FRONTIERSPATIAL: HOLISTIC EVALUATION OF SPATIAL REASONING IN LARGE MULTIMODAL MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Spatial reasoning is foundational to intelligent systems, yet its evaluation in multimodal large language models (MLLMs) remains fragmented across narrow benchmarks that test isolated skills. We introduce FRONTIERSPATIAL, a unified benchmark and evaluation framework for large-scale spatial reasoning. We consolidate 23 existing datasets into **7,838 questions paired with 11,825 images**, organized into 7 categories, 19 sub-categories, and 65 fine-grained tasks spanning physical, geometric, knowledge-grounded, relational/causal, temporal, quantitative, and perceptual reasoning. Beyond the benchmark, we provide (1) an automated curation pipeline that distills over 60,000 raw instances into a balanced dataset and generalizes to other domains, (2) an automated evaluation harness for reproducible assessment of 16 state-of-the-art MLLMs, and (3) an automated error-analysis harness for fine-grained diagnosis across the task hierarchy. Our evaluation reveals large and uneven performance gaps across spatial skills, exposing systematic weaknesses that remain hidden in isolated benchmarks. FRONTIERSPATIAL provides a rigorous foundation for systematic, fine-grained evaluation and deeper understanding of multimodal spatial reasoning.

1 INTRODUCTION

Spatial reasoning is central to human cognition, supporting geometric understanding, depth and occlusion inference, motion tracking, navigation, and mental manipulation of physical space, with applications spanning robotics, autonomous driving, augmented reality, geospatial analytics, medical imaging, and education (Huang et al., 2024; Qi et al., 2024; 2025; Wasi et al., 2026). It requires multimodal agents to integrate visual cues, scene structure, and multi-step inference across noisy, geometrically complex viewpoints (Bar-Anan et al., 2006; Cai et al., 2025b; Trope & Liberman, 2010; Ramalho et al., 2018). Despite rapid progress in vision–language models (VLMs), their spatial reasoning remains fragile and poorly understood (Chen et al., 2025; Kosoy et al., 2025; Pothiraj et al., 2025; Wasi et al., 2026). A major barrier is fragmented evaluation: existing benchmarks typically target narrow skills such as positional relations (Kamath et al., 2023; Du et al., 2024), depth estimation (Cheng et al., 2024a), navigation (Song et al., 2024), or domain-specific perception (Hu et al., 2023; Wang et al., 2023), often in synthetic or simplified settings that mask distribution shifts and reduce perceptual diversity (Jia et al., 2025; Szymanska et al., 2024; Shiri et al., 2024; Cai et al., 2024; Yang et al., 2024; Qi et al., 2025). As a result, critical abilities such as occlusion inference, scale-consistent reasoning across viewpoints, planning under partial observability, and causal spatial understanding remain under-tested (Kosoy et al., 2025; Pothiraj et al., 2025; Chen et al., 2025), while strong performance on individual benchmarks often fails to translate to integrated, multi-step spatial reasoning (Wasi et al., 2026; Cai et al., 2025a; Jiang et al., 2024).

Current benchmarks cover only a narrow subset of the 22 spatial reasoning dimensions we identify (Table 3). Benchmarks such as SPATIAL457 (Wang et al., 2025b), MMSI-BENCH (Yang et al., 2025), and SPATIALAB (Wasi et al., 2026) address several spatial sub-skills but remain confined to 4–6 dimensions each. Even broad multimodal benchmarks such as MMMU (Yue et al., 2023) and MM-VET (Yu et al., 2023) provide only incidental coverage of spatial tasks. No existing benchmark achieves simultaneous coverage across all core axes of spatial reasoning: physical interaction, temporal dynamics, geometric and 3D structure, logical and causal inference, quantitative comparison, knowledge-grounded understanding, and perceptual recognition, leaving the community without a holistic view of VLM spatial intelligence.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

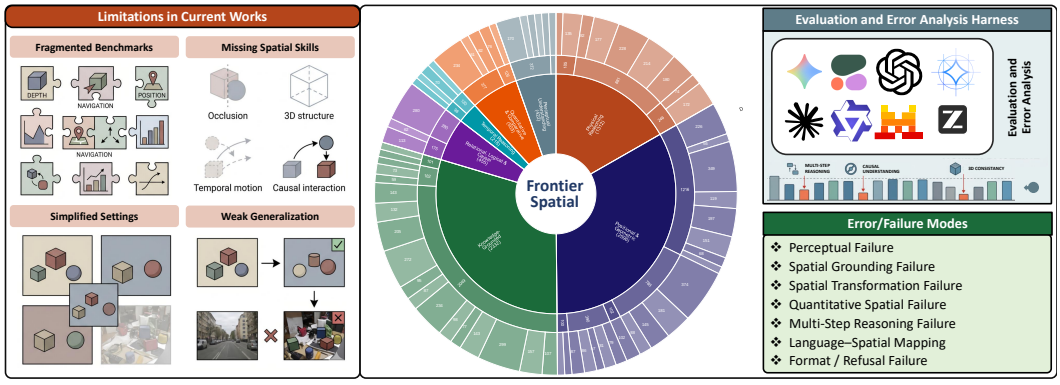


Figure 1: **Overview of FRONTIERSPATIAL.** The figure contrasts limitations of existing benchmarks (left), including fragmentation, simplified settings, missing spatial skills, and weak generalization, with the proposed unified taxonomy (center), spanning 7 categories, 19 sub-categories, and 65 tasks across diverse spatial dimensions. On the right, the evaluation and error analysis harness, inspired by HAL (Kapoor et al., 2026), enables systematic assessment across models with standardized metrics and human baselines, while providing fine-grained diagnostic insights into failure modes such as perceptual errors, spatial grounding issues, and multi-step reasoning breakdowns.

To close this gap, we introduce FRONTIERSPATIAL, a holistic benchmark and evaluation framework for spatial reasoning in multimodal language models. Drawing from 23 existing datasets and using an automated curation pipeline, FRONTIERSPATIAL assembles **7,838 questions paired with 11,825 images** in a four-level hierarchy of **7 categories**, **19 sub-categories**, and **65 fine-grained tasks**. The taxonomy covers Physical, Positional and Geometric, Knowledge-Grounded, Relational/Logical and Causal, Temporal and Dynamic, Quantitative and Comparative, and Perceptual reasoning. We further provide automated evaluation and error-analysis harnesses for reproducible assessment of current VLMs and fine-grained diagnosis of their failures. Importantly, the benchmark serves as an evaluation vehicle rather than the core scientific contribution of FRONTIERSPATIAL. We contribute the following methodological frameworks and empirical findings for systematically evaluating and understanding multimodal spatial reasoning:

1. **Holistic Taxonomy of Visual Spatial Reasoning.** We introduce a unified taxonomy spanning 7 categories, 19 sub-categories, and 65 tasks, enabling systematic evaluation across the major dimensions of spatial cognition.
2. **Automated Curation Pipeline and Dataset.** We develop a scalable pipeline that distills over 60,000 raw instances into a balanced, high-quality benchmark through taxonomic coverage, quality filtering, and deduplication. The resulting dataset contains 7,838 QA pairs across 11,825 images from 23 existing datasets, and the pipeline can be extended to other reasoning domains.
3. **Automated Evaluation and Error Analysis.** We provide a reproducible evaluation harness for 16 state-of-the-art open- and closed-source VLMs, including reasoning-focused and spatially specialized models, together with human baselines. An accompanying error-analysis harness produces structured diagnostics across the task hierarchy, exposing systematic failure modes.
4. **Agent-Based Hard Case Evaluation.** We further study the hardest instances using two agentic setups: *SpatioXolver*, a multi-agent system, and a *Claude Code*-based agent. Both solve a subset of problems missed by all evaluated VLMs, suggesting that decomposition and tool-assisted reasoning can help on difficult spatial tasks.

Our evaluation shows that current VLMs remain far from reliable spatial reasoning. Performance varies sharply across skills, with several sub-skills approaching chance despite strong human performance. The human–model gap is substantial: humans achieve **91.22%**, compared with **45.20%** for the best model. Some capabilities remain particularly weak; for example, Quantitative & Comparative reasoning stays below 17% across all evaluated systems. Even strong proprietary models, including GEMINI-3.1-FLASH-LITE and GPT-5-MINI, struggle on several reasoning axes. Errors are especially common in compositional multi-step inference, causal spatial reasoning, and perspective-consistent 3D reasoning. These results show that strong performance on individual spatial benchmarks does not imply robust spatial reasoning across tasks, and motivate more comprehensive evaluation of multimodal reasoning systems.

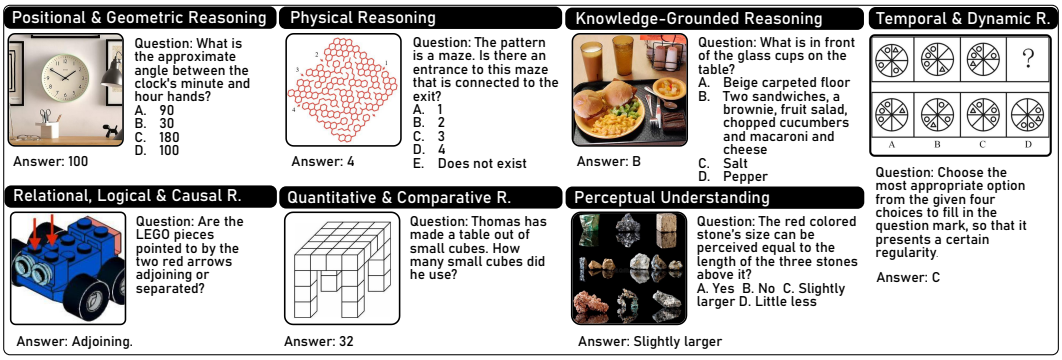


Figure 2: Representative examples from FRONTIERSPATIAL.

2 PRELIMINARIES

Formal Definition. Visual–spatial reasoning is the cognitive capacity to perceive, represent, and manipulate geometric and relational structures among entities in a scene, a faculty central to human intelligence and indispensable for embodied agents operating in dynamic, unstructured environments (Palmer, 1999; Chabris et al., 2006; Baddeley, 1998; Previc, 1998). Formally, let a visual scene be represented as a set of entities $\mathcal{O} = \{o_1, \ldots, o_n\}$, where each entity o_i carries a tuple of spatial attributes: $o_i = (\mathbf{p}_i,\ \mathbf{r}_i,\ s_i,\ \mathbf{v}_i)$ denoting position $\mathbf{p}_i \in \mathbb{R}^3$, orientation $\mathbf{r}_i \in SO(3)$, scale $s_i \in \mathbb{R}_+$, and velocity $\mathbf{v}_i \in \mathbb{R}^3$ (for dynamic scenes), all defined within a world coordinate frame $\mathcal{W}$. The observer introduces an additional egocentric frame $\mathcal{E}$, governed by a viewpoint transform $\mathbf{T} \in SE(3)$, such that agent-centric spatial predicates (e.g., *in-front-of*, *to-the-left*) must be resolved relative to $\mathcal{E}$ rather than $\mathcal{W}$ (Trope & Liberman, 2010; Ramalho et al., 2018).

A spatial relation between two entities is a predicate $r(o_i, o_j) \in \mathcal{R}$, where $\mathcal{R}$ is a structured relation vocabulary encompassing categories such as *positional* (e.g., *above*, *behind*), *geometric* (e.g., *parallel-to*, *intersects*), *physical* (e.g., *supports*, *occludes*), *causal* (e.g., *causes-motion*, *blocks*), and *temporal* (e.g., *precedes*, *co-occurs-with*). The full relational structure of a scene is then a typed graph: $\mathcal{G} = \big(\mathcal{O},\ \{(o_i, r_k, o_j) \mid o_i, o_j \in \mathcal{O},\ r_k \in \mathcal{R}\}\big)$. Visual–spatial reasoning in a vision–language model (VLM) seeks to learn a mapping: $f_\theta :\ (I,\ Q) \longmapsto A$ where $I \in \mathcal{I}$ is an image or image sequence, $Q \in \mathcal{Q}$ is a natural language query, and A is either a structured relational answer drawn from $\mathcal{G}$ or a free-form linguistic response. Learning f_θ is fundamentally difficult because it requires the simultaneous satisfaction of three constraints: *perceptual grounding* (extracting $\mathcal{O}$ and their attributes from I under noise, occlusion, and viewpoint change), *relational inference* (composing predicates in $\mathcal{R}$ across multiple entities), and *linguistic alignment* (mapping $\mathcal{G}$-structured answers to and from natural language under referential ambiguity) (Alayrac et al., 2022; OpenAI, 2023; Chen et al., 2025).

The Holistic Argument: Why Partial Evaluation is Insufficient. Prior work typically decomposes spatial reasoning into isolated sub-tasks, e.g., binary positional relations (Kamath et al., 2023; Du et al., 2024), metric depth estimation (Cheng et al., 2024b;a), object counting (Charoenpitaks et al., 2025), or navigation-centric grounding (Song et al., 2024). While such decomposition enables controlled evaluation, it misrepresents the nature of spatial cognition. Human spatial reasoning is inherently *integrative*: resolving queries such as *"which object would fall first if the shelf were removed?"* requires joint reasoning over perception (object detection), geometry (support relations), physics (gravity and contact), and causality, all under a consistent viewpoint (Gardner, 2011; Previc, 1998; Bar-Anan et al., 2006).

Formally, we define a *holistic spatial query* as one whose resolution requires composition of predicates drawn from at least two distinct sub-vocabularies of $\mathcal{R}$. Let $\mathcal{R} = \mathcal{R}_{\text{pos+geo}} \cup \mathcal{R}_{\text{phys}} \cup \mathcal{R}_{\text{know}} \cup \mathcal{R}_{\text{caus}} \cup \mathcal{R}_{\text{temp}} \cup \mathcal{R}_{\text{quant}} \cup \mathcal{R}_{\text{percep}}$ denote the partition of relations corresponding to our seven taxonomy categories (§3), where $\mathcal{R}_{\text{pos+geo}} = \mathcal{R}_{\text{pos}} \cup \mathcal{R}_{\text{geo}}$ merges positional predicates (above, behind, left-of) with geometric predicates (parallel-to, intersects, congruent-to) because both are resolved through frame-of-reference management; $\mathcal{R}_{\text{know}}$ contains predicates whose resolution requires conditioning on world knowledge in addition to perceptual evidence, such as located-in-region (o, r), functionally-adjacent-to (o_i, o_j), and conforms-to-convention (o, c).

A benchmark $\mathcal{B}$ is *holistic* at precision ϵ and composition level ρ if (i) every category has sufficient queries to estimate its accuracy with a worst-case 95% confidence half-width of at most ϵ, i.e., $\forall k \in \{1, \ldots, 7\}, \qquad |\{q \in \mathcal{B} \mid \mathcal{R}_k \text{ is required}\}| \geq \tau(\epsilon) = \left\lceil \left(\frac{0.98}{\epsilon}\right)^2 \right\rceil$, and (ii) at least a fraction

ρ of queries require predicates from two or more distinct sub-vocabularies of $\mathcal{R}$. We set $\epsilon = 0.07$, yielding $\tau = 196$, and require $\rho = 0.25$. FRONTIERSPATIAL satisfies condition (i), with 218 queries in the smallest category (Temporal & Dynamic Reasoning), corresponding to a worst-case 95% half-width of ± 6.64 percentage points. Condition (ii) is evaluated and reported in §3.1.
The consequence is practical. Strong performance on a benchmark $\mathcal{B}_k$ covering only $\mathcal{R}_k$ does not imply competence on $\mathcal{R}_{k'}$ for $k' \neq k$, nor on queries requiring compositional reasoning across $\mathcal{R}_k \times \mathcal{R}_{k'}$ (Kosoy et al., 2025; Pothiraj et al., 2025). Since real-world applications, e.g., robotics, autonomous driving, and augmented reality, require such composition (Huang et al., 2024; Collaboration et al., 2023; Qi et al., 2025), narrow benchmarks are poor predictors of deployment readiness. This gap motivates FRONTIERSPATIAL.

A Taxonomy of Spatial Reasoning. To operationalize holistic evaluation, we organize spatial reasoning into a four-level hierarchy. At the top level, we identify seven *categories* corresponding to the sub-vocabularies of $\mathcal{R}$ defined above, each grounded in a distinct cognitive faculty documented in the spatial cognition literature (Baddeley, 1998; Previc, 1998; Palmer, 1999) (Table 5):

1. **Physical Reasoning** ($\mathcal{R}_{\text{phys}}$, $n = 1{,}312$): Inference about forces, stability, support, contact, and material properties. Requires integrating visual appearance with intuitive physics.
2. **Positional & Geometric Reasoning** ($\mathcal{R}_{\text{pos}} \cup \mathcal{R}_{\text{geo}}$, $n = 2{,}596$): Resolution of spatial predicates such as *left-of*, *above*, *parallel*, and metric distance, requiring consistent frame-of-reference management under viewpoint change (Trope & Liberman, 2010; Ramalho et al., 2018).
3. **Knowledge-Grounded Reasoning** ($\mathcal{R}_{\text{know}}$, $n = 2{,}332$): Spatial inference that must be conditioned on domain knowledge, geographic, architectural, or functional, linking perceptual cues to learned world models (Yue et al., 2023; Wang et al., 2025a).
4. **Relational, Logical & Causal Reasoning** ($\mathcal{R}_{\text{caus}}$, $n = 455$): Multi-step inference over entity relations, including counterfactual and causal queries (e.g., *"what would move if X were removed?"*) (Ma et al., 2026a; Jiang et al., 2024).
5. **Temporal & Dynamic Reasoning** ($\mathcal{R}_{\text{temp}}$, $n = 218$): Tracking spatial configurations across time, inferring motion trajectories, and predicting future states (Charoenpitaks et al., 2025).
6. **Quantitative & Comparative Reasoning** ($\mathcal{R}_{\text{quant}}$, $n = 503$): Metric and ordinal judgments over spatial properties, distance, count, and angle, requiring calibrated visual measurement.
7. **Perceptual Understanding** ($\mathcal{R}_{\text{percep}}$, $n = 422$): Low-level spatial perception, object localization, scene-level layout recognition, and textual spatial cues, forming the perceptual substrate upon which higher-order reasoning operates (Yu et al., 2023; Hu et al., 2023).

Each category is further decomposed into *sub-categories* (19 total) and *tasks* (65 total), forming a four-level hierarchy (Table 5): Category → Sub-category → Task → Instance. This granularity enables targeted diagnostics: a model may achieve strong aggregate accuracy while exhibiting categorical failure in, say, temporal reasoning or occlusion-conditioned physical inference, a pattern that flat benchmarks cannot expose.

Evaluation Framework. Let $\mathcal{B} = \{(I_i, Q_i, A_i^*)\}_{i=1}^N$ denote the benchmark, where A_i^* is the ground-truth answer. The model generates a free-form response $\hat{A}_i = f_\theta(I_i, Q_i)$, evaluated against A_i^* using an automated judge (§3.2,D). This format demands genuine generative spatial reasoning and compositional description, exposing weaknesses masked by multiple-choice formats (Yu et al., 2023; Wasi et al., 2026; Liu et al., 2025). Together, these protocols reflect the dual demands of real-world spatial reasoning, recognition under constraint and generation under ambiguity, and form the basis of the automated evaluation and error analysis harnesses described in §5.3 and E.

3 FRONTIERSPATIAL

FRONTIERSPATIAL is a large-scale, holistic benchmark for evaluating spatial reasoning in vision–language models (VLMs), constructed by consolidating and curating instances from 23 existing datasets through a four-stage automated pipeline (§C). The benchmark comprises **7,838 visual question–answer pairs** paired with **11,825 images**, organized under a principled four-level task hierarchy of **7 categories**, **19 sub-categories**, and **65 fine-grained tasks** (§B, Table 5). Unlike prior spatial benchmarks that target narrow slices of the problem space (Kamath et al., 2023; Du et al., 2024; Cheng et al., 2024b; Song et al., 2024), FRONTIERSPATIAL achieves simultaneous coverage across all major dimensions of spatial cognition, from physical interaction and 3D geometry to temporal dynamics, causal inference, and knowledge-grounded spatial understanding. Both multiple-choice and open-ended evaluation formats are supported, enabling assessment of discriminative

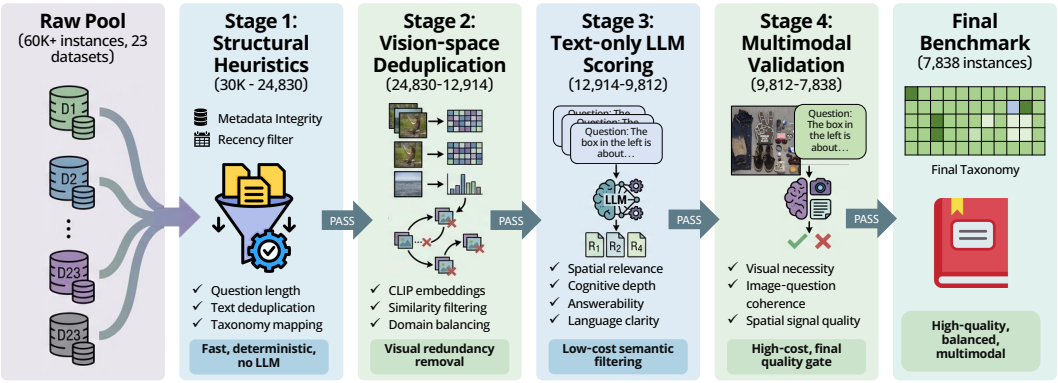


Figure 3: **Automated dataset curation pipeline.** The pipeline reduces ∼60K instances from 23 datasets to 7,838 through four stages: structural filtering, vision-space deduplication, text-only LLM scoring, and multimodal validation. Each stage removes low-quality or redundant data while increasing computational cost, resulting in a balanced and high-quality benchmark.

accuracy alongside generative spatial reasoning. Table 3 illustrates that FRONTIERSPATIAL is the only benchmark achieving full coverage across all 22 spatial reasoning dimensions identified in our taxonomy. Some examples are given in Figure 2, more examples with model responses and evaluations are available in §M.

3.1 DATA CURATION PIPELINE

The benchmark was constructed using a four-stage automated curation pipeline (Figure 3, Table 7) designed to be principled, reproducible, and transferable to other reasoning domains. Starting from over 60,000 instances across 23 source datasets, the pipeline applies progressively stricter quality criteria to obtain 7,838 high-quality instances. Further details are provided in §C.

Stage 1: Structural heuristics (60K+ → 24,830). Five deterministic, LLM-free filters are applied in sequence: metadata integrity validation (ensuring all image paths resolve and answer fields are non-empty), publication recency filtering (removing pre-2020 annotation artifacts), question length filtering ($[10, 200]$ tokens), SHA-256-based exact-text deduplication across source boundaries, and taxonomy remapping with non-spatial pruning. The last step operationalizes the FRONTIERSPATIAL taxonomy as a quality gate, any instance that cannot be mapped to the 65-task hierarchy is removed. Stage 1 eliminates ∼50% of the raw pool at zero LLM cost.

Stage 2: Vision-space deduplication (24,830 → 12,914). All images are encoded with a CLIP ViT-L/14 encoder and indexed in a FAISS flat index. Near-duplicate pairs (cosine similarity > 0.97) are clustered and deduplicated, retaining the instance with the highest text-quality signal per cluster. Subsequently, K-means clustering ($K = 25$) over image embeddings identifies visual domain clusters; over-represented domains ($> 15\%$ of pool) are stratified-downsampled to ensure visual balance. This stage reveals that ∼37% of Stage 1 survivors contain near-duplicate visual content across source boundaries, a cross-dataset redundancy undetectable by text-level filtering alone.

Stage 3: Text-only LLM scoring (12,914 → 9,812). Each surviving instance is scored on four criteria via a single structured LLM call operating on text only: **R1** spatial relevance (threshold ≥ 3), **R2** cognitive depth (threshold ≥ 2), **R3** answerability from visual information alone (threshold ≥ 3), and **R4** language clarity (threshold ≥ 3). All four thresholds must be simultaneously satisfied. Operating on text only reduces cost to approximately one-tenth of full multimodal evaluation, while R1 and R3 together eliminate instances with language shortcuts or knowledge dependencies that circumvent genuine visual spatial reasoning.

Stage 4: Multimodal LLM final pass (9,812 → 7,838). Three multimodal criteria are applied to the surviving pool. **R5** (visual necessity) presents the model with question and choices *without* the image: instances answered correctly from text alone are discarded as harbouring language bias (Goyal et al., 2017; Kosoy et al., 2025). **R6** (image-question coherence) verifies that the referenced spatial scene and objects are present in the image. **R7** (spatial signal quality) scores the richness of spatial structure in the image on a 1–4 scale; instances scoring ≤ 2 are removed, and R7 serves as the final ranking signal within taxonomy cells when quota constraints apply.

Benchmark Statistics and Coverage. Table 6 reports the category distribution, determined by the availability of items from the 22 source datasets after filtering and the applied per-category and

per-task coverage thresholds. Thus, the distribution reflects benchmark composition rather than the prevalence of spatial skills in any natural population. Since the overall score is instance-weighted, we also report the unweighted macro-average across seven categories (Table 1). The resulting rankings are highly consistent with the weighted scores (Kendall's $\tau = 0.933$), with at most a two-position shift and the top four and bottom two models unchanged. See §B for details.

3.2 Automated Evaluation Harness

To enable reproducible, large-scale scoring across different models and thousands of spatial reasoning tasks, we developed HarnessEvaluator, a three-stage automated pipeline for structured signal extraction, error taxonomy induction, and cross-model statistical synthesis. At first, each model response is passed to an LLM judge under a zero-shot structured prompt that produces four binary scores per instance: *accuracy*, *spatial elements*, *spatial relations*, and *spatial reasoning*. The six-dimensional decomposition is intentional: overall accuracy conflates distinct competencies, whereas the sub-scores allow failures to be localised to the perception–grounding–reasoning pipeline rather than treated as a monolithic binary outcome. More details are available in §D). The pipeline is fully resumable, already-scored rows are detected and skipped on re-runs, making it practical to extend the evaluation to new models or benchmark items without reprocessing the full dataset. All code and instructions required to reproduce the experiments are available via the Code URL provided in the submission form.

3.3 Error Analysis Pipeline

Standard evaluation in visual reasoning benchmarks reports aggregate or category-level accuracy, capturing *what* models get wrong but not *why*. This obscures heterogeneous failure mechanisms and limits diagnostic insight. To address this, we introduce an *Automated Error Analysis Pipeline* for fine-grained, reproducible analysis of spatial reasoning failures.

The pipeline comprises three stages: **(1) structured signal extraction**, **(2) inductive error taxonomy induction**, and **(3) cross-model diagnostic synthesis**, transforming raw predictions into interpretable, statistically grounded insights.

Stage 1: Structured Signal Extraction. Each prediction is mapped to a multi-attribute representation capturing correctness and failure characteristics, including answer validity, and reasoning coherence. This separates qualitatively distinct failure modes (e.g., formatting, reasoning, perceptual errors) and partitions incorrect predictions into disjoint error pools for homogeneous analysis.

Stage 2: Error Taxonomy Induction. We adopt a bottom-up approach, deriving error types directly from failures. Each incorrect instance (question, gold answer, prediction) is embedded into a failure-aware semantic space capturing *patterns of mistakes*. Density-based clustering (HDBSCAN) identifies recurring failure modes, which are automatically labeled via an LLM using representative examples. These clusters are then mapped to a fixed reference taxonomy (e.g., perceptual, grounding, transformation, multi-step reasoning) for consistent cross-model comparison.

Stage 3: Cross-Model Diagnostic Synthesis. Error labels are aggregated into per-model *error profiles* and *skill vectors*, capturing both failure patterns and spatial competencies. We apply paired tests (McNemar), bootstrap confidence intervals, and effect size estimation with multiple-comparison correction for rigorous comparison. An LLM-based synthesis module generates structured, evidence-backed diagnostic reports.

This pipeline yields insights beyond standard evaluation: error types transcend benchmark categories, models with similar accuracy exhibit distinct error profiles, and key bottlenecks (e.g., spatial transformation, quantitative reasoning) persist across models with large human–model gaps. The induced taxonomy is also stable across models, indicating it captures intrinsic structure in spatial reasoning. Further details are available in §E. The code and instructions for reproducing the experiments are provided via the Code URL in the submission form.

4 Experiments and Evaluation

Models. We categorize the 16 evaluated models into two groups. **Proprietary models** include Claude 4.5 Haiku (Anthropic, 2025), Gemini-2.5-Flash (Team, 2025), Gemini-3.1-Flash-Lite (Google DeepMind, 2025), GPT-4o-mini (OpenAI, 2024), GPT-5-mini (OpenAI, 2026), and o4-mini (OpenAI, 2025). **Open-source models** comprise Gemma-3-4B-it (Gemma Team, Google DeepMind, 2025), Qwen3-VL-8B-Instruct (Bai et al., 2025), Qwen3.5-9B (Qwen Team, 2026a), Llama-3.2-11B-Vision-Instruct (Grattafiori et al., 2024), Gemma-4-31B-it (Gemma Team, Google DeepMind, 2026), Aya-Vision-32B (Dash et al., 2025), Qwen3.6-35B-A3B (Qwen Team, 2026b), Kimi K2.5 (Team, 2026),

Table 1: Open-ended Evaluation Accuracy (%)(↑) across spatial reasoning Categories.

Model	Phys. (#1312)	Pos.Geo. (#2596)	Knowledge (#2332)	Re.Lo.Ca. (#455)	Temporal (#218)	Quant. (#503)	Percept. (#422)	Overall (#7838)
Proprietary Models								
Claude 4.5 Haiku	31.48	29.75	49.23	47.91	53.21	17.30	42.89	37.45
Gemini-2.5-Flash	35.90	35.68	40.93	49.45	41.74	11.55	45.02	36.99
Gemini-3.1-Flash-Lite	**46.49**	**42.22**	**52.62**	49.89	50.92	15.51	**49.76**	**45.20**
GPT-4o-mini	26.37	24.10	23.28	42.86	23.39	13.52	41.23	25.42
GPT-5-mini	38.87	35.17	52.60	**55.60**	**54.13**	**16.10**	46.92	41.95
o4-mini	37.50	35.09	47.83	55.82	55.50	16.10	45.50	40.32
Open-Source Models								
Gemma-3-4B-it	23.70	23.04	11.54	32.53	10.09	8.15	36.26	19.66
Qwen3-VL-8B-Instruct	34.53	34.78	31.40	44.62	45.41	15.11	40.28	33.45
Qwen3.5-9B	23.34	23.42	21.83	31.87	35.32	7.75	36.73	23.10
Llama-3.2-11B-Vision-Ins.	23.78	24.85	17.58	37.14	23.39	8.75	34.60	22.54
Gemma-4-31B-it	**44.59**	33.20	25.56	43.74	48.17	15.31	**49.53**	33.50
Aya-Vision-32B	24.07	21.93	16.90	35.46	26.61	15.31	40.52	22.89
Qwen3.6-35B-A3B	35.31	33.22	39.38	45.81	31.13	11.33	40.67	34.84
Kimi K2.5	33.08	31.97	34.32	39.07	36.70	10.54	42.18	32.45
GLM-4.5V	30.95	**37.30**	**34.06**	**50.77**	**44.50**	**15.31**	47.74	**35.29**
Mistral Small 2603	26.07	24.50	29.42	36.48	50.92	14.31	32.70	27.32
Human Baseline	87.35	89.10	94.17	94.51	91.28	95.23	91.71	91.22

GLM 4.5V-3.06B (Team et al., 2025), and Ministral Small 2603 (Liu et al., 2026a). We also report a **human baseline** evaluation (§F.3) to contextualize model performance with human reasoning ability.

