientation.

A.6 Confounding Factors

In our dataset we have carefully designed our questions for both the prompt and image to reduce the amount of different confounding factors that are present in past datasets. We include aspects like: defining axes, defining what the canonical view is, and we direct the readers to section Sec. A.4 & Sec. E which showcases these elements present in our questions.

We can see in examples like Fig. 11 we utilize a bounding box which helps the model/rater determine which object to focus on; this helps reduce object recognition difficulty. For scene clutter we can look at Fig. 21 which focuses on particular objects without any objects being present in the image. Fig. 13 showcases an example of reducing linguistic ambiguity, by describing what the MLLM should look at and identify its front facing surface, additionally objects that have ambiguous front facing surface like a table we have labelled those as



Ground Truth:

C: Facing away from the camera

Gemini_1.5_pro Answer:

"answer": "B",

"reasoning": "Object A is a person. The key structural feature for a person is their face. We can see her face clearly and we can tell her body is oriented towards the camera, therefore Object A is facing directly towards the camera."

Prompt:

TASK: Determine which way Object A's front is facing relative to the camera.

CONTEXT: Object A is marked with a bounding box in the image. An object is considered 'front facing' when its inherent structural features (e.g., a person's face, car headlights, TV screen) are visible from the camera.

Please analyze this image using these steps:
1. Identify Object A and its key structural features 2. Determine the orientation of its front surface relative to the camera 3. Compare with the provided options 4. Select the most appropriate answer 5. Explain your reasoning with specific visual evidence

OPTIONS:
A. Not visible enough to determine B. Directly facing the camera
C. Facing away from the camera

FORMAT YOUR RESPONSE AS:
"answer": "Letter/number of chosen option",
"reasoning": "Step-by-step explanation citing specific visual evidence"

EXAMPLES:
- A person whose body is directed towards the camera would be "facing toward the camera"

Fig. 10: An example VQA from the **View Parallelism orientation task** illustrating a failure case on a coarse-level question in DORI, with a prediction from Gemini 1.5 Pro. Many models failed this question, highlighting a common challenge among MLLMs in understanding fundamental scene geometry. Specifically, in reasoning whether an object's front-facing surface is oriented toward, away from, or perpendicular to the camera plane.

"Cannot be determined". For contextual distractions we can also look to Fig. 16 which contains backgrounds that do not influence the exact orientation of the object.



Ground Truth:

A : 135 degrees to 180 degrees

Gemini-2.0-Flash Answer:

"Answer": "B",

"Reasoning": "Object A's back is to the camera. Therefore, the deviation from the camera plane is between 0 degrees and 15 degrees."

Prompt:

TASK: Determine how much Object A's front surface deviates from being parallel to the camera plane.

CONTEXT: Object A is marked with a bounding box in the image. You'll be assessing the angle between Object A's front surface and the camera plane (the camera taking this photo). Please analyze this image using these steps: 1. Identify Object A and establish its front-facing surface 2. Estimate the angular deviation between this surface and the camera plane 3. Compare with the provided options 4. Select the most appropriate answer 5. Explain your reasoning with specific visual evidence

OPTIONS: A. 135 degrees to 180 degrees B. 0 degrees to 15 degrees C. Cannot be determined D. 65 degrees to 95 degrees

FORMAT YOUR RESPONSE AS: "answer": "Letter/number of chosen option", "reasoning": "Step-by-step explanation citing specific visual evidence" EXAMPLES: - A person looking directly at the camera: 0-15 degrees - A person is perpendicular to the camera: 65-95 degrees - A car with its rear-end to the camera: 135-180 degrees

Fig. 11: An example VQA from the **View Parallelism orientation task** highlighting a failure case on a granular-level question in DORI, with a response from Gemini 2.0 Flash. Although the model confidently selects an answer, its reasoning reflects a fundamental misunderstanding of the object's orientation relative to the camera. While the ground truth indicates the object is turned between 135° and 180° away, the model incorrectly infers a near-parallel alignment, underscoring limitations in fine-grained spatial perception among current MLLMs.



Prompt:

TASK: Determine which direction Object A's front-facing surface is oriented from the camera's viewpoint.

CONTEXT: Object A is marked with a bounding box in the image. You will identify its front-facing surface orientation based on inherent structural features.Please analyze this image using these steps:1. Identify Object A and establish its front-facing surface 2. Determine the direction this surface is oriented relative to the camera3. Compare with the provided options 4. Select the most appropriate answer 5. Explain your reasoning with specific visual evidence

OPTIONS:{'A': 'Toward the camera' 'B': "Can't be determined" 'C': 'Away from the camera' 'D': 'Rightward relative to the camera plane' 'E': 'Leftward relative to the camera plane'}

FORMAT YOUR RESPONSE AS: "answer": "Letter/number of chosen option" "reasoning": "Step-by-step explanation citing specific visual evidence"

EXAMPLES:- If you see a car with its headlights directly pointing toward the camera select "Toward the camera"- If you see a person with their back to the camera select "Away from the camera"

Ground Truth:

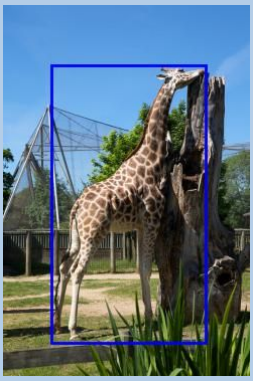
E: Leftward relative to the camera plane

Gpt_4_1 Answer:

Answer: "B"

"reasoning": "The person inside the bounding box (Object A) is facing away from the camera. This is evident because the back of their head and body are visible while their face is not. The orientation of the shoulders and the direction of the arms also indicate that the front-facing surface (the person's face and torso) is directed away from the camera."

Fig. 12: A coarse-level VQA from the **Directional Facing task** in DORI, illustrating a failure case by GPT-4-1. While the ground truth indicates a leftward orientation, the model predicts 'away,' misreading body posture and head direction. Notably, 13 out of 15 models failed this question, underscoring a widespread struggle with coarse directional inference.



Prompt:

TASK: Determine which direction Object A's front-facing surface is oriented from the camera's viewpoint.

CONTEXT: Object A is marked with a bounding box in the image. You will identify its front-facing surface orientation based on inherent structural features. Please analyze this image using these steps: 1. Identify Object A and establish its front-facing surface 2. Determine the direction this surface is oriented relative to the camera 3. Compare with the provided options 4. Select the most appropriate answer 5. Explain your reasoning with specific visual evidence

OPTIONS: A. Can't be determined B. Rightward relative to the camera plane C. Away from the camera D. Toward the camera E. Leftward relative to the camera plane FORMAT YOUR RESPONSE AS: "answer": "Letter/number of chosen option", "reasoning": "Step-by-step explanation citing specific visual evidence"

EXAMPLES: - If you see a car with its headlights directly pointing toward the camera, select "Toward the camera" - If you see a person with their back to the camera, select "Away from the camera"

Ground Truth:

B : Rightward relative to the camera plane

Gemini-2.0-Flash Answer:

"Answer": "E",

"Reasoning": "The giraffe in the image (Object A) is oriented so that its front-facing surface (as indicated by its face, neck, and body) is primarily directed leftward relative to the camera's perspective. This is because the giraffe is largely facing towards the left side of the image."

Fig. 13: A coarse-level VQA from the **Directional Facing task** in DORI, showing a failure case by Gemini-2.0-Flash. While the ground truth identifies the giraffe's orientation as rightward, the model incorrectly infers a leftward direction. This highlights the difficulty in interpreting animal pose and orientation cues, a challenge shared by the majority of models in this example.



Ground Truth:

E: 180 degrees

GPT-4O Answer:

"Answer": "E"
"reasoning": "The object marked as 'A' is a chair and it does not have a clear front-facing surface that can be definitively identified. Without a clear indication of its front orientation it is impossible to determine how many degrees of clockwise rotation would be needed for it to face the camera. Therefore the correct answer is 'Cannot be determined."

Prompt:

TASK: Determine the closest clockwise rotation needed for Object A to face the camera.

CONTEXT: Object A is marked with a bounding box in the image. Objects can rotate clockwise or counterclockwise around a vertical axis (running from top to bottom).Please analyze this image using these steps: 1. Identify Object A and establish its front-facing surface 2. Determine its current orientation relative to the camera3. Calculate how many degrees of clockwise rotation would be needed to make it face the camera4. Compare with the provided options 5. Select the most appropriate answer 6. Explain your reasoning with specific visual evidence

OPTIONS:{'A': '0 degrees' 'B': '90 degrees' 'C': '45 degrees' 'D': '135 degrees' 'E': '180 degrees' 'F': 'Cannot be determined'} FORMAT YOUR RESPONSE AS:"answer": "Letter/number of chosen option" "reasoning": "Step-by-step explanation citing specific visual evidence"

EXAMPLES:- If the object is already facing the camera the answer is "0 degrees"- If the object's front is facing to your right a "90 degrees" clockwise rotation would make it face the camera- If the object's front is at a 45-degree angle to your right a "45 degrees" clockwise rotation would make it face the camera- If the object's front is at a 135-degree angle to your right a "135 degrees" clockwise rotation would make it face the camera- If the object's front is facing away from the camera a "180 degrees" clockwise rotation would make it face the camera- If the object has no clear front or its orientation cannot be determined select "Cannot be determined"

Fig. 14: A granular-level VQA from the **Single-Axis Rotation** task in DORI, featuring a case where GPT-4o selects the correct answer (180° rotation) but provides flawed reasoning. While the model concludes with the correct choice, its explanation suggests uncertainty due to perceived ambiguity in object orientation. This mismatch between prediction and rationale underscores gaps in visual reasoning, with only 2 out of 15 models answering correctly.



Ground Truth:

E: 180 degrees

Gemini-2.0-Flash Answer:

"Answer": "B",

"Reasoning": "Object A is currently oriented approximately 90 degrees clockwise from facing the camera. A clockwise rotation of 90 degrees would be needed to make it directly face the camera."

Prompt:

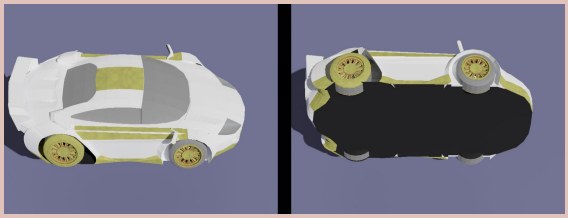
TASK: Determine the closest clockwise rotation needed for Object A to face the camera.

CONTEXT: Object A is marked with a bounding box in the image. Objects can rotate clockwise or counterclockwise around a vertical axis (running from top to bottom). Please analyze this image using these steps: 1. Identify Object A and establish its front-facing surface 2. Determine its current orientation relative to the camera 3. Calculate how many degrees of clockwise rotation would be needed to make it face the camera 4. Compare with the provided options 5. Select the most appropriate answer 6. Explain your reasoning with specific visual evidence

OPTIONS: A. 0 degrees B. 90 degrees C. 45 degrees D. Cannot be determined E. 180 degrees F. 135 degrees FORMAT YOUR RESPONSE AS: "answer": "Letter/number of chosen option", "reasoning": "Step-by-step explanation citing specific visual evidence"

EXAMPLES: - If the object is already facing the camera, the answer is "0 degrees" - If the object's front is facing to your right, a "90 degrees" clockwise rotation would make it face the camera - If the object's front is at a 45-degree angle to your right, a "45 degrees" clockwise rotation would make it face the camera - If the object's front is at a 135-degree angle to your right, a "135 degrees" clockwise rotation would make it face the camera - If the object's front is facing away from the camera, a "180 degrees" clockwise rotation would make it face the camera - If the object has no clear front or its orientation cannot be determined, select "Cannot be determined"

Fig. 15: A granular-level VQA from the **Single-Axis Rotation** in DORI, showing a failure case with a prediction from Gemini 2.0 Flash. While the correct answer is a 180° rotation, the model incorrectly selects 90° (option B), suggesting a misjudgment of the object's current orientation relative to the camera. This highlights challenges models face in reasoning about precise rotational alignment.



Gemini-2.0-Flash Answer:

"Answer": "C",

"Reasoning": "The car in the first image is upright with the wheels on the bottom. In the second image, the car is upside down with the undercarriage showing. This indicates a rotation around a horizontal axis."

Prompt:

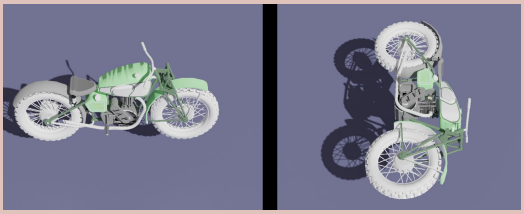
TASK: Determine what type of rotation the object has undergone between the two images.

CONTEXT: Two images of the same object are shown side by side. The object can rotate around two axes: Horizontal (like a cartwheel coming out of the screen) and Vertical (like a ballerina spinning clockwise). Please analyze these images using these steps: 1. Identify the object in both images and establish its key features 2. Compare the object's orientation in both images 3. Determine which axes the object has rotated around 4. Compare with the provided options 5. Select the most appropriate answer 6. Explain your reasoning with specific visual evidence

OPTIONS: A. Cannot be determined B. Only Vertical C. Only Horizontal D. Both Horizontal and Vertical E. No Change FORMAT YOUR RESPONSE AS: "answer": "Letter/number of chosen option", "reasoning": "Step-by-step explanation citing specific visual evidence"

EXAMPLES: - If an object appears to have moved to face left or right between images, select "Only Vertical" - If an object appears to have moved up or down (like nodding), select "Only Horizontal" - If an object has rotated both horizontally and vertically, select "Both Horizontal and Vertical" - If the object appears identical in both images with no rotation, select "No Change" - If you cannot clearly determine the type of rotation, select "Cannot be determined"

Fig. 16: A coarse-level VQA from the **Compound Rotation task** in DORI, with a correct prediction from Gemini 2.0 Flash. The model identifies a horizontal axis rotation, but its reasoning overstates the change, describing the car as 'upside down' when it is only partially inverted.



Ground Truth:
E: 90 Degrees Horizontal, 90 Degrees Vertical

Gemini-2.0-Flash Answer:

"Answer": "E",

"Reasoning": "The motorcycle in the first image is upright. In the second image, the motorcycle is rotated 90 degrees horizontally, leaning to the right. Additionally, it has been rotated 180 degrees vertically, as evidenced by the seat and rear tire now facing upward."

Prompt:

TASK: Determine the exact rotation angles the object has undergone between the two images.