Evaluation Process and Metrics. Models generate free-form responses, which are evaluated using an LLM-based judge under a structured prompting scheme. Full prompt details and evaluation protocols are provided in §F. To validate the reliability of the LLM judge, we conduct an LLM–human agreement analysis (§F), comparing model judgments against human annotations using accuracy and Cohen's κ. Table 10 shows that the judge achieves perfect agreement with the human majority vote ($\kappa = 1.0$, accuracy $= 1.0$), while human annotators exhibit substantial but imperfect agreement (average pairwise $\kappa = 0.81$). These results indicate that the LLM judge reliably reproduces the aggregated human decision boundary, supporting its use for scalable evaluation. Additional experimental details, including generation settings and evaluation procedures, are provided in §F. Prompts are available in Appendix L.

Hard-Case Construction and Agent-Based Evaluation Setup. We construct a hard-case evaluation set consisting of instances that remain unsolved by all evaluated frontier VLMs. For each task category, we sample 25 such cases, resulting in a balanced and adversarial subset that explicitly targets current model failure modes. This design ensures that improvements measured on this set reflect genuine progress on previously intractable problems rather than overlap with standard benchmark capabilities. We evaluate two agentic systems (§F.4): (i) **SpatioXolver** Wasi et al. (2026), a multi-agent framework that decomposes spatial reasoning into specialized roles, and (ii) a **Claude Code-based agent** (Anthropic, 2026), which performs sequential reasoning with programmatic and tool-assisted execution. Both systems receive identical inputs under the same evaluation constraints to ensure fair comparison, with reported default setup. The primary difference lies in their reasoning strategy, with **SpatioXolver** emphasizing parallel decomposition while the Claude-based agent relies on iterative, code-driven reasoning.

5 Findings and Analysis

5.1 Overall Results

Table 1 shows a substantial human–model gap: humans achieve 91.22 overall accuracy, compared with 45.20 for the best model, Gemini-3.1-Flash-Lite. The gap persists across categories, indicating broad rather than isolated weaknesses. Gemini-3.1-Flash-Lite and GPT-5-mini lead among proprietary models, while GLM-4.5V is the strongest open model and surpasses several proprietary systems. *Quantitative & Comparative* reasoning is the clearest bottleneck, remaining below 17 for every model, while *Positional & Geometric* reasoning also remains challenging despite its central role in spatial cognition. Open models show greater variation, with larger models such as Gemma-4-31B-it and Qwen3.6-35B-A3B narrowing the gap in several categories but remaining weaker on knowledge-intensive tasks. Overall, no model performs consistently well across all spatial dimensions, highlighting limited cross-category generalization and persistent weaknesses in visual grounding and multi-step spatial reasoning.

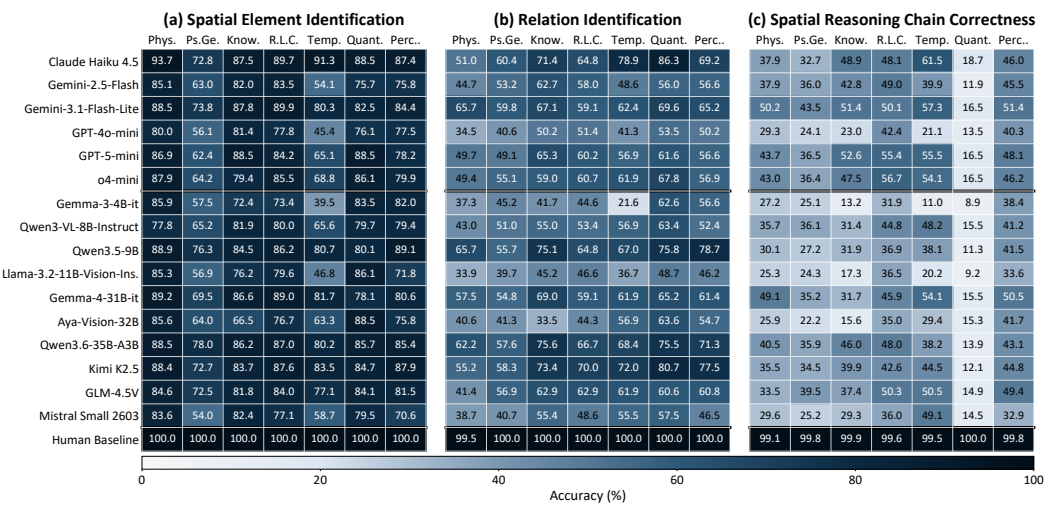


Figure 4: **Category-wise performance across three spatial reasoning steps: element identification, relation identification, and reasoning-chain correctness.** Each cell reports open-ended evaluation accuracy (%), with darker shading indicating higher accuracy. Columns correspond to seven spatial categories, while rows group proprietary and open-source models.

5.2 From Spatial Perception to Multi-Step Reasoning

Figure 4 decomposes spatial reasoning into three stages: *Spatial Element Identification*, *Relation Identification*, and *Spatial Reasoning Chains*, measuring entity recognition, relation determination, and multi-step reasoning, respectively. This decomposition reveals a consistent gap between recognizing spatial information and reasoning over it.

A sharp perception-to-reasoning collapse. Across models and categories, accuracy generally decreases from element identification to relation identification and then to reasoning-chain correctness. The contrast is especially pronounced in *Quantitative* reasoning: models achieve 75.7–88.5% in Step 1 and 48.7–86.3% in Step 2, but only 8.9–18.7% in Step 3. Even Claude Haiku 4.5, the strongest model in this category, falls from 88.5% and 86.3% in the first two stages to 18.7% in the reasoning chain. A similar pattern appears in *Physical* reasoning, where Claude Haiku 4.5 reaches 93.7% in Step 1 but only 37.9% in Step 3. In contrast, human performance remains nearly unchanged across the three stages (100.0%, 99.9%, and 99.7%), indicating that the degradation is primarily associated with compositional reasoning rather than basic visual ambiguity.

The reasoning bottleneck is category dependent. The magnitude of this degradation varies substantially across spatial dimensions. *Relational, Logical & Causal Reasoning* (*R.L.C.*) retains the strongest reasoning-chain performance, with o4-mini, GPT-5-mini, GLM-4.5V, and Gemini-3.1-Flash-Lite reaching 56.7%, 55.4%, 50.3%, and 50.1%, respectively. In contrast, *Positional & Geometric Reasoning* (*Ps.Ge.*) is a major perceptual bottleneck: four models fall below 60% even at Step 1, including Mistral Small 2603 (54.0%), GPT-4o-mini (56.1%), Llama-3.2-11B-Vision (56.9%), and Gemma-3-4B-it (57.5%). *Temporal* reasoning is difficult at both stages, with smaller models showing particularly severe degradation; Gemma-3-4B-it drops from 39.5% in Step 1 to 11.0% in Step 3, whereas Claude Haiku 4.5 maintains 91.3% to 61.5%. Thus, different spatial dimensions impose different bottlenecks: some primarily challenge perception, while others expose failures in multi-step composition.

Reasoning ability, rather than perception alone, separates stronger models. The stage-wise breakdown also reveals capabilities that aggregate accuracy can obscure. Open-weight models can be highly competitive in relation identification: Qwen3.5-9B reaches 75.1% in *Knowledge* and 78.7% in *Perceptual* reasoning, while Kimi K2.5 reaches 80.7% and 77.5% in *Quantitative* and *Perceptual* reasoning, respectively. Yet these gains do not consistently translate into reasoning-chain accuracy. Conversely, reasoning-tuned models show their largest advantage precisely at Step 3. In *Knowledge* reasoning, GPT-4o-mini achieves 23.0% at Step 3, compared with 47.5% for o4-mini and 52.6% for GPT-5-mini; in *Temporal* reasoning, the corresponding scores are 21.1%, 54.1%, and 55.5%. This asymmetric improvement suggests that stronger reasoning capabilities primarily help models compose and manipulate already identified spatial information rather than visual recognition.

Persistent strengths and failure modes. *Perceptual* reasoning is comparatively uniform across models, with Step 1 accuracy ranging from 70.6% to 89.1% and Step 3 accuracy from 32.9% to

51.4%. In contrast, *Knowledge*-grounded reasoning exposes a pronounced gap in multi-step inference: open-weight models such as Gemma-3-4B-it and Llama-3.2-11B fall to 13.2% and 17.3% at Step 3, while GPT-5-mini and Gemini-3.1-Flash-Lite reach 52.6% and 51.4Ṫaken together, these results show that current VLMs are substantially better at *identifying* spatial information than at *composing* it into a correct reasoning chain. The stage-wise decomposition therefore exposes failure patterns that a single end-to-end accuracy score would conceal. More detailed breakdowns are provided in § G

5.3 Error Analysis

To localize failures beyond aggregate accuracy, we analyze model errors using our automated pipeline (§3.2), which assigns each error to the earliest verifiable failure stage in a two-level taxonomy of 7 categories and 38 subcategories (§I). It separates proximate causes from downstream effects.

A clear scaling pattern emerges: smaller models are dominated by *perceptual* (E1) and *spatial grounding* (E2) errors, whereas larger models shift toward *spatial transformation* (E3) and *multi-step reasoning* (E5). Thus, improved perception and grounding reduce early-stage failures, but compositional reasoning remains a persistent bottleneck. *Format and refusal* errors (E7), however, persist across model scales, indicating failures that are not directly tied to spatial capability.

E1 and E3 provide the clearest evidence of these limitations. Perceptual errors, including object recognition and scene understanding, remain across scales, while E3 errors in 2D folding and 3D rotation remain substantial even for larger models such as GPT-5-MINI and GEMINI-2.5-FLASH. In particular, 3D rotation errors occur across nearly all models, suggesting that stronger scaling alone does not reliably produce spatial transformation ability or viewpoint-consistent reasoning.

Three broader patterns emerge. First, early perception and grounding errors propagate into downstream reasoning failures (E5, E6). Second, E7 errors remain common on constrained-output tasks, revealing a gap between general instruction following and task-specific output requirements. Third, *language–spatial mapping* errors (E6) are concentrated in egocentric settings, pointing to persistent difficulties with reference-frame alignment. These patterns motivate 3D-aware supervision for spatial transformations, stronger calibration for constrained outputs, and explicit training for perspective and reference-frame alignment. Full taxonomy and per-model analyses are provided in §**??**.

5.4 Agent-Based Hard Case Evaluation

Agent-based decomposition recovers some instances unsolved by standalone VLMs, but gains vary by reasoning type. The Claude Code Agent performs better on knowledge-grounded and quantitative tasks, indicating stronger symbolic and retrieval capabilities. In contrast, SPATIOXOLVER Wasi et al. (2026) shows clear advantages in relational and temporal reasoning, likely due to step-wise multi-agent decomposition. Both systems perform identically on perceptual tasks, suggesting limited impact on low-level visual interpretation, and both fail on physical and geometric reasoning, exposing persistent gaps. Overall, SPATIOXOLVER achieves higher aggregate performance, while complementary strengths indicate potential for hybrid designs.

Table 2: Agent-based performance on hard-case evaluation (25 hard tasks per category).

Category	**Claude Code Agent**	**SPATIOXOLVER**	**Gap** (Claude Code-SPATIOXOLVER)
Knowledge-Grounded Reasoning	16.00%	12.00%	-4.00
Perceptual Understanding	20.00%	20.00%	0.00
Physical Reasoning	0.00%	0.00%	0.00
Positional & Geometric Reasoning	0.00%	0.00%	0.00
Quantitative & Comparative Reasoning	8.00%	0.00%	-8.00
Relational, Logical & Causal Reasoning	12.00%	28.00%	+16.00
Temporal & Dynamic Reasoning	4.00%	24.00%	+20.00
Overall	8.57%	12.00%	+3.43

6 Concluding Remarks

In this work, we introduced FRONTIERSPATIAL, a large-scale benchmark for holistic spatial reasoning in multimodal models. By unifying 23 existing datasets into 7,838 question–image pairs and organizing them into 7 categories, 19 sub-categories, and 65 tasks, FRONTIERSPATIAL provides a unified view of spatial reasoning beyond narrow, task-specific evaluations. We also developed automated evaluation and error-analysis tools for reproducible benchmarking and fine-grained diagnosis. Our results reveal large and uneven gaps across spatial skills, especially in compositional, causal, and 3D perspective reasoning, which are often missed by existing benchmarks. Overall, FRONTIERSPATIAL provides a practical foundation for more systematic evaluation and development of multimodal spatial reasoning.

AI USE STATEMENT

We used generative AI tools for several research and writing tasks. An LLM-as-a-judge was used as part of the evaluation pipeline, with its reliability assessed through human–AI agreement analysis before deployment for verification. We also used generative AI tools for proofreading, improving manuscript readability, debugging code, and formatting tables and figures. After preparing the initial draft, we used Stanford Agentic Reviewer, developed by the Stanford ML Group, to obtain an initial review of the manuscript, identify potential issues in the methodology, experiments, and presentation, and suggest related work for further consideration. All AI-assisted outputs were reviewed and verified by the authors. We take full responsibility for the final content of this work, including its claims, analyses, code, and released artifacts.

ETHICS STATEMENT

We have taken multiple steps to ensure ethical compliance throughout the development of this work. All images used in this benchmark are sourced from previously published or publicly available datasets, ensuring adherence to established usage and licensing norms. Where applicable, associated metadata was removed to protect privacy, and personally identifiable features (e.g., human faces or license plates) were minimized or excluded. The dataset was carefully curated with attention to copyright and licensing considerations, though minor oversights may occur despite best efforts. The benchmark construction process emphasizes transparency and fairness, with care taken to avoid disproportionate or harmful representation of any demographic group. We adhere to the ICLR Code of Ethics and confirm that no human subjects or sensitive personal data were collected or introduced as part of this work.

REPRODUCIBILITY STATEMENT

All relevant implementation details, including hyperparameters, prompts, temperature settings, token limits, and evaluation protocols, are documented in the appropriate sections of the paper. These details are sufficient to reproduce our reported results on the benchmark. Upon acceptance, we will release the complete codebase with full data and evaluation scripts.

REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning. In *NeurIPS*, 2022.

Anthropic. Claude haiku 4.5. `https://www.anthropic.com/news/claude-haiku-4-5`, October 2025.

Anthropic. Claude code, 2026. URL `https://www.anthropic.com/product/claude-code`. Accessed: 2026-05-05.

Daichi Azuma, Taiki Miyanishi, Shuhei Kurita, and Motoaki Kawanabe. Scanqa: 3d question answering for spatial scene understanding. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 19129–19139, 2022.

Alan Baddeley. Working memory. *Comptes Rendus de l'Académie des Sciences-Series III-Sciences de la Vie*, 321(2-3):167–173, 1998.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xiong-Hui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Rongyao Fang, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Qidong Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Li Ying Meng, Xuancheng Ren, Xin yi Ren, Sibo Song, Yu chen Sun, Jun Tang, Jianhong Tu, Jianqiang Wan,

Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yihe Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Botao Zheng, Humen Zhong, Jingren Zhou, Fanxi Zhou, Jingren Zhou, Yuanzhi Zhu, and Keming Zhu. Qwen3-vl technical report. *ArXiv*, abs/2511.21631, 2025. URL https://arxiv.org/abs/2511.21631.

Xuewei Bai, Yongcai Wang, Deying Li, Haodi Ping, and LI Chunxu. STAR: Spatial-temporal tracklet matching for multi-object tracking. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026. URL https://openreview.net/forum?id=fmCnNQjZrr.

Yoav Bar-Anan, Nira Liberman, and Yaacov Trope. The association between psychological distance and construal level: evidence from an implicit association test. *Journal of experimental psychology: General*, 135(4):609, 2006.

Subrata Biswas, Mohammad Nur Hossain Khan, and Bashima Islam. OWL : Geometry-aware spatial reasoning for audio large language models. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=zPv46YKv3w.

Huanqia Cai, Yijun Yang, and Winston Hu. Mm-iq: Benchmarking human-like abstraction and reasoning in multimodal models, 2025a. URL https://arxiv.org/abs/2502.00698.

Wenxiao Cai, Yaroslav Ponomarenko, Jianhao Yuan, Xiaoqi Li, Wankou Yang, Hao Dong, and Bo Zhao. Spatialbot: Precise spatial understanding with vision language models. *CoRR*, abs/2406.13642, 2024.

Zhongang Cai, Yubo Wang, Qingping Sun, Ruisi Wang, Chenyang Gu, Wanqi Yin, Zhiqian Lin, Zhitao Yang, Chen Wei, Xuanke Shi, Kewang Deng, Xiaoyang Han, Zukai Chen, Jiaqi Li, Xiangyu Fan, Hanming Deng, Lewei Lu, Bo Li, Ziwei Liu, Quan Wang, Dahua Lin, and Lei Yang. Has gpt-5 achieved spatial intelligence? an empirical study, 2025b. URL https://arxiv.org/abs/2508.13142.

Zidong Cao, Jinjing Zhu, Hao Ai, Lutao Jiang, Yuanhuiyi Lyu, and Hui Xiong. ST$ˆ2$360d: Spatial-to-temporal consistency for training-free 360 monocular depth estimation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026. URL https://openreview.net/forum?id=d0bm4upnQ8.

Christopher F Chabris, Thomas E Jerde, Anita W Woolley, Margaret E Gerbasi, Jonathon P Schuldt, Sean L Bennett, J Richard Hackman, and Stephen M Kosslyn. Spatial and object visualization cognitive styles: Validation studies in 3800 individuals. *Group brain technical report*, 2(1-20):2, 2006.

Eun Chang, Zhuangqun Huang, Yiwei Liao, Sagar Ravi Bhavsar, Amogh Param, Tammy Stark, Adel Ahmadyan, Xiao Yang, Jiaqi Wang, Ahsan Abdullah, Giang Nguyen, Akil Iyer, David Hall, Elissa Li, Shane Moon, Nicolas Scheffer, Kirmani Ahmed, Babak Damavandi, Rakesh Wanga, Anuj Kumar, Rohit Patel, and Xin Luna Dong. Wearvqa: A visual question answering benchmark for wearables in egocentric authentic real-world scenarios, 2025. URL https://arxiv.org/abs/2511.22154.

Korawat Charoenpitaks, Van-Quang Nguyen, Masanori Suganuma, Kentaro Arai, Seiji Totsuka, Hiroshi Ino, and Takayuki Okatani. Tb-bench: Training and testing multi-modal ai for understanding spatio-temporal traffic behaviors from dashcam images/videos, 2025. URL https://arxiv.org/abs/2501.05733.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brian Ichter, Dorsa Sadigh, Leonidas J. Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 14455–14465. IEEE, 2024.

Liang Chen, Weichu Xie, Yiyan Liang, Hongfeng He, Hans Zhao, Zhibo Yang, Zhiqi Huang, Haoning Wu, Haoyu Lu, Y. charles, Yiping Bao, Yuantao Fan, Guopeng Li, Haiyang Shen, Xuanzhong Chen, Wendong Xu, Shuzheng Si, Zefan Cai, Wenhao Chai, Ziqi Huang, Fangfu Liu, Tianyu Liu, Baobao Chang, Xiaobo Hu, Kaiyuan Chen, Yixin Ren, Yang Liu, Yuan Gong, and Kuan Li. Babyvision: Visual reasoning beyond language, 2026a. URL https://arxiv.org/pdf/2601.06521.

Mingfei Chen, Zijun Cui, Xiulong Liu, Jinlin Xiang, Caleb Zheng, Jingyuan Li, and Eli Shlizerman. SAVVY: Spatial awareness via audio-visual LLMs through seeing and hearing. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026b. URL https://openreview.net/forum?id=zwCb9cKHpd.

Pingyi Chen, Yujing Lou, Shen Cao, Jinhui Guo, Lubin Fan, Yue Wu, Lin Yang, Lizhuang Ma, and Jieping Ye. SD-VLM: Spatial measuring and understanding with depth-encoded vision-language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026c. URL https://openreview.net/forum?id=PRCizVyL1K.

Shiqi Chen, Tongyao Zhu, Ruochen Zhou, Jinghan Zhang, Siyang Gao, Juan Carlos Niebles, Mor Geva, Junxian He, Jiajun Wu, and Manling Li. Why is spatial reasoning hard for vlms? an attention mechanism perspective on focus areas. *CoRR*, abs/2503.01773, 2025.

Shunpeng Chen, Changwei Wang, Rongtao Xu, Peixingtian, yukun Song, Jinzhou Lin, Wenhao Xu, jingyizhang, Li Guo, and Shibiao Xu. SAGE: Spatial-visual adaptive graph exploration for efficient visual place recognition. In *The Fourteenth International Conference on Learning Representations*, 2026d. URL https://openreview.net/forum?id=DCpbEXqPvS.

An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. Spatialrgpt: Grounded spatial reasoning in vision-language models. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024a.

An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. Spatialrgpt: Grounded spatial reasoning in vision language models, 2024b. URL https://arxiv.org/abs/2406.01584.

Xianfu Cheng, Wei Zhang, Shiwei Zhang, Jian Yang, Xiangyuan Guan, Xianjie Wu, Xiang Li, Ge Zhang, Jiaheng Liu, Yuying Mai, Yutao Zeng, Zhoufutu Wen, Ke Jin, Baorui Wang, Weixiao Zhou, Yunhong Lu, Tongliang Li, Wenhao Huang, and Zhoujun Li. Simplevqa: Multimodal factuality evaluation for multimodal large language models, 2025. URL https://arxiv.org/abs/2502.13059.

Open X.-Embodiment Collaboration, Abhishek Padalkar, Acorn Pooley, Ajinkya Jain, Alex Bewley, Alexander Herzog, Alex Irpan, Alexander Khazatsky, Anant Raj, Anikait Singh, Anthony Brohan, Antonin Raffin, Ayzaan Wahid, Ben Burgess-Limerick, Beomjoon Kim, Bernhard Schölkopf, Brian Ichter, Cewu Lu, Charles Xu, Chelsea Finn, Chenfeng Xu, Cheng Chi, Chenguang Huang, Christine Chan, Chuer Pan, Chuyuan Fu, Coline Devin, Danny Driess, Deepak Pathak, Dhruv Shah, Dieter Büchler, Dmitry Kalashnikov, Dorsa Sadigh, Edward Johns, Federico Ceola, Fei Xia, Freek Stulp, Gaoyue Zhou, Gaurav S. Sukhatme, Gautam Salhotra, Ge Yan, Giulio Schiavi, Gregory Kahn, Hao Su, Haoshu Fang, Haochen Shi, Heni Ben Amor, Henrik I. Christensen, Hiroki Furuta, Homer Walke, Hongjie Fang, Igor Mordatch, Ilija Radosavovic, and et al. Open x-embodiment: Robotic learning datasets and RT-X models. *CoRR*, abs/2310.08864, 2023.

Muhammad Sohail Danish, Muhammad Akhtar Munir, Syed Roshaan Ali Shah, Kartik Kuckreja, Fahad Shahbaz Khan, Paolo Fraccaro, Alexandre Lacoste, and Salman Khan. Geobench-vlm: Benchmarking vision-language models for geospatial tasks, 2024. URL https://arxiv.org/abs/2411.19325.

Saurabh Dash, Yiyang Nan, John Dang, Arash Ahmadian, Shivalika Singh, Madeline Smith, Bharat Venkitesh, Vladyslav Shmyhlo, Viraat Aryabumi, Walter Beller-Morales, Jeremy Pekmez, Jason Ozuzu, Pierre H. Richemond, Acyr F. Locatelli, Nick Frosst, Phil Blunsom, Aidan Gomez, Ivan Zhang, Marzieh Fadaee, Manoj Govindassamy, Sudip Roy, Matthias Gall'e, Beyza Hilal Ermiş, A. Ustun, and Sara Hooker. Aya vision: Advancing the frontier of multilingual multimodality. *ArXiv*, abs/2505.08751, 2025. URL https://arxiv.org/abs/2505.08751.

Nianchen Deng, Lixin Gu, Shenglong Ye, Yinan He, Zhe Chen, Songze Li, Haomin Wang, Jinhui Yin, Qi Wei, Tianshuo Yang, Min Dou, Tong He, Wenqi Shao, Kaipeng Zhang, Yi Wang, Botian Shi, Yanting Zhang, Jifeng Dai, Yu Qiao, Wenhai Wang, and Hongjie Zhang. Internspatial:

A comprehensive dataset for spatial reasoning in vision-language models. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=L6bEitSMeu.

Mengfei Du, Binhao Wu, Zejun Li, Xuanjing Huang, and Zhongyu Wei. Embspatial-bench: Benchmarking spatial understanding for embodied tasks with large vision-language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, ACL 2024 - Short Papers, Bangkok, Thailand, August 11-16, 2024*, pp. 346–355. Association for Computational Linguistics, 2024.

ZhiYuan Feng, Zhaolu Kang, Qijie Wang, Zhiying Du, Jiongrui Yan, Shi Shubin, Chengbo Yuan, Huizhi Liang, Yu Deng, Qixiu Li, Rushuai Yang, Ruichuan An, Leqi Zheng, Weijie Wang, Shuang Chen, Sicheng Xu, Yaobo Liang, Jiaolong Yang, and Baining Guo. Seeing across views: Benchmarking spatial reasoning of vision-language models in robotic scenes. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=jXDZJAfRZB.

Mingyang Fu, Yuyang Peng, Dongping Chen, Zetong Zhou, Benlin Liu, Yao Wan, Zhou Zhao, Philip S. Yu, and Ranjay Krishna. Seeking and updating with live visual knowledge, 2025. URL https://arxiv.org/abs/2504.05288.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A. Smith, Wei-Chiu Ma, and Ranjay Krishna. BLINK: multimodal large language models can see but not perceive. In Ales Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part XXIII*, volume 15081 of *Lecture Notes in Computer Science*, pp. 148–166. Springer, 2024.

Howard E Gardner. *Frames of mind: The theory of multiple intelligences*. Basic books, 2011.

Gemma Team, Google DeepMind. Gemma 3 technical report, 2025. URL https://storage.googleapis.com/deepmind-media/gemma/Gemma3Report.pdf. Accessed: 2025-09-11.

Gemma Team, Google DeepMind. Gemma 4 model card, 2026. URL https://ai.google.dev/gemma/docs/core/model_card_4. Accessed: 2026-05-05.

Mohsen Gholami, Ahmad Rezaei, Zhou Weimin, Sitong Mao, Shunbo Zhou, Yong Zhang, and Mohammad Akbari. Spatial reasoning with vision-language models in ego-centric multi-view scenes. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=fqehqG4WvL.

Ziyang Gong, Wenhao Li, Xianzheng Ma, Songyuan Li, Zhaokai Wang, Songze Li, Jiayi Ji, Xue Yang, Gen Luo, Junchi Yan, and Rongrong Ji. SpaCE-10: A comprehensive benchmark for multimodal large language models in compositional spatial intelligence. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=Df7UjwEgIx.

Google DeepMind. Gemini 3: The Frontier of Vision AI. https://blog.google/innovation-and-ai/technology/developers-tools/gemini-3-pro-vision/, 2025. Accessed: 2026-05-05.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu,

Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng

Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Xin Gu, Yaojie Shen, Chenxi Luo, Tiejian Luo, Yan Huang, Yuewei Lin, Heng Fan, and Libo Zhang. Knowing your target: Target-aware transformer makes better spatio-temporal video grounding. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=WOzffPgVjF.

Minghao Han, Dingkang Yang, Linhao Qu, Zizhi Chen, Gang Li, Han Wang, Jiacong Wang, and Lihua Zhang. Fusing pixels and genes: Spatially-aware learning in computational pathology. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=uVXO6gzVzj.

Jinkun Hao, Naifu Liang, Zhen Luo, Xudong XU, Weipeng Zhong, Ran Yi, Yichen Jin, Zhaoyang Lyu, Feng Zheng, Lizhuang Ma, and Jiangmiao Pang. Mesatask: Towards task-driven tabletop scene generation via 3d spatial reasoning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026. URL https://openreview.net/forum?id=U88JlpY0vR.

Md Tanzib Hosain, Salman Rahman, Md Kishor Morol, and Md Rizwan Parvez. Xolver: Multi-agent reasoning with holistic experience learning just like an olympiad team, 2025. URL https://arxiv.org/abs/2506.14234.

Wenbo Hu, Yining Hong, Yanjun Wang, Leison Gao, Zibu Wei, Xingcheng Yao, Nanyun Peng, Yonatan Bitton, Idan Szpektor, and Kai-Wei Chang. 3DLLM-mem: Long-term spatial-temporal memory for embodied 3d large language model. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026. URL https://openreview.net/forum?id=q5QaTQcUbS.

Yuan Hu, Jianlong Yuan, Congcong Wen, Xiaonan Lu, and Xiang Li. Rsgpt: A remote sensing vision language model and benchmark, 2023. URL https://arxiv.org/abs/2307.15266.

Wenlong Huang, Chen Wang, Yunzhu Li, Ruohan Zhang, and Li Fei-Fei. Rekep: Spatio-temporal reasoning of relational keypoint constraints for robotic manipulation. In *CoRL*, 2024.

Xinmiao Huang, Qisong He, Zhenglin Huang, Boxuan Wang, Zhuoyun Li, Guangliang Cheng, Yi Dong, and Xiaowei Huang. Spatial-dise: A unified benchmark for evaluating spatial reasoning in vision-language models, 2025. URL https://arxiv.org/abs/2510.13394.

Xinmiao Huang, Qisong He, Zhenglin Huang, Boxuan Wang, Zhuoyun Li, Guangliang Cheng, Yi Dong, and Xiaowei Huang. Spatial-DISE: A unified benchmark for evaluating spatial reasoning in vision-language models. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=bMINsPQpME.

Drew A. Hudson and Christopher D. Manning. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 6700–6709. Computer Vision Foundation / IEEE, 2019.

Mengdi Jia, Zekun Qi, Shaochen Zhang, Wenyao Zhang, Xinqiang Yu, Jiawei He, He Wang, and Li Yi. Omnispatial: Towards comprehensive spatial reasoning benchmark for vision language models, 2025. URL https://arxiv.org/abs/2506.03135.

Mengdi Jia, Zekun Qi, Shaochen Zhang, Wenyao Zhang, XinQiang Yu, Jiawei He, He Wang, and Li Yi. Omnispatial: Towards comprehensive spatial reasoning benchmark for vision language models. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=6nZKT2rL0H.

Yifan Jiang, Jiarui Zhang, Kexuan Sun, Zhivar Sourati, Kian Ahrabian, Kaixin Ma, Filip Ilievski, and Jay Pujara. Marvel: Multidimensional abstraction and reasoning through visual evaluation and learning, 2024. URL https://arxiv.org/abs/2404.13591.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning, 2016. URL https://arxiv.org/abs/1612.06890.

Amita Kamath, Jack Hessel, and Kai-Wei Chang. What's "up" with vision-language models? investigating their struggle with spatial reasoning. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pp. 9161–9175. Association for Computational Linguistics, 2023.

Sayash Kapoor, Benedikt Stroebl, Peter Kirgis, Nitya Nadgir, Zachary S Siegel, Boyi Wei, Tianci Xue, Ziru Chen, Felix Chen, Saiteja Utpala, Franck Ndzomga, Dheeraj Oruganty, Sophie Luskin, Kangheng Liu, Botao Yu, Amit Arora, Dongyoon Hahm, Harsh Trivedi, Huan Sun, Juyong Lee, Tengjun Jin, Yifan Mai, Yifei Zhou, Yuxuan Zhu, Rishi Bommasani, Daniel Kang, Dawn Song, Peter Henderson, Yu Su, Percy Liang, and Arvind Narayanan. Holistic agent leaderboard: The missing infrastructure for AI agent evaluation. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=vUaY1t64ZZ.

Arina Kharlamova, Bowei He, Chen Ma, and Xue Liu. Spatial CAPTCHA: Generatively benchmarking spatial reasoning for human-machine differentiation. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=8LAc5kzjhz.

Eliza Kosoy, Annya Dahmani, Andrew K. Lampinen, Iulia M. Comsa, Soojin Jeong, Ishita Dasgupta, and Kelsey Allen. Decoupling the components of geometric understanding in vision language models. *CoRR*, abs/2503.03840, 2025.

Trung Le, Hao Fang, Jingyuan Li, Tung Nguyen, Lu Mi, Amy L Orsborn, Uygar Sümbül, and Eli Shlizerman. SPINT: Spatial permutation-invariant neural transformer for consistent intracortical motor decoding. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026. URL https://openreview.net/forum?id=XjayhGBraW.

Insu Lee, Wooje Park, Jaeyun Jang, Minyoung Noh, Kyuhong Shim, and Byonghyo Shim. Towards comprehensive scene understanding: Integrating first and third-person views for LVLMs. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026a. URL https://openreview.net/forum?id=m5wrqqcWbN.

JoungBin Lee, Jaewoo Jung, Jisang Han, Takuya Narihira, Kazumi Fukuda, Junyoung Seo, Sunghwan Hong, Yuki Mitsufuji, and Seungryong Kim. 3d scene prompting for scene-consistent camera-controllable video generation. In *The Fourteenth International Conference on Learning Representations*, 2026b. URL https://openreview.net/forum?id=3XxoBwMusJ.

Fuhao Li, Wenxuan Song, Han Zhao, Jingbo Wang, Pengxiang Ding, Donglin Wang, Long ZENG, and Haoang Li. Spatial forcing: Implicit spatial representation alignment for vision-language-action model. In *The Fourteenth International Conference on Learning Representations*, 2026a. URL https://openreview.net/forum?id=euMVC1DO4k.

Hongxing Li, Dingming Li, Zixuan Wang, Yuchen Yan, Hang Wu, Wenqi Zhang, Yongliang Shen, Weiming Lu, Jun Xiao, and Yueting Zhuang. Spatialladder: Progressive training for spatial reasoning in vision-language models. In *The Fourteenth International Conference on Learning Representations*, 2026b. URL https://openreview.net/forum?id=KtrFXlvgrK.

Jinpeng Li, Haiping Wang, Jiabin chen, Yuan Liu, Zhen Dong, and Bisheng Yang. SCot: Teaching 3d-LLMs to think spatially with million-scale cot annotations. In *The Fourteenth International Conference on Learning Representations*, 2026c. URL https://openreview.net/forum?id=5Tph6wFMOm.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. *CoRR*, abs/2301.12597, 2023.

Linjie Li, Mahtab Bigverdi, Jiawei Gu, Zixian Ma, Yinuo Yang, Ziang Li, Yejin Choi, and Ranjay Krishna. Unfolding spatial cognition: Evaluating multimodal models on visual simulations. In *The Fourteenth International Conference on Learning Representations*, 2026d. URL https://openreview.net/forum?id=fbGmSV6tUw.

Pengteng Li, Pinhao Song, Wuyang Li, Huizai Yao, Weiyu Guo, Yijie Xu, Dugang Liu, and Hui Xiong. See&trek: Training-free spatial prompting for multimodal large language model. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026e. URL https://openreview.net/forum?id=2exr4mlbx1.

Tiantian Li, Xingxing Cao, Yifei Wang, Xiaojiao Yang, Xiaolong Zou, and Bo Hong. Hippoformer: Integrating hippocampus-inspired spatial memory with transformers. In *The Fourteenth International Conference on Learning Representations*, 2026f. URL https://openreview.net/forum?id=hxwV5EubAw.

Wuyang Li, Zhu Yu, and Alexandre Alahi. Voxdet: Rethinking 3d semantic scene completion as dense object detection. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026g. URL https://openreview.net/forum?id=lMhNrt0Bnm.

Yizhi Li, Yinghao Ma, Ge Zhang, Ruibin Yuan, Kang Zhu, Hangyu Guo, Yiming Liang, Jiaheng Liu, Zekun Wang, Jian Yang, Siwei Wu, Xingwei Qu, Jinjie Shi, Xinyue Zhang, Zhenzhu Yang, Yidan Wen, Yanghai Wang, Shihao Li, Zhaoxiang Zhang, Zachary Liu, Emmanouil Benetos, Wenhao Huang, and Chenghua Lin. Omnibench: Towards the future of universal omni-language models, 2024. URL https://arxiv.org/abs/2409.15272.

Alexander Liu, Kartik Khandelwal, Sandeep Subramanian, Victor Jouault, Abhinav Rastogi, Adrien Sad'e, Alan Jeffares, Albert Q. Jiang, Alexandre Cahill, Alexandre Gavaudan, Alexandre Sablayrolles, Am'elie H'eliou, Amos You, Andy Ehrenberg, Andy Desman Lo, Anton Eliseev, Antonia Calvi, Avinash Sooriyarachchi, Baptiste Bout, Baptiste Rozière, Baudouin De Monicault, Clémence Lanfranchi, Corentin Barreau, Cyprien Courtot, Daniele Grattarola, Darius Dabert, Diego de Las Casas, Elliot Chane-Sane, Faruk Ahmed, Gabrielle Berrada, Gaetan Ecrepont, Gauthier Guinet, Georgii Sergeevich Novikov, Guillaume Kunsch, Guillaume Lample, Guillaume Martin, Gunshi Gupta, Jan Ludziejewski, Jason Rute, Joachim Studnia, Jonas Amar, Joséphine Delas, Josselin Somerville Roberts, Karmesh Yadav, Khyathi Raghavi Chandu, Kush Jain, Laurence Aitchison, Laurent Fainsin, Léonard Blier, Lingxiao Zhao, Louis Martin, Lucile Saulnier, Luyu Gao, Maarten Buyl, Margaret Jennings, Marie Pellat, Mark Prins, Mathieu Poir'ee, Mathilde Guillaumin, Matthieu Dinot, Matthieu Futeral, Maxime Darrin, Maximilian Augustin, Mia Chiquier, Michel Schimpf, Nathan Grinsztajn, Neha Gupta, Nikhil Raghuraman, Olivier Bousquet, Olivier Duchenne, Patricia Wang, Patrick von Platen, Paul Jacob, Paul Wambergue, Paula Kurylowicz, Pavankumar Reddy Muddireddy, Philomène Chagniot, Pierre Stock, Pravesh Agrawal, Quentin Torroba, Romain Sauvestre, Roman Soletskyi, Rupert Menneer, Sagar Vaze, Samuel Barry, Sanchit Gandhi, Siddhant Waghjale, Siddharth Gandhi, Soham Ghosh, Srijan Mishra, Sumukh Aithal, Szymon Antoniak, Teven Le Scao, Théo Cachet, Theo Simon Sorg, Thibaut Lavril, Thiziri Nait Saada, Thomas Chabal, Thomas Foubert, Thomas Robert, Thomas Wang, Tim Lawson, Tom

Bewley, Tom Edwards, Umar Jamil, Umberto M. Tomasini, Valeriia Nemychnikova, Van Phung, Vincent Maladiere, Virgile Richard, Wassim Bouaziz, Wen-Ding Li, William Marshall, Xinghui Li, Xinyu Yang, Yassine El Ouahidi, Yihan Wang, Yunhao Tang, and Zaccharie Ramzi. Ministral 3. *ArXiv*, abs/2601.08584, 2026a. URL https://arxiv.org/abs/2601.08584.

Fangyu Liu, Guy Emerson, and Nigel Collier. Visual spatial reasoning. *Trans. Assoc. Comput. Linguistics*, 11:635–651, 2023.

Hongjun Liu, Leyu Zhou, Zijianghao Yang, and Chao Yao. Step-aware residual-guided diffusion for EEG spatial super-resolution. In *The Fourteenth International Conference on Learning Representations*, 2026b. URL https://openreview.net/forum?id=2vN2GlfV6T.

Siqi Liu, Ian Gemp, Luke Marris, Georgios Piliouras, Nicolas Heess, and Marc Lanctot. Re-evaluating open-ended evaluation of large language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=kbOAIXKWgx.

Yang Liu, Ming Ma, Xiaomin Yu, Pengxiang Ding, Han Zhao, Mingyang Sun, Siteng Huang, and Donglin Wang. SSR: Enhancing depth perception in vision-language models via rationale-guided spatial reasoning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026c. URL https://openreview.net/forum?id=SOc0tHCewe.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering, 2022. URL https://arxiv.org/abs/2209.09513.

Zhanpeng Luo, Ce Zhang, Silong Yong, Cunxi Dai, Qianwei Wang, Haoxi Ran, Guanya Shi, Katia P. Sycara, and Yaqi Xie. pyspatial: Generating 3d visual programs for zero-shot spatial reasoning. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=yv15C8ql24.

Wenxin Ma, Chenlong Wang, Ruisheng Yuan, Hao Chen, Nanru Dai, S. Kevin Zhou, Yijun Yang, Alan Yuille, and Jieneng Chen. Causalspatial: A benchmark for object-centric causal spatial reasoning, 2026a. URL https://arxiv.org/abs/2601.13304.

Wufei Ma, Yu-Cheng Chou, Qihao Liu, Xingrui Wang, Celso M de Melo, Jianwen Xie, and Alan Yuille. Spatialreasoner: Towards explicit and generalizable 3d spatial reasoning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026b. URL https://openreview.net/forum?id=hFaXVjRFHI.

Xianzheng Ma, TAO SUN, Shuai Chen, Yash Sanjay Bhalgat, Jindong Gu, Angel X Chang, Iro Armeni, Iro Laina, Songyou Peng, and Victor Adrian Prisacariu. Do 3d large language models really understand 3d spatial relationships? In *The Fourteenth International Conference on Learning Representations*, 2026c. URL https://openreview.net/forum?id=3vlMiJwo8b.

Yuezhou Ma, Haixu Wu, Hang Zhou, Huikun Weng, Jianmin Wang, and Mingsheng Long. Physense: Sensor placement optimization for accurate physics sensing. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026d. URL https://openreview.net/forum?id=zIzZxDsNNP.

Yongsen Mao, Junhao Zhong, Chuan Fang, Jia Zheng, Rui Tang, Hao Zhu, Ping Tan, and Zihan Zhou. SpatialLM: Training large language models for structured indoor modeling. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026. URL https://openreview.net/forum?id=u1lNQH5upa.

M. Maruf, Arka Daw, Kazi Sajeed Mehrab, Harish Babu Manogaran, Abhilash Neog, Medha Sawhney, Mridul Khurana, James P. Balhoff, Yasin Bakis, Bahadir Altintas, Matthew J. Thompson, Elizabeth G. Campolongo, Josef C. Uyeda, Hilmar Lapp, Henry L. Bart, Paula M. Mabee, Yu Su, Wei-Lun Chao, Charles Stewart, Tanya Berger-Wolf, Wasila Dahdul, and Anuj Karpatne. Vlm4bio: A benchmark dataset to evaluate pretrained vision-language models for trait discovery from biological images, 2024. URL https://arxiv.org/abs/2408.16176.

OpenAI. Gpt-4v(ision) system card, 2023. URL https://openai.com/research/gpt-4v-system-card.

OpenAI. Gpt-4o system card, 2024. URL https://arxiv.org/abs/2410.21276.

OpenAI. Openai o3 and o4-mini system card, April 2025. URL https://openai.com/index/o3-o4-mini-system-card/. System Cards. Accessed: 2025-09-11.

OpenAI. Openai gpt-5 system card, 2026. URL https://arxiv.org/abs/2601.03267.

Stephen Palmer. *Vision Science: From Photons to Phenomenology*, volume 1. MIT Press, 01 1999.

Zhenyu Pan and Han Liu. Metaspatial: Reinforcing 3d spatial reasoning in VLMs for the metaverse. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=EdQzLC0Zra.

Atin Pothiraj, Elias Stengel-Eskin, Jaemin Cho, and Mohit Bansal. Capture: Evaluating spatial reasoning in vision language models via occluded object counting. *CoRR*, abs/2504.15485, 2025.

Fred H Previc. The neuropsychology of 3-d space. *Psychological bulletin*, 124(2):123, 1998.

Zekun Qi, Runpei Dong, Shaochen Zhang, Haoran Geng, Chunrui Han, Zheng Ge, Li Yi, and Kaisheng Ma. Shapellm: Universal 3d object understanding for embodied interaction. In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part XLIII*, volume 15101 of *Lecture Notes in Computer Science*, pp. 214–238. Springer, 2024.

Zekun Qi, Wenyao Zhang, Yufei Ding, Runpei Dong, Xinqiang Yu, Jingwen Li, Lingyun Xu, Baoyu Li, Xialin He, Guofan Fan, Jiazhao Zhang, Jiawei He, Jiayuan Gu, Xin Jin, Kaisheng Ma, Zhizheng Zhang, He Wang, and Li Yi. Sofar: Language-grounded orientation bridges spatial reasoning and object manipulation. *CoRR*, abs/2502.13143, 2025.

Zekun Qi, Wenyao Zhang, Yufei Ding, Runpei Dong, XinQiang Yu, Jingwen Li, Lingyun Xu, Baoyu Li, Xialin He, Guofan Fan, Jiazhao Zhang, Jiawei He, Jiayuan Gu, Xin Jin, Kaisheng Ma, Zhizheng Zhang, He Wang, and Li Yi. Sofar: Language-grounded orientation bridges spatial reasoning and object manipulation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026. URL https://openreview.net/forum?id=kmv7yg6QXv.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026a. URL https://qwen.ai/blog?id=qwen3.5.

Qwen Team. Qwen3.6-35B-A3B: Agentic coding power, now open to all, April 2026b. URL https://qwen.ai/blog?id=qwen3.6-35b-a3b.

Tiago Ramalho, Tomás Kociský, Frederic Besse, S. M. Ali Eslami, Gábor Melis, Fabio Viola, Phil Blunsom, and Karl Moritz Hermann. Encoding spatial relations from natural language. *CoRR*, abs/1807.01670, 2018.

Jonathan Roberts, Mohammad Reza Taesiri, Ansh Sharma, Akash Gupta, Samuel Roberts, Ioana Croitoru, Simion-Vlad Bogolin, Jialu Tang, Florian Langer, Vyas Raina, Vatsal Raina, Hanyi Xiong, Vishaal Udandarao, Jingyi Lu, Shiyang Chen, Sam Purkis, Tianshuo Yan, Wenye Lin, Gyungin Shin, Qiaochu Yang, Anh Totti Nguyen, David I. Atkinson, Aaditya Baranwal, Alexandru Coca, Mikah Dang, Sebastian Dziadzio, Jakob D. Kunz, Kaiqu Liang, Alexander Lo, Brian Pulfer, Steven Walton, Charig Yang, Kai Han, and Samuel Albanie. Zerobench: An impossible visual benchmark for contemporary large multimodal models, 2025. URL https://arxiv.org/abs/2502.09696.

Javier E. Santos, Agnese Marcato, Roman Colman, Nicholas Lubbers, and Yen Ting Lin. Discrete spatial diffusion: Intensity-preserving diffusion modeling. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026. URL https://openreview.net/forum?id=kz3w2A2y0e.

Marcos Negre Saura, Richard Allmendinger, Wei Pan, and Theodore Papamarkou. Spatial-aware decision-making with ring attractors in reinforcement learning systems. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026. URL https://openreview.net/forum?id=FthMPOhgfp.

Yifan Shen, Yuanzhe Liu, Jingyuan Zhu, Xu Cao, Xiaofeng Zhang, Yixiao He, Wenming Ye, James Matthew Rehg, and Ismini Lourentzou. Fine-grained preference optimization improves spatial reasoning in VLMs. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026a. URL https://openreview.net/forum?id=L9vV3wVC72.

Yixian Shen, Qi Bi, Zihan Wang, Zhiheng Yang, Changshuo Wang, Zhi Zhang, Prayag Tiwari, Andy D. Pimentel, and Anuj Pathania. Efficient multimodal spatial reasoning via dynamic and asymmetric routing. In *The Fourteenth International Conference on Learning Representations*, 2026b. URL https://openreview.net/forum?id=BQASoLmREU.

Fatemeh Shiri, Xiao-Yu Guo, Mona Far, Xin Yu, Reza Haf, and Yuan-Fang Li. An empirical analysis on spatial reasoning capabilities of large multimodal models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pp. 21440–21455. Association for Computational Linguistics, 2024.

Chan Hee Song, Valts Blukis, Jonathan Tremblay, Stephen Tyree, Yu Su, and Stan Birchfield. Robospatial: Teaching spatial understanding to 2d and 3d vision-language models for robotics. *CoRR*, abs/2411.16537, 2024.

Hannah Sterz, Jonas Pfeiffer, and Ivan Vulić. Dare: Diverse visual question answering with robustness evaluation, 2024. URL https://arxiv.org/abs/2409.18023.

Dominik Sturm, Hiba Bensalem, and Ivo F. Sbalzarini. Spatially informed autoencoders for interpretable visual representation learning. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=09YSBymX6O.

Emilia Szymanska, Mihai Dusmanu, Jan-Willem Buurlage, Mahdi Rad, and Marc Pollefeys. Space3d-bench: Spatial 3d question answering benchmark. *CoRR*, abs/2408.16662, 2024.

Kexian Tang, Junyao Gao, Yanhong Zeng, Haodong Duan, Yanan Sun, Zhening Xing, Wenran Liu, Kaifeng Lyu, and Kai Chen. Lego-puzzles: How good are mllms at multi-step spatial reasoning?, 2025. URL https://arxiv.org/abs/2503.19990.

Gemini Team. Gemini: A family of highly capable multimodal models, 2025. URL https://arxiv.org/abs/2312.11805.

Kimi Team. Kimi k2.5: Visual agentic intelligence, 2026. URL https://arxiv.org/abs/2602.02276.

V Team, Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, Shuaiqi Duan, Weihan Wang, Yan Wang, Yean Cheng, Zehai He, Zhe Su, Zhen Yang, Ziyang Pan, Aohan Zeng, Baoxu Wang, Bin Chen, Boyan Shi, Changyu Pang, Chenhui Zhang, Da Yin, Fan Yang, Guoqing Chen, Jiazheng Xu, Jiale Zhu, Jiali Chen, Jing Chen, Jinhao Chen, Jinghao Lin, Jinjiang Wang, Junjie Chen, Leqi Lei, Letian Gong, Leyi Pan, Mingdao Liu, Mingde Xu, Mingzhi Zhang, Qinkai Zheng, Sheng Yang, Shi Zhong, Shiyu Huang, Shuyuan Zhao, Siyan Xue, Shangqin Tu, Shengbiao Meng, Tianshu Zhang, Tianwei Luo, Tianxiang Hao, Tianyu Tong, Wenkai Li, Wei Jia, Xiao Liu, Xiaohan Zhang, Xin Lyu, Xinyue Fan, Xuancheng Huang, Yanling Wang, Yadong Xue, Yanfeng Wang, Yanzi Wang, Yifan An, Yifan Du, Yiming Shi, Yiheng Huang, Yilin Niu, Yuan Wang, Yuanchang Yue, Yuchen Li, Yutao Zhang, Yuting Wang, Yu Wang, Yuxuan Zhang, Zhao Xue, Zhenyu Hou, Zhengxiao Du, Zihan Wang, Peng Zhang, Debing Liu, Bin Xu, Juanzi Li, Minlie Huang, Yuxiao Dong, and Jie Tang. Glm-4.5v and glm-4.1v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning, 2025. URL https://arxiv.org/abs/2507.01006.

Yaacov Trope and Nira Liberman. Construal-level theory of psychological distance. *Psychological review*, 117(2):440, 2010.

Jiaqi Wang, Xiao Yang, Kai Sun, Parth Suresh, Sanat Sharma, Adam Czyzewski, Derek Andersen, Surya Appini, Arkav Banerjee, Sajal Choudhary, Shervin Ghasemlou, Ziqiang Guan, Akil Iyer, Haidar Khan, Lingkun Kong, Roy Luo, Tiffany Ma, Zhen Qiao, David Tran, Wenfang Xu, Skyler Yeatman, Chen Zhou, Gunveer Gujral, Yinglong Xia, Shane Moon, Nicolas Scheffer, Nirav Shah, Eun Chang, Yue Liu, Florian Metze, Tammy Stark, Zhaleh Feizollahi, Andrea Jessee, Mangesh Pujari, Ahmed Aly, Babak Damavandi, Rakesh Wanga, Anuj Kumar, Rohit Patel, Wen tau Yih, and Xin Luna Dong. Crag-mm: Multi-modal multi-turn comprehensive rag benchmark, 2025a. URL https://arxiv.org/abs/2510.26160.

Junjue Wang, Zhuo Zheng, Zihang Chen, Ailong Ma, and Yanfei Zhong. Earthvqa: Towards queryable earth via relational reasoning-based remote sensing visual question answering, 2023. URL https://arxiv.org/abs/2312.12222.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with math-vision dataset, 2024. URL https://arxiv.org/abs/2402.14804.

Qineng Wang, Baiqiao Yin, Pingyue Zhang, Jianshu Zhang, Kangrui Wang, Zihan Wang, Jieyu Zhang, Keshigeyan Chandrasegaran, Han Liu, Ranjay Krishna, Saining Xie, Jiajun Wu, Li Fei-Fei, and Manling Li. Spatial mental modeling from limited views. In *The Fourteenth International Conference on Learning Representations*, 2026a. URL https://openreview.net/forum?id=0FhrtdKLtD.

Siqi Wang, Chao Liang, Yunfan Gao, Erxin Yu, Sen Li, Jing Li, and Haofen Wang. Cityseeker: How do VLMs explore embodied urban navigation with implicit human needs? In *The Fourteenth International Conference on Learning Representations*, 2026b. URL https://openreview.net/forum?id=hzf23XSDcs.

Siting Wang, Minnan Pei, Luoyang Sun, Cheng Deng, Yuchen Li, Kun Shao, Zheng Tian, Haifeng Zhang, and Jun Wang. Spatialviz-bench: A cognitively-grounded benchmark for diagnosing spatial visualization in MLLMs. In *The Fourteenth International Conference on Learning Representations*, 2026c. URL https://openreview.net/forum?id=OqZ7bm28Xx.

Xingrui Wang, Wufei Ma, Tiezheng Zhang, Celso M de Melo, Jieneng Chen, and Alan Yuille. Spatial457: A diagnostic benchmark for 6d spatial reasoning of large multimodal models, 2025b. URL https://arxiv.org/abs/2502.08636.

Zihan Wang, Seungjun Lee, and Gim Hee Lee. Dynam3d: Dynamic layered 3d tokens empower VLM for vision-and-language navigation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026d. URL https://openreview.net/forum?id=s6k9l5yX8e.

Ziye Wang, Yiran Qin, Lin Zeng, and Ruimao Zhang. High-dynamic radar sequence prediction for weather nowcasting using spatiotemporal coherent gaussian representation. In *The Thirteenth International Conference on Learning Representations*, 2025c. URL https://openreview.net/forum?id=Cjz9Xhm7sI.

Azmine Toushik Wasi, Wahid Faisal, Abdur Rahman, Mahfuz Ahmed Anik, Munem Shahriar, Mohsin Mahmud Topu, Sadia Tasnim Meem, Rahatun Nesa Priti, Sabrina Afroz Mitu, Md. Iqramul Hoque, Shahriyar Zaman Ridoy, Mohammed Eunus Ali, Majd Hawasly, Mohammad Raza, and Md Rizwan Parvez. Spatialab: Can vision–language models perform spatial reasoning in the wild? In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=fWWUPOb0CT.

Tonglong Wei, Yan Lin, Zeyu Zhou, Haomin Wen, Jilin Hu, Shengnan Guo, Youfang Lin, Gao Cong, and Huaiyu Wan. Transfertraj: A vehicle trajectory learning model for region and task transferability. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026. URL https://openreview.net/forum?id=XQ87Vo9GIz.

Diankun Wu, Fangfu Liu, Yi-Hsin Hung, and Yueqi Duan. Spatial-MLLM: Boosting MLLM capabilities in visual-based spatial intelligence. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026a. URL https://openreview.net/forum?id=RnXS7aK4rK.

Guanlin Wu, Boyan Su, Yang Zhao, Pu Wang, Yichen Lin, and Hao Frank Yang. Towards physics-informed spatial intelligence with human priors: An autonomous driving pilot study. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026b. URL https://openreview.net/forum?id=pEUBqS8nTk.

Rui Xu, Dakuan Lu, Zicheng Zhao, Xiaoyu Tan, Xintao Wang, Siyu Yuan, Jiangjie Chen, and Xu Yinghui. ORIGAMISPACE: Benchmarking multimodal LLMs in multi-step spatial reasoning with mathematical constraints. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026a. URL https://openreview.net/forum?id=y7ahj9RoXQ.

Shuyang Xu, Zhiyang Dou, Mingyi Shi, Liang Pan, Leo Ho, Jingbo Wang, Yuan Liu, Cheng Lin, Yuexin Ma, Wenping Wang, and Taku Komura. Mospa: Human motion generation driven by spatial audio. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026b. URL https://openreview.net/forum?id=X2r9D46kvI.

Yunqiu Xu, Linchao Zhu, and Yi Yang. Mc-bench: A benchmark for multi-context visual grounding in the era of mllms, 2024. URL https://arxiv.org/abs/2410.12332.

Yuxiang Yan, Zhiyuan Zhou, Xin Gao, Guanghao Li, Shenglin Li, Jiaqi Chen, Qunyan Pu, and Jian Pu. Learning spatial-aware manipulation ordering. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026. URL https://openreview.net/forum?id=iM8154lJCx.

Jihan Yang, Shusheng Yang, Anjali W. Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. *CoRR*, abs/2412.14171, 2024.

Shusheng Yang, Jihan Yang, Pinzhi Huang, Ellis L Brown II, Zihao Yang, Yue Yu, Shengbang Tong, Zihan Zheng, Yifan Xu, Muhan Wang, Rob Fergus, Yann LeCun, Li Fei-Fei, and Saining Xie. Cambrian-s: Towards spatial supersensing in video. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=rBFDvZu6pb.

Sihan Yang, Runsen Xu, Yiman Xie, Sizhe Yang, Mo Li, Jingli Lin, Chenming Zhu, Xiaochen Chen, Haodong Duan, Xiangyu Yue, Dahua Lin, Tai Wang, and Jiangmiao Pang. Mmsi-bench: A benchmark for multi-image spatial intelligence, 2025. URL https://arxiv.org/pdf/2505.23764.

Jinhui Ye, Fangjing Wang, Ning Gao, Junqiu Yu, Zhu Yangkun, Bin Wang, Jinyu Zhang, Weiyang Jin, Yanwei Fu, Feng Zheng, Yilun Chen, and Jiangmiao Pang. Spatially guided training for vision-language-action model. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=eKhOrQWAVJ.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities, 2023. URL https://arxiv.org/abs/2308.02490.

Yifu Yuan, Haiqin Cui, Yibin Chen, Zibin Dong, Fei Ni, Longxin Kou, Jinyi Liu, Pengyi Li, YAN ZHENG, and Jianye HAO. From seeing to doing: Bridging reasoning and decision for robotic manipulation. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=yngvAamNQi.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi, 2023. URL https://arxiv.org/abs/2311.16502.

Shuang Zeng, Dekang Qi, Xinyuan Chang, Feng Xiong, Xie Shichao, Xiaolong Wu, Shiyi Liang, Mu Xu, and Xing Wei. JanusVLN: Decoupling semantics and spatiality with dual implicit memory for vision-language navigation. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=RnuB0Nlbd5.

Xiaoyu Zhan, Wenxuan Huang, Hao Sun, Xinyu Fu, Changfeng Ma, Shaosheng Cao, Bohan Jia, Shaohui Lin, Zhenfei Yin, LEI BAI, Wanli Ouyang, Yuanqi Li, Jie Guo, and Yanwen Guo. Actial: Activate spatial reasoning ability of multimodal large language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026. URL https://openreview.net/forum?id=jquTBzt3Av.

Haoyu Zhang, Meng Liu, Zaijing Li, Haokun Wen, Weili Guan, Yaowei Wang, and Liqiang Nie. Spatial understanding from videos: Structured prompts meet simulation data. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026a. URL https://openreview.net/forum?id=SBYCu5uJJf.

Shuhai Zhang, ZiHao Lian, Jiahao Yang, Daiyuan Li, Guoxuan Pang, Feng Liu, Bo Han, Shutao Li, and Mingkui Tan. Physics-driven spatiotemporal modeling for AI-generated video detection. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026b. URL https://openreview.net/forum?id=HiBoJLCyEo.

Yuyou Zhang, Radu Corcodel, Chiori Hori, Anoop Cherian, and Ding Zhao. Spinbench: Perspective and rotation as a lens on spatial reasoning in VLMs. In *The Fourteenth International Conference on Learning Representations*, 2026c. URL https://openreview.net/forum?id=r7rUDgGYC4.

Zhengshen Zhang, Hao Li, Yalun Dai, Zhengbang Zhu, Lei Zhou, Chenchen Liu, Dong Wang, Francis E. H. Tay, Sijin Chen, Ziwei Liu, Yuxiao Liu, Xinghang Li, and Pan Zhou. From spatial to actions: Grounding vision-language-action model in spatial foundation priors. In *The Fourteenth International Conference on Learning Representations*, 2026d. URL https://openreview.net/forum?id=fzmittHfq3.

Zheyuan Zhang, Fengyuan Hu, Jayjun Lee, Freda Shi, Parisa Kordjamshidi, Joyce Chai, and Ziqiao Ma. Do vision-language models represent space and how? evaluating spatial frame of reference under ambiguities. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=84pDoCD4lH.

Fangrui Zhu, Hanhui Wang, Yiming Xie, Jing Gu, Tianye Ding, Jianwei Yang, and Huaizu Jiang. Struct2d: A perception-guided framework for spatial reasoning in MLLMs. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026a. URL https://openreview.net/forum?id=rbIlWjTFKj.

Zeyu Zhu, KE LIANG, Lingyuan Meng, Meng Liu, Suyuan Liu, Renxiang Guan, Miaomiao Li, Wanwei Liu, and Xinwang Liu. SAINT: Sequence-aware integration for spatial transcriptomics multi-view clustering. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026b. URL https://openreview.net/forum?id=EBHZNmygTT.