CONTEXT: Two images of the same object are shown side by side. The object can rotate around two axes: Horizontal (like a cartwheel coming out of the screen) and Vertical (like a ballerina spinning clockwise). Please analyze these images using these steps: 1. Identify the object in both images and establish its key features 2. Compare the object's orientation in both images 3. Determine the horizontal rotation angle first, then the vertical rotation angle 4. Compare with the provided options 5. Select the most appropriate answer 6. Explain your reasoning with specific visual evidence

OPTIONS: A. 270 degrees Horizontal then 180 degrees Vertical B. 180 degrees Horizontal then 0 degrees Vertical C. Cannot be determined D. 180 degrees Horizontal then 90 degrees Vertical E. 90 degrees Horizontal then 180 degrees Vertical F. 0 degrees Horizontal then 0 degrees Vertical G. 270 degrees Horizontal then 90 degrees Vertical H. 270 degrees Horizontal then 270 degrees Vertical I. 0 degrees Horizontal then 90 degrees Vertical J. 180 degrees Horizontal then 270 degrees Vertical K. 0 degrees Horizontal then 180 degrees Vertical L. 0 degrees Horizontal then 270 degrees Vertical M. 90 degrees Horizontal then 0 degrees Vertical N. 90 degrees Horizontal then 270 degrees Vertical O. 90 degrees Horizontal then 90 degrees Vertical P. 180 degrees Horizontal then 180 degrees Vertical Q. 270 degrees Horizontal then 0 degrees Vertical

FORMAT YOUR RESPONSE AS: "answer": "Letter/number of chosen option", "reasoning": "Step-by-step explanation citing specific visual evidence"

EXAMPLES: - If an upright cup in the first image appears upside-down in the second image, this might be "180 degrees Horizontal then 0 degrees Vertical" - If a car facing left in the first image appears to be tilted upward and facing the camera in the second image, this might be "90 degrees Horizontal then 90 degrees Vertical" - If you cannot clearly determine the type of rotation, select "Cannot be determined"

Fig. 17: A granular-level VQA from the **Compound Rotation task** in DORI showing a failure case from Gemini 2.0 Flash. While the model selects the correct option, its reasoning misidentifies the vertical rotation, describing a 180° flip when the ground truth indicates a 90° transformation.



Ground Truth:
B: Partially facing the same direction

Gpt-4-1 Answer:

Answer: "C"

"reasoning": "...the most appropriate answer is 'Partially facing the opposite directions' as their front surfaces are somewhat oriented toward each other but not directly..."

Prompt:

TASK: Determine if objects A and B are facing each other from their own perspectives.

CONTEXT: Objects A and B are marked with red dots in the image. You will analyze their relationship from the objects' perspectives not from the camera view.Please analyze this image using these steps: 1. Identify objects A and B and establish their front-facing surfaces 2. Mentally position yourself at each object's location 3. Determine the relative directional orientation between the objects 4. Compare with the provided options 5. Select the most appropriate answer 6. Explain your reasoning with specific visual evidence

OPTIONS:{'A': 'Cannot be determined' 'B': 'Partially facing the same direction' 'C': 'Partially facing the opposite directions' 'D': 'They are facing opposite directions' 'E': 'They are facing perpendicular to each other' 'F': 'They are facing the same direction'}

FORMAT YOUR RESPONSE AS:"answer": "Letter/number of chosen option" "reasoning": "Step-by-step explanation citing specific visual evidence"

EXAMPLES:- Two people looking at each other would be "facing opposite directions" (as their faces point toward each other)- Two cars with headlights pointing the same way would be "facing the same direction"- A car and a person where the car's side faces the person's front would be "facing perpendicular to each other"- Two objects with front surfaces angled slightly toward each other would be "partially facing opposite directions"- Two objects with front surfaces angled slightly away from each other would be "partially facing the same direction"

Fig. 18: A coarse-level VQA from the **Inter-object Direction task** in DORI showing a failure case with GPT-4-1. Although the correct answer is 'Partially facing the same direction,' the model incorrectly selects 'Partially facing opposite directions,' misjudging the relative orientations. Notably, 13 out of 15 models failed this example, highlighting a shared difficulty in reasoning about partially aligned object directions.



Ground Truth:
D: 46 to 90 degrees clockwise

Gemini_1.5_pro Answer:

"Answer": "F",
"reasoning": "Object A (the black chair) is facing the coffee table in the center of the room. Object B (the sofa) is currently facing the wall with the two black canvases. To face the coffee table and therefore object A, object B needs to rotate clockwise a small amount, placing it between 0 and 45 degrees. "

Prompt:
TASK: Determine how much Object B would need to rotate to face Object A.

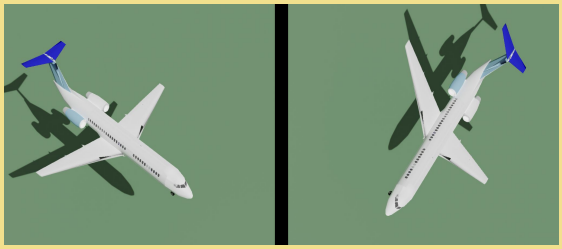
CONTEXT: Objects A and B are marked with red dots in the image. You will analyze their relationship from the objects' perspectives, not from the camera view. Please analyze this image using these steps:
1. Identify objects A and B and establish their front-facing surfaces
2. Mentally position yourself at each object's location
3. Determine how much and in which direction Object B would need to rotate to face Object A
4. Compare with the provided options
5. Select the most appropriate answer
6. Explain your reasoning with specific visual evidence

OPTIONS:
A. Cannot be determined B. No rotation C. 180 degrees in either direction
D. 46 to 90 degrees clockwise E. 46 to 90 degrees counterclockwise
F. 0 to 45 degrees clockwise G. 91 to 135 degrees clockwise
H. 0 to 45 degrees counterclockwise I. 91 to 135 degrees counterclockwise

FORMAT YOUR RESPONSE AS:
"answer": "Letter/number of chosen option",
"reasoning": "Step-by-step explanation citing specific visual evidence"

EXAMPLES:
- If Object B needs to turn slightly right to match Object A's direction, select "0 to 45 degrees clockwise"
- If Object B is already facing the same direction as Object A, select "0 degrees or no rotation"
- If Object B needs to turn completely around to match Object A, select "180 degrees in either direction"

Fig. 19: A granular-level VQA from the **Inter-object Direction task** in DORI illustrating a failure case with Gemini 1.5 Pro. While the ground truth indicates a required clockwise rotation between 46° and 90° for alignment, the model underestimates this angle. Its reasoning misinterprets the spatial alignment between the chair and sofa. All 15 models failed this example, underscoring the challenge of inter-object directional understanding.



Gemini-2.0-Flash Answer:

"Answer": "C",

"Reasoning": "The airplane rotates 90 degrees clockwise. In the first picture, the plane's nose points towards the right, in the second picture, it points to the bottom."

Prompt:

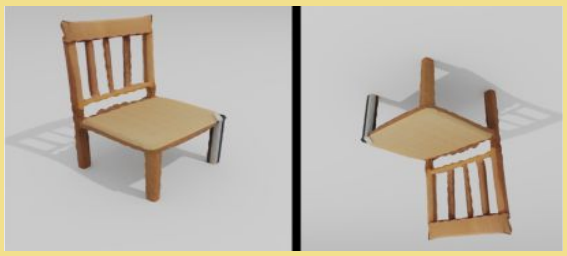
TASK: Determine how many degrees clockwise the object has rotated between the two images.

CONTEXT: Two images of the same object are shown side by side. Assume the camera remains stationary, and the object can rotate around a vertical axis. Please analyze these images using these steps: 1. Identify the object in both images and establish its key features 2. Compare the object's orientation in both images 3. Determine the precise clockwise rotation angle around the vertical axis 4. Compare with the provided options 5. Select the most appropriate answer 6. Explain your reasoning with specific visual evidence

OPTIONS: A. Cannot be determined B. 180 degrees C. 90 degrees D. 270 degrees E. 0 degrees

FORMAT YOUR RESPONSE AS: "answer": "Letter/number of chosen option", "reasoning": "Step-by-step explanation citing specific visual evidence" EXAMPLES: - Objects that are symmetrical like a ball or a box would be considered "Cannot be determined

Fig. 20: Viewer-Scene Direction. Example success case on a granular-level DORI question with Gemini-2.0-Flash. While the model correctly selects the 90-degree rotation, it inaccurately describes the directional shift as 'right' to 'bottom' instead of the more precise 'bottom-right' to 'bottom-left,' reflecting a partial misunderstanding of fine-grained orientation.



Ground Truth:

A: 180 degrees

Gpt-4-1 Answer:

Answer: "A"

"reasoning": "...Comparing the two images the chair has rotated 90 degrees clockwise around the vertical axis: the backrest moves from facing upper left to facing right ..."

Prompt:

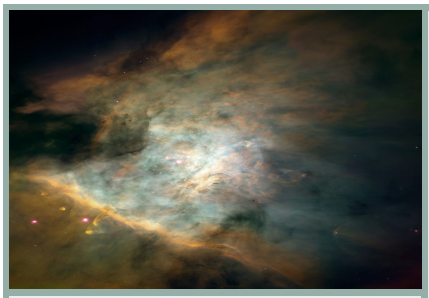
TASK: Determine how many degrees clockwise the object has rotated between the two images.

CONTEXT: Two images of the same object are shown side by side. Assume the camera remains stationary and the object can rotate around a vertical axis.Please analyze these images using these steps: 1. Identify the object in both images and establish its key features 2. Compare the object's orientation in both images3. Determine the precise clockwise rotation angle around the vertical axis 4. Compare with the provided options 5. Select the most appropriate answer 6. Explain your reasoning with specific visual evidence

OPTIONS:{'A': '180 degrees' 'B': '0 degrees' 'C': '270 degrees' 'D': '90 degrees' 'E': 'Cannot be determined'}

FORMAT YOUR RESPONSE AS:"answer": "Letter/number of chosen option" "reasoning": "Step-by-step explanation citing specific visual evidence" EXAMPLES:- Objects that are symmetrical like a ball or a box would be considered "Cannot be determined"

Fig. 21: Viewer-Scene Direction. Example failure case on a granular-level DORI question with GPT-4-1. Although the model selects the correct answer of 180 degrees, its reasoning mistakenly describes a 90-degree rotation, highlighting a disconnect between answer selection and spatial understanding. Notably, 13 out of 15 models failed this case.



Ground Truth:

B. Cannot be determined

Gemini-2.0-Flash Answer:

"Answer": "A",

"Reasoning": "The image depicts a nebula. There is no single canonical orientation for nebulae as they are typically viewed from various angles. Therefore, the image is most likely in its canonical orientation."

Prompt:

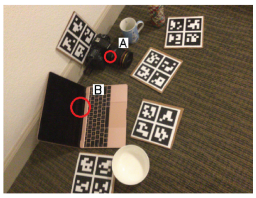
TASK: Determine if the image is in its canonical orientation.

CONTEXT: The image may be flipped or rotated or both. Canonical orientation refers to the standard way objects are typically viewed (e.g., people standing upright, cars with wheels on the ground, text reading left to right).An image that has been mirrored horizontally or vertically is considered flipped. An image that has been turned from its normal viewing angle is considered rotated. Please analyze this image using these steps: 1. Identify the main objects/elements in the image 2. Determine their expected canonical orientation 3. Compare with their current orientation in the image 4. Compare with the provided options 5. Select the most appropriate answer 6. Explain your reasoning with specific visual evidence

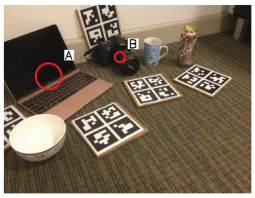
OPTIONS: A. Yes, the image is in the canonical orientation B. Cannot be determined C. No, the image is not in the canonical orientation

FORMAT YOUR RESPONSE AS: "answer": "Letter/number of chosen option", "reasoning": "Step-by-step explanation citing specific visual evidence"

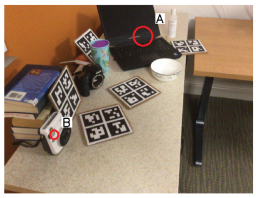
Fig. 22: Canonical Orientation. Example failure case on a coarse-level DORI question with Gemini-2.0-Flash. While the ground truth is 'Cannot be determined,' the model incorrectly selects a definitive orientation. Its reasoning contradicts the inherent ambiguity it acknowledges, exposing uncertainty in handling objects with no fixed canonical pose.



0 to 45 degrees clockwise



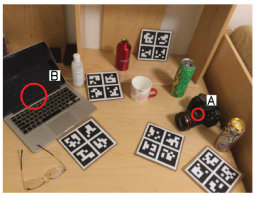
0 to 45 degrees counter-clockwise



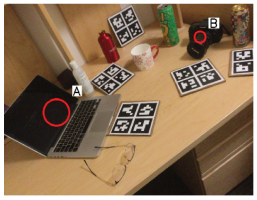
46 to 90 degrees clockwise



46 to 90 degrees counter-clockwise



91 to 135 degrees clockwise



91 to 135 degrees counter-clockwise

Fig. 23: An example of clockwise and counter-clockwise rotations for the Inter-object direction perception task, we utilized samples from the NOCS REAL dataset

B Additional Analyses of Model Performance

To deepen our understanding of current Multimodal Large Language Models (MLLMs) on DORI, we present a series of additional analyses that dissect performance across key axes of task structure and model behavior.

Fig. 24 examines the relationship between answer set size and model accuracy. We observe a peak in performance at 3-option questions (42.5%), with accuracy declining markedly as the number of candidate answers increases. At 5 and 6 options, performance drops to 26% and 19%, respectively, with a further decline to 13% at 9 options, and a minimum of 6% at 16 options. This degradation illustrates the difficulty MLLMs face when navigating more complex decision spaces, suggesting that increasing output space size strains their orientation reasoning capabilities.

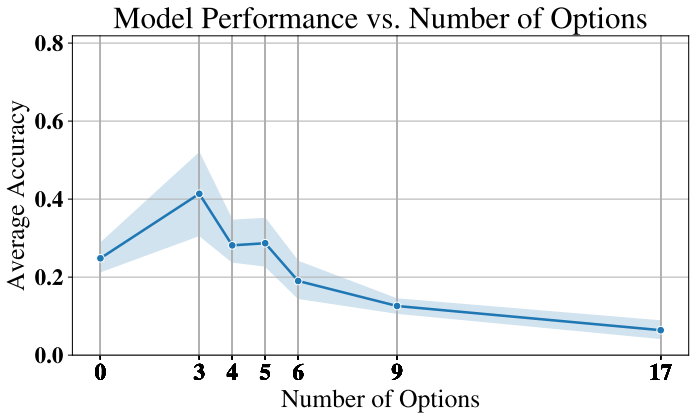
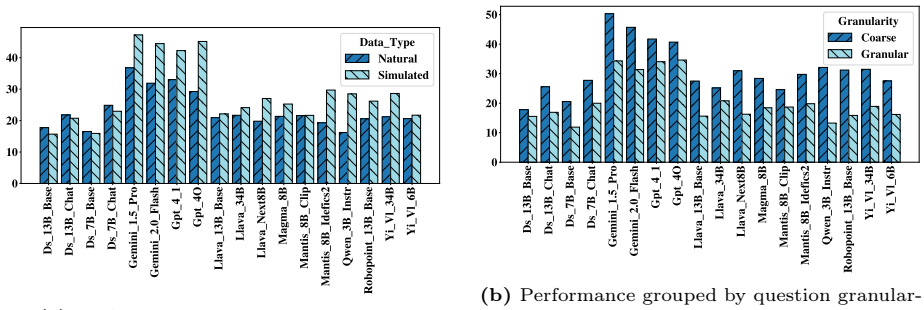


Fig. 24: Model accuracy as a function of answer set size. Accuracy peaks at 42.5% for 3-option questions, then declines steadily, dropping to about 25% at 5 options, 19% at 6, about 13% at 9 options, and reaching a low of about 6% at 16 options, indicating increasing difficulty with larger candidate sets.

In Fig. 25, we break down model performance across simulated vs. natural imagery and annotation granularity levels. Gemini and GPT variants dominate across the board. Yet, all models exhibit significantly higher accuracy on simulated datasets and coarse-type questions, revealing a persistent challenge in transferring orientation understanding to more fine-grained and realistic visual contexts.