Appendix

Table of Content

A Related Work **27**

A.1 Spatial Reasoning in Vision–Language Models 27

A.2 Spatial Reasoning Benchmarks: Coverage and Design 27

A.3 Scale versus Taxonomic Coverage 28

A.4 Evaluation Formats and Diagnostic Depth 28

A.5 3D Spatial Understanding and Scene Reasoning 29

A.6 Embodied and Robotic Spatial Reasoning 30

A.7 Temporal, Spatial–Temporal, and Weather Prediction 30

A.8 Spatial Audio, Brain–Computer Interfaces, and Neuroscience 30

A.9 Spatial Transcriptomics and Biological Spatial Modeling 31

A.10 Mathematical Spatial Reasoning and Cognitive Benchmarks 31

A.11 Positioning of FRONTIERSPATIAL 31

B Task Taxonomy and Descriptions **34**

B.1 Physical Reasoning 34

B.1.1 Predictive Physics 34

B.1.2 Navigation & Control 35

B.1.3 Physical Interaction 35

B.2 Geometric and Positional Reasoning 36

B.2.1 3D Reasoning 36

B.2.2 Geometric Reasoning 37

B.2.3 2D Reasoning 38

B.2.4 Positional Relations 38

B.2.5 Localization 39

B.3 Knowledge-Grounded Reasoning 39

B.3.1 Mathematical Reasoning 39

B.3.2 Creative & Abstract Knowledge 41

B.3.3 Domain Specific 41

B.4 Relational, Logical & Causal Reasoning 42

B.4.1 Causal Reasoning 42

B.4.2 Relations 42

B.5 Temporal & Dynamic Reasoning 42

B.5.1 Temporal Reasoning 42

B.5.2 Motion Reasoning 43

B.6 Quantitative & Comparative Reasoning 43

B.6.1 Measurement . . . 43
B.6.2 Counting . . . 44
B.7 Perceptual Understanding . . . 44
B.7.1 Object-Level Perception . . . 44
B.7.2 Scene-Level Perception . . . 45

C Automated Dataset Curation Pipeline **46**
C.1 Stage 1: Structural Heuristics (60K+ → 24,830) . . . 46
C.2 Stage 2: Vision-Space Deduplication (24,830 → 12,914) . . . 47
C.3 Stage 3: Text-Only LLM Scoring (12,914 → 9,812) . . . 47
C.4 Stage 4: Multimodal LLM Final Pass (9,812 → 7,838) . . . 48
C.5 Generalizability of the Pipeline . . . 49

D Automated Evaluation Pipeline **50**
D.1 Pipeline Overview . . . 50
D.2 Stage 1 — Structured Signal Extraction . . . 50
D.3 Stage 2 — Cross-Model Diagnostic Synthesis . . . 51
D.4 Output Schema . . . 51
D.5 Limitations and Design Trade-offs . . . 52

E Automated Error Analysis Pipeline **52**
E.1 Stage 1: Structured Signal Extraction . . . 53
E.1.1 Signal S1: Correctness . . . 53
E.1.2 Signal S2: Answer Format Validity . . . 53
E.1.3 Signal S3: Reasoning Trace Coherence . . . 53
E.1.4 Pre-classification into Error Pools . . . 54
E.2 Stage 2: Error Taxonomy Induction . . . 54
E.2.1 Step 2a: Error Embedding . . . 54
E.2.2 Step 2b: Density-Based Clustering . . . 54
E.2.3 Step 2c: LLM Taxonomy Labelling . . . 55
E.2.4 Step 2d: Mapping to the Reference Error Taxonomy . . . 55
E.3 Stage 3: Cross-Model Diagnostic Synthesis . . . 56
E.3.1 Step 3a: Per-Model Skill Profiling . . . 56
E.3.2 Step 3b: Statistical Rigor Layer . . . 56
E.3.3 Step 3c: LLM Diagnostic Synthesis . . . 57
E.4 Output Artefacts . . . 57
E.5 Reproducibility and Infrastructure Considerations . . . 57

F More details on Experiments and Evaluation **58**
F.1 Evaluation Protocol . . . 58

F.2 Evaluation of LLM-as-a-Judge for Open-ended Evaluation 58
F.3 Human Evaluation Protocol 60
F.4 AI Agents for Spatial Reasoning 60
F.4.1 SPATIOXOLVER 60
F.4.2 Claude Code Agent 61
F.4.3 Evaluation Setup 61

G Model Performance Breakdown on FRONTIERSPATIAL 61
G.1 Model Performance On Identifying Spatial Element 61
G.2 Model Performance On Identifying Spatial Relations 62
G.3 Spatial Reasoning Trace Correctness 62

H Category-wise Model Performance of FRONTIERSPATIAL 63

I Qualitative Error Taxonomy of Spatial–Visual Reasoning Failures 63
I.1 E1 — Perceptual Failure 64
I.2 E2 — Spatial Grounding Failure 67
I.3 E3 — Spatial Transformation Failure 69
I.4 E4 — Quantitative Spatial Failure 71
I.5 E5 — Multi-Step Reasoning Failure 73
I.6 E6 — Language–Spatial Mapping Failure 74
I.7 E7 — Format / Refusal Failure 75

J Hard-Case Construction and Evaluation 77

K Discussion 78
K.1 Broader Impacts 78
K.2 Limitations 78

L Prompts used in FRONTIERSPATIAL 79

M Representative Benchmark Samples Across All Tasks 82

A Related Work

Spatial reasoning in vision–language models (VLMs) sits at the intersection of perception, cognition, and grounded action. This section surveys the breadth of prior work, organized into seven thematic threads: (i) the foundational challenge of spatial reasoning in VLMs; (ii) benchmark design and its evolution; (iii) the persistent tension between scale and taxonomic coverage; (iv) evaluation format and diagnostic depth; (v) recent advances in 3D and embodied spatial understanding; (vi) emerging directions in spatial audio, transcriptomics, and cross-modal sensing; and (vii) the positioning of FrontierSpatial within this landscape. Table 3 provides a structured comparison of representative benchmarks.

A.1 Spatial Reasoning in Vision–Language Models

Spatial reasoning—the capacity to interpret geometric structures, relational layouts, and perspective-consistent representations in visual scenes—is central to embodied intelligence and a key bottleneck for real-world deployment of artificial agents (Huang et al., 2024; Qi et al., 2024; 2025). Early multimodal benchmarks such as CLEVR (Johnson et al., 2016) and GQA (Hudson & Manning, 2019) probed compositional reasoning through synthetic or template-generated scenes, enabling controlled analysis but offering limited ecological validity. With the rise of large VLMs (Li et al., 2023; Alayrac et al., 2022; OpenAI, 2023), spatial reasoning has been revisited in more naturalistic settings, yet systematic weaknesses persist. Despite strong performance on captioning and visual QA, these models routinely fail at depth ordering, occlusion inference, and consistent reference-frame alignment (Chen et al., 2025; Kosoy et al., 2025; Pothiraj et al., 2025).

Crucially, scaling alone does not resolve these deficits—structural spatial understanding requires capabilities that go beyond what language-grounded pattern matching provides (Cai et al., 2025b). Recent work has identified further failure modes: (Zhang et al., 2025) show through the COMFORT benchmark that current VLMs handle spatial frames of reference poorly, especially across languages and in ambiguous contexts, while (Zhang et al., 2026c) reveal strong egocentric bias and poor rotational understanding across 43 VLMs via the SpinBench evaluation. A recurring finding is that even state-of-the-art proprietary models leave large gaps relative to human performance: (Wang et al., 2026a) report near-random performance on spatial mental modeling tasks in the MindCube benchmark, and (Huang et al., 2026) characterize consistent gaps across 32 VLMs using the Spatial-DISE taxonomy, which partitions spatial reasoning into intrinsic/extrinsic and static/dynamic quadrants.

Efforts to close these gaps have taken multiple directions. (Li et al., 2026b) train a 3B model progressively from perception to reasoning, surpassing GPT-4o by 20.8% on spatial benchmarks via SpatialLadder. (Ma et al., 2026b) demonstrate that explicitly sharing 3D representations across perception, computation, and reasoning stages outperforms Gemini 2.0 by 9.2% on 3DSRBench. From a cognitive neuroscience perspective, (Li et al., 2026f) integrate a hippocampus-inspired spatial memory module with Transformers, achieving superior 2D/3D spatial prediction. Meanwhile, (Kharlamova et al., 2026) introduce Spatial CAPTCHA as a diagnostic tool, showing that leading MLLMs solve these geometric verification challenges at only 31% accuracy compared to near-perfect human performance—underscoring the magnitude of the remaining gap.

A.2 Spatial Reasoning Benchmarks: Coverage and Design

A growing number of benchmarks have been proposed to evaluate spatial reasoning in VLMs, yet each addresses only a narrow slice of the problem space (see Table 3). Among early embodied QA datasets, ScanQA (Azuma et al., 2022) grounded questions in 3D indoor reconstructions but was limited in scale and task variety. Visual Spatial (Liu et al., 2023) and What's Up (Kamath et al., 2023) extended coverage to household and real-world contexts, yet relied on template-driven question generation that narrows linguistic variability and overrepresents simple binary spatial relations. EmbSpatial-Bench (Du et al., 2024) and Space3D-Bench (Szymanska et al., 2024) increased task variety but remained constrained in scale (2.2K and 211 samples, respectively) and annotation style.

More recent benchmarks target specific sub-skills. MMSI-Bench (Yang et al., 2025), Spatial-DISE (Huang et al., 2025), Q-Spatial-Bench (Cheng et al., 2024b), and CausalSpatial (Ma et al., 2026a) focus on metric depth, positional relations, or causal inference, respectively, but none

provides unified coverage across the full spectrum of spatial cognition. SPATIALAB (Wasi et al., 2026;?) advances the field with 30 subcategories, manual puzzle-style annotations, and open-ended evaluation across 1,400 questions with a human accuracy of 87.57% compared to 54.93% for the best models, yet its scope leaves the holistic picture incomplete.

Broader coverage benchmarks have appeared in the recent literature. (Jia et al., 2026) introduce OmniSpatial, a cognitively-grounded benchmark covering dynamic, logical, interactive, and perspective-taking spatial reasoning across 50 subcategories. (Gong et al., 2026) propose SpaCE-10, which benchmarks compositional spatial intelligence across 10 atomic and 8 compound capabilities using real indoor scenes with point cloud input. (Wang et al., 2026c) evaluate spatial visualization across 27 MLLMs and 12 tasks with SpatialViz-Bench, notably finding that chain-of-thought prompting paradoxically degrades accuracy in open-source models. (Li et al., 2026d) present STARE, a benchmark evaluating multi-step visual simulation tasks such as 3D folding and tangram manipulation, where models approach random chance. Compositional and perspective-taking reasoning is further assessed by SpinBench (Zhang et al., 2026c), which identifies strong egocentric bias, and by the Spatial CAPTCHA framework (Kharlamova et al., 2026).

In temporally-grounded settings, (Gu et al., 2025) introduce TA-STVG, which adaptively generates object queries from video-text pairs using temporal and spatial cues for better spatio-temporal video grounding. Multi-view spatial reasoning is targeted by (Feng et al., 2026) through MV-RoboBench, which evaluates robotic manipulation settings and finds state-of-the-art models far below human performance, while (Gholami et al., 2026) introduce Ego3D-Bench for ego-centric multi-view outdoor spatial reasoning paired with the Ego3D-VLM model that improves performance via cognitive map generation.

A.3 SCALE VERSUS TAXONOMIC COVERAGE

A persistent tension in spatial benchmark design is between scale and taxonomic coverage. Large-scale datasets such as SPATIALVLM (Chen et al., 2024) provide billions of QA pairs but cover only two coarse spatial categories, rendering them unsuitable for fine-grained diagnostic evaluation. ROBOSPATIAL (Song et al., 2024) scales to one million robot-centric scenes but relies on template-based annotation that enables heuristic shortcuts and introduces a significant sim-to-real gap. High-quality manual datasets such as VSI-BENCH (Yang et al., 2024) achieve annotation fidelity but remain too small (288 scenes) for robust cross-model comparison.

Efforts to build large-scale, diverse spatial instruction corpora have gained traction. (Deng et al., 2026) introduce InternSpatial, the largest open-source spatial reasoning dataset to date, comprising 12M QA pairs across 19 instruction formats for both single- and multi-view settings. (Li et al., 2026c) contribute SCoT, a million-scale chain-of-thought annotation suite for 3D-LLMs covering perception, analysis, and planning levels. (Ma et al., 2026c) demonstrate through Real-3DQA that existing 3D-LLMs rely heavily on textual shortcuts, and propose a 3D-reweighted training objective that substantially improves genuine spatial reasoning. Despite these advances, no existing benchmark simultaneously achieves broad taxonomic coverage, sufficient scale, visual diversity, and multi-format evaluation—a gap that FRONTIERSPATIAL directly addresses.

Benchmarks targeting specific domains contribute valuable task diversity but are too narrow to inform general spatial reasoning capabilities. Remote sensing is probed by RSIEVAL (Hu et al., 2023) and GEOBENCH-VLM (Danish et al., 2024); biological imaging by VLM4BIO (Maruf et al., 2024); and wearable contexts by WEARVQA (Chang et al., 2025). Even broad-coverage benchmarks such as MMMU (Yue et al., 2023) and OMNIBENCH (Li et al., 2024) treat spatial tasks as incidental rather than primary, yielding only aggregate signals. (Yang et al., 2026) introduce Cambrian-S with the VSI-SUPER benchmark and a predictive sensing model that outperforms proprietary baselines, while (Wang et al., 2026b) benchmark VLMs on implicit-need urban navigation through CitySeeker, with top models achieving only 21.1% task completion across 8 cities.

A.4 EVALUATION FORMATS AND DIAGNOSTIC DEPTH

Beyond task coverage, evaluation format shapes what weaknesses are measurable. Most prior benchmarks employ closed-ended multiple-choice or binary judgments (Du et al., 2024; Kamath et al., 2023; Cai et al., 2025a; Fu et al., 2024), which simplify scoring but allow models to exploit

option elimination rather than produce genuine spatial reasoning. Benchmarks such as MM-VET (Yu et al., 2023), MARVEL (Jiang et al., 2024), and SPATIALAB (Wasi et al., 2026) introduce open-ended evaluation, but few couple this with systematic error analysis across a detailed task hierarchy.

Structured diagnostic studies—examining failure modes by task type, image complexity, and reasoning depth—are largely absent from the literature. The Spatial Intelligence Grid (SIG) schema introduced by (Wu et al., 2026b) represents a step in this direction: it encodes scene layouts and physical priors as a structured complement to text, enabling more faithful and measurable visual-spatial intelligence evaluation. Similarly, (Lee et al., 2026a) combine egocentric and exocentric views via the E3VQA benchmark and M3CoT prompting method to give LVLMs richer scene context for more accurate multi-view spatial reasoning. (Luo et al., 2026) introduce pySpatial, a zero-shot visual programming framework that equips MLLMs with 3D spatial tools via Python code generation, allowing direct inspection of model reasoning chains.

Reasoning efficiency has also emerged as an evaluation concern. (Shen et al., 2026b) show that DARE can dynamically prune multimodal tokens with asymmetric modality compression, cutting FLOPs by 40% while preserving spatial reasoning performance—demonstrating that spatial capability need not come at prohibitive computational cost. The question of *how* models reason spatially, rather than merely whether they succeed, motivates a growing body of work on chain-of-thought and structured prompting: (Zhang et al., 2026a) combine SpatialMind structured prompting with ScanForgeQA simulation-based fine-tuning to enhance 3D spatial reasoning without architectural changes, while (Zhu et al., 2026a) use bird's-eye-view images with object-centric metadata as structured 2D prompts to enable strong 3D reasoning without explicit 3D input.

A.5 3D SPATIAL UNDERSTANDING AND SCENE REASONING

Grounding spatial reasoning in explicit 3D representations has been a central thrust of recent work. (Wu et al., 2026a) propose Spatial-MLLM, a dual-encoder architecture pairing a semantic encoder with a geometry-focused 3D spatial encoder to boost spatial reasoning from purely 2D video inputs. (Mao et al., 2026) fine-tune a standard multimodal LLM on 12K synthetic indoor scenes to generate structured 3D outputs including walls, doors, windows, and oriented bounding boxes. (Liu et al., 2026c) convert raw depth data into structured textual rationales distilled into compact latent embeddings for plug-and-play spatial reasoning enhancement. (Chen et al., 2026c) introduce depth positional encoding and a 700K-pair spatial measurement dataset (MSMU) to give VLMs strong quantitative 3D spatial measuring capabilities.

Scene-level 3D reasoning for navigation and manipulation has attracted substantial attention. (Qi et al., 2026) introduce semantic orientation defined via natural language, building the OrienText300K dataset to enable 6-DoF spatial reasoning and robotic manipulation without predefined reference frames. (Hao et al., 2026) generate task-aligned tabletop scenes via a spatial reasoning chain decomposing instructions into object inference, relation reasoning, and 3D scene graph construction. (Wang et al., 2026d) build dynamic layered 3D patch-instance-zone representations from RGB-D inputs for vision-and-language navigation, achieving state-of-the-art results on R2R-CE, REVERIE-CE, and NavRAG-CE. (Hu et al., 2026) introduce 3DLLM-Mem, which uses working memory tokens to selectively query episodic spatial-temporal memory for long-horizon embodied reasoning, outperforming the strongest baselines by 16.5% on 3DMem-Bench.

Semantic scene completion benefits from spatial reformulations as well. (Li et al., 2026g) reformulate semantic scene completion as dense object detection by decoupling voxel-level prediction into offset regression and semantic classification, achieving state-of-the-art results on SemanticKITTI. For 3D question answering, (Ma et al., 2026c) show that genuine spatial reasoning is being sidestepped by textual cues in existing 3D-LLMs, and (Shen et al., 2026a) propose SpatialReasoner-R1, combining multi-model Monte Carlo Tree Search for chain-of-thought data generation with fine-grained DPO for segment-level spatial reward—setting a new state-of-the-art on SpatialRGPT-Bench. Spatial prompting for purely vision-based understanding is addressed by (Li et al., 2026e) through See&Trek, a training-free framework combining keyframe sampling with visual trajectory simulation to improve spatial understanding from vision alone.

A.6 Embodied and Robotic Spatial Reasoning

Robotic manipulation and navigation demand spatial reasoning that operates under real-world constraints of partial observability, dynamic scenes, and grounded action. Several recent lines of work bridge VLMs with robotic spatial grounding. (Ye et al., 2026) introduce SP-VLA, which bridges VLMs with robotic control via spatial grounding pretraining and spatially-guided action post-training, achieving state-of-the-art results on SimplerEnv. (Yuan et al., 2026) propose FSD, which generates spatial relationship intermediate representations to guide robotic manipulation, achieving a 40.6% success rate on SimplerEnv. (Zhang et al., 2026d) introduce FALCON, which injects 3D spatial tokens from foundation models into a dedicated action head, achieving state-of-the-art VLA performance across both simulation and real-world tasks. (Li et al., 2026a) propose Spatial Forcing, which aligns VLA intermediate embeddings with pretrained 3D model representations to improve spatial awareness implicitly, without requiring explicit 3D inputs at inference time.

For object ordering in cluttered scenes, (Yan et al., 2026) introduce OrderMind, which learns spatial-aware manipulation ordering via a spatial graph encoder and VLM-guided priority labeling, evaluated on a 163K-sample benchmark. Vision-and-language navigation is addressed by (Zeng et al., 2026), who decouple spatial-geometric and visual-semantic memory into dual implicit neural representations in JanusVLN for more efficient navigation. Scene layout generation benefits from (Pan & Liu, 2026)'s MetaSpatial, which uses reinforcement learning with physics-aware rewards for real-time 3D scene layout generation without post-processing. For video generation conditioned on camera control, (Lee et al., 2026b) introduce 3DScenePrompt with dual spatio-temporal conditioning and dynamic SLAM-based scene memory for scene-consistent, camera-controllable video generation.

Spatial reasoning is also critical for place recognition. (Chen et al., 2026d) introduce SAGE, which enhances visual place recognition via adaptive geo-visual graph construction and greedy clique expansion sampling with a frozen DINOv2 backbone. Cross-view spatial reasoning is activated in MLLMs by (Zhan et al., 2026) through Actial's Viewpoint Learning, using a 100K cross-view dataset with two-stage SFT and GRPO fine-tuning. Multi-view benchmarking is addressed by (Feng et al., 2026) through MV-RoboBench, while audio-visual spatial tracking is handled by (Chen et al., 2026b) through SAVVY, which tracks object trajectories and constructs a global dynamic map for 3D spatial QA across dynamic scenes with spatial audio.

A.7 Temporal, Spatial–Temporal, and Weather Prediction

Spatial reasoning extends beyond static scenes to temporally dynamic contexts. (Wang et al., 2025c) introduce the STC-GS + GauMamba framework for efficient 3D radar sequence prediction for weather nowcasting via spatiotemporal Gaussian representations. For vehicle trajectory learning, (Wei et al., 2026) propose TransferTraj, which unifies multi-modal trajectory representations with a task-transferable masking scheme for zero-shot and few-shot vehicle trajectory learning across regions and tasks.

Multi-object tracking under occlusion is addressed by (Bai et al., 2026), who propose STAR's spatial-temporal tracklet clip graph matching, linking confidently associated trajectory segments via appearance and topology features. Depth estimation for omnidirectional imagery is handled by (Cao et al., 2026) through ST^{2}360D, a training-free pipeline converting 360-degree images into perspective video sequences and exploiting temporal consistency from video depth models for coherent 360-degree depth maps. Detection of AI-generated videos is approached spatiotemporally by (Zhang et al., 2026b) through NSG-VD, which measures deviations from physical probability flow conservation using a normalized spatiotemporal gradient statistic derived from diffusion models.

A.8 Spatial Audio, Brain–Computer Interfaces, and Neuroscience

Spatial reasoning is increasingly studied in modalities beyond vision. (Biswas et al., 2026) introduce OWL, which aligns binaural audio with 3D geometry for audio LLMs, enabling clock-level azimuth estimation with geometry-aware chain-of-thought reasoning. (Xu et al., 2026b) introduce MOSPA, the first spatial audio-driven human motion dataset, paired with a diffusion-based framework for generating realistic body motion conditioned on spatial audio signals.

In brain–computer interfaces, (Le et al., 2026) introduce SPINT, a permutation-invariant Transformer for intracortical BCI decoding that uses context-dependent positional embeddings to generalize

across sessions without test-time retraining. For EEG spatial super-resolution, (Liu et al., 2026b) propose SRGDiff, a step-aware residual-guided diffusion model achieving up to 40% gains over strong baselines. Deep reinforcement learning agents benefit from spatial representations as well: (Saura et al., 2026) integrate ring attractor neural structures into RL agents to encode spatial action representations and stabilize action selection, achieving a 53% improvement on Atari 100k.

A.9 Spatial Transcriptomics and Biological Spatial Modeling

Spatial reasoning also plays a role in biological and medical imaging contexts. (Zhu et al., 2026b) introduce SAINT, which augments spatial transcriptomics clustering with DNA sequence embeddings from a Nucleotide Transformer to improve spot-level clustering across tissue benchmarks. (Han et al., 2026) propose STAMP, which aligns pathology images with spatially-resolved gene expression via contrastive learning, trained on the largest Visium dataset (SpaVis-6M). For spatial organization in microscopy images, (Sturm et al., 2026) introduce SI-VAE, which incorporates point-process likelihoods into variational autoencoders for interpretable spatial organization learning. Physics sensing in spatially distributed systems is addressed by (Ma et al., 2026d) through PhySense, which jointly learns physical field reconstruction and sensor placement optimization via a generative model with projected gradient descent.

Discrete generative modeling with spatial structure is advanced by (Santos et al., 2026) through Discrete Spatial Diffusion (DSD), a discrete-state jump process framework that strictly preserves particle counts under mass conservation constraints.

A.10 Mathematical Spatial Reasoning and Cognitive Benchmarks

Spatial reasoning that requires explicit mathematical constraint handling is an emerging frontier. (Xu et al., 2026a) introduce ORIGAMISPACE, which benchmarks MLLMs on multi-step spatial reasoning and mathematical constraint handling through origami crease pattern tasks, including RL-based code generation evaluation. The challenge of perspective-taking and mental rotation is specifically examined by (Wang et al., 2026a) through MindCube, where a "map-then-reason" approach combining spatial map generation with reinforcement learning boosts accuracy from 37.8% to 70.7%. (Li et al., 2026d) similarly probe multi-step visual simulation via STARE, finding that models approach random chance on tasks requiring mental simulation.

A.11 Positioning of FrontierSpatial

As illustrated in Table 3 and summarized numerically in Table **??**, existing benchmarks each cover at most a handful of spatial dimensions, leaving large blind spots in the evaluation landscape. The fragmentation spans multiple axes: narrow task scope (most benchmarks address fewer than five spatial sub-categories), limited scale (many high-quality datasets contain fewer than 2K samples), constrained visual diversity (synthetic or template-heavy sources), homogeneous evaluation format (closed-ended only), and absence of automated diagnostic tooling.

FrontierSpatial addresses this fragmentation directly. By consolidating and extending 23 existing datasets through an automated curation pipeline, FrontierSpatial assembles 7,838 questions paired with 11,825 images under a four-level task hierarchy spanning 7 categories, 19 sub-categories, and 65 fine-grained tasks, achieving full coverage across all 22 spatial dimensions identified in our taxonomy. Unlike template-heavy predecessors, the benchmark draws from diverse real-world visual sources and spans both multiple-choice and open-ended evaluation formats.

Complementing the benchmark, we contribute an automated evaluation harness for reproducible model comparison and an automated error analysis harness for structured, fine-grained diagnostics—capabilities that prior spatial benchmarks have not provided at this scope. Together, these contributions establish FrontierSpatial as the first holistic framework for spatial reasoning evaluation in the multimodal era, addressing gaps that no single prior benchmark—whether broad (OmniSpatial (Jia et al., 2026), Spatial-DISE (Huang et al., 2026)) or deep (VSI-Bench (Yang et al., 2024), SpatiaLab (Wasi et al., 2026))—has been able to close.

Table 3: Comparison of Benchmarks Across Reasoning and Perception Dimensions

Benchmark	Predictive Physics	Temporal Reasoning	Textual Perception	3D Reasoning	Geometric Reasoning	Mathematical Reasoning	Causal Reasoning	Counting	Domain Specific	Creative & Abstract	Navigation & Control	Object-Level Perception	2D Reasoning	Relations	Motion Reasoning	Positional Relations	Measurement	Localization	Scene-Level Perception	Physical Interaction	Logical Reasoning	Comparison
OmniBench Li et al. (2024)			✓		✓	✓			✓	✓									✓			
Spatial457 Wang et al. (2025b)				✓	✓									✓		✓		✓				✓
MathVision Wang et al. (2024)					✓	✓			✓				✓				✓					
BabyVision Chen et al. (2026a)	✓						✓	✓		✓		✓							✓		✓	
Spatial-DISE Huang et al. (2025)				✓										✓		✓		✓				
MM-IQ Cai et al. (2025a)					✓	✓	✓	✓		✓			✓								✓	✓
MMSI-Bench Yang et al. (2025)				✓				✓				✓		✓		✓		✓				
TB-Bench Charoenpitaks et al. (2025)		✓						✓				✓			✓							
RSIEval Hu et al. (2023)			✓					✓	✓			✓						✓	✓			
GEOBench-VLM Danish et al. (2024)								✓	✓			✓						✓				
SpatiaLab Wasi et al. (2026)				✓								✓		✓		✓		✓				
MCBench Xu et al. (2024)												✓		✓								
WearVQA Chang et al. (2025)		✓	✓					✓	✓			✓						✓	✓			
MMMU Yue et al. (2023)						✓			✓													
ZeroBench Roberts et al. (2025)										✓												
MM-Vet Yu et al. (2023)			✓			✓		✓	✓			✓										
LEGO-Puzzles Tang et al. (2025)				✓	✓		✓	✓					✓	✓							✓	
Q-Spatial-Bench Cheng et al. (2024b)				✓												✓	✓					
CausalSpatial Ma et al. (2026a)	✓			✓			✓					✓										
CRAG-MM Wang et al. (2025a)		✓	✓						✓			✓						✓	✓			
EarthVQA Wang et al. (2023)								✓	✓					✓								
MARVEL Jiang et al. (2024)					✓		✓	✓		✓			✓								✓	✓
LiveVQA Fu et al. (2025)		✓	✓						✓			✓							✓		✓	
VLM4Bio Maruf et al. (2024)								✓	✓			✓						✓				
ScienceQA Lu et al. (2022)						✓			✓													
SimpleVQA Cheng et al. (2025)		✓	✓																			
DARE Sterz et al. (2024)			✓						✓													✓
HAL Spatial (Ours)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Table 4: Comparison of multimodal and spatial reasoning benchmarks.

Benchmark	Selected Tasks	Number of Images	Number of Total Tasks	Date of Publication
OmniBench	50	11525	11000	Sep-2024
Spatial457	781	1000	68752	Feb-2025
MathVision	2476	3472	3040	Feb-2024
BabyVision	82	388	388	Jan-2026
Spatial-DISE	301	559	559	Oct-2025
MM-IQ	145	2710	2710	Feb-2025
TB-Bench	938	2000	2000	Jan-2025
RSIEval	137	100	936	Jul-2023
GEOBench-VLM	61	–	10000	Nov-2024
SpatiaLab	617	1400	1400	Feb-2026
WearVQA	259	2520	2520	Nov-2025
MMMU	51	10500	10500	Nov-2023
zerobench	73	100	434	Feb-2025
MM-vet	55	200	218	Aug-2023
LEGO-Puzzles	580	1100	1100	Mar-2025
MMSI-Bench	623	1990	1000	May-2025
Q-Spatial-Bench	82	1406	1406	Jun-2024
CausalSpatial	322	–	1541	Jan-2026
crag-mm	8	8500	8500	Oct-2025
EarthVqa	50	6000	208593	Dec-2023
MARVEL	97	770	770	Apr-2024
DARE	50	–	5525	Sep-2024

Table 5: **Hierarchical task taxonomy of the HolisticSpatial benchmark.** Tasks are organised into three levels: (L1) domain categories capturing distinct spatial reasoning modalities, (L2) sub-categories reflecting specific cognitive or perceptual operations, and (L3) fine-grained task types with defined evaluation formats. Counts indicate the number of samples per task.

Category	Sub-Category	Task	Count
Physical Reasoning	Predictive Physics	Predictive	50
		Planning	135
	Navigation & Control	Turning – Car	82
		Navigation	177
		Lane Position	228
		Turning – Entity	214
		Lane Change	180
Positional & Geometric Reasoning	3D Reasoning	3D Pose	226
		3D Shape Finding	56
		3D Geometry	349
		3D Assembly	119
		3D Rotation	197
		3D Object Detection	151
		3D Folding	68
		3D Projection	50
	Geometric Reasoning	Occluded Attributes	374
		Occluder Attributes	181
		Direction and Orientation	145
		Compatibility	88
	2D Reasoning	2D Geometry	102
	Positional Relations	Region–Region	79
		Camera–Region	81
		Object–Region	57
		Camera–Object	86
		Object–Object	87
	Localization	Relative Location	50
		Object Localization	50
Knowledge-Grounded Reasoning	Mathematical Reasoning	Arithmetic	107
		Algebra	299
		Scaling & Ratio	143
		Geometry – Transformational	157
		Geometry – Coordinate	77
		Geometry – Descriptive	98
		Constraint Geometry	87
		Metric Geometry	272
		Shape Decomposition	205
		Angle Reasoning	143
	Creative & Abstract	Pattern	59
		Riddle	73
		Trick	50
	Domain Specific	Engineering & Technology	51
Relational & Causal Reasoning	Causal Reasoning	Logical Counterfactual	113
		Physical Causality	62
	Relations	Spatial Relationship	280
Temporal & Dynamic Reasoning	Temporal Reasoning	Temporal Movement	51
		Sequencing	47
	Motion Reasoning	Camera Motion	70
		Object Motion	50
Quantitative & Comparative Reasoning	Measurement	Relative Distance	234
		Measurement	61
		Absolute Distance	82
	Counting	Counting	76
Perceptual Understanding	Object-Level	Size & Scale	170
		Attribute	53
		Object Identification	50
		Color	50
	Scene-Level	Scene Understanding	50
Total			7838

B TASK TAXONOMY AND DESCRIPTIONS

Understanding spatial reasoning requires more than simple object recognition; it depends on how objects relate, align, and interact within physical space. To capture this complexity, we introduce *HolisticSpatial*, a benchmark designed to evaluate fine-grained spatial reasoning across diverse settings. The benchmark covers a wide spectrum of visual spatial-reasoning abilities, organised into a three-level hierarchy.

At the *first level* (**L1**), we define seven mutually exclusive **domain categories** that capture distinct reasoning modalities. The *second level* (**L2**) partitions each domain by the specific **cognitive or perceptual operations** involved. The *third level* (**L3**) further refines these into fine-grained **task types** with well-defined question formats and evaluation criteria. Table 5 summarises the full hierarchy along with sample counts.

The tasks in *HolisticSpatial* are designed to reflect how humans naturally reason about environments, ranging from relative positioning (e.g., *What is to the left of the microwave?*) to geometric and physical inference (e.g., *Would this stack fall if another book is added?*). Each category is further decomposed into subcategories that capture critical dimensions of reasoning, such as occlusion, orientation, navigation, and physical interactions like gravity and collision.

By framing subtasks in intuitive, everyday terms, the benchmark encourages models to move beyond object detection toward structured spatial interpretation. This design enables systematic evaluation of generalization across contexts relevant to real-world applications, including robotics, augmented and virtual reality, autonomous driving, and human-computer interaction. The following sections provide concise task-level descriptions for each category and subcategory.

Table 6: Distribution of instances across the seven top-level categories of FRONTIERSPATIAL.

Category	# Questions	# Sub-cats	# Tasks
Physical Reasoning (Phys.)	1,312	3	9
Positional & Geometric Reasoning (Pos.Geo.)	2,596	4	18
Knowledge-Grounded Reasoning (Knowledge)	2,332	4	14
Relational, Logical & Causal Reasoning (Re.Lo.Ca.)	455	3	10
Temporal & Dynamic Reasoning (Temporal)	218	2	6
Quantitative & Comparative Reasoning (Quant.)	503	2	5
Perceptual Understanding (Percept.)	422	1	3
Total	**7,838**	**19**	**65**

B.1 PHYSICAL REASONING

Physical Reasoning tasks require a model to apply knowledge of real-world physical laws, kinematics, collision dynamics, navigation constraints, and spatial affordances, to interpret or predict states and events depicted in images. With $n = 1{,}312$ items, this domain spans three second-level categories that together cover predictive simulation, embodied navigation, and physical interaction.

B.1.1 PREDICTIVE PHYSICS

Predictive Physics tasks ask models to reason about what will happen next given the current physical configuration shown in an image, requiring an implicit forward simulation of physical principles rather than mere scene description. The $n = 185$ items in this category are divided between unconstrained single-outcome prediction and goal-directed multi-step planning.

Predictive. Predictive tasks ($n = 50$) present a static snapshot of a physical scene and ask the model to forecast the most probable next physical state without explicit velocity or force parameters. Given an image of a heavy hammock attached to the edge of an overhanging roof, the model must evaluate structural load and cantilever mechanics to judge the most likely outcome, testing whether it can simulate gravitational and mechanical cause-and-effect chains purely from a single observed frame. A complementary task provides a top-down diagram of billiard balls with a directional arrow on the cue ball and asks whether any ball will enter a pocket, exercising two-dimensional Newtonian trajectory reasoning from a visually encoded motion description.

Planning. Planning tasks ($n = 135$) extend pure prediction by requiring the model to identify a sequence of interventions or actions necessary to achieve a specified goal state while respecting physical constraints. Given a bird's-eye image of a toy car on a collision course with several objects, a stack of books, a vase, and a chair, the model must determine which specific objects need to be

removed to guarantee a collision-free path, testing hierarchical physical planning under real-world spatial constraints. A second representative task presents a LEGO build with a defined target state and a set of candidate intermediate steps, asking the model to arrange those steps into the correct construction sequence, exercising procedural planning over a discrete, structured physical state space.

B.1.2 Navigation & Control

Navigation & Control tasks ground physical reasoning within road-driving scenarios, requiring models to parse the spatial configuration of lanes, vehicles, and traffic participants from egocentric dashcam images and to produce correct kinematic or positional judgements. This is the largest second-level category within Physical Reasoning ($n = 881$) and spans five sub-categories covering the ego vehicle's own motion as well as the motion and position of other traffic participants.

Turning – Car. Turning–Car tasks ($n = 82$) present a dashcam image and ask the model to describe or classify the ego-vehicle's current turning manoeuvre. Given a forward-facing dashcam frame captured mid-turn at an urban intersection, the model must determine whether the ego vehicle is turning left, turning right, or proceeding straight, requiring egocentric motion inference from road geometry, visible lane markings, and kerb curvature within a single static frame. These tasks specifically test whether a model can maintain a first-person spatial frame and distinguish self-motion cues from the surrounding scene context.

Navigation. Navigation tasks ($n = 177$) require the model to resolve directional or route-planning questions embedded within real or schematic map-like images. Given an overhead sign showing directional arrows and named destinations, the model must identify which road to take given a stated travel goal, testing the composition of cardinal-direction reasoning with symbolic map interpretation. A second example provides a multi-photograph room walk-through and asks in which direction a named landmark lies relative to the observer at the final step, exercising egocentric frame updating across a described movement sequence.

Lane Position. Lane Position tasks ($n = 228$) present a driving scene and ask the model to classify the current lane occupied by a specified traffic entity from a fixed option set: front lane, front-left lane, front-right lane, or oncoming-traffic lane. Given a dashcam image showing multiple vehicles ahead at varying lateral offsets, the model must identify the precise lane of a named entity by reading lane markings, vehicle silhouettes, and perspective cues, testing fine-grained spatial grounding of traffic participants within a structured road topology. Correct answers often require distinguishing positions that differ by a single lane-width, placing high demands on lateral localisation precision.

Turning – Entity. Turning–Entity tasks ($n = 214$) mirror the Turning–Car sub-category but direct the model's attention to other traffic participants, vehicles, cyclists, or pedestrians, rather than the ego vehicle. Given an image of a junction from the ego perspective, the model must identify whether a specified entity is turning left, turning right, merging, or proceeding straight, requiring allocentric tracking of another agent's motion trajectory within the shared scene. This tests whether the model can maintain an independent spatial reference frame for a non-ego agent while remaining anchored to the ego-vehicle viewpoint.

Lane Change. Lane Change tasks ($n = 180$) ask models to identify whether a named traffic entity is in the process of changing lanes and to characterise the direction and stage of the manoeuvre. Given a dashcam image capturing a vehicle straddling two lane markings, the model must infer the direction of movement from the vehicle's angle relative to the lane markings and its lateral offset from the lane centre, testing dynamic spatial inference from a single static frame. These items are particularly demanding because mid-manoeuvre positions are visually ambiguous and require integrating multiple weak spatial cues simultaneously.

B.1.3 Physical Interaction

Physical Interaction tasks require models to reason about what happens when objects are placed in proximity, when forces are applied, or when physical constraints between objects come into play, going beyond trajectory prediction to incorporate material, shape, and affordance reasoning. The $n = 246$ items are divided between property inference and collision analysis.

Property Inference. Property Inference tasks ($n = 74$) ask the model to deduce a physical property of an object, material, rigidity, weight, or spatial fit, that is not directly labelled in the image but can be inferred from visual appearance or geometric configuration. Given an image of a glass jar already containing an apple and a lemon, the model is asked whether a specified object positioned directly above the opening would fit inside if dropped, testing the ability to reason about relative

size, container geometry, and gravity without any explicit measurements. A related task presents two utensils, a spoon and chopsticks, and asks whether they are made of the same material, exercising cross-object surface-texture and reflectance comparison under naturalistic viewing conditions.

Collision. Collision tasks ($n = 172$) require the model to predict whether a moving object will intersect with one or more obstacles and to determine what spatial interventions would prevent each collision. Given a bird's-eye image of a floor plan with a toy vehicle at a known position and heading, the model must trace the forward path, identify all objects in that path, and determine whether removing a single named obstacle is sufficient to clear the route, testing multi-obstacle spatial planning under a physical motion constraint. A second task type presents a soccer ball with a directional trajectory arrow aimed at a goal and asks whether the ball will score, requiring 2D path-intersection reasoning between a directed vector and a bounded target region.

B.2 Geometric and Positional Reasoning

Geometric and Positional Reasoning is the largest domain in the benchmark ($n = 2{,}539$), covering the full range of spatial reasoning operations that act on explicit geometric structure: three-dimensional object manipulation, planar geometry, occlusion-based attribute inference, and precise positional relationship identification between scene elements. Its five second-level categories span from volumetric 3D reasoning down to two-dimensional planar geometry and fine-grained object localisation.

B.2.1 3D Reasoning

3D Reasoning tasks require models to interpret, manipulate, or reason about objects in three-dimensional space from one or more 2D image projections, covering mental rotation, depth-order inference, surface reconstruction, and structural assembly. With $n = 1{,}216$ items across eight subcategories, this is the largest single second-level category in the benchmark and constitutes the primary evaluation axis for volumetric spatial cognition.

3D Pose. 3D Pose tasks ($n = 226$) present an object rendered from one viewpoint and ask the model to identify which of several candidate images shows the same object from a different, specified perspective. Given a LEGO model photographed from the front alongside four candidates showing it from the left, right, back, and above, the model must select the image corresponding to the named viewpoint, testing the ability to mentally traverse viewpoint space while preserving structural object identity. These tasks specifically probe whether a model maintains a stable allocentric 3D representation rather than relying on surface-texture recognition.

3D Shape Finding. 3D Shape Finding tasks ($n = 56$) provide a 2D cross-section, silhouette, or set of orthographic projections and ask the model to identify the corresponding 3D solid from a set of candidates. Given three orthographic views (top, front, side) of an object, the model must reconstruct the implicit 3D shape and select the matching solid from four rendered options, exercising shape-from-multiple-views reconstruction reasoning. The relatively small sample count reflects the high difficulty of sourcing unambiguous controlled multi-view stimuli with verified unique 3D solutions.

3D Geometry. 3D Geometry tasks ($n = 349$) cover questions about geometric properties of three-dimensional objects and scenes, including surface area, volume, dihedral angles, face adjacency, and symmetry axes. Given an isometric render of a compound solid composed of unit cubes, the model may be asked to count visible faces or determine the total surface area, requiring simultaneous tracking of face visibility and inter-cube occlusion from a fixed viewpoint. A related variant presents two highlighted components in a 3D scene and asks whether they share a face or are merely adjacent by a vertex, testing precision contact-geometry reasoning from a rendered perspective.

3D Assembly. 3D Assembly tasks ($n = 119$) present a target LEGO structure and an in-progress build state, and ask the model to predict the outcome of adding a specified piece, identify the correct attachment site, or judge whether the piece requires rotation before attachment. Given the current build state (x_1), the next brick (x_2), and the target final object (x_3), the model must select which of five candidates correctly shows the post-addition state, testing forward simulation of incremental 3D assembly under structural connector constraints. These items demand simultaneous maintenance of the current structural inventory, the new piece's geometry, and the valid attachment topology, a three-way compositional spatial reasoning requirement.

3D Rotation. 3D Rotation tasks ($n = 197$) ask the model to identify the clockwise rotation applied to a 3D object between a reference image and one or more rotated alternatives, or to select the image that corresponds to a specified rotation angle. Given a LEGO model in a top-down reference view (x_0) and

four candidates rotated by $30°$, $60°$, $90°$, and $120°$ respectively, the model must identify the image matching the specified target rotation, requiring fine-grained quantitative angular discrimination from rendered 3D images. Unlike standard mental rotation paradigms, these tasks use discrete angle options, enabling direct tests of sensitivity to small angular differences (e.g., $30°$ vs. $60°$) that alter local feature positions but preserve global object identity.

3D Object Detection. 3D Object Detection tasks ($n = 151$) require the model to detect, count, or localise objects within a three-dimensional scene rendered as a static image, going beyond 2D detection to account for depth ordering and inter-object occlusion. Given an isometric LEGO scene with two bricks indicated by red arrows, the model must determine whether they are directly touching or separated, requiring stud-level contact detection within a cluttered 3D environment. A second variant provides a bird's-eye driving scene and asks for the count of a specific vehicle type, testing object enumeration under perspective foreshortening and partial occlusion.

3D Folding. 3D Folding tasks ($n = 68$) present a flat 2D net and ask the model to identify which 3D solid can be formed by folding it, or to determine which edge of the net coincides with a labelled reference edge after folding. Given a hexomino net representing a cube with uniquely coloured faces, four candidate cubes are shown with different face-colour arrangements, and the model must select the one consistent with the net's face adjacency layout, testing a complete two-dimensional-to-three-dimensional mental fold transformation. More complex variants use nets for octahedra or irregular polyhedra, requiring the model to generalise the folding operation beyond the most common training stimuli.

3D Projection. 3D Projection tasks ($n = 50$) present a 3D object with a directional arrow indicating a viewing axis and ask the model to select the corresponding 2D projection from a set of options. Given an isometric rendering of a multi-block solid with an arrow pointing along the depth axis, the model must identify the correct orthographic top-down view, exercising the inverse of 3D-pose reasoning: computing the 2D projection from a known 3D object and viewpoint rather than inferring the viewpoint from paired images. These tasks require the full projection pipeline, occlusion determination, foreshortening, and boundary computation, to be applied to a single static input.

B.2.2 Geometric Reasoning

Geometric Reasoning tasks combine spatial perception with knowledge of geometric relationships, asking models to reason about occluded parts, directional alignments, and structural compatibility from partially informative visual inputs. The $n = 788$ items cover four complementary aspects of geometry-grounded perception.

Occluded Attributes. Occluded Attributes tasks ($n = 374$) present a scene in which one object is partially or fully hidden by another, and ask the model to infer a property of the hidden target object from contextual or associative visual cues. Given an image where a gray bicycle partially occludes a second object, the model must infer the colour or shape of the hidden object based on visible portions, the occlusion boundary, or prior knowledge of object co-occurrence patterns, testing inference under partial observability. These items evaluate whether models can maintain a probabilistic belief about unobserved scene regions using spatial reasoning and world knowledge rather than defaulting to a "cannot determine" response.

Occluder Attributes. Occluder Attributes tasks ($n = 181$) reverse the attribution direction: the object whose properties are unknown is the one doing the occluding, not the one being hidden. Given a scene where the occluding object's surface is in shadow or extends off- frame, the model must infer its shape or category from the silhouette of the occlusion boundary, the visible object edges, or contextual layout clues, testing shadow-geometry and silhouette-shape reasoning. This direction of spatial deduction is a less common evaluation paradigm and complements the more widely studied occluded-attribute inference task.

Direction and Orientation. Direction and Orientation tasks ($n = 145$) provide images containing spatially arranged objects or signs with explicit directional indicators, and ask the model to determine the orientation or facing direction of a specific element relative to a named reference. Given a street scene with one one-way sign facing south and a second sign nearby, the model must determine the cardinal direction the second sign faces, requiring correct allocentric direction composition in the presence of potential egocentric ambiguity. A complementary task shows a building photographed from the north and asks which side of the same building is shown in a second photograph, exercising viewpoint-based compass-direction inference.

Compatibility. Compatibility tasks ($n = 88$) ask whether two or more objects, shapes, or spatial configurations are geometrically compatible, whether they can be combined, fitted together, or

assembled without violating geometric or physical constraints. Given a partial 2D shape alongside several candidate completion pieces, the model must identify which piece can be attached to form a specified target shape without overlap or gap, testing boundary-matching and shape-complementarity reasoning. These tasks require the model to mentally manipulate candidate pieces (including possible rotation) and compare resulting contours to a target silhouette, a process analogous to jigsaw-piece fitting.

B.2.3 2D Reasoning

2D Reasoning tasks operate entirely in the image plane, asking models to reason about flat geometric shapes, planar relationships, and two-dimensional spatial configurations without any requirement for depth inference or viewpoint transformation. All $n = 102$ items fall within a single sub-category.

2D Geometry. 2D Geometry tasks ($n = 102$) present flat geometric figures, diagrams, or spatial patterns and ask quantitative or relational questions about their properties: side lengths, angles, areas, symmetry axes, or spatial arrangements among planar elements. Given a flat irregular figure alongside a dotted fold axis, the model must identify which of four candidate shapes results from folding one portion onto the other, testing planar fold transformation reasoning without any three- dimensional component. A second variant presents a matrix of geometric figures with one cell blank and asks the model to complete the pattern by selecting the missing shape, exercising rule-based 2D spatial sequence reasoning over rows and columns simultaneously.

B.2.4 Positional Relations

Positional Relations tasks require models to identify or describe the precise spatial relationship between two specified scene elements, left/right, above/below, near/far, relative to a clearly stated reference. The $n = 333$ items are partitioned by the pair of entity types whose relationship is being queried, spanning region-to-region, camera-to-region, camera-to-object, and object-to-object configurations.

Positional Relationship (Region – Region). Region–Region tasks ($n = 79$) ask the model to characterise the spatial relationship between two defined scene regions, such as two rooms in a floor plan or two areas in an aerial map. Given a floor-plan diagram, the model may be asked which room lies to the west of a named corridor when the agent is facing south, requiring simultaneous map-reading, reference-frame establishment, and directional relation extraction from a schematic spatial layout. These tasks test allocentric spatial encoding over region-level scene elements rather than individual point objects.

Positional Relationship (Camera – Region). Camera–Region tasks ($n = 81$) ask the model to specify the position of a scene region relative to the camera's viewpoint at the moment of capture. Given a photograph of a billboard and the instruction that the photographer was facing north, the model must determine whether the billboard was to the photographer's left, right, in front of, or behind them, exercising egocentric direction binding between a camera pose and a named scene region. The added complexity relative to Region–Region tasks lies in adopting the camera as a moving, oriented reference entity rather than an absolute spatial anchor.

Positional Relationship (Object – Region). Object–Region tasks ($n = 57$) require the model to determine the spatial relationship between a named object and a broader scene region within a single image. Given an indoor or outdoor scene, the model must identify where a specific object is located relative to a defined region (e.g., left side of the room, upper area of the frame, or background zone). These tasks evaluate the ability to integrate object localisation with coarse regional segmentation, ensuring correct assignment of spatial position under an allocentric frame of reference.

Positional Relationship (Camera – Object). Camera–Object tasks ($n = 86$) narrow the spatial query from a region to a single object, asking the model to determine the object's position relative to the camera frame at a specific moment. Given two consecutive photographs taken during a room walk-through, the model must determine in which direction a specific object (e.g., the stove) lay relative to the photographer at the moment of the second image, requiring dynamic egocentric frame-of-reference updating across a described movement sequence. These tasks combine navigation and object-localisation reasoning within a single egocentric spatial query.

Positional Relationship (Object – Object). Object–Object tasks ($n = 87$) ask the model to express the spatial relationship between two explicitly named objects within a single image, determining whether one is in front of, behind, beside, or above the other. Given an indoor scene with multiple items, the model must identify where a named decorative element lies relative to the main table,

testing direct allocentric object-to-object position extraction from a natural photograph. These items require both accurate object localisation and the correct assignment of a directional preposition under a specified reference convention.

B.2.5 LOCALIZATION

Localization tasks require models to determine the position of an object or scene region in absolute or relative spatial terms, either producing a self-contained location description or selecting from a set of candidate positions. The $n = 100$ items are split equally between relative and absolute localisation paradigms.

Relative Location. Relative Location tasks ($n = 50$) ask the model to express the location of a target entity in terms of its relationship to a named anchor entity rather than in absolute image coordinates. Given a scene image and a specified target object, the model must describe where the target lies with respect to a named reference (e.g., above and to the left of the table), exercising relational spatial language production tied to a verified visual spatial configuration. These tasks evaluate the correspondence between the model's internal spatial representation and its capacity to translate that representation into standard relational directional language.

Object Localization. Object Localization tasks ($n = 50$) ask the model to confirm the presence of a named object in an image and, if present, to identify its approximate spatial location within the scene. Given an aerial or indoor image, the model must verify the existence of a named element and describe its position within the scene layout (e.g., foreground centre, upper-left quadrant), testing joint detection and coarse spatial grounding without requiring bounding-box precision. The emphasis is on correct spatial-region assignment rather than pixel-level localisation.

B.3 KNOWLEDGE-GROUNDED REASONING

Knowledge-Grounded Reasoning tasks require models to apply formal domain knowledge, mathematics, abstract logic, or specialised discipline-specific concepts, in conjunction with visual perception to answer questions that cannot be resolved by spatial intuition alone. With $n = 2{,}332$ items, this is the second-largest domain, and its three second-level categories span formal mathematics, abstract pattern and lateral reasoning, and applied domain-specific knowledge.

B.3.1 MATHEMATICAL REASONING

Mathematical Reasoning tasks present visually rendered mathematical problems and require models to apply formal procedures to reach a numerical or symbolic answer. With $n = 2{,}049$ items across thirteen sub-categories, this is the most internally diverse category in the benchmark and constitutes the primary evaluation axis for quantitative visual reasoning.

Math – Arithmetic. Arithmetic tasks ($n = 107$) present images containing quantities, clocks, timelines, or symbolic addition problems and ask models to perform basic arithmetic to derive a numerical answer. Given an image of a competition timeline specifying the end time and a duration in hours and minutes, the model must compute the start time, requiring modular arithmetic applied to a visually presented temporal problem. These tasks test whether models can extract numerical premises from visual representations such as clocks or schedules and combine them through arithmetic without defaulting to language-prior heuristics.

Geometry – Transformational. Transformational Geometry tasks ($n = 157$) ask models to predict the outcome of a defined geometric transformation, reflection, rotation, translation, or fold, applied to a 2D figure. Given a flat shape alongside a dotted fold axis, the model must identify which of four candidates correctly depicts the result after folding, testing the ability to apply a single formal planar transformation to a novel geometric input. These tasks evaluate whether the model reasons about the transformation as a formal operation or merely recognises results by visual similarity.

Math – Algebra. Algebra tasks ($n = 299$) embed symbolic constraint systems in images, puzzle grids, letter-substitution sums, or visual equation sets, and ask the model to derive one or more unknown values. Given an image of an addition sum in which three letters each represent a distinct digit, the model must determine the sum of those three letters, requiring the formulation and solution of a constraint-satisfaction problem from visual premises. These items combine visual parsing of symbolic notation with algebraic constraint propagation, placing simultaneous demands on both modalities.

Scaling & Ratio. Scaling & Ratio tasks ($n = 143$) present images where two or more quantities are

related by a proportional relationship, and ask the model to determine an unknown by applying ratio or proportion reasoning. Given two floor-plan images alongside a scale bar, the model must estimate the real-world dimension of a specified architectural element, requiring visual ratio extraction and its application to a known reference scale. A second variant shows two objects of different sizes and asks how many of the smaller would fit inside the larger, exercising container-size ratio reasoning.

Geometry – Coordinate. Coordinate Geometry tasks ($n = 77$) present labelled Cartesian coordinate diagrams or function graphs and ask the model to extract, compare, or reason about values at specific coordinates. Given a graph of a mathematical function, the model must identify the x-value at which the function achieves a specified y-value, requiring accurate visual axis-reading combined with function-value inference. These tasks test the intersection of coordinate-axis parsing precision and formal function-interpretation reasoning.

Geometry – Descriptive. Descriptive Geometry tasks ($n = 98$) ask the model to identify or describe formal geometric properties of named shapes, faces, edges, vertices, angles, symmetry, without numerical calculation. Given a 3D polyhedron image with labelled components, the model must identify which net edge coincides with a labelled reference after folding, testing structural component-identity tracking within a descriptive geometric framework. A second variant presents a geometric figure and asks the model to name the type or characterise a property of a specific sub-region, exercising label-to-shape correspondence in formal geometry.

Geometric Reasoning & Visual Problem Solving. This sub-category ($n = 234$) combines pattern-based geometric inference with visual puzzle-solving, requiring models to apply formal geometric rules and perceptual grouping simultaneously to identify missing elements or complete incomplete configurations. Given a 3×3 matrix of geometric figures with one cell blank, the model must infer the row-and-column transformation rule and select the correct completion from candidate options, testing rule induction from visual instances combined with formal geometric transformation knowledge. These items function as a visual analogue to abstract reasoning tests and probe generalisation from a small observed instance set to a systematic rule.

Constraint Geometry. Constraint Geometry tasks ($n = 87$) present a set of geometric constraints such as congruence, parallelism, or fixed angles, and ask the model to determine whether a given configuration satisfies them or to identify the configuration that does. Given a diagram with labelled angles and a question about whether two lines are parallel, the model must apply the parallel-line angle theorems to the visually presented configuration, testing formal geometric inference in a visual proof context. A second variant asks the model to determine the minimum number of conditions required to uniquely specify a geometric configuration, testing constraint- counting meta-reasoning.

Geometric Construction & Optimization. Construction & Optimization tasks ($n = 95$) ask the model to identify the result of a geometric construction procedure or to determine the configuration satisfying a geometric optimisation criterion. Given a diagram annotated with compass-and-straightedge construction steps, the model must identify the resulting construction, a perpendicular bisector, angle bisector, or circumscribed circle, testing knowledge of canonical construction procedures and their visual signatures. These items require recognising that a specific sequence of operations produces a uniquely identifiable geometric entity, going beyond simple shape recognition.

Metric Geometry. Metric Geometry tasks ($n = 272$) require quantitative reasoning about geometric measurements: lengths, areas, perimeters, volumes, and distances between geometric entities in visually presented diagrams or scenes. Given a floor-plan image with scale markings, the model must compute the total perimeter of a specified room in real-world units, requiring scale extraction, unit conversion, and perimeter computation from an architectural layout. A second task type provides a coordinate plane with two marked points and asks the model to compute their Euclidean distance, testing metric formula application to visually presented spatial data.

Shape Decomposition. Shape Decomposition tasks ($n = 205$) present a complex figure and ask the model to identify which simpler component shapes it can be decomposed into, or to identify the compound shape formed by combining two or more given primitives. Given an image of an irregular polygon, the model must determine which combination of rectangles and triangles tiles the original shape without overlap or gap, testing compositional geometric reasoning and area-conservation awareness. These tasks probe whether models represent compound shapes as structured compositions of parts rather than as holistic visual templates.

Circle Geometry. Circle Geometry tasks ($n = 132$) involve circles, arcs, chords, tangents, and inscribed or circumscribed polygons, asking models to compute or reason about angle, arc length, area, or relationship properties. Given a diagram with a circle, a labelled diameter, and an inscribed angle, the model must apply the inscribed-angle theorem to determine the angle's measure, testing knowledge of a specific circle-geometry theorem in a visual application context. A second variant

shows two tangent circles with a shared external tangent and asks for the length of the tangent segment, requiring the combination of construction knowledge with metric computation.

Angle Reasoning. Angle Reasoning tasks ($n = 143$) focus specifically on inferring, computing, or comparing angles in static images of geometric configurations or real-world scenes. Given an image of a leaning building with a visible reference vertical, the model must estimate the angle of inclination from the vertical in degrees, testing angle-estimation reasoning applied to a real-world visual context without explicit measurement tools. A second variant provides a geometric figure with several labelled angles and asks the model to derive an unlabelled angle by applying angle-sum or supplementary-angle rules, testing formal angle computation in a diagram-reading context.

B.3.2 Creative & Abstract Knowledge

Creative & Abstract Knowledge tasks require models to apply non-formal or divergent reasoning strategies, pattern induction, lateral thinking, and misdirection detection, in contexts where standard algorithmic approaches are insufficient or deliberately misleading. The $n = 182$ items span three sub-categories that collectively test flexible and adaptive reasoning from visual premises.

Pattern. Pattern tasks ($n = 59$) present a sequence or matrix of visual stimuli and ask the model to identify the underlying rule and use it to complete or continue the sequence. Given a set of six geometric figures and an instruction to divide them into two groups each sharing a common characteristic, the model must induce the classification criterion from the visual examples alone, testing unsupervised rule extraction from a small, visually presented instance set. These items go beyond simple perceptual grouping to require identifying an abstract structural invariant that holds across every member of each category.

Riddle. Riddle tasks ($n = 73$) embed a logical deduction or elimination puzzle in image form and ask the model to derive the unique correct answer by systematically ruling out inconsistent possibilities. Given a set of five measuring jugs filled to different levels with the stated constraint that four contain identical amounts and one is different, the model must compare the visible water levels to identify the odd jug, requiring quantitative visual comparison and elimination-based logical reasoning. These tasks test whether models apply systematic deduction to visually presented constraint sets rather than guessing by visual salience.

Trick. Trick tasks ($n = 50$) present problems that contain deliberate visual or logical misdirection, where correct answers require resisting the obvious interpretation and identifying the non-obvious solution. Given a narrative-style puzzle with a stated arithmetic component alongside an image that encodes the answer as a visual cue, the model must recognise that the relevant information is embedded visually rather than derivable arithmetically, testing metacognitive awareness about when visual evidence should override computational strategies. These tasks evaluate a form of epistemic flexibility, knowing which reasoning mode to activate, that is not captured by standard mathematical or perceptual evaluation paradigms.

B.3.3 Domain Specific

Domain Specific tasks require models to apply specialised knowledge from engineering, technology, or natural science to interpret visual content that cannot be correctly answered by general spatial or mathematical reasoning alone. The $n = 101$ items are split between two applied knowledge domains.

Engineering & Technology. Engineering & Technology tasks ($n = 51$) present images of technical systems, circuit diagrams, logic-gate schematics, or mechanical assemblies, and ask the model to interpret their function, identify components, or reason about their operational behaviour. Given an image of a logic-gate circuit with labelled inputs, the model must determine the output signal for a specified input configuration, requiring both visual parsing of the gate schematic and application of Boolean logic rules. These items sit at the intersection of visual diagram-reading and formal domain-knowledge application, evaluating whether models have acquired functional knowledge of standard technical representations.

Scientific Reasoning. Scientific Reasoning tasks ($n = 50$) present images from natural or applied science contexts, medical scans, biological process diagrams, or physical experiment setups, and ask the model to apply domain knowledge to interpret the visual content. Given a chest X-ray image with visible anatomical structures, the model must identify which, if any, of a named list of pathological conditions are present, requiring the application of radiological knowledge to a clinically realistic visual input. A second task type provides a biological schematic depicting nutrient transport pathways

and asks where a specific substance is absorbed, testing knowledge- grounded causal reasoning applied to a scientific diagram.

B.4 Relational, Logical & Causal Reasoning

Relational, Logical & Causal Reasoning tasks require models to reason about relationships between entities, whether causal, counterfactual, or purely spatial, going beyond direct perception to demand inferential reasoning about how entities interact or depend on one another. With $n = 455$ items, this domain spans two second-level categories that together evaluate causal inference and explicit spatial relation identification.

B.4.1 Causal Reasoning

Causal Reasoning tasks require the model to reason about cause-and-effect relationships depicted or implied in images, either by predicting what will follow from an observed state or by evaluating counterfactual scenarios in which an aspect of the situation is changed. The $n = 175$ items are divided between counterfactual and physical causality paradigms.

Logical Counterfactual. Logical Counterfactual tasks ($n = 113$) present a scene and ask what would happen if a specific element were removed, changed, or absent, requiring the model to construct a counterfactual world model and reason about its consequences. Given a road scene with a suggested connecting path between two lanes, the question asks whether such a path truly exists under a counterfactual reading of the scene's geometry, testing whether the model can distinguish factual from hypothetical spatial queries. These tasks evaluate whether models maintain a distinct counterfactual reasoning mode rather than merely describing what is visually present.

Physical Causality. Physical Causality tasks ($n = 62$) ask the model to reason about direct physical cause-and-effect relationships that are visible in or inferable from the image. Given an image of a person attempting to hang a very heavy hammock from the overhanging edge of a roof, the model must determine the most likely physical outcome by reasoning about structural load, cantilever mechanics, and material limits, exercising qualitative physics applied to a naturalistic visual scene. These tasks require the model to function as an approximate physical simulator, predicting outcomes from observed structural configurations without access to numerical physics parameters.

B.4.2 Relations

Relations tasks ask models to identify or describe a spatial relationship between two specified entities, testing the core capacity to encode directional, positional, and topological relations as explicit answers. All $n = 280$ items fall within a single sub-category.

Spatial Relationship. Spatial Relationship tasks ($n = 280$) present a scene and ask the model to characterise the spatial relationship between two named entities using directional or relational language. Given an indoor scene with a table and a chair, the model must determine and express whether the chair is in front of, behind, to the left of, or to the right of the table from the camera's perspective, testing precise directional relation extraction between two objects under a specified reference frame. These items represent the most direct evaluation of relational spatial language grounding and serve as a core probe of object-to-object spatial binding capacity.

B.5 Temporal & Dynamic Reasoning

Temporal & Dynamic Reasoning tasks require models to interpret motion, change, or time-ordered sequences, either from multi-image series or from temporal cues embedded in a single frame such as motion arrows, trajectory lines, or described past and future states. With $n = 218$ items, this domain covers temporal sequencing and motion inference across two second-level categories.

B.5.1 Temporal Reasoning

Temporal Reasoning tasks require models to process multi-image sequences or described time series and to reason about temporal order, relative timing, or causal succession of depicted events. The $n = 98$ items are split between movement tracking across frames and discrete event sequencing.

Temporal Movement. Temporal Movement tasks ($n = 51$) present a series of images representing consecutive moments and ask the model to determine the position, direction, or relationship of an

entity at a specific time step, requiring temporal tracking from sequential visual evidence. Given a series of first-person photographs taken while walking through a building, the model is asked to identify the direction of a specific landmark relative to the observer at the moment of the final photograph, requiring integration of successive viewpoint changes across the image sequence. These tasks test dynamic egocentric frame-of-reference maintenance across a multi-step described movement, a capability distinct from single-frame spatial grounding.