B.1 Performance by Common Categories

We can see in Fig. 26 we look at the performance for each of the models based on common categories. We see a similar trend for each model across each of



(a) Performance grouped by data type.

(b) Performance grouped by question granularity.

Fig. 25: Performance of 18 leading Multimodal Large Language Models (MLLMs) across data types **(a)** and annotation granularity levels **(b)**. **Gemini** and **GPT** models lead overall. Models perform noticeably better on simulated datasets and coarse-type questions, revealing a gap in generalizing to natural images and more fine-grained questions.

the different categories with Food being the one that appears to standout. In this instance it seems that only the GPT variants perform better with GPT-4o performing the best. In terms of datasets that contain food in them, this would be attributed to the COCO images for the Directional Facing question and OmniObject-3D images for Compound Rotation and View-scene direction questions. Where some of these questions tend to have "Cannot be determined" since it involves items like Pizza or Cake, GPT-4o appears to be able to determine these sort of questions. This could be attributed to the training data that was used for the GPT family of models since GPT-4-1 also is shown to do well for the Food category. For the person category, however, most models tend to perform best as is expected as most models would have been exposed to persons in their training data, for instance if we look at the most popular category in COCO [63] it is persons.

B.2 Performance by Data-Source

When looking at Fig. 27 we see the main data-sources present in our dataset. We see that for more simple synthetic data models tended to perform better like in JTA, ShapeNet and Get3D; however for natural datasets like Objectron and NOCS_Real, both the open and closed source models tended to have a drop in performance. This could be first attributed to the dense nature of the number of objects found in the NOCS_Real dataset which includes a number of objects like cars and persons, the same can be said for Cityscapes which is also densely populated. For Objectron which can look at an object at different angles this can also prove difficult for most open-source models, with the closed source models tending to performing better which could be attributed to the pretraining data as these closed-source models are exposed to a large set of data with object potentially at different views making it more robust to object rotational views.

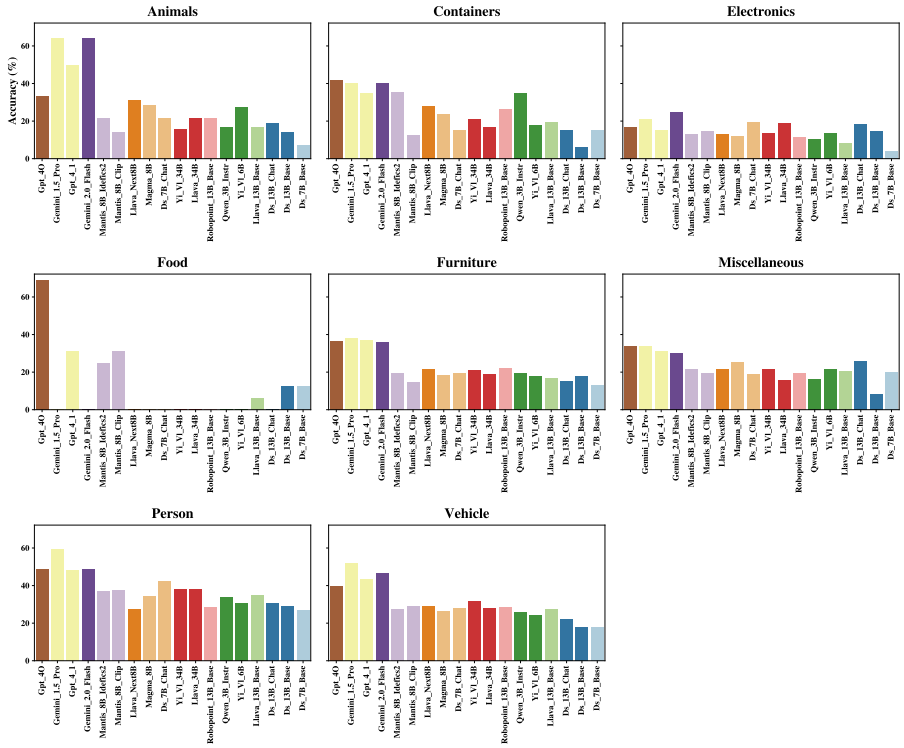


Fig. 26: Model accuracy for the 18 models across the broad categories described in SectionSec. A.3 with most categories tending to performing similarly except for Food which tends to have lower performances except for GPT-4o. Which highlights a gap for orientation related questions related to food.

B.3 Task Complexity Hierarchy

Our multi-dimensional analysis reveals a consistent difficulty pattern across orientation reasoning tasks. This trend is broadly aligned with prior cognitive literature, which distinguishes basic orientation perception from more demanding spatial transformation and reference-frame reasoning processes [74, 97, 99]. We use this literature to motivate DORI's task organization, rather than to claim that model behavior directly mirrors human cognitive development. In our experiments, models perform most competently on frontal alignment tasks, particularly view parallelism: the top open-source models (Qwen3-32B-Inst. and Qwen3-8B-Inst. in Tab. 3 of the main paper) achieve up to 75.8% coarse accuracy, while strong models in Tab. 4 reach 67.5–68.5%. Performance generally drops as tasks require more complex transformations. For compound rotation, open-source models peak at 49.0% coarse accuracy (Qwen3-8B-Inst., Tab. 3), while many models score substantially lower on granular compound-rotation

questions; for example, GPT-5-mini reaches 22.6% on this split (Tab. 4 of the main paper), with several models scoring well below 15

The models' difficulty with compound rotations suggests that current MLLMs do not reliably preserve or manipulate the geometric structure needed to track objects through multi-step transformations. This is further supported by the component-wise error decomposition in Tab. 13, which shows that approximately 60% of predictions fail on *both* rotation components simultaneously. Rather than making small errors on one axis, models often fail to recover the combined transformation, indicating that compound rotation remains a particularly challenging form of object-centric spatial reasoning.

A second large performance gap appears in tasks requiring different forms of relative orientation reasoning, specifically viewer-scene direction perception and inter-object direction perception in Tabs. 3 and 4 of the main paper. Viewer-scene direction perception elicits relatively strong performance across model families, with top open-source models reaching up to 94.3% coarse accuracy (IntVL-14B-Inst., Tab. 3) and strong models in Tab. 4 reaching around 90% or higher. By contrast, inter-object direction tasks are substantially more difficult: the strongest open-source models in Tab. 3 achieve only around 20% coarse accuracy, and many models in Tab. 4 remain far below their viewer-scene direction performance. This disparity suggests that models are better at recognizing orientation relative to a viewer or scene than at reasoning about how two objects are oriented relative to one another. This limitation is important for embodied AI applications such as robotics and navigation, where agents must reason about object relationships from multiple viewpoints.

Canonical orientation understanding, which requires determining whether objects appear in their expected or "natural" orientation, shows highly variable performance across models (Tabs. 3 and 4 of the main paper). Open-source models span a wide range, from near-chance performance (DS-1.3B-Base at 1.8% coarse accuracy, Tab. 3) to stronger performance (Mantis-Idfs-8B at 51.2% coarse accuracy, Tab. 3). Closed-source models also vary substantially, with several achieving high coarse canonical accuracy while still showing large coarse-to-granular gaps. This variability suggests that recognizing canonical object orientations depends on category-level visual cues, world knowledge, and training data coverage that are unevenly captured across architectures and training regimes.

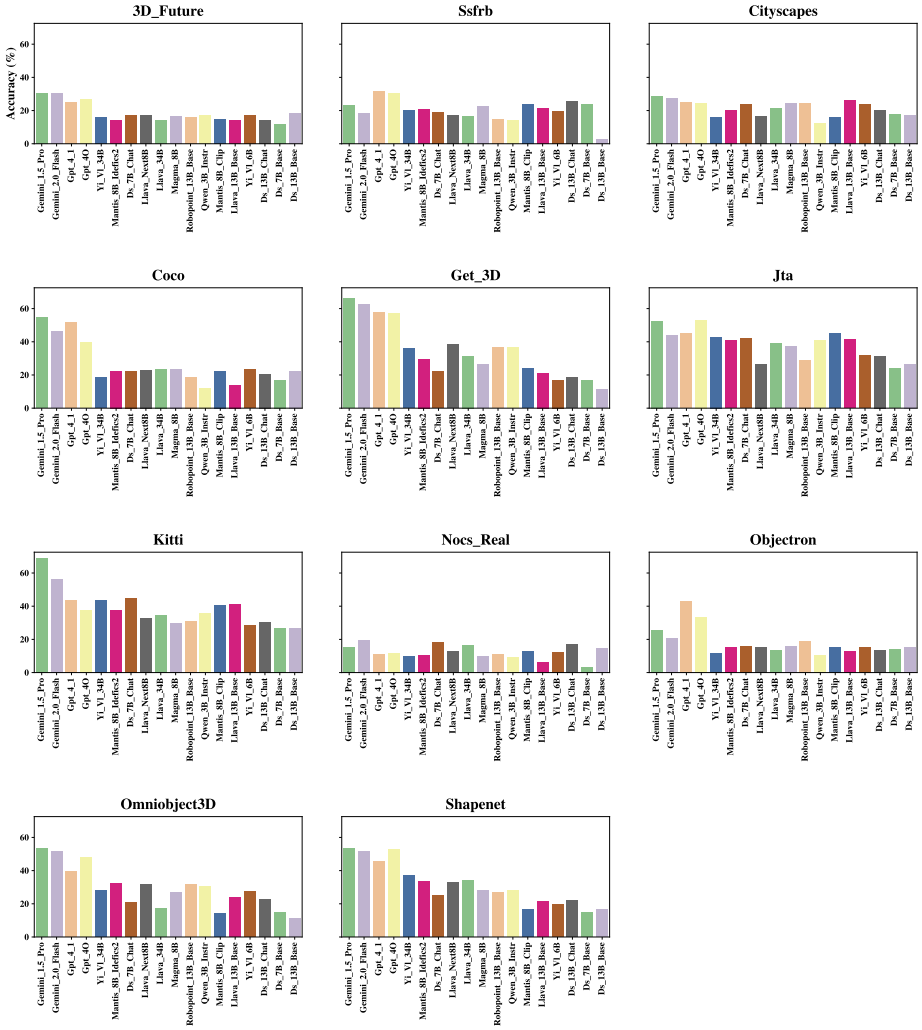
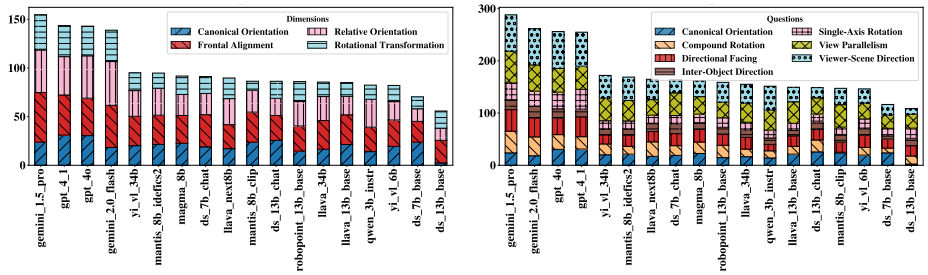


Fig. 27: Model accuracy for the 18 models across the data-sources present in DORI. We see that overall the closed source models tend to perform better than the open source models. With more real and smaller datasets like NOCS_Real and Objectron proving to be challenging for most of the models.



(a) Performance per model across the four core dimensions of orientation.

(b) Performance per model across the different questions present in DORI.

Fig. 28: Performance of 18 leading Multimodal Large Language Models (MLLMs) different levels of our questions hierachy with **(left)** being the being the four core dimensions of orientation and **(right)** being the different questions present in DORI

C Additional Error Analysis

We employ a systematic approach to identify consistent failure patterns rather than random prediction mistakes by analyzing the geometric relationship between predicted and ground truth orientations. Our methodology focuses on identifying repeatable, model-agnostic spatial reasoning deficits that indicate fundamental architectural limitations.

C.1 Systematic Error Pattern Identification

We define systematic failure patterns as specific "Ground Truth → Predicted Answer" confusion pairs that satisfy following criteria: (1) occur in more than 5% of a model's incorrect predictions, and (2) are observed across at least 3 different model architectures. This threshold-based approach filters out infrequent, stochastic mistakes to focus on consistent failure modes that indicate underlying spatial reasoning deficits rather than random prediction errors. We categorize confusion patterns by their geometric and cognitive properties:

- **Perpendicular Confusion**: Systematic misclassification between parallel (0°–15°) and perpendicular (65°–95°) orientations, indicating categorical spatial representation
- **Directional Reversal**: Confusion between opposite directions (*e.g.*, left↔right, toward↔away), suggesting directional processing failures
- **Angle Compression**: Tendency to predict intermediate angles when extreme angles are correct, indicating quantization of continuous spatial information
- **Frontal Bias**: Over-prediction of frontal/canonical orientations regardless of true orientation, likely reflecting training data distribution bias
- **Rotational Symmetry Confusion**: Systematic confusion between rotational equivalents (*e.g.*, 90°↔270°, quarter vs. three-quarter turns)
- **Uncertainty Cascade**: Inappropriate uncertainty responses when spatial ambiguity should trigger systematic reasoning attempts

Below, Tab. 12 presents the complete systematic error analysis across all spatial reasoning dimensions evaluated in DORI. The patterns reveal fundamental limitations in how current MLLMs process and represent spatial information.

Results: Our analysis reveals six systematic error categories affecting all evaluated models. The most severe is perpendicular confusion, where models systematically misclassify parallel (0°–15°) and perpendicular (65°–95°) orientations with 20.5% and 10.3% error rates respectively, indicating coarse categorical rather than continuous angular encoding. Models also demonstrate consistent directional processing failures including left-right confusion (8.3%), front-back reversal (14.7%), and clockwise-counterclockwise errors (5.7%), suggesting fundamental directional encoding limitations. Additional patterns include angle compression where extreme positions are systematically predicted as intermediate

Table 12: Systematic Error Patterns Across All DORI Questions

Question Type	True Answer	Predicted Answer	Error (%)	Pattern Type
View Parallelism	0°–15°	65°–95°	20.5	Perpendicular Confusion
View Parallelism	65°–95°	0°–15°	10.3	Perpendicular Confusion
Directional Facing	30° left	30° right	8.3	Directional Reversal
Directional Facing	180° (away)	0° (facing camera)	14.7	Directional Reversal
Inter-object Direction	46–90° CCW	46–90° CW	5.7	Directional Reversal
View Parallelism	135°–180°	65°–95°	15.8	Angle Compression
Single-axis Rotation	90°	45°	7.6	Angle Compression
Directional Facing	30° left	0° (facing camera)	15.6	Frontal Bias
Directional Facing	30° right	0° (facing camera)	13.7	Frontal Bias
Canonical Orientation	Cannot determine	No change needed	19.5	Frontal Bias
Viewer-scene Rotation	270°	90°	13.1	Rot. Symmetry Confusion
Viewer-scene Rotation	180°	90°	11.6	Rot. Symmetry Confusion
Single-axis Rotation	0°	90°	9.2	Rot. Symmetry Confusion
View Parallelism	0°–15°	Cannot determine	12.3	Uncertainty Cascade
Canonical Orientation	Cannot determine	UNKNOWN	23.5	Uncertainty Cascade
Viewer-scene Rotation	270°	Cannot determine	12.0	Uncertainty Cascade

values (7.6–15.8% error rates), frontal bias reflecting training data distribution effects (13.7–19.5% over-prediction of frontal orientations), rotational symmetry confusions particularly between quarter and three-quarter turns (9.2–13.1% error rates), and uncertainty cascade failures where models inappropriately handle spatial ambiguity (12.0–23.5% error rates). These systematic patterns across diverse architectures indicate fundamental limitations in current MLLM spatial processing mechanisms rather than model-specific deficits.