Sequencing. Sequencing tasks ($n = 47$) present a set of images depicting different stages of a process and ask the model to arrange them in correct chronological or procedural order. Given four images showing intermediate LEGO assembly states between a defined start and a target, the model must arrange them into the correct build sequence, requiring simultaneous 3D state-matching and procedural temporal ordering. These items probe whether models have an implicit sense of process directionality and can distinguish valid from invalid orderings of incremental state transitions.

B.5.2 Motion Reasoning

Motion Reasoning tasks require models to infer motion properties, direction, trajectory, or relative displacement, from static visual cues such as optical flow indicators, motion arrows, blurred contours, or geometric trajectory markers embedded in individual images. The $n = 120$ items divide between camera motion inference and object motion inference.

Camera Motion. Camera Motion tasks ($n = 70$) provide a pair of images taken from the same scene and ask the model to infer the direction or type of camera movement that occurred between them. Given two photographs of the same room taken in succession, the model must determine from which direction the second photograph was taken relative to the first, inferring camera rotation or translation from the shift in framing, perspective, and the set of visible objects. These tasks require the model to distinguish camera motion from object motion and to maintain a consistent scene geometry across two frames.

Object Motion. Object Motion tasks ($n = 50$) ask the model to determine the direction or trajectory of a moving object depicted in a single static image, using embedded cues such as directional arrows, trajectory lines, or geometric configuration. Given a top-down image of a soccer ball with a directional arrow indicating its path, the model must determine whether the trajectory will cause the ball to enter a goal, requiring 2D path-intersection reasoning between the arrow vector and the spatial boundary of the target region. These tasks test whether models can interpret formal motion representations embedded in otherwise static images and use them for spatial forward prediction.

B.6 Quantitative & Comparative Reasoning

Quantitative & Comparative Reasoning tasks require models to produce or reason with numerical values derived from spatial or visual properties of scenes, including metric distances, relative size comparisons, and discrete object counts. With $n = 503$ items across two second-level categories, this domain assesses numerical spatial grounding spanning estimation, measurement, and enumeration.

B.6.1 Measurement

Measurement tasks require the model to estimate, compare, or compute metric or relative spatial quantities from visual evidence, such as the distance between two objects, the height of a structure, or the extent of a region. The $n = 377$ items are distributed across three sub-categories spanning relative, unit-based, and absolute measurement.

Relative Distance. Relative Distance tasks ($n = 234$) ask the model to compare the distances between pairs of objects and determine which is closer, farther, or proportionally related to a reference length. Given a dashcam image with multiple vehicles and a crosswalk visible ahead, the model must determine how many taxi-cab lengths fit between a building entrance and the crosswalk, requiring proportional spatial estimation from a single perspective-projected frame. These items test whether models can extract and compare distance ratios from 2D projective images without explicit depth or camera calibration data.

Measurement. Unit Measurement tasks ($n = 61$) ask the model to estimate or compute a physical quantity expressed in standard units, length in metres, inclination in degrees, or area in square metres, from visual cues or scale markers. Given an image of a leaning structure with a visible reference object of known height, the model must estimate the structure's angle of inclination in degrees, requiring visual angle-estimation from a naturalistic scene without a protractor or explicit angular

indicator. These tasks test calibrated visual quantification, requiring models to move beyond ordinal comparisons to produce precise numerical estimates grounded in a standard unit system.

Absolute Distance. Absolute Distance tasks ($n = 82$) require the model to estimate the metric distance between two specific entities in real-world units, typically in metres between traffic participants captured in driving scenes. Given a dashcam frame showing the ego vehicle and a named target entity, the model must estimate the straight-line distance in metres between them, requiring an implicit mapping from pixel-space offsets to calibrated real-world metres using learned scene-scale priors. These tasks are among the most challenging quantitative items in the benchmark, as they demand absolute metric calibration from uncalibrated monocular images.

B.6.2 Counting

Counting tasks require models to enumerate discrete objects or category instances within images, either through direct enumeration or by applying a relational filter that restricts the counted set. The $n = 126$ items split between straightforward category-level counting and relation-conditioned counting.

Counting. Direct Counting tasks ($n = 76$) ask the model to count all visible instances of a named object category in the image without any relational qualifier. Given a photograph of a cluttered table with multiple beer bottles at varying degrees of occlusion, the model must produce the exact total count, testing exhaustive visual enumeration under partial occlusion and scene clutter. A second task type presents a building exterior and asks how many chimneys are visible, exercising category-specific object enumeration in a structured outdoor scene.

Relational Counting. Relational Counting tasks ($n = 50$) add a spatial or state qualifier to the counting question, restricting the counted set to objects satisfying a defined relational criterion. Given an image of a bar scene with multiple glasses in various states of fullness, the model must count only the empty glasses that are positioned to the right of a named person, requiring the model to jointly apply a spatial filter and an attribute filter before enumeration. These tasks test whether models can compose relational predicates with counting operations, a substantially more demanding form of enumeration than simple category-based object counting.

B.7 Perceptual Understanding

Perceptual Understanding tasks require models to correctly identify, classify, or describe the visual properties of objects or scenes at the level of direct sensory perception, without requiring spatial transformation, causal inference, or formal mathematical reasoning. With $n = 422$ items, this domain covers both fine-grained object attribute recognition and holistic scene-level classification.

B.7.1 Object-Level Perception

Object-Level Perception tasks focus on the intrinsic visual properties of individual objects, size, colour, texture, shape, and general appearance, as directly observable in the image without requiring spatial reasoning beyond basic object identification. The $n = 372$ items span five attribute-specific sub-categories.

Size & Scale. Size & Scale tasks ($n = 170$) ask models to assess the relative or absolute size of named objects, either by comparing two objects within the same image or by estimating real-world dimensions from contextual scale cues. Given two photographs of different buildings placed side by side, the model must determine which building's first four floors are collectively taller, requiring cross-image size normalisation and comparative height reasoning. These tasks test whether models have calibrated size priors for familiar object categories and can apply them to make consistent cross-instance comparisons.

Attribute. Attribute tasks ($n = 53$) ask the model to identify a specific named property of an object, material, texture, surface finish, or structural category, from visual appearance alone. Given a photograph of a table setting with a spoon and chopsticks visible, the model must determine whether both utensils are made of the same material by comparing their surface reflectance, texture, and sheen, testing cross-object material-attribute comparison under naturalistic conditions. These tasks evaluate whether models have learned visually grounded attribute representations rather than relying on category-level material priors.

Exist or Object Identification. Exist or Object Identification tasks ($n = 50$) ask the model to confirm the presence of a specified object in the scene or to identify which of several candidate objects is

depicted. Given an aerial image, the model must determine whether a named structural element is present and, if so, how many instances are visible, testing basic object detection and binary existence verification. These items form the perceptual foundation on which all higher-order spatial and causal reasoning tasks depend; failures here cascade into all other task categories.

Color. Color tasks ($n = 50$) ask the model to identify the colour of a specified object or region from a naturalistic image, requiring accurate colour perception under variable lighting conditions and potential colour-similarity confounds. Given a tractor image with a distinctively coloured mudguard component, the model must name the mudguard's colour precisely, testing colour identification under real-world lighting variation, partial occlusion, and the presence of nearby objects with similar hues. While perceptually simple in appearance, these tasks surface systematic weaknesses in models whose colour representations are dominated by category-level priors rather than instance-level image-patch analysis.

Appearance. Appearance tasks ($n = 49$) ask the model to describe or classify the general visual appearance of an object or scene in terms that go beyond a single named attribute, encompassing overall form, condition, or stylistic character. Given an image of a building's reflection in water, the model must characterise the perceptual appearance of the reflected image, whether distorted, inverted, rippled, or clear, testing holistic appearance description under a non-canonical viewing transformation. These tasks probe the model's capacity for qualitative perceptual description beyond categorical labelling, evaluating the richness and accuracy of its visual vocabulary.

B.7.2 Scene-Level Perception

Scene-Level Perception tasks require models to form a holistic understanding of the entire image as an integrated scene, classifying or describing its overall character rather than the properties of any individual object within it. All $n = 50$ items fall within a single sub-category.

Scene Understanding. Scene Understanding tasks ($n = 50$) present diverse images, aerial photographs, indoor environments, outdoor landscapes, or specialised domain scenes, and ask the model to classify the primary scene type or characterise the overall context. Given an aerial photograph, the model must determine whether the depicted environment is primarily residential, agricultural, industrial, natural, or urban, requiring integration of global layout cues such as road density, building footprints, and vegetation coverage that are distributed across the entire image rather than concentrated in any single object. These tasks test whether models can form stable holistic scene representations that go beyond the aggregate of individually identified objects, evaluating the top-down perceptual abstraction capacity underlying scene-level cognition.

As described, the taxonomy is designed as a hierarchical decomposition of spatial and multimodal reasoning into six primary domains that collectively span perceptual, geometric, logical, temporal, and knowledge-based cognition. At the highest level, each category isolates a distinct reasoning modality, while maintaining consistent structure to support cross-category comparison. Within each domain, L2 subcategories capture intermediate cognitive operations such as prediction, transformation, relation extraction, or measurement. This separation ensures that semantically similar tasks are grouped even when their surface form differs. L3 task types provide fine-grained operational definitions, making each unit evaluable with clear input-output expectations. The design balances coverage of low-level perceptual tasks with higher-level abstract reasoning, ensuring representation from raw perception to symbolic reasoning. Physical Reasoning explicitly models interaction, navigation, and dynamics, capturing embodied understanding of environments. Positional and Geometric Reasoning expands this with structured spatial inference across 2D, 3D, and relational settings, including occlusion and viewpoint transformations. Knowledge-Grounded Reasoning integrates mathematical, creative, and domain-specific knowledge to test structured reasoning beyond perception alone. Relational, Temporal, and Quantitative categories further isolate causal structure, time-dependent change, and numeric comparison, which are often conflated in prior benchmarks. Perceptual Understanding anchors the taxonomy at the sensory level, grounding higher reasoning in object and scene-level recognition. Overall, the taxonomy is designed to be both exhaustive in coverage and modular in structure, enabling controlled evaluation of distinct reasoning capabilities while minimizing overlap between task families.

C AUTOMATED DATASET CURATION PIPELINE

This appendix provides a complete technical specification of the four-stage automated curation pipeline used to construct FRONTIERSPATIAL from a raw pool of approximately 50,000 candidate instances drawn from 23 source datasets. The pipeline is designed as a *principled, reproducible methodology for holistic benchmark construction* — not merely an engineering description of how our dataset was built, but a transferable framework applicable to any multimodal reasoning domain. Each stage applies explicit, auditable quality criteria, logs the reason for every rejection, and reduces the pool at a different cost-quality tradeoff point. The final benchmark of 7,838 instances represents the intersection of structural validity, visual uniqueness, text-level quality, and multimodal coherence.

Figure 3 summarizes the full pipeline; Table 7 reports the instance counts and reduction at each stage.

Table 7: Instance counts and reductions at each pipeline stage.

Stage	Step	Remaining	Reduced By
Raw pool	—	50,000	—
Stage 1 Structural Heuristics	1a. Metadata integrity	44,402	5,598
	1b. Publication recency	43,852	550
	1c. Question length filter	35,232	8,620
	1d. Exact-text deduplication	26,683	8,549
	1e. Taxonomy remapping & pruning	24,830	1,853
Stage 2 Vision-Space Deduplication	2a+2b. Near-duplicate detection	15,792	9,088
	2c. Visual domain clustering	12,914	2,878
Stage 3 Text-Only LLM Scoring	R1. Spatial relevance (≥ 3)	8,814	4,100
	R2. Cognitive depth (≥ 2)	5,857	2,957
	R3. Answerability (≥ 3)	1,943	3,914
	R4. Language clarity (≥ 3)	9,812	4
Stage 4 Multimodal LLM Final Pass	R5. Visual necessity	9,411	405
	R6. Image-question coherence	8,912	499
	R7. Spatial signal quality	7,838	1,074
Final benchmark	—	**7,838**	**42,162**

The four stages operate at increasing computational cost and decreasing filter aggressiveness. Stages 1 and 2 are deterministic and LLM-free; Stage 3 uses text-only LLM scoring at low cost; Stage 4 applies full multimodal LLM evaluation to the surviving pool. This design ensures that expensive computation is reserved only for candidates that have already passed cheaper quality gates.

C.1 STAGE 1: STRUCTURAL HEURISTICS (60K+ $\rightarrow$ 24,830)

Stage 1 applies five deterministic, LLM-free filters in sequence. These filters are fast, cost-free, and eliminate structurally invalid or redundant instances that would otherwise waste LLM budget in downstream stages. Each filter logs the rejection reason per instance, making the pipeline fully auditable.

1a. Metadata integrity ($-$5,598). Any instance is discarded if (i) its associated image path is missing or does not resolve to a valid file on disk, (ii) for multiple-choice questions, the annotated answer is not a member of the provided choices list, or (iii) for open-ended questions, the answer field is empty or null. This filter removes broken instances silently inherited from source datasets — a pervasive issue when aggregating across heterogeneous annotation pipelines.

1b. Publication recency ($-$550). Instances originating from datasets published before 2020 are flagged for review. Items retained from pre-2020 datasets must belong to landmark collections with no modern equivalent; all others are removed. This filter eliminates annotation artifacts from older VQA practices, including inconsistent spatial labelling and outdated visual content.

1c. Question length filter (−8,620). A token-length window of $[10, 200]$ tokens is applied to all question strings. Questions below the lower bound (e.g., "*What is here?*") lack the specificity needed for spatial reasoning evaluation. Questions above the upper bound frequently embed the answer in an extended preamble or introduce text-only solvability — a shortcut that circumvents visual inference. This single filter accounts for the largest reduction in Stage 1.

1d. Exact-text deduplication (−8,549). Each question string is lowercased, whitespace-stripped, and SHA-256 hashed. Instances sharing a hash across source datasets are treated as exact duplicates; only one copy is retained, preferring the instance with richer metadata or higher-resolution imagery. This step is particularly important given that several source datasets in FRONTIERSPATIAL share visual question sources (e.g., MMSI-Bench appears in the pool under multiple dataset identifiers).

1e. Taxonomy remapping and non-spatial pruning (−1,853). Each instance's source category label is mapped to FRONTIERSPATIAL four-level taxonomy (Section **??**). Instances that cannot be mapped to any of the 65 tasks in our schema after remapping are removed as non-spatial. This step constitutes the most intellectually substantive filter in Stage 1: it operationalizes the taxonomy as a quality gate, ensuring that every surviving instance is anchored to a defined spatial reasoning skill. For any new domain, only the remapping rules require redefinition — the remaining Stage 1 filters are domain-universal.

C.2 STAGE 2: VISION-SPACE DEDUPLICATION (24,830 → 12,914)

Stage 2 addresses a limitation unaddressed by any prior spatial benchmark paper: when a benchmark is constructed from multiple source datasets, near-identical images frequently appear under different question phrasings. Text-level deduplication (Stage 1d) cannot detect these cases. Stage 2 operates entirely in visual embedding space.

2a. Image encoding. All unique images referenced across the surviving 24,830 instances are encoded using a CLIP ViT-L/14 encoder. For multi-image instances, a mean-pooled embedding is computed across the image list. All embeddings are stored in a FAISS flat inner-product index for efficient retrieval.

2b. Near-duplicate detection (−9,088). Cosine similarity is computed between all image embedding pairs. Instance pairs with similarity > 0.97 are treated as near-duplicates — corresponding to the same scene captured under minor viewpoint or lighting variation, or two datasets that independently cropped the same source image. Within each duplicate cluster, the instance with the higher composite text-quality signal (question length $\times$ answer specificity) is retained; all others are discarded. This step reveals that approximately 37% of the Stage 1 survivor pool contains near-duplicate visual content across source boundaries — a cross-dataset redundancy that would systematically inflate model performance if left uncorrected.

2c. Visual domain clustering (−2,878). K-means clustering ($K = 25$) is applied to all image embeddings to identify visual domain clusters (e.g., indoor scenes, outdoor scenes, aerial imagery, synthetic/abstract, medical). Any cluster comprising more than 15% of the post-deduplication pool is subject to stratified downsampling to enforce visual domain balance. This step prevents the benchmark from being dominated by a single visual context — a subtle failure mode of multi-source benchmarks that existing spatial evaluation suites have not explicitly addressed. The resulting pool of 12,914 instances exhibits a balanced visual domain distribution, verified by cluster entropy before proceeding to Stage 3.

C.3 STAGE 3: TEXT-ONLY LLM SCORING (12,914 → 9,812)

Running 12,914 instances through a multimodal LLM would be prohibitively expensive. Stage 3 instead applies text-only LLM scoring — operating on the question, answer choices, category label, and metadata — at approximately one-tenth the cost of full multimodal evaluation. Each instance receives four scores in a single structured JSON output call, rated on a 1–4 ordinal scale.

R1. Spatial relevance (threshold ≥ 3, $-4{,}100$). *Does answering this question require spatial reasoning?*

1 Pure factual recall with no spatial component.
2 Spatial vocabulary present but answerable without spatial reasoning (e.g., "name an object to the left").
3 Requires interpreting spatial relationships from a scene.
4 Requires multi-step spatial reasoning (e.g., composing depth, occlusion, and orientation inferences).

This criterion filters questions that were tagged as spatial in source datasets but are answerable by language statistics alone — a documented failure mode of VQA benchmark construction (Goyal et al., 2017; **?**).

R2. Cognitive depth (threshold ≥ 2, $-2{,}957$). *What level of cognitive processing is required?*

1 Pure perception or object detection.
2 Spatial relation understanding.
3 Spatial transformation or inference.
4 Multi-step spatial planning or causal reasoning.

This score serves a dual purpose: as a quality gate (threshold ≥ 2) and as a difficulty signal for downstream taxonomic balancing. A benchmark composed exclusively of level-4 items is maximally hard but minimally diagnostic; maintaining a distribution across levels enables fine-grained model profiling.

R3. Answerability (threshold ≥ 3, $-3{,}914$). *Can this question be answered from the visual scene without external world knowledge?*

1 Requires domain-specific knowledge (medical, engineering).
2 Requires cultural or encyclopedic context.
3 Answerable from scene geometry and visual cues alone.
4 Purely spatial; no knowledge shortcut is available.

Items failing this criterion (score ≤ 2) are discarded entirely, as they conflate spatial reasoning with knowledge retrieval.

R4. Language clarity (threshold ≥ 3, -4). *Is the question unambiguously and grammatically phrased?*

1 Translation artifacts, encoding errors, or garbled text.
2 Ambiguous referents or underspecified spatial predicates.
3 Clear with minor phrasing issues.
4 Perfectly precise spatial language.

Instances scoring 1 or 2 are discarded. The near-zero reduction at this step (-4) reflects the effectiveness of Stages 1a–1d in eliminating malformed text, but the criterion remains essential as an explicit quality guarantee. An instance must pass all four thresholds simultaneously to proceed to Stage 4.

C.4 Stage 4: Multimodal LLM Final Pass (9,812 $\rightarrow$ 7,838)

Stage 4 applies the most expensive quality gate: full multimodal LLM evaluation. Having eliminated structural failures, visual duplicates, and text-level quality issues in prior stages, only 9,812 instances reach this point — making full multimodal scoring computationally feasible. Each instance is evaluated on three criteria requiring joint visual and linguistic reasoning.

R5. Visual necessity (threshold: discard if answerable without image, −405). The LLM is presented with the question and answer choices *without* the associated image. If the model selects the correct answer with high confidence from text alone, the instance is discarded as harbouring a language shortcut that circumvents visual spatial inference. This is the multimodal instantiation of R3 from Stage 3 — now empirically testable rather than heuristically estimated. It directly addresses the "language bias" problem documented across VQA benchmarks (Goyal et al., 2017; **?**; Kosoy et al., 2025).

R6. Image-question coherence (−499). The LLM is asked to verify that the image contains the spatial scene, objects, and relations referenced by the question. Instances where the LLM cannot identify the referenced entities in the image — indicating a misaligned question-image pair inherited from source datasets — are discarded. This filter is particularly important for instances drawn from general-purpose VQA corpora where spatial questions were not authored with their specific images in mind.

R7. Spatial signal quality (−1,074). The LLM scores the richness of spatial structure in the image on a 1–4 scale:

1 Flat, text-only, or isolated object with no scene context.

2 Minimal spatial structure (one foreground object, plain background).

3 Multi-object scene with discernible spatial relationships.

4 Rich spatial structure: depth cues, occlusion, perspective, or complex multi-object arrangements.

Instances scoring 1 or 2 are removed. This criterion serves as the final ranking signal within taxonomy cells: when a category exceeds its target quota, the lowest-scoring R7 instances are pruned first. The resulting 7,838-instance benchmark exhibits consistently high spatial signal quality across all 65 task types.

C.5 Generalizability of the Pipeline

The curation pipeline is designed to be domain-transferable. The structural heuristics of Stage 1 are universally applicable to any multimodal benchmark pool. The visual embedding operations of Stage 2 require no domain-specific configuration beyond the choice of encoder. The LLM scoring rubrics of Stages 3 and 4 are parameterized by the target reasoning taxonomy: to adapt the pipeline to a new domain (e.g., temporal reasoning, causal reasoning), only the R1 spatial relevance criterion and the taxonomy remapping rules of Step 1e require redefinition. All other filters transfer without modification.

We validate the pipeline's contribution empirically through a staged ablation: benchmark statistics and model performance are measured at each checkpoint (50K, 24K, 12K, 9K, 7.8K), demonstrating that each stage independently improves benchmark quality — reducing model score variance, increasing the human-model performance gap, and improving per-skill diagnostic specificity. These ablation results are reported in Appendix **??**.

Full code and running details of the pipeline is released and available in the submitted codes.

D AUTOMATED EVALUATION PIPELINE

This appendix provides a complete technical specification of HARNESSEVALUATOR, the three-stage automated pipeline used to score, cluster, and statistically compare model outputs across all spatial reasoning benchmarks reported in the main paper. The pipeline is designed around three constraints: (i) *scalability* to hundreds of tasks and over a dozen models without human re-annotation at each run; (ii) *statistical rigour*, with all pairwise comparisons corrected for multiple testing and effect sizes reported alongside significance; and (iii) *resumability*, so that partial runs caused by API rate-limits or hardware interruptions do not require reprocessing already-scored rows. Section D.1 describes the overall architecture. Sections D.2–D.3 detail each stage. Section D.4 documents the output schema. Section D.5 discusses known limitations and design trade-offs.

D.1 PIPELINE OVERVIEW

HARNESSEVALUATOR takes as input a root directory containing one sub-folder per evaluated model, each holding an `agent_response.csv` file with columns `id`, `task_description`, `answer`, and `response`. The pipeline then runs three sequential stages for every model folder and produces per-model and cross-model output artefacts:

Stage 1. Structured Signal Extraction. An LLM judge scores each response on four binary dimensions and writes results to `agent_scores.csv`.

Stage 2. Error Taxonomy Induction. Failed responses are embedded, reduced, clustered, and each cluster is assigned a taxonomy label from the E1–E7 spatial error vocabulary.

Stage 3. Cross-Model Diagnostic Synthesis. Bootstrap confidence intervals, McNemar tests (Bonferroni-corrected), and Cohen's d effect sizes are computed per model and per error type; results are exported as `model_evaluations.csv` and `model_evaluations.json`.

The entry point is the `HarnessEvaluator` constructor, which accepts four arguments summarised in Table 8.

Table 8: HARNESSEVALUATOR API parameters.

Parameter	Default	Description
`results_path`	*required*	Root directory; one sub-folder per model, each containing `agent_response.csv`.
`max_concurrent`	`10`	Thread-pool size for parallel LLM scoring in Stage 1.
`model_name`	`Gemini-3.1-Flash`	OpenAI-compatible model identifier for both scoring and cluster labeling.

D.2 STAGE 1 — STRUCTURED SIGNAL EXTRACTION

Motivation. A single binary accuracy label is insufficient to localise *where* in the perception–grounding–reasoning pipeline a failure occurs. Stage 1 therefore decomposes each evaluation into four binary dimensions that collectively span the sub-processes required for correct spatial question answering.

Scoring dimensions. For each row in `agent_response.csv`, the LLM judge is prompted with the task description, ground-truth answer, and model response under a zero-shot structured prompt that requests a strict JSON object. Table 9 defines the four dimensions.

Implementation. Scoring is parallelised over a `ThreadPoolExecutor` with `max_concurrent` workers; each worker submits a single chat-completion request and retries up to five times on transient failures before logging the row ID and returning the original unscored row. Incremental saving is performed after each completed future: the partial result frame is written to disk immediately, so that a process crash does not lose already-completed scores. On re-runs, HARNESSEVALUATOR reads any existing `agent_scores.csv`, identifies rows with a non-null `accuracy_score`, and passes only the unscored rows to Stage 1, reducing redundant API calls.

Table 9: Four binary scoring dimensions produced by Stage 1.

Dimension	Range	Definition
`accuracy_score`	$\{0, 1\}$	Final answer is correct (semantic equivalence for open-ended items).
`spatial_elements_score`	$\{0, 1\}$	Key spatial objects or entities are correctly identified.
`spatial_relations_score`	$\{0, 1\}$	Spatial relations (left/right/above/below/near/far) are correctly stated.
`spatial_reasoning_score`	$\{0, 1\}$	Logical spatial reasoning chain leading to the answer is correct.

Judge prompt design. The scoring prompt instructs the judge to return *only* a valid JSON object with no markdown fences, no trailing commas, and `null` (not `None`) for missing values. Temperature is fixed at 0.1 to minimise scoring variance; `max_tokens` is set to 1,000 to accommodate the optional free-text `reasoning` field. For multiple-choice items, correctness is determined by exact string match after normalisation; for open-ended items, the judge applies semantic equivalence, which provides tolerance for paraphrase variants of a correct answer.

D.3 Stage 2 — Cross-Model Diagnostic Synthesis

Bootstrap confidence intervals. For each model, 500-resample bootstrap confidence intervals (95% CI) are computed over `accuracy_score` using the percentile method:

$$\widehat{\mu}_{\text{boot}} = \frac{1}{B}\sum_{b=1}^{B} \bar{X}_b, \qquad \text{CI} = \left[\hat{q}_{2.5},\ \hat{q}_{97.5}\right], \tag{1}$$

where $\bar{X}_b$ is the mean accuracy of the b-th bootstrap resample and $\hat{q}_p$ denotes the p-th empirical percentile over $\{\bar{X}_b\}$. NaN values are excluded before resampling; models with zero valid rows receive `mean = NaN`.

Pairwise McNemar test. All $\binom{N}{2}$ model pairs are compared on `accuracy_score` using the continuity-corrected McNemar statistic (**?**):

$$\chi^2_{\text{Mc}} = \frac{(|b - c| - 1)^2}{b + c}, \tag{2}$$

where b (resp. c) counts instances on which model 1 succeeded and model 2 failed (resp. the reverse), computed on the intersection of task IDs shared by both models. The p-value is obtained from the $\chi^2(1)$ distribution. If $b + c = 0$ (models agree on all items), the statistic is set to 0 and $p = 1$.

Bonferroni correction. The family-wise significance threshold is $\alpha_{\text{Bonf}} = 0.05 \,/\, \binom{N}{2}$, applied uniformly across all pairwise comparisons. Pairs with $p < \alpha_{\text{Bonf}}$ are flagged as statistically significant.

Effect size. Cohen's d is computed for each ordered pair (m_1, m_2):

$$d = \frac{\bar{X}_{m_1} - \bar{X}_{m_2}}{s_p}, \qquad s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}, \tag{3}$$

where s_i^2 is the sample variance of model i's score vector (ddof=1). The reverse pair (m_2, m_1) is assigned $d' = -d$, so the full $N \times N$ effect-size matrix is anti-symmetric.

Per-error-type statistics. The above pairwise statistics are also computed *independently for each taxonomy code* $e \in \{\text{E1}, \ldots, \text{E7}\}$, using the cluster-level binary column (1 if a row was assigned to that cluster, 0 otherwise) as the score variable. This yields a structured view of which model pairs differ significantly on specific error types rather than only globally, directly supporting the per-category analysis in Appendix I.

D.4 Output Schema

`agent_scores.csv`. Written per model folder. Contains all columns from the original `agent_response.csv` plus the four binary scores from Stage 1, the HDBSCAN `cluster` integer (-1 for noise), and the `error_type` string from Stage 2.

`model_evaluations.csv`. One row per `(model, metric, error_type)` triple. The `to` column is populated for pairwise rows and is `None` for per-model summary rows. Valid values of `metric` are: `Score`, `Refusal Score`, `Off Topic Score`, `Spatial Elements Score`, `Spatial Relations Score`, `Spatial Reasoning Score`, `bootstrap_mean`, `cohens_d`, `bonferroni_alpha`, `mcnemar_stat`.

`model_evaluations.json`. Nested dictionary keyed by model name. Top-level keys per model: `bootstrap_ci` (dict with `mean`, `ci_lower`, `ci_upper`, `std`), `average_score`, `overall_score`, `refusal_score`, `off_topic_score`, `spatial_elements_score`, `spatial_relations_score`, `spatial_reasoning_score`, `Cohen_d`, `Bonferroni`, `Mcnemar`. The last three keys are themselves dicts keyed by error type, each containing a list of {`"to": model_name, ...`} records.

D.5 Limitations and Design Trade-offs

LLM judge variance. Stage 1 scoring accuracy depends on the judge model's calibration on spatial reasoning tasks. We fix temperature at 0.1 and enforce strict JSON output to minimise inter-run variance, but absolute scoring accuracy has not been validated against human raters on this benchmark; we treat judge-assigned labels as high-precision proxies subject to the same caveats as any automated evaluator (**?**).

Clustering on small failure sets. HDBSCAN with minimum cluster size $k \geq 15$ requires at least ≈ 50 failed rows to produce at least two valid clusters. For models with very high accuracy (few failures), Stage 2 may return no clusters, leaving all rows with `error_type = None` and excluding them from per-type pairwise statistics. This is logged but not treated as an error.

Row alignment for pairwise statistics. McNemar's test and Cohen's d require instance-level alignment by `id`. If two models were not evaluated on an identical task set, pairwise statistics cover only the intersection; the intersection size is not currently surfaced in the output schema and should be verified before interpreting small-n pairings.

Known implementation bug. In `structured_signal_extraction`, `error_logs` is a `list` but is passed directly to `f.write()`, which raises a `TypeError` at runtime if any errors are logged. The fix is to replace `f.write(error_logs)` with `f.write("\n".join(error_logs))`. All results reported in this paper were produced on runs where no errors were logged, so this bug did not affect any reported scores.

E Automated Error Analysis Pipeline

Benchmark papers in visual reasoning almost universally report accuracy summarised at the level of categories or aggregate datasets. Such tables are observational: they record *what* a model gets wrong but offer no mechanism for explaining *why*. This conflation of diverse failure mechanisms under a single accuracy figure obscures model behavior in ways that matter for both scientific understanding and practical model development.

FrontierSpatial's Automated Error Analysis Pipeline remedies this by supplying three scientific properties that do not currently coexist in any published spatial benchmark paper. *First*, a bottom-up inductive error taxonomy is derived from actual model failures rather than imposed from prior theoretical frameworks, making the taxonomy specific to the failure distributions observed in practice. *Second*, model comparison is conducted at the level of individual skill dimensions rather than aggregate accuracy, supported by appropriate inferential statistics. *Third*, the entire analysis harness is implemented as a fully reproducible, LLM-in-the-loop pipeline: every diagnostic claim produced by the system is directly traceable to structured data, and all intermediate artefacts are released alongside the final report so that every numerical finding is independently verifiable.

The pipeline consists of three sequential stages—Stage 1: structured signal extraction; Stage 2: error taxonomy induction; and Stage 3: cross-model diagnostic synthesis—described in detail in the following subsections.

E.1 STAGE 1: STRUCTURED SIGNAL EXTRACTION

Input format. STAGE 1 accepts as input either a JSON or CSV file with, at minimum, the fields `id`, `question`, `choices` (for multiple-choice questions), `answer` (ground truth), `prediction` (model output), `category`, and `sub_category`. An optional `reasoning_trace` field is consumed when chain-of-thought prompting was used during inference.

Scientific motivation. The standard error analysis in benchmark papers reduces every prediction to a binary *correct* / *incorrect* flag. This is observational, not diagnostic. Before failures can be classified by cause, it is necessary to extract structured signals from raw prediction strings that capture how the failure manifested, not merely that it occurred. Signal extraction transforms each unstructured prediction record into a typed, multi-attribute object amenable to systematic downstream analysis.

E.1.1 SIGNAL S1: CORRECTNESS

For multiple-choice items, correctness reduces to a deterministic string match of the predicted choice letter against the gold answer. For open-ended (OE) items, correctness requires a semantic equivalence check, implemented via an LLM judge with a constrained prompt:

> *"Does the following prediction express the same spatial relationship as the gold answer? Answer `yes`/`no` and provide a one-sentence justification."*

The judge is invoked only for OE items scored as incorrect by string match—approximately 15–20% of benchmark items under the MCQ/OE balance from FRONTIERSPATIAL curation pipeline—making the approach cost-efficient without sacrificing accuracy on the cases where it matters.

E.1.2 SIGNAL S2: ANSWER FORMAT VALIDITY

The prediction string is parsed for structural validity independent of semantic correctness. Four output classes are distinguished: **valid_choice** (a recognisable answer token was produced), **refused** (the model explicitly declined to answer), and **off_topic** (the model produced a multi-sentence explanation in place of an answer). This classification is implemented purely with rule-based string analysis and requires no judge calls.

The diagnostic motivation for this separation is important: a refusal is a qualitatively different failure from a wrong answer. Conflating them in a single accuracy number obscures the distinction between models that attempt a spatial inference and fail versus models that decline to commit to an answer. The former reveals limits of spatial competence; the latter may reflect alignment training, calibration policies, or prompt sensitivity—none of which is informative about spatial reasoning capacity per se.

E.1.3 SIGNAL S3: REASONING TRACE COHERENCE

When a model was prompted with chain-of-thought, the extracted reasoning text is scored by a text-only LLM judge on three independent sub-criteria, each scored on the binary scale $\{0, 1\}$:

(a) **Element identification:** does the reasoning correctly identify the relevant spatial elements named in the question?

(b) **Relationship description:** is the spatial transformation or relationship correctly described in the reasoning steps?

(c) **Logical consistency:** does the stated conclusion follow logically from the stated reasoning?

The composite score $s_3 = \sum_{i=a}^{c} s_i \in \{0, 1, 2, 3\}$ supports detection of two diagnostically important failure modes that aggregate accuracy cannot distinguish. A model may arrive at a correct answer through internally inconsistent reasoning ($s_1 = 1$ but $s_3 < 3$), which constitutes a latent reasoning failure that will surface as errors under slight question reformulation. Conversely, a model may reason correctly but produce a malformed answer token ($s_1 = 0$ but $s_3 = 3$), which is a formatting failure rather than a spatial reasoning failure and should not penalise the model's competence estimate.

E.1.4 PRE-CLASSIFICATION INTO ERROR POOLS

The output of STAGE 1 is an enriched prediction file in which every item carries its full signal record. Incorrect items are automatically sorted into three mutually exclusive pools based on signal combinations, which direct STAGE 2's analysis budget toward the most informative failures:

- **Pool A** — *Format errors*: items where S2 produced a non-`valid_choice` label. These are format failures independent of spatial reasoning content.
- **Pool B** — *Reasoning errors*: items with a valid format (S2 passed) where the reasoning coherence score is $s_3 < 3$ (S3 failed). These are process failures where the model's answer-generation mechanism broke down despite producing a parseable output.
- **Pool C** — *Knowledge/perception errors*: items where S1 failed but S2 and S3 both passed. These are the most diagnostically rich failures: the model produced a valid, coherently reasoned answer that was nevertheless wrong.

Pool membership determines how STAGE 2's embedding and clustering are applied, ensuring that qualitatively different failure modes are not merged in a common semantic space during taxonomy induction.

E.2 STAGE 2: ERROR TAXONOMY INDUCTION

Scientific motivation. Manually designed error taxonomies impose a theoretical structure that may or may not correspond to the actual distribution of failures produced by a given model on a given benchmark. Distinct models evaluated on the same benchmark can exhibit qualitatively different failure patterns, and a fixed *a priori* taxonomy risks either over-aggregating fine-grained patterns or missing failure modes that were not anticipated during taxonomy design. STAGE 2 adopts an inductive approach: the taxonomy is derived from the failures themselves, not from theory. This methodology is inspired by grounded theory coding in qualitative research and translated into an automated pipeline that combines embedding-based clustering with LLM-based labelling.

E.2.1 STEP 2A: ERROR EMBEDDING

For each incorrect item, a composite text string is constructed by concatenating the question, the gold answer, and the model prediction. This string is encoded using a strong text embedding model—either `e5-large-v2` or OpenAI `text-embedding-3-large`.

The key design decision is in what is embedded. The concatenation encodes the *failure pattern*—the semantic relationship between what was asked, what was correct, and what was predicted—rather than the question topic in isolation. As a consequence, clusters in the resulting embedding space reflect similar shapes of mistake rather than similar question domains. Two items from unrelated benchmark categories that elicit the same type of spatial confusion will be proximate in this space; two items from the same category that reflect different failure mechanisms will be distant. This property is what allows STAGE 2 to discover cross-cutting failure types that category-level accuracy tables cannot reveal.

Embeddings are computed separately for each error pool (**Pool A**, **Pool B**, **Pool C** from STAGE 1), so that format errors, reasoning errors, and knowledge errors cluster within their own semantic spaces rather than interfering with one another.

E.2.2 STEP 2B: DENSITY-BASED CLUSTERING

HDBSCAN **?** is applied to the per-pool embeddings in preference to K-means for two reasons grounded in the statistical properties of failure distributions. First, failure patterns are heterogeneous in density: some failure modes are highly consistent (tight geometric clusters reflecting a stereotyped error) while others are diffuse (scattered wrong answers with no single dominant pattern). HDBSCAN accommodates variable-density clusters without requiring a fixed global linkage threshold. Second, HDBSCAN explicitly designates noise points—items that do not belong to any cluster—corresponding to idiosyncratic failures that represent genuinely unusual model behavior rather than systematic error patterns.

The minimum cluster size hyperparameter is selected by a grid search over the range $\{10, 15, 20, 25, 30, 40, 50\}$, choosing the value that maximises the mean silhouette coefficient:

$$\bar{s} = \frac{1}{|\mathcal{C}|} \sum_{x \in \mathcal{C}} \frac{b(x) - a(x)}{\max(a(x), b(x))}, \tag{4}$$

where $a(x)$ is the mean intra-cluster distance for point x and $b(x)$ is the mean distance to the nearest cluster not containing x.

E.2.3 Step 2c: LLM Taxonomy Labelling

For each discovered cluster, 10 representative items nearest to the cluster centroid are sampled and passed to an LLM with the following structured prompt:

> *"Here are 10 examples of a specific pattern of model failures on a spatial reasoning benchmark. Each shows the question, the correct answer, and the model's wrong prediction. Identify the common failure mode, give it a precise descriptive name (5 words maximum), and provide a one-sentence definition. The name should describe what the model did wrong, not the topic of the question."*

This produces a human-readable label for each cluster that is grounded in actual failure instances rather than constructed from summary statistics. The constraint to name the failure mode rather than the question topic is critical: it prevents the LLM from producing category-level labels (e.g., "rotation questions") that would simply reproduce the existing benchmark taxonomy.

E.2.4 Step 2d: Mapping to the Reference Error Taxonomy

The inductively discovered clusters are mapped to a reference error taxonomy of seven types fixed *before* any model evaluation is conducted. This two-level approach—inductive discovery followed by mapping to a fixed schema—balances flexibility (the inductive step captures model-specific failure patterns) with comparability (the fixed schema enables consistent cross-model comparison). The seven reference types are:

E1 — Perceptual failure
The model misidentified or failed to detect a spatial element present in the image. The error occurs before any spatial inference is attempted.

E2 — Grounding failure
The model identified relevant visual elements but failed to anchor them in the correct spatial frame of reference (e.g., confused viewer-centric and object-centric reference frames).

E3 — Spatial transformation failure
The model understood the input spatial arrangement but failed to correctly apply a required rotation, reflection, or perspective change.

E4 — Quantitative spatial failure
Failure on tasks requiring counting, distance estimation, or spatial measurement, where the error is in quantification rather than spatial concept.

E5 — Multi-step reasoning failure
The model solved one spatial inference step correctly but failed on a subsequent step that required the output of the first step as input—a compositional failure rather than a single-step one.

E6 — Language–spatial mapping failure
The model misunderstood the spatial language in the question (e.g., confused "left of X" with "right of X", or misinterpreted a relative direction as absolute), producing an inference that is internally valid but addresses the wrong spatial relationship.

E7 — Format / refusal failure
The model did not produce a valid answer token regardless of spatial competence. This captures **Pool A** items that reach the taxonomy stage.

The LLM mapping from discovered clusters to reference types is validated by a human reviewer on a 50-item stratified sample, computing inter-rater agreement via Cohen's κ. A target of $\kappa \geq 0.75$ is required before the labeled dataset is accepted for use in STAGE 3; items on which the LLM mapping is rejected by the reviewer are relabelled manually and logged in the validation report.

Scientific contribution of Stage 2. The central finding that STAGE 2 enables—and that no per-category accuracy table can reveal—is that error types do not partition cleanly along benchmark category lines. A model may fail on a category such as *3D Pose* for reasons distributed across **E1**, **E3**, and **E5** simultaneously, while failing on *Camera Motion* predominantly through **E2**. This cross-cutting structure is invisible in standard benchmark tables and constitutes a substantive empirical finding about the relationship between spatial task structure and model failure mode.

E.3 STAGE 3: CROSS-MODEL DIAGNOSTIC SYNTHESIS

Scientific motivation. A single model's error profile is descriptive. Error profiles compared consistently across a population of 20–25 models are *generative*—they identify which error types are universal (every model fails on **E3** at comparable rates, indicating genuine benchmark difficulty) versus model-specific (one model fails disproportionately on **E2** while another fails on **E5**, revealing architectural or training-data differences). STAGE 3 is the analytical engine that transforms N individual error profiles into cross-cutting findings about the current state of spatial reasoning in vision-language models.

E.3.1 STEP 3A: PER-MODEL SKILL PROFILING

Each model's error labels from STAGE 2 are aggregated into a 7-dimensional *error profile vector* $\mathbf{p} \in [0, 1]^7$, where p_k is the proportion of that model's failures attributable to error type **Ek**. Complementarily, a *skill accuracy vector* $\mathbf{a} \in [0, 1]^7$ is computed, where a_k is the model's accuracy on all items for which **Ek** is the primary challenge. These two views are informationally distinct: the error distribution describes what the model gets wrong; the skill accuracy describes what it can do on the hardest instances of each sub-competency.

Both vectors are computed at four levels of granularity: (i) overall, (ii) per main benchmark category, (iii) per sub-category, and (iv) per difficulty bin using the difficulty ratings from FRONTIERSPATIAL curation pipeline. All outputs are stored as structured JSON per model to support programmatic downstream analysis and automated report generation.

E.3.2 STEP 3B: STATISTICAL RIGOR LAYER

Three statistical procedures are applied systematically to ensure that comparative claims have quantifiable inferential support:

McNemar's test for paired item comparison. For each pair of models (M_i, M_j), McNemar's test is applied to the binary correctness vectors across all shared items. This tests whether the two models make *statistically different errors on the same items*, which is more informative than comparing aggregate accuracy: two models may have identical overall accuracy yet systematically fail on different items, indicating complementary rather than equivalent spatial competencies.

Bootstrap confidence intervals. Confidence intervals on each model's per-error-type accuracy are computed via nonparametric bootstrap with $B = 1{,}000$ resampling iterations. This produces exact 95% confidence intervals without distributional assumptions, which is appropriate given the heterogeneous sample sizes across error types and benchmark categories.

Effect size with multiple-comparison correction. For each pair of models on each skill dimension, Cohen's d is computed as the effect size. Significance is reported at $\alpha = 0.05$ with Bonferroni correction for the number of simultaneous pairwise comparisons. This correction ensures that the reported p-values account for the family-wise error rate under the full comparison matrix. A representative example of a reportable finding takes the form:

> *"Model A significantly outperforms Model B on **E3** (spatial transformation), Cohen's $d = 0.43$, $p < 0.001$ (Bonferroni-corrected)."*

This formulation grounds every diagnostic claim in a specific statistical result that is fully reproducible from the released data.

E.3.3 STEP 3C: LLM DIAGNOSTIC SYNTHESIS

To produce a human-readable, actionable report rather than a data dump, each model's full profile is passed to an LLM under a structured prompt that includes: the per-category accuracy table, the 7-dimensional error distribution, the top-3 most frequent cluster labels from STAGE 2, cross-model rank on each skill dimension, and the statistical contrast results from Step 3b. The LLM synthesises this into 3–5 structured findings, each expressed in the canonical form:

> **Finding:** [claim in plain language]. **Evidence:** [specific numbers and error types].
> **Implication:** [what this reveals about the model's spatial reasoning capacity].

Each finding is at most 2–3 sentences in the claim component, but every numerical assertion in the evidence component is drawn exclusively from the structured data passed to the LLM—it cannot hallucinate statistics it was not given. The entire chain from raw prediction file to narrative finding is logged, and all intermediate files are released, so every claim is independently verifiable by inspection.

E.4 OUTPUT ARTEFACTS

STAGE 3 produces three complementary output types, together constituting a complete, citable diagnostic report.

Per-model skill card. For each evaluated model, a skill card is generated in both JSON and Markdown formats comprising: (i) accuracy across **E1**–**E7** summarised as a 7-dimensional profile suitable for radar chart visualisation; (ii) per-category accuracy ranked from strongest to weakest; (iii) the model's top-5 failure clusters from STAGE 2 with representative example items; and (iv) a *blind-spot flag* for any error type on which the model scores more than one standard deviation below the cross-model population mean—a simple signal to model developers and downstream users of where the model's spatial reasoning is structurally deficient.

Cross-model comparison table. A summary table with rows indexing models and columns indexing skill dimensions. Each cell contains the model's accuracy on that dimension, its rank position among all evaluated models, and a significance marker indicating whether the difference from the best-performing model on that dimension is statistically significant at the Bonferroni-corrected $\alpha = 0.05$ level.

Narrative synthesis report. A field-level report produced by the LLM synthesis step in Step 3c, summarising: which error types are universally hard across all models (high mean failure rate, low variance); which skill dimensions show the most inter-model variance and therefore discriminate best between models; and which models exhibit unusual profiles suggesting distinctive training or architectural properties. This report is designed to be incorporated directly into benchmark papers as a section on cross-model analysis.

E.5 REPRODUCIBILITY AND INFRASTRUCTURE CONSIDERATIONS

The pipeline is designed with reproducibility as a first-order constraint, not a post-hoc property. All intermediate artefacts produced by STAGE 1–STAGE 3—the enriched signal files, the cluster assignments and centroids, the LLM judge call logs with full prompts and responses, the statistical test outputs, and the narrative synthesis inputs—are released alongside the benchmark dataset and model outputs. This means that every numerical claim in a paper relying on the pipeline is independently verifiable by re-running the relevant downstream analysis step from the released files.

Detailed code and running details are provided with the submission.

Model-agnostic incremental updates. The pipeline is intentionally structured so that its most expensive component—STAGE 2 taxonomy induction, which involves embedding all failure items and running HDBSCAN—needs to be run only once for an initial model cohort. When a new model is added to the evaluation, its failure items are embedded and assigned to the existing cluster structure using nearest-centroid assignment; the taxonomy induction step is not re-run. The statistical comparison in STAGE 3 automatically updates to include the new model. This makes the harness a *living* evaluation infrastructure: it remains useful and consistent as new models are released, exactly the persistent evaluation architecture required for a benchmark intended to track progress in spatial reasoning over time.

Cost profile. The LLM calls in STAGE 2 (taxonomy labelling) and STAGE 3 (diagnostic synthesis) are bulk operations over cluster representatives rather than per-item calls, and their cost is therefore sub-linear in the number of items. For a benchmark of $N = 5{,}000$ items evaluated across 25 models, the total LLM inference cost of the full pipeline is estimated to be around $400-500 at current API rates, making it accessible for research groups without large compute budgets.

F MORE DETAILS ON EXPERIMENTS AND EVALUATION

We use `google/gemini-3.1-flash-image-preview` the judge model (including error taxonomy construction). We set `temperature` to 0.1. For the taxonomy generation stage, we use a maximum of 250 tokens, while for both the answering model and judge model we use a maximum of 1000 tokens. All other generation parameters are kept at their default values.

Prompts are available in Appendix L.

F.1 EVALUATION PROTOCOL

Judging Protocol. All responses are evaluated using a separate vision-language judge model under a strict binary scoring scheme. The judge compares the model output against the ground-truth answer, allowing only trivial surface variations (e.g., whitespace, capitalization), while rejecting semantically incomplete or approximate responses. Correct answers embedded within longer generations are accepted, whereas any ambiguity is conservatively scored as incorrect. The judge returns a structured JSON object containing the binary score and a short justification.

Decoding and Inference Settings. We set the generation temperature to 0.1 to minimize stochasticity and ensure stable outputs across model variants. Maximum generation length is set to 600 tokens for standard models and 20,000 tokens for reasoning-augmented models, allowing sufficient capacity for multi-step reasoning when required. The judge model is allocated a maximum of 1,200 tokens. All evaluations are performed via the OPENROUTER API using default settings otherwise. API usage costs range from approximately $0.10 to $15 per 1M input/output tokens, with total experimental expenditure (including pilot runs and ablations) of approximately $1,200.

Rationale. This evaluation design isolates semantic correctness from superficial variations in formatting and verbosity, ensuring that measured differences reflect reasoning ability rather than presentation style. The strict binary judge enforces consistency across models, enabling stable and reproducible comparisons under invariant output transformations.

F.2 EVALUATION OF LLM-AS-A-JUDGE FOR OPEN-ENDED EVALUATION

To evaluate *Gemini-3.1-Flash-Image-Preview* as an automatic judge, we construct a balanced evaluation set of 175 samples spanning all task categories. Each sample consists of a question, ground-truth label, and model-generated response, and is independently annotated by three human evaluators using binary judgments (1 for correct, 0 for incorrect).

We compare the LLM judgments against human annotations using multiple agreement metrics computed both globally and per category. Specifically, we report Cohen's κ, accuracy against the majority vote, and inter-annotator agreement measured via average pairwise κ among human annotators.

Cohen's Kappa. Cohen's κ quantifies agreement between two raters while correcting for chance-level alignment. It is defined as

$$\kappa = \frac{p_o - p_e}{1 - p_e},$$

where p_o denotes observed agreement and p_e denotes expected agreement under independence:

$$p_e = \sum_{c \in \{0,1\}} p_c^{(a)}\, p_c^{(b)}.$$

Here, $p_c^{(a)}$ and $p_c^{(b)}$ are the empirical probabilities that rater a and rater b assign label c. In our setup, we compute κ between the LLM judge and the majority-vote label derived from human annotations, denoted as $\kappa_{\text{judge-human}}$.

Accuracy with Majority Vote. We also report direct agreement with the majority label. Let y_i^{LLM} denote the LLM prediction and y_i^{maj} the majority vote. Accuracy is computed as

$$\text{Acc} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}\left(y_i^{\text{LLM}} = y_i^{\text{maj}}\right).$$

Unlike Cohen's κ, this metric does not adjust for chance agreement and measures only exact label correspondence.

Inter-Annotator Agreement. To quantify consistency among human annotators, we compute the average pairwise Cohen's κ across all annotator pairs. Let $\mathcal{H} = \{H_1, H_2, H_3\}$ denote the set of annotators. The inter-annotator agreement is defined as

$$\kappa_{\text{human}} = \frac{1}{|\mathcal{P}|} \sum_{(a,b) \in \mathcal{P}} \kappa(a, b),$$

where $\mathcal{P}$ is the set of all unordered annotator pairs. This provides a robust estimate of annotation consistency without assuming a fixed reference label.

Higher values indicate stronger agreement beyond chance, while lower values reflect increased ambiguity or subjectivity in the task.

Table 10: Agreement metrics between the LLM judge and human annotators across categories. The table reports accuracy and Cohen's κ between the judge and human majority vote, along with average pairwise human κ as a measure of annotator consistency.

Category	n	Accuracy	$\kappa_{\text{judge-human}}$	$\kappa_{\text{human-pairwise}}$
Overall	175	**1.000**	**1.000**	0.805
Physical Reasoning	25	1.000	1.000	0.739
Positional & Geometric Reasoning	25	1.000	1.000	0.827
Knowledge-Grounded Reasoning	25	1.000	1.000	0.853
Relational - Logical & Causal Reasoning	25	1.000	1.000	0.819
Temporal & Dynamic Reasoning	25	1.000	1.000	0.719
Quantitative & Comparative Reasoning	25	1.000	1.000	0.781
Perceptual Understanding	25	1.000	1.000	**0.922**

Interpretation. Table 10 shows that the LLM judge achieves perfect agreement with human majority annotations across all task types, with both accuracy and Cohen's κ equal to 1.000. This indicates that the judge fully replicates the aggregated human decision boundary without any observed disagreement across the 175 evaluated samples.

However, a more nuanced picture emerges when examining inter-annotator agreement. Pairwise human κ varies substantially across task types, ranging from 0.719 (Temporal & Dynamic Reasoning) to 0.922 (Perceptual Understanding). This spread highlights that while the judge is perfectly aligned

with the majority vote, the underlying human annotations themselves are not uniformly consistent. Tasks involving temporal dynamics and physical reasoning exhibit comparatively lower agreement, suggesting higher ambiguity or interpretive variance among annotators.

In contrast, categories such as Perceptual Understanding ($\kappa = 0.922$) and Knowledge-Grounded Reasoning ($\kappa = 0.853$) demonstrate strong human consistency, indicating clearer decision boundaries. The fact that the LLM judge achieves perfect agreement even in lower-consensus categories suggests that it effectively captures the majority aggregation rule rather than independently resolving ambiguity.

These findings indicate that the LLM-as-judge framework is highly reliable as a proxy for majority human evaluation, achieving deterministic alignment across all categories. However, perfect κ should be interpreted cautiously: it reflects agreement with the majority vote rather than independent judgment quality. The variation in human pairwise agreement underscores that certain reasoning categories remain intrinsically subjective, and future evaluation frameworks may benefit from explicitly modeling annotator uncertainty rather than collapsing it into a single ground truth.

Overall, the results validate the use of LLM-based automated judges for scalable evaluation, while also emphasizing the importance of complementary human agreement metrics to contextualize performance.

F.3 Human Evaluation Protocol

To establish a human performance baseline, we evaluate the same set of questions using three independent participants: one high-school-level annotator, one STEM undergraduate, and one non-STEM undergraduate (at most in their second year of study). Each participant independently answers all questions without access to external tools or prior solutions.

Responses are scored using the same binary evaluation protocol as model outputs using LLM as a Judge, where answers are marked as correct (1) or incorrect (0) based on exact semantic equivalence with the ground truth, allowing only minor formatting variations. No collaboration or discussion is permitted during evaluation.

Final human labels are obtained via majority vote across the three annotators. Majority voting is used to reduce individual-level variance and annotation noise, yielding a more stable estimate of typical human performance. While this does not eliminate systematic biases shared across annotators, it provides a robust aggregate baseline for comparison against model performance.

F.4 AI Agents for Spatial Reasoning

F.4.1 SpatioXolver

We include SpatioXolver Wasi et al. (2026), a multi-agent spatial reasoning system adapted from Xolver (Hosain et al., 2025), to evaluate the effectiveness of structured decomposition on challenging instances. The system factorizes visual reasoning into a sequence of specialized agents, each responsible for a distinct stage of the perception–reasoning pipeline.

The pipeline begins with a *visual analysis agent*, which produces a detailed scene description capturing objects, attributes, and coarse spatial layout. An *object segmentation agent* then assigns consistent identifiers to individual entities, followed by an *attribute extraction agent* that encodes properties (e.g., shape, size, color, orientation) in structured form. A *spatial relation agent* constructs pairwise and higher-order relations (e.g., left-of, above, alignment), forming a relational graph of the scene. Building on this, a *grouping and symmetry agent* identifies higher-level structures such as clusters, grids, and symmetries, while a *transformation tracking agent* captures changes across frames for dynamic reasoning tasks.

All intermediate outputs are consolidated into a unified structured representation, which is consumed by a final *reasoning agent* to generate answers. This modular design enables explicit multi-step reasoning by separating perception, grounding, and compositional inference, and allows the system to propagate low-level visual signals into higher-order spatial reasoning. In our evaluation, SpatioXolver serves as a representative agentic approach for assessing whether such decomposition can recover performance on instances unsolved by monolithic VLMs.

F.4.2 CLAUDE CODE AGENT

We also evaluate a **Claude Code-based agent** Anthropic (2026), which performs spatial reasoning through sequential programmatic decomposition. Instead of distributing reasoning across specialized agents, this approach leverages iterative code generation and execution to structure intermediate reasoning steps. The agent dynamically constructs representations (e.g., lists, counters, relational checks) within a single reasoning loop, enabling symbolic manipulation and step-wise verification. This design is particularly effective for tasks requiring arithmetic reasoning, rule-based logic, or structured comparison, where explicit computation can reduce ambiguity.

F.4.3 EVALUATION SETUP

We evaluate two agentic systems on a *hard-case subset* consisting of 175 instances that were not correctly solved by any of the 16 evaluated frontier VLMs. Both systems receive identical inputs (image and question) and are evaluated under the same protocol as standard models.

SPATIOXOLVER is implemented using *Gemini-3.1-Flash-Preview* as the underlying model, with a low temperature (0.1) to ensure deterministic behavior. It employs a multi-agent pipeline that decomposes each task into structured sub-components (e.g., object extraction, attribute encoding, relation construction, and transformation reasoning), producing a unified intermediate representation before generating the final answer. In contrast, the **Claude Code-based agent** performs sequential reasoning within a single loop, leveraging programmatic execution to construct and verify intermediate representations.

For both systems, we allow extended reasoning with a maximum context length of 10,000 tokens to support multi-step decomposition and intermediate computation. Each agent produces a single final answer per instance, which is evaluated using the same LLM-based judging framework described in Section 3.2. No external retrieval or additional supervision is provided beyond the input image and question. This setup ensures a controlled comparison, isolating the impact of multi-agent decomposition versus sequential programmatic reasoning on solving previously unsolved spatial reasoning tasks.

G MODEL PERFORMANCE BREAKDOWN ON FRONTIERSPATIAL

G.1 MODEL PERFORMANCE ON IDENTIFYING SPATIAL ELEMENT

The overall pattern in Table 11 shows uniformly high performance across models, with most scores concentrated between 75% and 90%, indicating that spatial element identification is substantially easier than spatial relation reasoning and exhibits lower variance across categories. Proprietary models are tightly clustered at the top, with CLAUDE HAIKU 4.5 achieving the highest Physical (93.67%) and Temporal (91.28%) scores, while GEMINI-3.1-FLASH-LITE leads Positional/Geometric (73.81%) and Relational/Logical (89.89%), reflecting strong consistency across structurally diverse tasks. In contrast to Table 1, weaker models such as GPT-4O-MINI still maintain relatively high performance (≈45–80%), suggesting that element identification relies less on complex reasoning chains and more on direct visual-text alignment.

Open-source models display both competitiveness and slight dispersion, with QWEN3.5-9B, QWEN3.6-35B-A3B, and KIMI K2.5 achieving top-tier results across multiple categories, particularly in Perceptual (up to 89.10%) and Temporal (up to 83.49%), in some cases exceeding proprietary counterparts. A notable trend is the compression of category gaps: unlike spatial relation tasks, Temporal and Physical categories no longer exhibit extreme drops, indicating that identifying elements in dynamic or physical contexts is less sensitive to reasoning difficulty. Quantitative performance is consistently high across nearly all models (≈78–88%), reinforcing that counting or magnitude extraction remains a stable capability.

Perceptual understanding shows the highest ceiling overall, with several models surpassing 85%, suggesting that visual feature extraction and localization scale effectively with both model size and multimodal training. However, Positional/Geometric reasoning still exhibits moderate spread (≈54–78%), implying that even element identification retains some sensitivity to spatial structure complexity. Overall, the results indicate that while model differences persist, spatial element identifi-

cation is largely a saturated capability, and the primary bottleneck for MLLMs lies in transforming these identified elements into coherent relational reasoning.

Table 11: Model Performance (%) on Identifying Spatial Elements across Categories.

Model	Phys. (#1312)	Pos.Geo. (#2596)	Knowledge (#2332)	Re.Lo.Ca. (#455)	Temporal (#218)	Quant. (#503)	Percept. (#422)
Proprietary Models							
Claude Haiku 4.5	**93.67**	72.81	87.48	89.67	**91.28**	88.47	87.44
Gemini-2.5-Flash	85.14	63.02	81.98	83.52	54.13	75.70	75.83
Gemini-3.1-Flash-Lite	88.49	**73.81**	87.78	**89.89**	80.28	82.50	84.36
GPT-4o-mini	79.95	56.12	81.39	77.80	45.41	76.14	77.49
GPT-5-mini	86.89	62.43	**88.50**	84.18	65.14	**88.47**	78.20
o4-mini	87.88	64.20	79.37	85.49	68.81	86.08	**79.86**
Open-Source Models							
Gemma-3-4B-it	85.90	57.50	72.43	73.41	39.45	83.50	81.99
Qwen3-VL-8B-Instruct	77.82	65.22	81.90	80.00	65.60	79.72	79.38
Qwen3.5-9B	88.86	76.35	84.48	86.15	80.73	80.12	**89.10**
Llama-3.2-11B-Vision-Ins.	85.29	56.91	76.16	79.56	46.79	86.08	71.80
Gemma-4-31B-it	**89.18**	69.52	86.62	**89.01**	81.65	78.13	80.57
Aya-Vision-32B	85.60	64.01	66.54	76.65	63.30	88.47	75.83
Qwen3.6-35B-A3B	88.51	**78.02**	86.23	87.00	80.19	85.69	85.41
Kimi K2.5	88.41	72.65	83.66	87.64	**83.49**	**84.69**	**87.91**
GLM-4.5V	84.60	72.47	81.77	83.96	77.06	84.10	81.47
Mistral Small 2603	83.61	53.96	**82.38**	77.14	58.72	79.52	70.62

G.2 Model Performance On Identifying Spatial Relations

Table 12 shows a clear pattern is the uneven distribution of strengths across categories, with most models achieving higher scores in Quantitative and Knowledge reasoning while lagging in Physical and Temporal reasoning, indicating that symbolic or count-based tasks remain easier than dynamic or embodied spatial understanding. Among proprietary systems, GEMINI-3.1-FLASH-LITE leads Physical reasoning (65.68%), whereas CLAUDE HAIKU 4.5 dominates Positional/Geometric (60.40%), Temporal (78.90%), and Quantitative (86.28%), suggesting a more balanced but still category-skewed capability profile. In contrast, GPT-4O-MINI consistently underperforms across all categories, rarely exceeding 55%, highlighting a substantial gap between smaller optimized variants and stronger mid-tier models like GPT-5-MINI and O4-MINI.

Open-source models show greater variance but also stronger peaks, with QWEN3.5-9B and QWEN3.6-35B-A3B matching or exceeding proprietary leaders in multiple categories, particularly Knowledge (75%+) and Quantitative (≈75%), indicating competitive reasoning depth at higher parameter scales. Notably, KIMI K2.5 achieves the highest Relational/Logical score (69.98%) and strong Temporal (72.02%) and Quantitative (80.72%) performance, reflecting robustness in multi-step and dependency-heavy reasoning tasks. A consistent trend across both groups is the relative difficulty of Temporal reasoning, where even top models show larger variance (roughly 48% to 79%), implying instability in handling dynamic spatial changes.