C.2 Component-wise Error Decomposition

Compound rotation tasks (Q4 in DORI) present the most cognitively demanding spatial reasoning challenge, requiring models to track sequential 3D transformations around multiple axes. To understand the specific failure modes, we decompose compound rotation errors into orthogonal components that isolate different aspects of spatial transformation understanding.We analyze three distinct error categories that provide insight into different failure modes:

1. **Order Swap Errors**: Models correctly identify both rotation components but reverse their sequence (*e.g.*, ground truth "90° Horizontal then 180° Vertical" predicted as "180° Horizontal then 90° Vertical"). This isolates sequence understanding from content understanding.
2. **Component Accuracy**: Percentage of predictions where individual rotation components (horizontal or vertical) are correctly identified regardless of the accuracy of the other component. This measures partial understanding capabilities.
3. **Complete Joint Failure**: Percentage of predictions where both rotation components are incorrectly predicted, indicating total breakdown of 3D spatial reasoning.

Table 13: Component-wise Error Decomposition for Compound Rotations (Q4)

Model	Order Swap (%)	Horizontal Acc. (%)	Vertical Acc. (%)	Both Wrong (%)
Qwen-3B-instr	0.2 ± 0.1	22.7 ± 2.1	21.4 ± 1.9	55.9 ± 3.2
LLaVA-13B-base	3.8 ± 0.7	21.3 ± 1.8	19.5 ± 1.6	59.2 ± 2.9
DeepSeek-7B-chat	4.6 ± 0.8	20.4 ± 1.7	20.5 ± 1.7	59.1 ± 2.8
DeepSeek-1.3B-chat	4.2 ± 0.7	20.4 ± 1.6	19.7 ± 1.5	59.9 ± 2.7
LLaVA-Next-8B	4.9 ± 0.9	20.7 ± 1.8	19.6 ± 1.6	59.6 ± 2.9
Magma-8B	4.6 ± 0.8	20.1 ± 1.6	20.0 ± 1.6	59.9 ± 2.8
DeepSeek-1.3B-base	4.7 ± 0.8	19.9 ± 1.5	20.0 ± 1.6	60.1 ± 2.8
DeepSeek-7B-base	1.2 ± 0.3	19.7 ± 1.5	20.1 ± 1.6	60.3 ± 2.8
Average	3.5 ± 1.7	20.6 ± 0.9	20.1 ± 0.6	59.3 ± 1.4

Tab. 13 presents the complete component-wise error decomposition across all evaluated models, revealing distinct failure patterns and architectural effects.

Results: The component-wise decomposition reveals a fundamental dissociation between sequence processing and spatial reasoning in MLLMs. While order swap errors remain low across all models (mean: 3.5%), indicating competent instruction following, approximately 60% of predictions fail on both rotation components, representing near-complete breakdown of 3D spatial processing. . Models demonstrate axis-agnostic processing (horizontal vs. vertical accuracy differential: 0.7%), unlike human embodied cognition, suggesting identical mechanisms for all rotational transformations rather than specialized processing pathways. Notably, instruction-tuned models significantly outperform larger base models. Qwen-3B-instr achieves superior performance across all metrics despite smaller parameter count, and DeepSeek-1.3B-chat outperforms DeepSeek-7B-base with 5.4x fewer parameters. These findings indicate three critical limitations: (1) models can manipulate rotation symbols but lack geometric transformation mechanisms, (2) the consistent 60% joint failure rate across diverse architectures suggests fundamental rather than model-specific deficits, and (3) training methodology appears more crucial than parameter scaling for spatial reasoning capabilities, suggesting targeted training approaches may be more effective than architectural scaling.

C.3 Soft Accuracy Calculation

Soft accuracy metrics were introduced to provide a more nuanced evaluation of models' orientation understanding capabilities and help distinguish between models that are completely wrong versus those that have an approximate understanding of orientation concepts. Unlike standard binary accuracy that only awards points for exact matches, soft accuracy awards half points (0.5) for answers that are partially correct or adjacent to the ground truth. To calculate

such accuracies, we implement carefully designed spatial tolerance thresholds and logical equivalences. Soft accuracy is only calculated for the fine-grained questions that typically demand a precise metric response.

- For **View Parallelism**, Predictions within $\pm 45°$ of the ground truth angle receive partial credit. This threshold captures predictions in adjacent sectors while excluding opposed orientations.
- For **Directional Facing**, half points are awarded exclusively for mirror-image confusions between "30 degrees left" and "30 degrees right". We intentionally do not extend partial credit to other angular errors, preserving the specificity of directional understanding assessment.
- For **Single-axis Rotatoin**, a $45°$ tolerance window applies to predicted rotations $(0°, 45°, 90°, 135°, 180°)$. This allows credit for adjacent discrete positions while maintaining distinction between major orientation categories. For instance, predicting $45°$ when the correct answer is $90°$ would not qualify, but a $135°$ prediction for $180°$ ground truth would receive partial credit.
- In **Compound Rotation**, partial credit is awarded if either the horizontal or vertical rotation component is correct in multi-axis transformations. In this "X then Y" rotation sequence response, we parse both components separately. For example, prediction of "$90°$ horizontal then $0°$ vertical" would receive 0.5 points for either correct component when compared to the ground truth "$90°$ horizontal then $180°$ vertical"
- For **Inter-object Direction**, half points are given for adjacent magnitude ranges, but only if the direction (clockwise vs. counterclockwise) matches. For instance, if the ground truth was "0 to 45 degrees clockwise" and the prediction was "46 to 90 degrees clockwise", we award 0.5 points. However, if the prediction was "46 to 90 degrees counterclockwise," it would earn 0 points despite similar magnitude
- For **Viewer-scene direction**, the soft accuracy specifically addresses confusion between opposite rotational directions. For example, if an object has rotated $90°$ clockwise between images, but the model reports $270°$ clockwise (which is equivalent to $90°$ counterclockwise), it receives 0.5 points. No partial credit is given for other angle confusions.
- For **Canonical Orientation**, the soft accuracy addresses confusion in the order of operations. For example, if an image requires rotation followed by flipping to restore its canonical orientation, but the model suggests flipping followed by rotation, it receives 0.5 points.

Results. Fig. 29,Fig. 30,Fig. 31,Fig. 32,Fig. 33,Fig. 34,Fig. 35, compares standard (hard) vs. soft accuracy on DORI questions.

Fig. 29 and Fig. 30 reveal that for View Parallelism and Directional Facing tasks, soft accuracy provides no benefit over standard accuracy, with identical performance metrics across all models. This indicates that when models err on these fundamental orientation tasks, they tend to make categorical mistakes rather than near-miss approximations. The lack of improvement suggests that errors in these tasks stem from fundamental misunderstandings rather than subtle misjudgments.

In contrast, the **Single-axis Rotation** task (Fig. 31) shows the most substantial gains under soft accuracy metrics, with improvements ranging from 8.6% to 17.7% across models. DeepSeek-7B-Chat achieves the most dramatic improvement, a remarkable 17.7% gain. LLaVA-Next-8B improves of +11.5%, while Qwen-3B-Instruct shows improvement from + 16.4%. These substantial gains suggest that while models often fail to identify the exact rotational angle, they frequently select adjacent angular categories, demonstrating partial understanding of rotational relationships.

For Compound Rotation (Fig. 32), all models except Qwe2.5-3B-Instruct (0.5% improvement) show notable improvements under soft accuracy metrics. Both LLaVA-7B-Chat and LLava-Next-8B improves from by +14.6%, indicating that models often correctly identify one of the two rotation components (horizontal or vertical) while missing the other. This partial success highlights both the inherent complexity of multi-axis rotations and the models' fragmentary grasp of compound transformations.

The Inter-object Direction task (Fig. 33) shows similar amount of soft accuracy gains across all models, with improvements ranging from 7.7% to 17.1%. DeepSeek-7B-Chat improves from 12.7% to 29.8%, more than doubling its effective performance. This suggests that models often select directionally appropriate answers that fall into adjacent angular ranges, indicating a coarse understanding of relative orientations despite lacking precise angular discrimination.

For Viewer-scene Direction (Fig. 34), soft accuracy provides moderate improvements, with DeepSeek-7B-Chat showing the largest gain. The comparatively smaller improvements here suggest that models correctly identify rotational changes, with fewer "near miss" responses than in other tasks.

Canonical Orientation (Fig. 35) shows minimal improvement under soft accuracy metrics. The small gains observed suggest that models are rarely confused in the order of operations. When they fail on canonical orientation tasks, they typically misidentify the necessary operations entirely rather than simply reversing their order. This indicates a more fundamental gap in understanding canonical object positioning rather than mere sequencing errors.

Fig. 36 summarizes the relative gains of soft accuracy across all tasks and models. Notably, this more lenient metric only helps some questions (Fig. 31 to Fig. 35). The heatmap reveals that Single-axis Rotation, Compound Rotation, and Inter-object Direction tasks benefit most from soft accuracy metrics, with improvements frequently exceeding 10%. This pattern suggests that rotational and relational orientation understanding in current MLLMs exists on a spectrum rather than in binary states of correctness. In addition, the highest gain reported was for Single-axis Rotation, but is significantly below human performance using standard accuracy (18% vs. avg of about 30%). This shows that even when given an advantage, these models still fall significantly below human ability.

C.4 Performance Error Bars

The error bars in Fig. 37,Fig. 38,Fig. 39,Fig. 40,Fig. 41,Fig. 42, and Fig. 43 report the mean and standard deviation of various models on DORI ques-

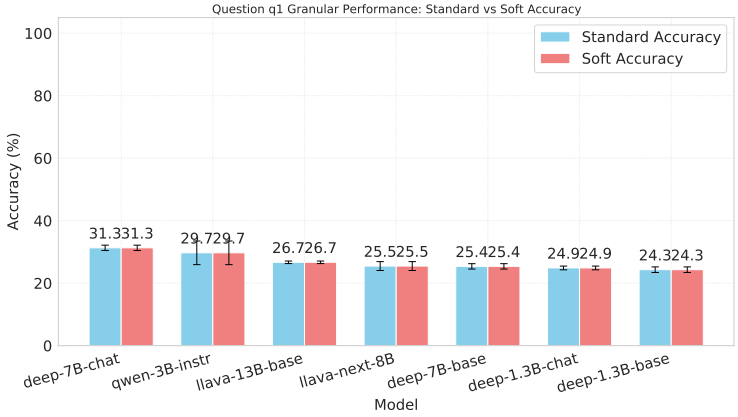


Fig. 29: Comparing the mean and standard deviation of soft vs. standard (hard) accuracy on View Parallelism task. See discussion Sec.Sec. C.3.

tions. For each model and question type combination, the formula $error = std_accuracy/sqrt(seed_count)$ was applied. We used 3 different seeds: $[42, 1998, 107983]$.

Looking at the **View Parallelism** task (Fig. 37), we observe relatively narrow error bars for most models, indicating consistent performance across different seeds. However, both DeepSeek base models and the LLaVa-13 B-base models display wider error bars on coarse questions (approximately $\pm 3-4\%$), suggesting that their performance comes with greater variability. Despite LLaVa-13B-base demonstrating better performance than the other 2 base models, it's clear that all three base models' performance is more sensitive to initialization conditions. Qwen2.5-3B-instruct model demonstrates notable initialization variability for both coarse and granular questions.

The **Directional Facing** task (Fig. 38) reveals generally smaller error bars across all models, with most variations under $\pm 2\%$, indicating that performance on cardinal direction assessment remains relatively stable regardless of initialization. However, all models perform notably worse on this task compared to View Parallelism, with even the best model (DeepSeek-7B-Chat) achieving only 32.5% accuracy on coarse questions. LLaVA-Next-8B shows consistent but lower performance, while Qwen2.5-3B-Instruct has low performance with tight error bars.

The **Single-axis Rotation** task (Fig. 39) is more uniform across models, with most showing variations of $\pm 2-3\%$. The comparable error bar sizes across models suggest that this task presents similar levels of difficulty for all architectures, with no model demonstrating significantly more stable performance than others. This uniformity in variability indicates that improvements in this task may require fundamental architectural innovations rather than just parameter tuning.

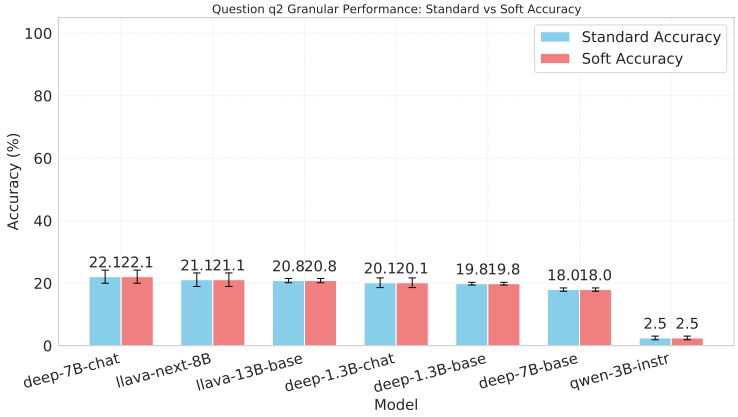


Fig. 30: Comparing the mean and standard deviation of soft vs. standard (hard) accuracy on Directional Facing task. See discussion Sec.Sec. C.3.

The **Compound Rotation** task (Fig. 40) shows relatively narrow error bars across most models, indicating that performance on this task is stable across initialization seeds rather than highly variable. However, the coarse and granular results reveal a large coarse-to-granular gap: several models can recognize whether an image is broadly canonical, but still struggle to infer the precise transformations needed to restore canonical orientation. This suggests that models may learn category-level canonicality cues while lacking robust representations of the geometric operations underlying non-canonical object poses.

For **Inter-object Direction** (Fig. 41), we observe slightly asymmetric performance patterns between coarse and granular questions. While DeepSeek-7B-Chat achieves the highest coarse accuracy, all models show substantially lower performance on granular questions (generally below 15%). The error bars appear to be tight for coarse and granular questions, indicating consistent performance across different trials. The performance gap between coarse and granular questions suggests that while models can sometimes succeed at basic relational orientation tasks (determining if objects face the same/opposite directions), they systematically fail when asked to make precise angular judgments about inter-object relationships. Interestingly, for LLaVA-13B-base model, the granular performance slightly exceeds its coarse performance, running counter to the typical pattern observed in other tasks and models. This anomaly may indicate that LLaVA's training regime potentially encodes some specific features that assist with fine-grained angular estimations between objects, though its overall performance remains well below human capabilities on these tasks.

Similarly, the **Viewer-Scene Direction** task (Fig. 42) reveals intriguing performance inversions between coarse and granular questions for certain models. Qwen-3B-Instruct shows the highest coarse accuracy with poor granular performance, while DeepSeek-7B-Chat demonstrates the opposite pattern. These inversions, coupled with wide error bars, indicate that different model archi-

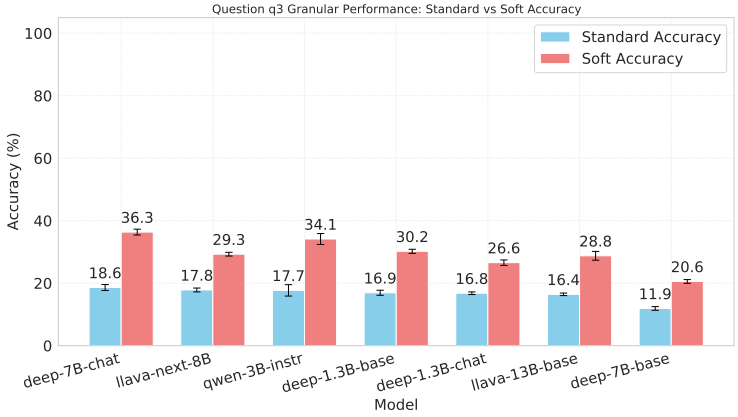


Fig. 31: Comparing the mean and standard deviation of soft vs. standard (hard) accuracy on Single-axis Rotation task. See discussion Sec.Sec. C.3.

tectures encode rotation perception in fundamentally different ways. This task exposes fundamental inconsistencies in how current MLLMs process orientation changes, suggesting that rotation tracking may rely on different computational mechanisms than static orientation perception, with these mechanisms developing unevenly across model architectures and training regimes.

The **Canonical Orientation** task (Fig. 43) doesn't exhibit any high variability among the error bars across coarse and granular performance. The error bars remain relatively narrow for most models ($\pm 1 - 3\%$), indicating that performance limitations on this task are consistent across initialization seeds rather than highly variable. This consistency, coupled with generally poor performance, suggests that canonical orientation understanding, which requires both world knowledge about natural object positions and spatial transformation reasoning, represents a fundamental capability gap in current MLLMs. The results indicate that models lack robust internal representations of how objects "should" appear in the world, a crucial component for embodied navigation and manipulation tasks where recognizing and correcting non-canonical orientations is essential.

Cross-Task Insights Across all tasks, our error bar analysis reveals several critical insights about MLLMs' orientation reasoning capabilities. First, model performance stability varies substantially across tasks, with simpler perception tasks (View Parallelism, Directional Facing) showing more consistent performance across initializations compared to complex reasoning tasks (Compound Rotation, Viewer-scene direction, Canonical Orientation). Moreover, the consistently tight error bars on poor-performing granular questions, particularly for rotational tasks, indicate systematic limitations rather than chance variability, suggesting architectural rather than parametric constraints. We also observe that Chat-tuned models generally demonstrate more stable performance than their base counterparts, suggesting that instruction tuning not only improves accu-

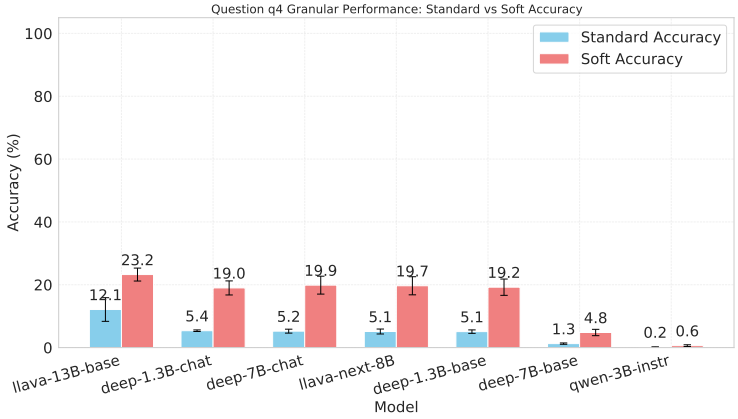


Fig. 32: Comparing the mean and standard deviation of soft vs. standard (hard) accuracy on Compound Rotation task. See discussion Sec.Sec. C.3.

racy but also reduces initialization sensitivity. Finally, the performance inversions observed between coarse and granular questions for some models highlight the disconnect between categorical and precise metric orientation understanding, a fundamental challenge that persists across all model families and architectures evaluated.

C.5 Out-Of-Pretraining Analysis

Most MLLMs have closed pretraining datasets, *i.e.*, it is not possible to be completely sure what they were trained on. However, to provide insight into what the MLLM model might have learned (*i.e.*, if it saw similar images in the pretraining dataset) we use the approach described in [92] to filter images in COCO and Cityscapes into those that were likely seen during pretraining. Specifically, we measured the cosine similarity between text features representing the object our question referred to (*e.g.*, "a photo of a person" for a question asking about the orientation of a person) and the image using LLaVa-13B. Those with high similarity (using a 0.19 threshold for COCO and 0.18 for Cityscapes) were removed, leaving only images that were not learned well by the model during pretraining. On COCO we evaluated performance on person images, which accounted for 547 images in the Directional Facing questions, of which 253 were removing (leaving 294 images). On Cityscapes Directional Facing, we filtered based on car questions, resulting in removing 122 of 342 car images, leaving 220 images. Cars and person categories were selected as they were common objects in their respective datasets. We refer to these splits as ALL, which includes questions that have every car or person image, and Out-Of-Pretraining (OOP) for the images that remained after our filtering process. We then compare the performance on a number of models armed with these splits, as seen in Tab. 14.

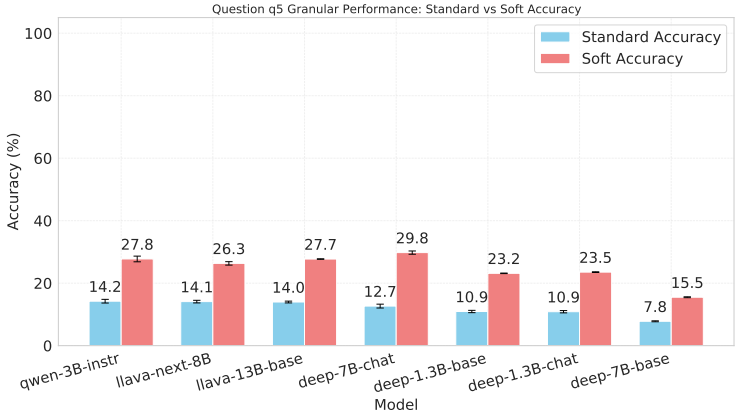


Fig. 33: Comparing the mean and standard deviation of soft vs. standard (hard) accuracy on Inter-object Direction task. See discussion Sec.Sec. C.3.

Armed with these splits, we compared performance on LLava-13B, LLava-34B, Yi-VL-6B and Deepseek-7B, shown below. Interestingly, the OOP results are generally better than ALL, which seems counterintuitive at first glance. However, we suspect that this is an effect of being pretrained with a different objective than one focused on orientation. For example, if a model was pretrained to align images to alt-text, which give a high level description of the image that often does not include orientation information, then when the model saw similar images, it naturally assumes that the goal is to match to a high-level description and extracts features accordingly, essentially overfitting to that task. That said, there are many confounding factors that could provide alternative explanations, and is an interesting avenue for exploration in future work.

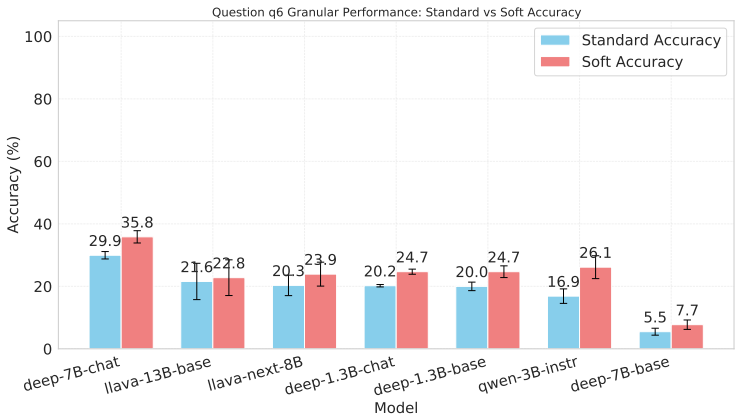


Fig. 34: Comparing the mean and standard deviation of soft vs. standard (hard) accuracy on Viewer-scene Direction task. See discussion Sec.Sec. C.3.

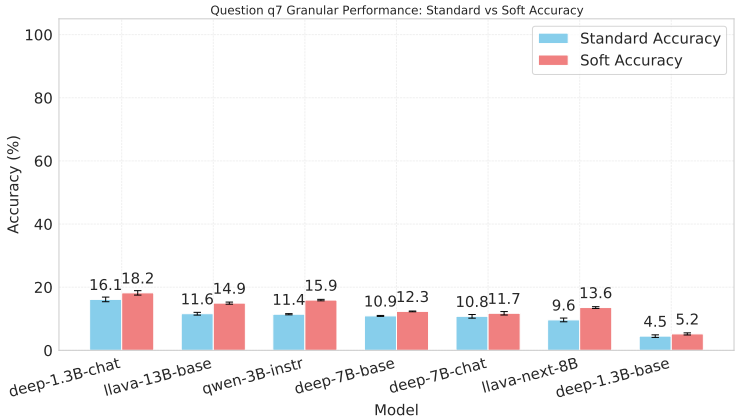


Fig. 35: Comparing the mean and standard deviation of soft vs. standard (hard) accuracy on Canonical Orientation task. See discussion Sec.Sec. C.3.

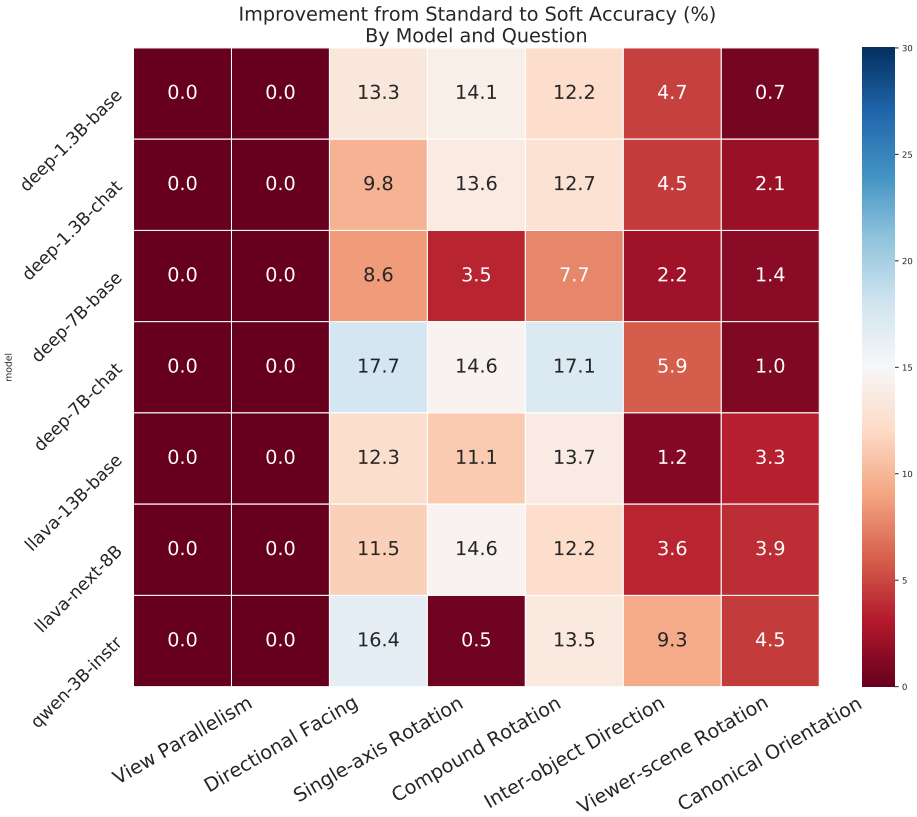


Fig. 36: Relative gain from using Soft accuracy rather than Standard (hard) accuracy per question. SeeSec. C.3 for discussion

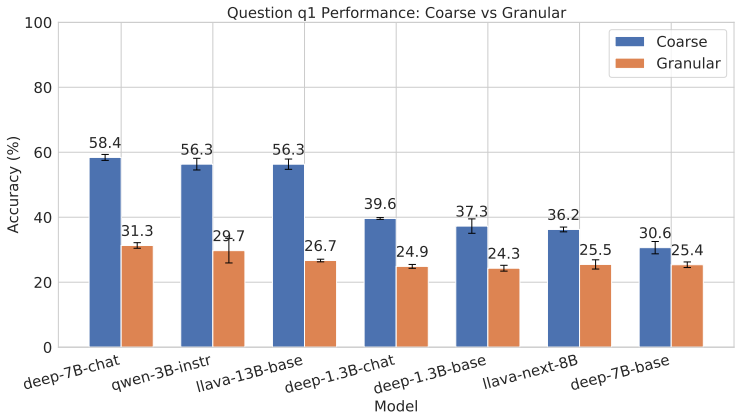


Fig. 37: Mean and Standard Deviation of various models on View Parallelism task

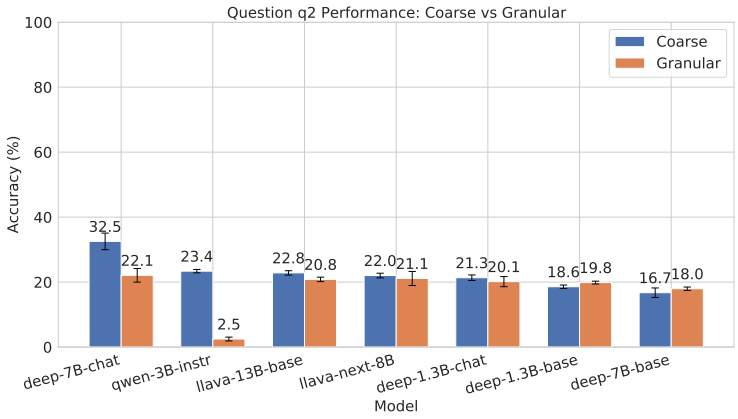


Fig. 38: Mean and Standard Deviation of various models on Directional Facing task

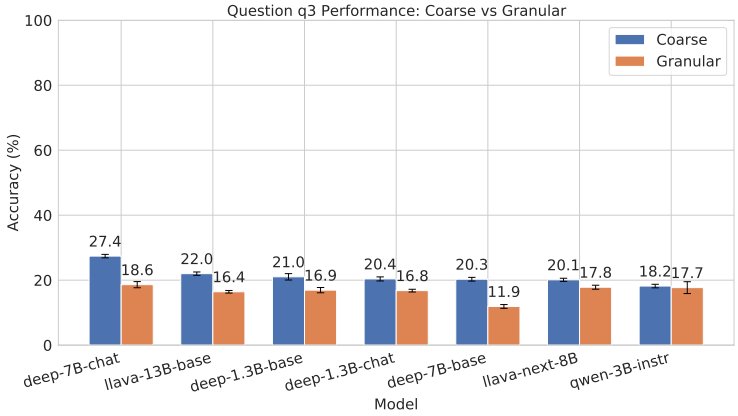


Fig. 39: Mean and Standard Deviation of various models on Single-axis Rotation task

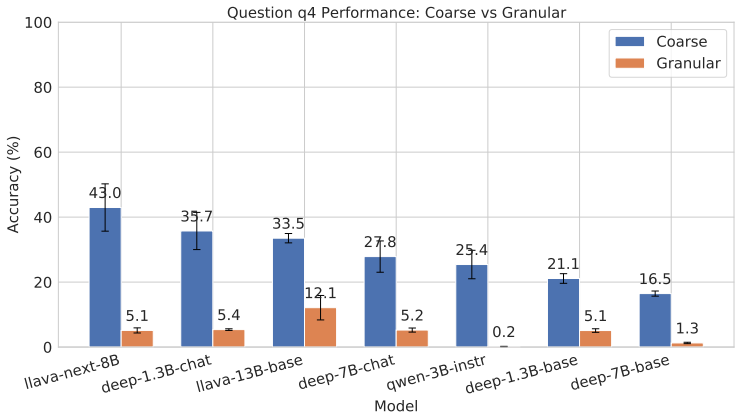


Fig. 40: Mean and Standard Deviation of various models on Compound Rotation task

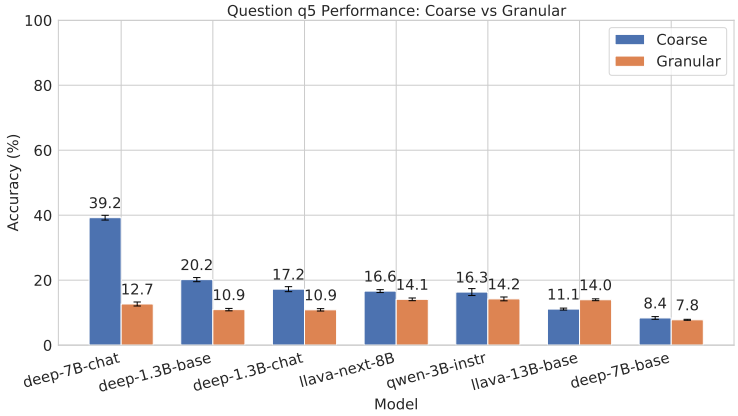


Fig. 41: Mean and Standard Deviation of various models on Inter-object direction.

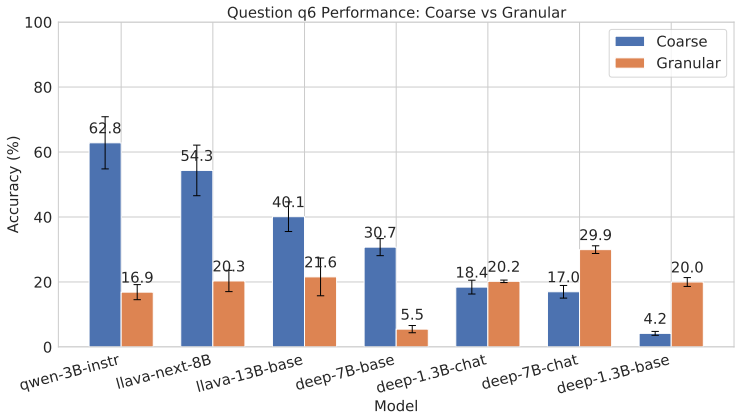


Fig. 42: Mean and Standard Deviation of various models on Viewer-Scene direction task.

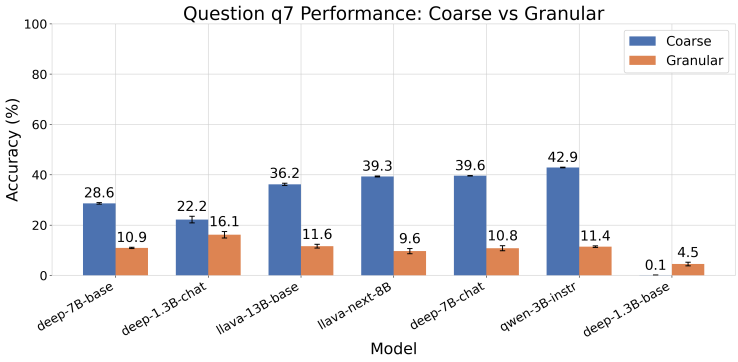


Fig. 43: Mean and Standard Deviation of various models on the Canonical Orientation task.

Table 14: Out-Of-Pretraining (OOP) performance when evaluated on the COCO and Cityscapes dataset.

COCO	All	OOP	All	OOP
Category (Person)	Coarse	Coarse	Granular	Granular
LLava-v1.6-13B [69]	23.0	25.1	20.8	21.0
LLava-v1.6-34B [69]	42.6	41.3	39.6	39.6
Yi-VL-6B [2]	30.3	28.6	33.5	33.1
Deepseek-Base-7B [7]	17.7	17.5	17.7	17.9
Cityscapes	All	OOP	All	OOP
Category (Car)	Coarse	Coarse	Granular	Granular
LLava-v1.6-13B [69]	22.6	25.7	16.7	18.1
LLava-v1.6-34B [69]	36.5	41.1	21.6	21.7
Yi-VL-6B [2]	25.6	29.1	21.6	22.1
Deepseek-Base-7B [7]	17.5	18.5	16.9	17.1

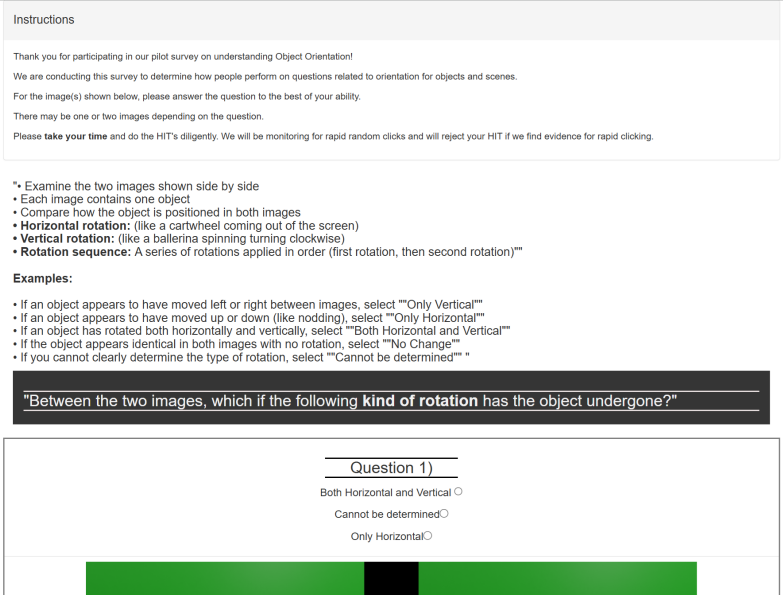


Fig. 44: An example of the high level instructions shown for the Human Evaluation

D Human Evaluation Interface

An example of the high level instructions shown for this task can be seen in Fig. 44, examples of questions are shown in Fig. 45 and Fig. 46

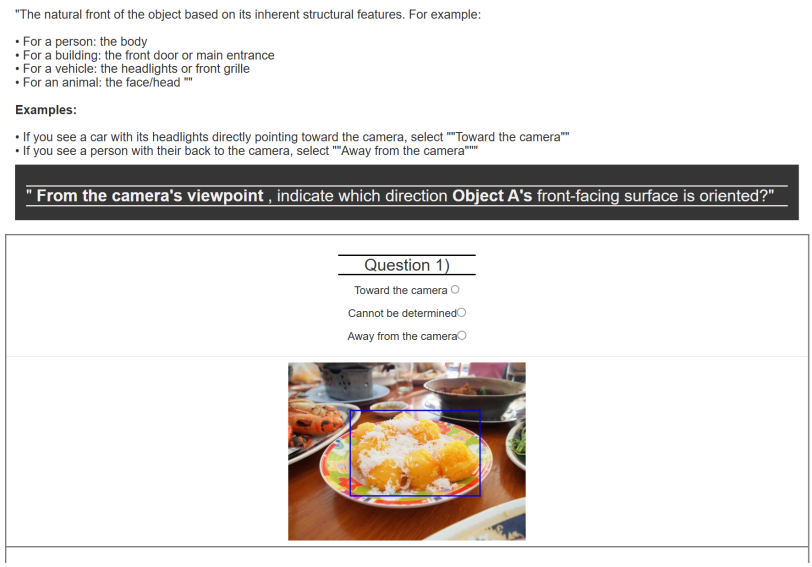


Fig. 45: An example of a sample for the coarse-level VQA for the Directional Facing task in DORI shown for the Human Evaluation

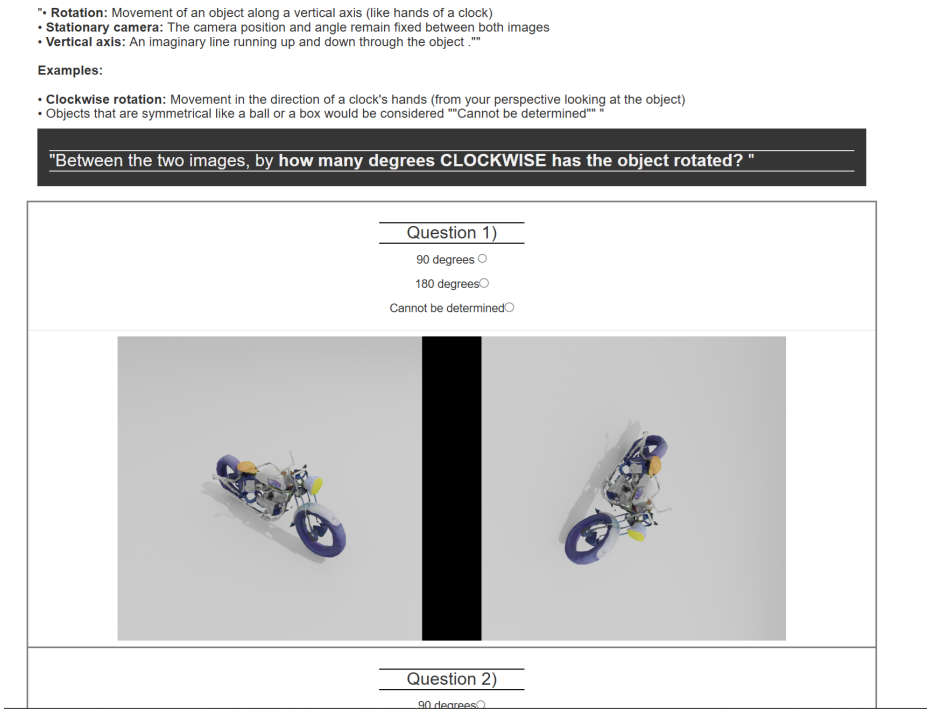


Fig. 46: An example of a sample for the coarse-level VQA for the Viewer-Scenen Direction task in DORI shown for the Human Evaluation

E Terminology Glossary

Front-facing surface / Frontal surface. The front-facing surface (or frontal surface) of an object is the side that carries its primary functional or semantic features, such as a person's face and torso, a car's headlights and grille, or a laptop's screen, and this surface is what is considered to be facing the camera when those features are visible and oriented toward the viewer (*e.g.*, a person standing and looking directly at the camera is presented with their frontal surface clearly visible).

Canonical orientation. Canonical orientation is the typical, upright, real-world pose in which an object is most commonly encountered or depicted, often aligned with gravity and human conventions (for instance, a car resting on its wheels on a horizontal road, a building with the roof at the top and the foundation at the bottom, or a mug standing upright with the handle to the side), and questions about canonical orientation ask whether the object is in this usual pose or has been rotated or flipped away from it (*e.g.*, a building image rotated 180° so the roof is at the bottom would be judged non-canonical).

Camera plane / Camera axis. The camera plane (or image plane) is the 2D plane onto which the 3D scene is projected by the camera, while the camera axis (optical axis) is the line perpendicular to this plane passing through the camera center, so that orientation relative to the camera is described by how an object's surfaces are angled with respect to this plane and axis (for example, a frontal surface parallel to the image plane and centered on the optical axis corresponds to an object facing straight toward the camera).

Camera Viewpoint vs. Object Viewpoint. The camera viewpoint is the position and orientation of the camera in the 3D scene, defining an egocentric frame where directions like "toward the camera" or "left of the camera" live, whereas the object viewpoint is the orientation and placement of the object relative to its own intrinsic front/back/left/right axes or relative to other objects, so a question might either ask how the object appears from the camera's perspective ("Is the car facing toward the camera?") or, alternatively, how the object is oriented if one imagines standing at the object and looking out ("From the car's viewpoint, is it facing toward or away from the truck?").

Rotation transformation. A rotation transformation changes an object's orientation by turning it around one or more axes while preserving its shape and size, such as rotating a mug 90° around the vertical axis to bring its handle from the right side to face the camera or rotating a picture frame 180° in the image plane so that it appears upside down relative to its original orientation.

Sequential transformations. Sequential transformations refer to applying multiple geometric operations in a specific order, typically rotations and flips, where order matters because compositions can yield different final poses, for example first rotating a building image 90° clockwise in the image plane and then flipping it horizontally produces a different orientation than performing the horizontal flip first and then rotating, even though the same primitive operations are used.

Vertical axis rotation. Vertical axis rotation is a 3D rotation around an axis that runs approximately up–down through the object and aligns with gravity (analogous to yaw), changing which horizontal direction the object's front faces, so turning a car so that it goes from facing east to facing north is a vertical axis rotation of about 90° and a person spinning in place to look from the camera to the right-hand side is another example of such a rotation.

Horizontal axis rotation. Horizontal axis rotation is a 3D rotation around an axis that runs roughly left–right through the object (analogous to pitch), changing how much the object is tilted toward or away from the camera, so when a person nods their head "yes" or a box tips forward so its top leans toward the camera, these are horizontal axis rotations that alter how the object's surfaces are foreshortened in the image.

Lateral Axis Rotation. Lateral axis rotation is a rotation around an axis that runs roughly front–back through the object (analogous to roll), causing the object to lean sideways relative to gravity and the image frame, so tilting your head so that your ear moves toward your shoulder or rolling a car so that it lies partially on its side are examples of lateral axis rotations that change which parts appear above or below in the image without changing which direction the object's front is pointing.

Clockwise rotation. Clockwise rotation is defined with respect to a chosen viewing direction and axis: in this context, for rotations around the vertical axis, it means that when viewed from above along the vertical axis, the object's front turns in the same direction as the hands of a clock (*e.g.*, a car that initially faces north and then rotates to face east has undergone a 90° clockwise rotation about the vertical axis), and in image-plane contexts, clockwise means the top of the object appears to rotate toward the right side of the image.

Counter-clockwise rotation. Counter-clockwise rotation is the opposite directional sense from clockwise around a specified axis: for vertical axis rotations observed from above, the object's front turns in the opposite direction to clock hands (*e.g.*, a car rotating from facing east to facing north has rotated 90° counter-clockwise about the vertical axis), and in the image plane, counter-clockwise means the top of the object appears to rotate toward the left side of the image. We include examples of what clockwise and counter-clockwise rotations looks like as seen in Fig. 23 for the Inter-object direction perception task, we utilize samples from the NOCS REAL dataset to highlight these different question samples.

Facing each other (objects' perspective). Two objects are said to be facing each other when each object's front-facing surface points approximately toward the other's position in the scene in an allocentric or object-centric frame, so that if you imagine rays extending from their fronts, those rays intersect between them (for example, two chairs arranged on opposite sides of a table with their seats and backs oriented toward the table center are facing each other).

Facing same/opposite directions. Objects face the same direction when their front-facing axes are approximately parallel in 3D space, meaning they are ori-

ented with similar yaw so they would move side-by-side if they advanced forward together, whereas they face opposite directions when their front-facing axes differ by about 180°, such as two cars parked in the same lane but one pointing north and the other pointing south along the road.

Perpendicular facing. Two objects are perpendicular in facing when the angle between their front-facing directions is approximately 90° in yaw, such that one object's front points roughly to the side of the other, for example, when one car is parked facing north and another is parked facing east at an intersection, so their fronts form an L-shape orientation relative to each other.

Flip / Mirror. A flip or mirror transformation reverses an image or object along a specified axis without changing its depth order, so a horizontal flip mirrors the scene left–right (like viewing it in a mirror placed vertically), turning a car that originally faces left into one that appears to face right, while a vertical flip mirrors the scene top–bottom, making a building appear upside down with the roof at the bottom of the image.

Ambiguous objects / Symmetrical objects. Ambiguous or symmetrical objects are those whose shape or appearance does not specify a unique front or canonical orientation because many rotations produce indistinguishable images, such as a perfect sphere, a uniform cylinder, or a highly symmetric abstract logo, so for such objects questions about "which way it is facing" or "whether it is upside down" may have no well-defined answer and are therefore treated as "Cannot be determined" cases in the dataset.

References

1. Ahmadyan, A., Zhang, L., Ablavatski, A., Wei, J., Grundmann, M.: Objectron: A large scale dataset of object-centric videos in the wild with pose annotations. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7822–7831 (2021)
2. AI, ., :, Young, A., Chen, B., Li, C., Huang, C., Zhang, G., Zhang, G., Li, H., Zhu, J., Chen, J., Chang, J., Yu, K., Liu, P., Liu, Q., Yue, S., Yang, S., Yang, S., Yu, T., Xie, W., Huang, W., Hu, X., Ren, X., Niu, X., Nie, P., Xu, Y., Liu, Y., Wang, Y., Cai, Y., Gu, Z., Liu, Z., Dai, Z.: Yi: Open foundation models by 01.ai (2024)
3. Alomari, M., Li, F., Hogg, D., Cohn, A.G.: Online perceptual learning and natural language acquisition for autonomous robots. Artificial Intelligence **303**, 103637 (2022). https://doi.org/10.1016/j.artint.2021.103637
4. 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)
5. Baker, T.J., Norcia, A.M., Candy, T.R.: Orientation tuning in the visual cortex of 3-month-old human infants. Vision Research **51**, 470–478 (2011). https://doi.org/10.1016/j.visres.2011.01.003
6. Ballaz, C., Boutsen, L., Peyrin, C., Humphreys, G.W., Marendaz, C.: Visual search for object orientation can be modulated by canonical orientation. Journal of Experimental Psychology: Human Perception and Performance **31**(1), 20 (2005)
7. Bi, X., Chen, D., Chen, G., Chen, S., Dai, D., Deng, C., Ding, H., Dong, K., Du, Q., Fu, Z., et al.: Deepseek llm: Scaling open-source language models with longtermism. arXiv preprint arXiv:2401.02954 (2024)
8. Blades, M., Spencer, C.: The development of children's ability to use spatial representations. Advances in child development and behavior **25**, 157–199 (1994)
9. Braddick, O., Birtles, D., Wattam-Bell, J., Atkinson, J.: Motion- and orientation-specific cortical responses in infancy. Vision Research **45**(25), 3169–3179 (2005). https://doi.org/https://doi.org/10.1016/j.visres.2005.07.021, https://www.sciencedirect.com/science/article/pii/S0042698905003846
10. Bramley, N.R., Gerstenberg, T., Tenenbaum, J.B., Gureckis, T.M.: Intuitive experimentation in the physical world. Cognitive psychology **105**, 9–38 (2018)
11. Castle, C.: Using design thinking to create human-centered assessments. In: NERA Conference Proceedings 2024 (2024)
12. Chavez, R.S., Guthrie, T.D., Kapustka, J.M.: Independent allocentric and egocentric reference frame encoding of person knowledge within the brains of social groups. bioRxiv pp. 2023–09 (2023)
13. Chen, B., Xu, Z., Kirmani, S., Ichter, B., Sadigh, D., Guibas, L., Xia, F.: Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 14455–14465 (June 2024)
14. Chen, Y.L., Cai, Y.R., Cheng, M.Y.: Vision-based robotic object grasping—a deep reinforcement learning approach. Machines **11**(2), 275 (2023)
15. 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. In: NeurIPS (2024)
16. 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 (2025)

17. Cohn, A.G.: Qualitative spatial representation and reasoning techniques. KI-97: Advances in Artificial Intelligence pp. 1–30 (1997). https://doi.org/10.1007/3540634932_1
18. Cong, Y., Chen, R., Ma, B., Liu, H., Hou, D., Yang, C.: A comprehensive study of 3-d vision-based robot manipulation. IEEE Transactions on Cybernetics **53**(3), 1682–1698 (2021)
19. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3213–3223 (2016)
20. Corsetti, J., Boscaini, D., Oh, C., Cavallaro, A., Poiesi, F.: Open-vocabulary object 6d pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 18071–18080 (June 2024)
21. Cortes, R.A., Colaizzi, G.A., Dyke, E.L., Peterson, E.G., Walker, D.L., Kolvoord, R.A., Uttal, D.H., Green, A.E.: Individual differences in parietal and premotor activity during spatial cognition predict figural creativity. Creativity Research Journal **35**, 23–32 (2022). https://doi.org/10.1080/10400419.2022.2049532
22. Delhaye, B.P., Long, K.H., Bensmaia, S.J.: Neural basis of touch and proprioception in primate cortex. Comprehensive Physiology **8**(4), 1575 (2018)
23. Deng, Q., Deb, O., Patel, A., Rupprecht, C., Torr, P., Trigoni, N., Markham, A.: Towards multi-modal animal pose estimation: An in-depth analysis. CoRR **abs/2410.09312** (2024)
24. Deng, X., Xiang, Y., Mousavian, A., Eppner, C., Bretl, T., Fox, D.: Self-supervised 6d object pose estimation for robot manipulation. In: 2020 IEEE International Conference on Robotics and Automation (ICRA). pp. 3665–3671. IEEE (2020)
25. Doganci, N., Iannotti, G.R., Coll, S.Y., Ptak, R.: How embodied is cognition? fmri and behavioral evidence for common neural resources underlying motor planning and mental rotation of bodily stimuli. Cerebral Cortex **33**(22), 11146–11156 (2023)
26. Du, G., Wang, K., Lian, S., Zhao, K.: Vision-based robotic grasping from object localization, object pose estimation to grasp estimation for parallel grippers: a review. Artif. Intell. Rev. **54**(3), 1677–1734 (Mar 2021)
27. Du, M., Wu, B., Li, Z., Huang, X., Wei, Z.: EmbSpatial-bench: Benchmarking spatial understanding for embodied tasks with large vision-language models. In: Ku, L.W., Martins, A., Srikumar, V. (eds.) Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). pp. 346–355. Association for Computational Linguistics, Bangkok, Thailand (Aug 2024). https://doi.org/10.18653/v1/2024.acl-short.33, https://aclanthology.org/2024.acl-short.33/
28. Fabbri, M., Lanzi, F., Calderara, S., Palazzi, A., Vezzani, R., Cucchiara, R.: Learning to detect and track visible and occluded body joints in a virtual world. In: Proceedings of the European conference on computer vision (ECCV). pp. 430–446 (2018)
29. Feng, Y., Lin, J., Dwivedi, S.K., Sun, Y., Patel, P., Black, M.J.: Chatpose: Chatting about 3d human pose. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2093–2103 (June 2024)
30. Frick, A., Möhring, W., Newcombe, N.S.: Development of mental transformation abilities. Trends in cognitive sciences **18**(10), 536–542 (2014)
31. Fu, H., Cohen-Or, D., Dror, G., Sheffer, A.: Upright orientation of man-made objects. In: ACM SIGGRAPH 2008 Papers. SIGGRAPH '08, Association for

Computing Machinery, New York, NY, USA (2008). https://doi.org/10.1145/1399504.1360641, https://doi.org/10.1145/1399504.1360641
32. Fu, H., Jia, R., Gao, L., Gong, M., Zhao, B., Maybank, S., Tao, D.: 3d-future: 3d furniture shape with texture. International Journal of Computer Vision **129**, 3313–3337 (2021)
33. 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)
34. 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)
35. Ganesan, M., Kandhasamy, S., Chokkalingam, B., Mihet-Popa, L.: A comprehensive review on deep learning-based motion planning and end-to-end learning for self-driving vehicle. IEEE Access (2024)
36. Gao, J., Sarkar, B., Xia, F., Xiao, T., Wu, J., Ichter, B., Majumdar, A., Sadigh, D.: Physically grounded vision-language models for robotic manipulation. In: 2024 IEEE International Conference on Robotics and Automation (ICRA). pp. 12462–12469 (2024). https://doi.org/10.1109/ICRA57147.2024.10610090
37. Gao, J., Shen, T., Wang, Z., Chen, W., Yin, K., Li, D., Litany, O., Gojcic, Z., Fidler, S.: Get3d: A generative model of high quality 3d textured shapes learned from images. Advances In Neural Information Processing Systems **35**, 31841–31854 (2022)
38. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the kitti vision benchmark suite. In: 2012 IEEE conference on computer vision and pattern recognition. pp. 3354–3361. IEEE (2012)
39. Goble, D.J., Coxon, J.P., Van Impe, A., Geurts, M., Van Hecke, W., Sunaert, S., Wenderoth, N., Swinnen, S.P.: The neural basis of central proprioceptive processing in older versus younger adults: an important sensory role for right putamen. Human brain mapping **33**(4), 895–908 (2012)
40. Golarai, G., Ghahremani, D.G., Whitfield-Gabrieli, S., Reiss, A.L., Eberhardt, J.L., Gabrieli, J.D.E., Grill-Spector, K.: Differential development of high-level visual cortex correlates with category-specific recognition memory. Nature Neuroscience **10**, 512–522 (2007). https://doi.org/10.1038/nn1865
41. Gontier, C., Jordan, J., Petrovici, M.A.: Delaunay: A dataset of abstract art for psychophysical and machine learning research. In: International Conference on Soft Computing and Pattern Recognition. pp. 134–143. Springer (2023)
42. Gramann, K., Onton, J., Riccobon, D., Mueller, H.J., Bardins, S., Makeig, S.: Human brain dynamics accompanying use of egocentric and allocentric reference frames during navigation. Journal of cognitive neuroscience **22**(12), 2836–2849 (2010)
43. Guan, T., Yang, Y., Cheng, H., Lin, M., Kim, R., Madhivanan, R., Sen, A., Manocha, D.: LOC-ZSON: language-driven object-centric zero-shot object retrieval and navigation. In: 2024 IEEE International Conference on Robotics and Automation (ICRA) (2024)
44. Guibas, L.J.: Collaborative Research: CI-P: ShapeNet: An Information-Rich 3D Model Repository for Graphics, Vision and Robotics Research. NSF Award Number 1729205. Directorate for Computer and Information Science and Engineering, Division Of Computer and Network Systems. 2017. (Sep 2017)

45. Hamrick, J., Battaglia, P., Tenenbaum, J.B.: Internal physics models guide probabilistic judgments about object dynamics. In: Proceedings of the 33rd annual conference of the cognitive science society. vol. 2. Cognitive Science Society Austin, TX (2011)
46. Harris, I.M.: Interpreting the orientation of objects: A cross-disciplinary review. Psychonomic Bulletin & Review **31**(4), 1503–1515 (2024)
47. Hewing, M., Leinhos, V.: The prompt canvas: a literature-based practitioner guide for creating effective prompts in large language models. arXiv preprint arXiv:2412.05127 (2024)
48. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. ICLR **1**(2), 3 (2022)
49. Huang, W., Liu, H., Guo, M., Gong, N.: Visual hallucinations of multi-modal large language models. In: Ku, L.W., Martins, A., Srikumar, V. (eds.) Findings of the Association for Computational Linguistics: ACL 2024. pp. 9614–9631. Association for Computational Linguistics, Bangkok, Thailand (Aug 2024). https://doi.org/10.18653/v1/2024.findings-acl.573, https://aclanthology.org/2024.findings-acl.573/
50. Janai, J., Güney, F., Behl, A., Geiger, A., et al.: Computer vision for autonomous vehicles: Problems, datasets and state of the art. Foundations and trends® in computer graphics and vision **12**(1–3), 1–308 (2020)
51. Jia, M., Qi, Z., Zhang, S., Zhang, W., Yu, X., He, J., Wang, H., Yi, L.: Omnispatial: Towards comprehensive spatial reasoning benchmark for vision language models. In: The Fourteenth International Conference on Learning Representations (2026), https://openreview.net/forum?id=6nZKT2rL0H
52. Jiang, D., He, X., Zeng, H., Wei, C., Ku, M.W., Liu, Q., Chen, W.: Mantis: Interleaved multi-image instruction tuning. Transactions on Machine Learning Research **2024** (2024), https://openreview.net/forum?id=skLtdUVaJa
53. Jung, J.H., Kim, E.T., Kim, S., Lee, J.H., Kim, B., Chang, B.: Is 'right'right? enhancing object orientation understanding in multimodal large language models through egocentric instruction tuning. In: 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 14257–14267. IEEE (2025)
54. Kallmayer, A., Võ, M.L.H., Draschkow, D.: Viewpoint dependence and scene context effects generalize to depth rotated three-dimensional objects. Journal of Vision **23**(10), 9–9 (2023)
55. Kostakos, K., Pliakopanou, A., Meimaridis, V., Galanou, O., Anagnostou, A., Sertidou, D., Katis, P., Anastasiou, P., Katsoulidis, K., Lykogiorgos, Y., Mytilinaios, D., Katsenos, A.P., Simos, Y.V., Bellos, S., Konitsiotis, S., Peschos, D., Tsamis, K.I.: Development of spatial memory: A behavioral study. Neurosci **5**, 713–728 (2024). https://doi.org/10.3390/neurosci5040050
56. Kourtzi, Z., Nakayama, K.: Distinct mechanisms for the representation of moving and static objects. Visual Cognition **9**(1-2), 248–264 (2002). https://doi.org/10.1080/13506280143000421
57. Krishnan, A., Kundu, A., Maninis, K.K., Hays, J., Brown, M.: Omninocs: A unified nocs dataset and model for 3d lifting of 2d objects. In: European Conference on Computer Vision. pp. 127–145. Springer (2024)
58. Lei, X., Yang, Z., Chen, X., Li, P., Liu, Y.: Scaffolding coordinates to promote vision-language coordination in large multi-modal models. In: Rambow, O., Wanner, L., Apidianaki, M., Al-Khalifa, H., Eugenio, B.D., Schockaert, S. (eds.) Proceedings of the 31st International Conference on Computational Linguistics. pp.

2886–2903. Association for Computational Linguistics, Abu Dhabi, UAE (Jan 2025), https://aclanthology.org/2025.coling-main.195/

59. Li, B., Ge, Y., Chen, Y., Ge, Y., Zhang, R., Shan, Y.: Seed-bench-2-plus: Benchmarking multimodal large language models with text-rich visual comprehension. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2024)
60. Li, F., Hogg, D.C., Cohn, A.G.: Reframing spatial reasoning evaluation in language models: a real-world simulation benchmark for qualitative reasoning. In: Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence. IJCAI '24 (2024). https://doi.org/10.24963/ijcai.2024/701, https://doi.org/10.24963/ijcai.2024/701
61. Li, H., Liu, N., Li, Y., Weidner, R., Fink, G.R., Chen, Q.: The simon effect based on allocentric and egocentric reference frame: Common and specific neural correlates. Scientific reports **9**(1), 13727 (2019)
62. Li, S., Koh, P.W., Du, S.S.: Exploring how generative MLLMs perceive more than CLIP with the same vision encoder. In: Che, W., Nabende, J., Shutova, E., Pilehvar, M.T. (eds.) Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 10101–10119. Association for Computational Linguistics, Vienna, Austria (Jul 2025). https://doi.org/10.18653/v1/2025.acl-long.499, https://aclanthology.org/2025.acl-long.499/
63. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. pp. 740–755. Springer (2014)
64. Lin, X., Dai, Z., Verma, A., Ng, S.K., Jaillet, P., Low, B.K.H.: Prompt optimization with human feedback (2025), https://openreview.net/forum?id=UW0zetsx8X
65. Lin, Z., Liu, D., Zhang, R., Gao, P., Qiu, L., Xiao, H., Qiu, H., Shao, W., Chen, K., Han, J., Huang, S., Zhang, Y., He, X., Qiao, Y., Li, H.: Sphinx: A mixer of weights, visual embeddings and image scales for multi-modal large language models. In: ECCV (62). pp. 36–55 (2024), https://doi.org/10.1007/978-3-031-73033-7_3
66. Ling, S., Pearson, J., Blake, R.: Dissociation of neural mechanisms underlying orientation processing in humans. Current Biology **19**(17), 1458–1462 (2009)
67. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 26296–26306 (June 2024)
68. 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/
69. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning (2023)
70. Liu, R., Yu, Z., Fan, Q., Sun, Q., Jiang, Z.: The improved method in fabric image classification using convolutional neural network. Multimedia Tools and Applications **83**(3), 6909–6924 (2024)
71. Liu, Y., Duan, H., Zhang, Y., Li, B., Zhang, S., Zhao, W., Yuan, Y., Wang, J., He, C., Liu, Z., Chen, K., Lin, D.: Mmbench: Is your multi-modal model an all-around player? In: Leonardis, A., Ricci, E., Roth, S., Russakovsky, O., Sattler, T., Varol, G. (eds.) Computer Vision – ECCV 2024. pp. 216–233. Springer Nature Switzerland, Cham (2025)

72. Loy, J.E., Demberg, V.: Individual differences in spatial orientation modulate perspective taking in listeners. Journal of Cognition **6** (2023). https://doi.org/10.5334/joc.321
73. Ma, W., Chen, H., Zhang, G., Chou, Y.C., de Melo, C.M., Yuille, A.: 3dsrbench: A comprehensive 3d spatial reasoning benchmark. In: International Conference on Computer Vision (ICCV) (2025)
74. Mallot, H.A.: From geometry to behavior : an introduction to spatial cognition. The MIT Press, Cambridge, 1st ed. edn. (2023)
75. Mezuman, E., Weiss, Y.: Learning about canonical views from internet image collections. Advances in neural information processing systems **25** (2012)
76. Mirzaee, R., Rajaby Faghihi, H., Ning, Q., Kordjamshidi, P.: SPARTQA: A textual question answering benchmark for spatial reasoning. In: Toutanova, K., Rumshisky, A., Zettlemoyer, L., Hakkani-Tur, D., Beltagy, I., Bethard, S., Cotterell, R., Chakraborty, T., Zhou, Y. (eds.) Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 4582–4598. Association for Computational Linguistics, Online (Jun 2021). https://doi.org/10.18653/v1/2021.naacl-main.364, https://aclanthology.org/2021.naacl-main.364/
77. Mu, Y., Zhang, Q., Hu, M., Wang, W., Ding, M., Jin, J., Wang, B., Dai, J., Qiao, Y., Luo, P.: Embodiedgpt: vision-language pre-training via embodied chain of thought. In: Proceedings of the 37th International Conference on Neural Information Processing Systems. NIPS '23, Curran Associates Inc., Red Hook, NY, USA (2023)
78. Muto, H.: Correlational evidence for the role of spatial perspective-taking ability in the mental rotation of human-like objects. Experimental Psychology **68**, 41–48 (2021). https://doi.org/10.1027/1618-3169/a000505
79. neelgajare: Rock images (2022), https://www.kaggle.com/datasets/neelgajare/rocks-dataset [Accessed: (01-29-2025)]
80. Peer, M., Salomon, R., Goldberg, I., Blanke, O., Arzy, S.: Brain system for mental orientation in space, time, and person. Proceedings of the National Academy of Sciences **112**(35), 11072–11077 (2015). https://doi.org/10.1073/pnas.1504242112, https://www.pnas.org/doi/abs/10.1073/pnas.1504242112
81. Pei, J., Viola, I., Huang, H., Wang, J., Ahsan, M., Ye, F., Jiang, Y., Sai, Y., Wang, D., Chen, Z., Ren, P., César, P.: Autonomous workflow for multimodal fine-grained training assistants towards mixed reality. In: Annual Meeting of the Association for Computational Linguistics (2024), https://api.semanticscholar.org/CorpusID:269982847
82. Ramakrishnan, S.K., Wijmans, E., Kraehenbuehl, P., Koltun, V.: Does spatial cognition emerge in frontier models? In: ICLR (2025)
83. Ranasinghe, K., Shukla, S.N., Poursaeed, O., Ryoo, M.S., Lin, T.Y.: Learning to localize objects improves spatial reasoning in visual-llms. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12977–12987 (2024)
84. Ray, A., Duan, J., II, E.L.B., Tan, R., Bashkirova, D., Hendrix, R., Ehsani, K., Kembhavi, A., Plummer, B.A., Krishna, R., Zeng, K.H., Saenko, K.: SAT: Dynamic spatial aptitude training for multimodal language models. In: Second Conference on Language Modeling (2025), https://openreview.net/forum?id=DW8U8ZWa1U
85. Schmithorst, V.J., Holland, S.K., Plante, E.: Object identification and lexical/semantic access in children: A functional magnetic resonance imaging study

of word-picture matching. Human Brain Mapping **28**, 1060–1074 (2006). https://doi.org/10.1002/hbm.20328
86. Shiri, F., Guo, X.Y., Far, M., Yu, X., Haf, R., Li, Y.F.: An empirical analysis on spatial reasoning capabilities of large multimodal models. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. pp. 21440–21455 (2024)
87. Song, C.H., Blukis, V., Tremblay, J., Tyree, S., Su, Y., Birchfield, S.: RoboSpatial: Teaching spatial understanding to 2D and 3D vision-language models for robotics. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2025), oral Presentation
88. Spinelli, D., Antonucci, G., Daini, R., Martelli, M.L., Zoccolotti, P.: Hierarchical organisation in perception of orientation. Perception **28**(8), 965–979 (1999)
89. Stogiannidis, I., McDonagh, S., Tsaftaris, S.A.: Mind the gap: Benchmarking spatial reasoning in vision-language models. In: Greeks in AI Symposium 2025 (2025), https://openreview.net/forum?id=BMl8OyGfo5
90. Team, Q.: Qwen2.5: A party of foundation models (September 2024), https://qwenlm.github.io/blog/qwen2.5/
91. Ter Horst, A.C., Van Lier, R., Steenbergen, B.: Mental rotation task of hands: differential influence number of rotational axes. Experimental Brain Research **203**, 347–354 (2010)
92. Teterwak, P., Saito, K., Tsiligkaridis, T., Plummer, B.A., Saenko, K.: Is large-scale pretraining the secret to good domain generalization? Advances in Neural Information Processing Systems (2025)
93. 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)
94. Tong, S., Liu, Z., Zhai, Y., Ma, Y., LeCun, Y., Xie, S.: Eyes wide shut? exploring the visual shortcomings of multimodal llms. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9568–9578 (2024)
95. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., Lample, G.: Llama: Open and efficient foundation language models. CoRR **abs/2302.13971** (2023)
96. Tuthill, J.C., Azim, E.: Proprioception. Current Biology **28**(5), R194–R203 (2018)
97. Tversky, B., Suwa, M.: Thinking with sketches. Tools for Innovation pp. 75–84 (2009). https://doi.org/10.1093/acprof:oso/9780195381634.003.0004
98. unsplash: The unsplash dataset (2020), https://github.com/unsplash/datasets?tab=readme-ov-file [Accessed: (01-29-2025)]
99. Vasilyeva, M., Lourenco, S.F.: Development of spatial cognition. Wiley Interdisciplinary Reviews: Cognitive Science **3**(3), 349–362 (2012)
100. Vieites, V., Nazareth, A., Reeb-Sutherland, B.C., Pruden, S.M.: A new biomarker to examine the role of hippocampal function in the development of spatial reorientation in children: a review. Frontiers in Psychology **Volume 6 - 2015** (2015). https://doi.org/10.3389/fpsyg.2015.00490, https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2015.00490
101. Viganò, S., Bayramova, R., Doeller, C.F., Bottini, R.: Mental search of concepts is supported by egocentric vector representations and restructured grid maps. Nature Communications **14**(1), 8132 (2023)

102. Wang, H., Sridhar, S., Huang, J., Valentin, J., Song, S., Guibas, L.J.: Normalized object coordinate space for category-level 6d object pose and size estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2642–2651 (2019)
103. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., Fan, Y., Dang, K., Du, M., Ren, X., Men, R., Liu, D., Zhou, C., Zhou, J., Lin, J.: Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191 (2024)
104. 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)
105. Wang, X., Kwon, T., Rad, M., Pan, B., Chakraborty, I., Andrist, S., Bohus, D., Feniello, A., Tekin, B., Frujeri, F.V., et al.: Holoassist: an egocentric human interaction dataset for interactive ai assistants in the real world. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 20270–20281 (2023)
106. Wang, X., Ma, W., Li, Z., Kortylewski, A., Yuille, A.L.: 3d-aware visual question answering about parts, poses and occlusions. Advances in Neural Information Processing Systems **36**, 58717–58735 (2023)
107. Wang, X., Ma, W., Zhang, T., de Melo, C.M., Chen, J., Yuille, A.: Spatial457: A diagnostic benchmark for 6d spatial reasoning of large multimodal models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2025)
108. 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: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 24669–24679 (2025)
109. Wei, Y., Wang, Z., Lu, Y., Xu, C., Liu, C., Zhao, H., Chen, S., Wang, Y.: Editable scene simulation for autonomous driving via collaborative llm-agents. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2024)
110. Weston, J., Bordes, A., Chopra, S., Mikolov, T.: Towards ai-complete question answering: A set of prerequisite toy tasks. In: Bengio, Y., LeCun, Y. (eds.) 4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings (2016)
111. Willett, K.W., Lintott, C.J., Bamford, S.P., Masters, K.L., Simmons, B.D., Casteels, K.R., Edmondson, E.M., Fortson, L.F., Kaviraj, S., Keel, W.C., et al.: Galaxy zoo 2: detailed morphological classifications for 304 122 galaxies from the sloan digital sky survey. Monthly Notices of the Royal Astronomical Society **435**(4), 2835–2860 (2013)
112. Wu, T., Zhang, J., Fu, X., Wang, Y., Ren, J., Pan, L., Wu, W., Yang, L., Wang, J., Qian, C., et al.: Omniobject3d: Large-vocabulary 3d object dataset for realistic perception, reconstruction and generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 803–814 (2023)
113. Xue, J., Li, C., Quan, C., Lu, Y., Yue, J., Zhang, C.: Uncovering the cognitive processes underlying mental rotation: An eye-movement study. Scientific Reports **7**(1), 10076 (2017)
114. Yamada, Y., Bao, Y., Lampinen, A.K., Kasai, J., Yildirim, I.: Evaluating spatial understanding of large language models. Transactions on Machine Learning Research (2024), https://openreview.net/forum?id=xkiflfKCw3

115. Yan, X., Yuan, Z., Du, Y., Liao, Y., Guo, Y., Cui, S., Li, Z.: Comprehensive visual question answering on point clouds through compositional scene manipulation. IEEE Transactions on Visualization & Computer Graphics pp. 1–13 (2023)
116. Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., Zheng, C., Liu, D., Zhou, F., Huang, F., Hu, F., Ge, H., Wei, H., Lin, H., Tang, J., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Zhou, J., Lin, J., Dang, K., Bao, K., Yang, K., Yu, L., Deng, L., Li, M., Xue, M., Li, M., Zhang, P., Wang, P., Zhu, Q., Men, R., Gao, R., Liu, S., Luo, S., Li, T., Tang, T., Yin, W., Ren, X., Wang, X., Zhang, X., Ren, X., Fan, Y., Su, Y., Zhang, Y., Zhang, Y., Wan, Y., Liu, Y., Wang, Z., Cui, Z., Zhang, Z., Zhou, Z., Qiu, Z.: Qwen3 technical report. arXiv preprint arXiv:2505.09388 (2025)
117. Yang, C., Xu, R., Guo, Y., Huang, P., Chen, Y., Ding, W., Wang, Z., Zhou, H.: Improving vision-and-language reasoning via spatial relations modeling. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 769–778 (2024)
118. Yang, J., Tan, R., Wu, Q., Zheng, R., Peng, B., Liang, Y., Gu, Y., Cai, M., Ye, S., Jang, J., et al.: Magma: A foundation model for multimodal ai agents. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 14203–14214 (2025)
119. 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)
120. Yang, S., Xu, R., Xie, Y., Yang, S., Li, M., Lin, J., Zhu, C., Chen, X., Duan, H., Yue, X., Lin, D., Wang, T., Pang, J.: Mmsi-bench: A benchmark for multi-image spatial intelligence. In: ICLR (2025)
121. Yeh, C.H., Wang, C., Tong, S., Cheng, T.Y., Wang, R., Chu, T., Zhai, Y., Chen, Y., Gao, S., Ma, Y.: Seeing from another perspective: Evaluating multi-view understanding in mllms. arXiv preprint arXiv:2504.15280 (2025)
122. Yin, H., Lin, Z., Liu, X., Sun, B., Li, K.: Do multimodal language models really understand direction? a benchmark for compass direction reasoning. In: ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5. IEEE (2025)
123. Yiu, E., Qraitem, M., Majhi, A.N., Wong, C., Bai, Y., Ginosar, S., Gopnik, A., Saenko, K.: KiVA: Kid-inspired visual analogies for testing large multimodal models. In: The Thirteenth International Conference on Learning Representations (2025), https://openreview.net/forum?id=vNATZfmY6R
124. Yuan, W., Duan, J., Blukis, V., Pumacay, W., Krishna, R., Murali, A., Mousavian, A., Fox, D.: Robopoint: A vision-language model for spatial affordance prediction in robotics. In: Agrawal, P., Kroemer, O., Burgard, W. (eds.) Proceedings of The 8th Conference on Robot Learning. Proceedings of Machine Learning Research, vol. 270, pp. 4005–4020. PMLR (06–09 Nov 2025), https://proceedings.mlr.press/v270/yuan25c.html
125. Zaehle, T., Jordan, K., Wüstenberg, T., Baudewig, J., Dechent, P., Mast, F.W.: The neural basis of the egocentric and allocentric spatial frame of reference. Brain research **1137**, 92–103 (2007)
126. Zheng, C., Wu, W., Chen, C., Yang, T., Zhu, S., Shen, J., Kehtarnavaz, N., Shah, M.: Deep learning-based human pose estimation: A survey. ACM Comput. Surv. **56**(1) (2023)

127. Zhong, Y., Yang, J., Zhang, P., Li, C., Codella, N., Li, L.H., Zhou, L., Dai, X., Yuan, L., Li, Y., et al.: Regionclip: Region-based language-image pretraining. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16793–16803 (2022)
128. Zhou, S., Vilesov, A., He, X., Wan, Z., Zhang, S., Nagachandra, A., Chang, D., Chen, D., Wang, E.X., Kadambi, A.: Vlm4d: Towards spatiotemporal awareness in vision language models. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 8600–8612 (2025)