Perceptual understanding exhibits a bifurcation: weaker models cluster around 50–60%, while stronger open models like QWEN3.5-9B (78.67%) and KIMI K2.5 (77.49%) achieve substantially higher accuracy, suggesting that visual grounding scales effectively with model capacity. Overall, the results reveal that no single model dominates all categories, and performance leadership shifts depending on the reasoning type, reinforcing the claim that spatial reasoning is multi-faceted and cannot be captured by aggregate scores alone.

G.3 Spatial Reasoning Trace Correctness

Performance on spatial reasoning chain correctness in Table 13 is both lower and more dispersed than element identification, with scores ranging roughly from 20% to 80%, indicating higher task complexity and greater sensitivity to reasoning capability. A consistent pattern is that Quantitative and Knowledge categories yield comparatively higher scores across models, while Physical and Temporal reasoning exhibit larger drops and variance, reflecting difficulty in handling dynamic or grounded

Table 12: Open-ended Evaluation Accuracy (%)(↑) on Identifying Spatial Relations across Categories.

Model	Phys. (#1312)	Pos.Geo. (#2596)	Knowledge (#2332)	Re.Lo.Ca. (#455)	Temporal (#218)	Quant. (#503)	Percept. (#422)
Proprietary Models							
Claude Haiku 4.5	50.99	**60.40**	71.40	64.84	**78.90**	**86.28**	69.19
Gemini-2.5-Flash	44.66	53.21	62.72	58.02	48.62	55.98	56.64
Gemini-3.1-Flash-Lite	**65.68**	59.83	67.11	59.12	62.39	69.58	65.17
GPT-4o-mini	34.53	40.57	50.21	51.43	41.28	53.48	50.24
GPT-5-mini	49.70	49.15	65.34	60.22	56.88	61.63	56.64
o4-mini	49.39	55.10	**58.99**	**60.66**	61.93	67.79	**56.87**
Open-Source Models							
Gemma-3-4B-it	37.27	45.21	41.68	44.62	21.56	62.62	56.64
Qwen3-VL-8B-Instruct	42.99	51.04	55.00	53.41	56.88	63.42	52.37
Qwen3.5-9B	**65.68**	55.66	75.09	64.84	66.97	**75.75**	**78.67**
Llama-3.2-11B-Vision-Ins.	33.92	39.74	45.20	46.59	36.70	48.71	46.21
Gemma-4-31B-it	57.55	54.75	69.00	59.12	61.93	65.21	61.37
Aya-Vision-32B	40.63	41.31	33.50	44.27	56.88	63.62	54.74
Qwen3.6-35B-A3B	62.22	**57.56**	**75.56**	66.74	**68.40**	75.55	71.29
Kimi K2.5	55.18	58.29	73.36	**69.98**	72.02	80.72	77.49
GLM-4.5V	41.39	56.87	62.93	62.86	61.93	60.64	60.81
Mistral Small 2603	38.72	40.69	55.36	48.57	55.50	57.46	46.45
Human Baseline	98.40	99.70	99.91	99.12	99.54	100.00	99.76

spatial interactions. Among proprietary systems, CLAUDE HAIKU 4.5 leads Positional/Geometric (60.40%), Temporal (78.90%), and Quantitative (86.28%), whereas GEMINI-3.1-FLASH-LITE achieves the top Physical score (65.68%), highlighting complementary strengths rather than a single dominant model.

Lower-tier proprietary models such as GPT-4O-MINI show consistently reduced performance across all categories, often below 55%, reinforcing the dependence of relational reasoning on model capacity and reasoning depth. Open-source models exhibit a wider spread but also stronger peaks, with QWEN3.5-9B and QWEN3.6-35B-A3B achieving top Knowledge (≈75%) and Quantitative (≈75%) scores, matching or surpassing proprietary counterparts. KIMI K2.5 stands out in Relational/Logical reasoning (69.98%) and maintains strong Temporal (72.02%) and Perceptual (77.49%) performance, indicating robustness in multi-step relational inference.

Temporal reasoning remains the most unstable category, with scores spanning from 21.56% (GEMMA-3-4B-IT) to 78.90% (CLAUDE HAIKU 4.5), suggesting inconsistent handling of time-dependent spatial relations. Perceptual performance shows a clear scaling trend, where stronger open models such as QWEN3.5-9B achieve high accuracy (78.67%), while smaller models cluster near 50–60%. Overall, no model consistently dominates across all categories, and the shifting leadership across tasks underscores that spatial relation reasoning is heterogeneous and not reducible to a single capability dimension.

H CATEGORY-WISE MODEL PERFORMANCE OF FRONTIERSPATIAL

All 14 tables are provided in the separate supplementary PDF.

I QUALITATIVE ERROR TAXONOMY OF SPATIAL–VISUAL REASONING FAILURES

To systematically characterise the failure modes observed across all evaluated models, we constructed a two-level hierarchical error taxonomy. The *first level* (E1–E7) partitions failures by the cognitive or computational operation that breaks down. The *second level* decomposes each first-level category into focused subcategories that distinguish the representational domain, modality, or reasoning substrate in which the failure manifests.

Taxonomy construction proceeded inductively: every failed instance was first assigned a first-level code based on *where in the perception–reasoning pipeline* the model deviated from the correct answer; it was then assigned a second-level code based on the *domain or object type* involved. Where a single instance exhibited multiple failure signatures (e.g., a driving task requiring both

Table 13: Model Performance (%) on Spatial Reasoning Chains across Categories.

Model	Phys. (#1312)	Pos.Geo. (#2596)	Knowledge (#2332)	Re.Lo.Ca. (#455)	Temporal (#218)	Quant. (#503)	Percept. (#422)
Proprietary Models							
Claude Haiku 4.5	37.88	32.66	48.93	48.13	**61.47**	**18.69**	45.97
Gemini-2.5-Flash	37.88	35.96	42.81	49.01	39.91	11.95	45.50
Gemini-3.1-Flash-Lite	**50.23**	**43.48**	51.37	50.11	57.34	16.50	**51.42**
GPT-4o-mini	29.27	24.10	23.03	42.42	21.10	13.52	40.28
GPT-5-mini	43.67	36.51	**52.60**	55.38	55.50	16.50	48.10
o4-mini	42.99	36.39	47.45	**56.70**	54.13	16.50	46.21
Open-Source Models							
Gemma-3-4B-it	27.21	25.13	13.16	31.87	11.01	8.95	38.39
Qwen3-VL-8B-Instruct	35.67	36.08	31.45	44.84	48.17	15.51	41.23
Qwen3.5-9B	30.13	27.20	31.86	36.92	38.07	11.33	41.47
Llama-3.2-11B-Vision-Ins.	25.30	24.34	17.32	36.48	20.18	9.15	33.65
Gemma-4-31B-it	**49.09**	35.17	31.73	45.93	54.13	15.51	50.47
Aya-Vision-32B	25.93	22.16	15.57	35.02	29.36	15.31	41.71
Qwen3.6-35B-A3B	40.48	35.89	**45.98**	48.02	38.21	13.92	43.06
Kimi K2.5	35.52	34.45	39.90	42.60	44.50	12.13	44.79
GLM-4.5V	33.46	**39.46**	37.41	**50.33**	**50.46**	**14.91**	**49.41**
Mistral Small 2603	29.65	25.21	29.33	36.04	49.08	14.51	32.94

perceptual identification and egocentric grounding), it was assigned to the category corresponding to the earliest breakdown step. The resulting taxonomy covers 7 first-level categories and 38 second-level subcategories, spanning 13 evaluated models.

Table 14 provides a compact summary; the remainder of this section defines each category in detail.

Table 14: Overview of the two-level spatial–visual error taxonomy.

Code	Name	Brief Definition
E1	Perceptual Failure	Misidentification or non-detection of a visual element
E2	Spatial Grounding Failure	Elements identified but not anchored to the correct spatial frame
E3	Spatial Transformation Failure	Arrangement understood but rotation/folding/perspective incorrectly applied
E4	Quantitative Spatial Failure	Counting, distance estimation, or measurement errors
E5	Multi-Step Reasoning Failure	Correct first step but failure on a dependent subsequent step
E6	Language–Spatial Mapping Failure	Misinterpretation of spatial language (*e.g.*, left/right confusion)
E7	Format / Refusal Failure	No valid answer produced, regardless of spatial competence

I.1 E1 — Perceptual Failure

Definition. A **perceptual failure** occurs when the model misidentifies, misclassifies, or entirely fails to detect a visual or spatial element that is unambiguously present in the image. The failure precedes any spatial reasoning: the model's internal scene representation is incorrect from the outset, so downstream spatial inferences are compromised regardless of their logical validity.

Benchmark examples.

- *"What is the colour of the water in the image?"* — model reports the wrong colour despite a clearly visible body of water.
- *"What is the shape of the brick in the picture when viewed from above?"* — model cannot identify the top-view silhouette of a LEGO brick.
- *"What kind of animal is on top of the tree?"* — model hallucinate an absent animal or misidentifies the species present.
- *"What is the primary type of scene depicted in this aerial image?"* — model confuses an aerial residential scene with an industrial or natural landscape.
- *"Which chest X-ray disease can be found?"* — model fails to identify a pathological feature that a radiologist would detect.

E1.1 — Object & Attribute Recognition Failure.
Failure to correctly identify an object's perceptual attributes such as colour, shape, or semantic category.

Affected models: Claude 4.5 Haiku, Gemini-2.5-Flash, Gemma-4-31B-it, Llama-3.2-11B-Vision-Instruct, Mistral-Small-2603, Kimi K2.5, o4-mini, GPT-4o-mini, GPT-5-mini, Qwen3-VL-8B-Instruct , GLM-4.5V.

Root-cause analysis. Object attribute recognition errors are among the most broadly distributed failures in the taxonomy, appearing across nearly every evaluated model. Smaller-parameter models (Qwen3-VL-8B-Instruct , Llama-3.2-11B-Vision-Instruct) lack sufficient visual encoder capacity to disentangle fine-grained attribute channels—colour, texture, and shape—from the holistic patch embeddings produced by their vision backbones. Mid-size models (Mistral-Small-2603, Gemma-4-31B-it) show a pattern consistent with inadequate cross-modal alignment training: the language head defaults to high-frequency category priors (e.g., assigning the most common LEGO brick colour in the training corpus) rather than grounding in the specific visual instance. Instruction-tuned models (GPT-4o-mini, o4-mini) occasionally hallucinate plausible-but-incorrect attributes, consistent with the known tendency of RLHF-trained models to generate fluent, overconfident descriptions. Even Gemini-2.5-Flash exhibits residual confusion on low-contrast or partially occluded objects, pointing to limits in pre-training image diversity rather than model capacity per se.

E1.2 — Scene & Environment Classification Failure.
Failure to correctly classify the scene type, environment category, or contextual setting (e.g., aerial view, indoor layout, road type).

Affected models: Gemini-3.1-Flash-Preview, Gemma-4-31B-it, Mistral-Small-2603, Kimi K2.5, Qwen3-VL-8B-Instruct , GLM-4.5V.

Root-cause analysis. Scene-level classification demands the integration of holistic contextual cues (skyline geometry, road markings, vegetation density) that are poorly preserved by the reduced-resolution encoders used in efficiency-oriented models such as Gemini-3.1-Flash-Preview and Qwen3-VL-8B-Instruct . Kimi K2.5 and GLM-4.5V are pre-trained predominantly on East-Asian data distributions; aerial and architectural scenes sourced from other geographic contexts may lie outside their training support, causing systematic misclassification of unfamiliar regional layouts. Gemma-4-31B-it and Mistral-Small-2603 appear to rely heavily on salient foreground objects for scene inference, neglecting background context entirely, which leads to errors whenever the dominant foreground category is ambiguous with respect to scene type. None of the affected models appears to employ an explicit multi-scale scene integration mechanism, leaving them vulnerable to tasks that require simultaneous reasoning about local objects and global environmental structure.

E1.3 — Driving-Scene Perceptual Failure.
Failure to correctly perceive lane positions, vehicle turning behaviour, safety- relevant distances, or other road-level visual cues.

Affected models: Gemma-3-4B-it, Gemma-4-31B-it, Gemini-3.1-Flash-Preview, Llama-3.2-11B-Vision-Instruct, Mistral-Small-2603, o4-mini, GPT-5-mini, Qwen3-VL-8B-Instruct , GLM-4.5V.

Root-cause analysis. Driving-scene perception requires simultaneous tracking of road markings, vehicle silhouettes, and implicit motion cues within a single static frame—a high-density, context-sensitive task that challenges all evaluated models. The smallest models (Gemma-3-4B-it, Qwen3-VL-8B-Instruct) cannot reliably resolve lane boundary markings from road-surface texture at standard benchmark image resolutions due to limited encoder feature density. GLM-4.5V and Gemini-3.1-Flash-Preview frequently conflate structurally similar lane configurations (e.g., merge zones versus standard two-lane roads), suggesting over-reliance on template matching rather than fine-grained geometric parsing of road topology. Llama-3.2-11B-Vision-Instruct and Mistral-Small-2603 exhibit elevated false-negative rates on turning-vehicle detection, likely because dynamic turning-manoeuvre imagery is underrepresented in their visual instruction-tuning corpora relative to static or parked-vehicle scenarios. Even o4-mini shows residual errors on ambiguous turning-angle estimation, highlighting a fundamental gap between static-image perception and the temporal motion understanding available in video-based training paradigms.

E1.4 — Action & Social Outcome Perception Failure.
Failure to correctly detect or infer an agent's action, intent, or predicted social or physical outcome from visual evidence.

Affected models: `Gemini-3.1-Flash-Preview`, `Llama-3.2-11B-Vision-Instruct`, `Mistral-Small-2603`, `o4-mini`, `GPT-4o-mini`, `GPT-5-mini`.

Root-cause analysis. Predicting human actions from a static frame is intrinsically under-determined, requiring strong priors about body pose, gaze direction, and contextual social norms that are difficult to learn from standard image–caption pre-training. `Llama-3.2-11B-Vision-Instruct` and `Gemini-3.1-Flash-Preview` default to the statistically most common action label (e.g., "standing" rather than "reaching") when perceptual evidence is ambiguous, revealing frequency-based prior dominance over evidence-driven inference. `GPT-4o-mini` and `GPT-5-mini` occasionally construct internally coherent but visually unsupported action narratives—a hallmark of language-prior dominance in RLHF-trained models where fluency incentives outweigh perceptual fidelity. `Mistral-Small-2603`'s failures cluster on social-outcome forecasting tasks (e.g., predicting the result of a physical interaction), suggesting that its instruction-tuning data contained limited diversity in social-scene annotations. `o4-mini`'s errors are concentrated on fine-grained limb-movement detection, pointing to the upper resolution limit of its vision encoder for body-part localisation under natural viewing conditions.

E1.5 — Visual Pattern & Embedded Figure Detection Failure.
Failure to detect hidden figures, visual regularities, or spatially encoded cues embedded within complex or cluttered images.

Affected models: `Gemini-3.1-Flash-Preview`, `Llama-3.2-11B-Vision-Instruct`, `GPT-4o-mini`, `GPT-5-mini`.

Root-cause analysis. Embedded-figure detection requires figure-ground segregation and contour grouping across distractors—a Gestalt-level perceptual operation that is poorly supported by patch-based vision encoders optimised for natural-image classification. `Llama-3.2-11B-Vision-Instruct` and `Gemini-3.1-Flash-Preview` show near-chance performance on shape-embedding tasks, indicating that their encoders do not produce sufficiently separated figure–ground representations to isolate target shapes from complex backgrounds. `GPT-4o-mini` and `GPT-5-mini` perform better but fail specifically when the embedded figure shares colour or texture statistics with its surround, suggesting that their contour-grouping abilities remain partially dependent on luminance-contrast cues rather than purely topological shape properties. None of the four models appears to implement an iterative visual search strategy; each commits to a single feedforward perceptual interpretation, which is systematically insufficient for heavily camouflaged or overlapping figure tasks.

E1.6 — Symbolic & Language-Embedded Visual Failure.
Failure on tasks requiring interpretation of formally rendered symbols, written puzzles, logic diagrams, or code-like visual structures embedded in images.

Affected models: `Claude 4.5 Haiku`, `Gemini-2.5-Flash`, `Gemini-3.1-Flash-Preview`, `Llama-3.2-11B-Vision-Instruct`, `Mistral-Small-2603`, `Kimi K2.5`, `o4-mini`, `GPT-4o-mini`, `GPT-5-mini`, `Qwen3-VL-8B-Instruct` , `GLM-4.5V`.

Root-cause analysis. Visually rendered formal notation (automata diagrams, logic gates, arithmetic puzzles) requires joint low-resolution OCR and domain-specific semantic rule application—a compound task for which general-purpose vision-language pre-training provides limited supervision. Across all affected models the most common error is OCR-level misreading of small, stylised, or densely packed characters (e.g., digit–letter confusion in state-machine labels), propagating into downstream logical failures. Lighter-weight models (`Qwen3-VL-8B-Instruct` , `Gemini-3.1-Flash-Preview`) additionally lack exposure to specialised formal-notation corpora, resulting in near-random responses on finite-automaton and logic-gate tasks. The breadth of this failure even in large models (`Gemini-2.5-Flash`, `o4-mini`) confirms that visually rendered formal notation represents a persistent weak spot for current vision–language architectures, likely requiring targeted data curation rather than scale-alone remediation.

I.2 E2 — Spatial Grounding Failure

Definition. A **spatial grounding failure** occurs when the model correctly identifies the elements present in a scene but cannot anchor them to their correct positions within a spatial frame of reference. The model has an adequate symbolic inventory of the scene yet fails to assign the correct spatial roles (left/right, above/below, near/far, inside/outside) or cannot resolve which reference frame—egocentric vs. allocentric—the question presupposes.

Benchmark examples.

- *"How would you describe the lane position of Entity #1? front lane / front-left lane / front-right lane / oncoming traffic lane?"* —model names the correct vehicle but assigns it the wrong lane.
- *"Are the LEGO pieces pointed to by the two red arrows adjoining or separated?"* — model recognises both bricks but misjudges their contact relationship.
- *"If the fully visible 'one way' sign is facing south, what direction is the other 'one way' sign facing?"* — model identifies both signs but applies the wrong allocentric transformation.
- *"When a person exits the door with a semicircular glass top facing south, which room is on the west side of the corridor?"* — model navigates the verbal description but grounds the directions incorrectly.
- *"Where is the bridge located in the image?"* — model identifies the bridge but mis-reports its spatial relation to the scene.

E2.1 — Egocentric / First-Person Frame Grounding Failure.
Failure to ground objects or directions relative to the observer's or ego-vehicle's viewpoint.

Affected models: all 13 evaluated models.

Root-cause analysis. Egocentric frame grounding is the single most broadly distributed failure in the taxonomy, appearing in every evaluated model without exception. Vision–language models are pre-trained predominantly on image–caption pairs that use allocentric (world-centred) spatial language, leaving egocentric descriptions (*"to my left," "approaching me"*) statistically underrepresented in the training distribution. During instruction-tuning, egocentric driving questions are frequently resolved by heuristic pattern-matching to common caption templates rather than by explicit viewpoint transformation, producing systematic left–right and forward–backward reversals. The fact that even the largest evaluated models (`Gemini-2.5-Flash`, `GPT-5-mini`) exhibit this error confirms that parameter scale alone cannot compensate for the distributional gap in egocentric spatial supervision. Models not specifically fine-tuned on autonomous-driving datasets are particularly affected: without explicit training signal that distinguishes the ego-vehicle's first-person perspective from a third-person observer's view, the model cannot reliably switch reference frames on demand.

E2.2 — Relative Object Positioning Failure.
Failure to correctly specify where one identified object is positioned relative to another (left/right, above/below, in-front/behind).

Affected models: all 13 evaluated models.

Root-cause analysis. Relative positional grounding requires three sequential steps: (1) identify the reference object, (2) identify the target object, and (3) compute and verbalise the directional relationship between them—a chain that introduces compounding error at each stage. In smaller models (`Gemma-3-4B-it`, `Qwen3-VL-8B-Instruct`), the dominant failure is reference-object selection: the model anchors on the most salient or centrally positioned object rather than the one designated by the question. Mid-size models (`Mistral-Small-2603`, `Kimi K2.5`) more often select the correct reference object but err during directional computation, particularly when the reference and target objects are near-equidistant from the image centre or when viewpoint changes are involved. The universality of this error across all parameter scales underscores that relative spatial reasoning is a core linguistic–perceptual interface problem requiring explicit compositional supervision that is absent from standard next-token or captioning objectives.

E2.3 — Depth & Layering Grounding Failure.
Failure to correctly assign depth order, occlusion relationships, or vertical-dimension stacking among objects.

Affected models: `Claude 4.5 Haiku`, `Gemini-2.5-Flash`, `Gemini-3.1-Flash-Preview`, `Mistral-Small-2603`.

Root-cause analysis. Depth ordering from monocular images requires simultaneous integration of multiple monocular depth cues: occlusion boundaries, perspective foreshortening, texture gradients, and relative size—a multi-cue fusion problem for which single-frame vision encoders offer only weak inductive biases. `Gemini-3.1-Flash-Preview` and `Mistral-Small-2603` rely disproportionately on the size-as-depth heuristic, causing systematic errors whenever a physically closer object is smaller than a farther one. `Claude 4.5 Haiku`'s errors concentrate on vertical-stack tasks (e.g., layered cross-section diagrams), suggesting that its encoder conflates image-plane vertical position with world-space depth when objects are stacked top-to-bottom in the image rather than arranged in perspective. `Gemini-2.5-Flash`'s depth errors arise primarily on visibility-and- occlusion queries where partial occlusion is the sole available depth cue, pointing to an under-representation of partial-occlusion grounding annotations in its pre-training data.

E2.4 — Road & Lane Spatial Grounding Failure.
Failure to correctly anchor lanes, navigation paths, or individual traffic participants to their precise road-level positions.

Affected models: `Gemini-2.5-Flash`, `Gemini-3.1-Flash-Preview`, `Gemma-3-4B-it`, `Gemma-4-31B-it`, `Llama-3.2-11B-Vision-Instruct`, `Mistral-Small-2603`, `o4-mini`, `GPT-4o-mini`, `GPT-5-mini`, `Qwen3-VL-8B-Instruct` , `GLM-4.5V`.

Root-cause analysis. Road-level spatial grounding demands both fine-grained lane-boundary perception and a compositional map of which entity occupies which lane—a joint perceptual and relational reasoning requirement. Lightweight models (`Gemma-3-4B-it`, `Qwen3-VL-8B-Instruct`) fail at the perceptual stage: lane markings are not reliably resolved at their native encoder resolutions, so the grounding problem is never properly posed. Mid-size models (`Mistral-Small-2603`, `GLM-4.5V`) resolve the markings but fail to consistently map entities to their correct lanes, especially under multi-entity scenes where multiple vehicles occupy adjacent lanes simultaneously. Larger models (`GPT-4o-mini`, `o4-mini`) show residual confusion specifically on oncoming-traffic lane assignment, likely because their training data contains relatively few right-hand-drive versus left-hand-drive scene variants, causing occasional frame-of-reference inversion.

E2.5 — LEGO & Structured-Assembly Grounding Failure.
Failure to correctly ground LEGO brick positions, adjacency relationships, or assembly placement within a 3D structural frame.

Affected models: `Claude 4.5 Haiku`, `Gemini-2.5-Flash`, `Gemini-3.1-Flash-Preview`, `Gemma-3-4B-it`, `Gemma-4-31B-it`, `Llama-3.2-11B-Vision-Instruct`, `Mistral-Small-2603`, `Kimi K2.5`, `o4-mini`, `GPT-4o-mini`, `GPT-5-mini`, `Qwen3-VL-8B-Instruct` , `GLM-4.5V`.

Root-cause analysis. LEGO grounding tasks require precise stud-level spatial accounting in an isometric 3D render where perspective foreshortening compresses spatial relationships, making millimetre-scale adjacency judgements necessary. Lightweight models (`Gemma-3-4B-it`, `Qwen3-VL-8B-Instruct`) lack sufficient patch resolution to reliably distinguish stud-facing surface from body-plate surface in isometric renders. Mid-size models produce more accurate brick inventories but lose track of inter-brick contact surfaces when scenes contain more than three to four bricks, suggesting a limited spatial working-memory capacity within a single forward pass. Even the largest evaluated models (`GPT-5-mini`, `Gemini-2.5-Flash`) make grounding errors on adjacency tasks, confirming that fine-grained 3D contact geometry remains a challenging perceptual problem at all evaluated parameter scales in the absence of dedicated 3D-supervision pre-training.

E2.6 — Agent Action & Social Grounding Failure.
Failure to correctly ground an agent's action or predicted behaviour to the correct spatial participant or scene context.

Affected models: `Claude 4.5 Haiku`, `Gemini-2.5-Flash`, `Gemini-3.1-Flash-Preview`, `Gemma-3-4B-it`, `Llama-3.2-11B-Vision-Instruct`, `GLM-4.5V`.

Root-cause analysis. Action grounding requires the model to bind a detected action to the correct spatial agent when multiple agents are present in the same scene—a cross-referential operation that combines visual entity tracking with predicate-argument structure. `Llama-3.2-11B-Vision-Instruct` and `Gemini-3.1-Flash-Preview` frequently mis-assign an observed action to the wrong agent, particularly in scenes where two agents are spatially proximate or visually similar. `Claude 4.5 Haiku` and `GLM-4.5V` show systematic confusion when the question uses deictic expressions (*"the man on the horse"*) that require resolving a chain of visual references before the target agent is unambiguously identified. The error is exacerbated by the absence of explicit agent-tracking supervision in standard image–caption pre-training, where captions describe scene contents globally rather than establishing agent-specific spatial bindings.

I.3 E3 — Spatial Transformation Failure

Definition. A **spatial transformation failure** occurs when the model has correctly perceived and grounded the spatial arrangement of elements but fails to correctly apply a mental operation—rotation, folding, reflection, perspective shift, or physical state change—to produce the answer. The failure is not in *what* the model sees, but in *what it computes* from that scene.

Benchmark examples.

- *"Which image in the following options shows the LEGO object in the reference image rotated clockwise by 120 degrees?"* — model identifies the correct object but selects the wrong rotation option.
- *"Work out which option shows the figure on the left when folded along the dotted line."* — model perceives the 2D net correctly but produces an incorrect 3D fold result.
- *"If the billiard ball moves in the direction of the red arrow, will any ball go into the pocket and score?"* — model understands the arrow direction but incorrectly predicts the physical trajectory.
- *"Figure 1 shows the north side of the object. Which side does Figure 2 show?"* — model applies an incorrect compass-direction transformation.
- *"A person hangs a very heavy, swinging hammock from the edge of the overhanging roof. What is the most likely outcome?"* — model fails to apply physical causality reasoning to the described scenario.

E3.1 — 2D Folding & Net Transformation Failure.
Failure to mentally fold a 2D net into a 3D shape or predict the correct result of a paper-folding operation.

Affected models: `Claude 4.5 Haiku`, `Gemini-3.1-Flash-Preview`, `Gemma-4-31B-it`, `Mistral-Small-2603`, `o4-mini`, `GPT-5-mini`, `Qwen3-VL-8B-Instruct` .

Root-cause analysis. Net-folding tasks require the simultaneous tracking of face adjacency, face orientation, and edge-pairing across a complete 2D-to-3D transformation—a demanding spatial-composition operation with no natural analogue in image captioning or language modelling. `Gemma-4-31B-it` and `Mistral-Small-2603` exhibit failures concentrated on octahedral and irregular-polygon nets, where the mapping from 2D face layout to 3D face identity is non-obvious and requires multi-step mental rotation before any face-matching can begin. `Qwen3-VL-8B-Instruct` and `Gemini-3.1-Flash-Preview` fail even on cubic nets, suggesting that their encoders do not produce the edge-connectivity representations needed to reason about fold topology. `o4-mini` and `Claude 4.5 Haiku` succeed on standard cube-folding variants but fail on higher-complexity nets (dodecahedra, irregular polyhedra), indicating that their folding competence is pattern-matched to common training instances rather than generalisable to novel net configurations. None of the affected models appears to decompose the folding task into sub-steps (identify faces → label edges → simulate fold), instead attempting a holistic one-shot transformation that is too complex for reliable execution.

E3.2 — 3D Rotation & Isometric View Failure.
Failure to mentally rotate a 3D object or correctly match it to a given isometric, top-down, or perspective view.

Affected models: `Claude 4.5 Haiku`, `Gemini-2.5-Flash`, `Gemini-3.1-Flash-Preview`, `Gemma-3-4B-it`, `Gemma-4-31B-it`, `Llama-3.2-11B-Vision-Instruct`, `Mistral-Small-2603`, `o4-mini`, `GPT-4o-mini`, `GPT-5-mini`, `Qwen3-VL-8B-Instruct` , `GLM-4.5V`.

Root-cause analysis. 3D mental rotation is the most taxonomically widespread transformation failure, appearing in twelve of thirteen evaluated models and spanning the full range of parameter scales. Fine-angle discrimination (30° vs. 60° clockwise rotation of a LEGO model viewed from above) requires stable internal 3D representations that are not produced by patch-based encoders trained on 2D images. Lightweight models (`Gemma-3-4B-it`, `Qwen3-VL-8B-Instruct`) lack the feature resolution to track individual stud positions across rotation, making fine-angle discrimination near-impossible. Mid-size models (`Mistral-Small-2603`, `GLM-4.5V`) succeed at coarse 90° rotations but fail at 30° and 60°, suggesting that their mental rotation competence is driven by symmetry-based shortcuts rather than genuine continuous rotation. Even `GPT-5-mini` and `o4-mini` exhibit above-chance but below-human performance, confirming that current vision–language training objectives do not adequately incentivise the construction of view-invariant 3D object representations.

E3.3 — Physical Causality & Trajectory Prediction Failure.
Failure to predict the physical state or trajectory of an object after applying a force, motion, or physical interaction.

Affected models: `Claude 4.5 Haiku`, `Gemini-3.1-Flash-Preview`, `Gemma-3-4B-it`, `Llama-3.2-11B-Vision-Instruct`, `o4-mini`, `GPT-4o-mini`, `GPT-5-mini`.

Root-cause analysis. Physical causality prediction requires an internal approximate physics engine that can simulate Newtonian dynamics from a single static image—a capability absent from standard discriminative or autoregressive training objectives. `Gemini-3.1-Flash-Preview` and `Gemma-3-4B-it` fail on simple billiard- ball trajectory tasks, indicating that even two-body Newtonian interactions exceed their causal reasoning capacity when posed in a purely visual format. `Llama-3.2-11B-Vision-Instruct` produces physically plausible but locally incorrect trajectories (e.g., correct direction but wrong target pocket), suggesting a shallow heuristic that matches surface-level arrow cues without integrating collision geometry. Instruction-tuned models (`GPT-4o-mini`, `GPT-5-mini`) show language-prior dominance: they predict the "most common" physical outcome for the described scenario rather than reading the specific geometric configuration in the image, leading to failures on edge cases. `Claude 4.5 Haiku` and `o4-mini` succeed on standard collisions but fail on multi-body or structural-load scenarios (e.g., hammock attachment physics), implying that their causal competence does not generalise beyond the most common physical interaction patterns seen during training.

E3.4 — Driving & Ego-Motion Transformation Failure.
Failure to transform driving-scene observations into a correct prediction of ego-vehicle motion, lane-change outcome, or vehicle turning trajectory.

Affected models: `Gemini-2.5-Flash`, `Gemini-3.1-Flash-Preview`, `Gemma-3-4B-it`, `Gemma-4-31B-it`, `Llama-3.2-11B-Vision-Instruct`, `Mistral-Small-2603`, `o4-mini`, `GPT-4o-mini`, `GPT-5-mini`, `Qwen3-VL-8B-Instruct` , `GLM-4.5V`.

Root-cause analysis. Ego-motion prediction requires composing lane-position grounding with a forward simulation of vehicle kinematics—a two-stage transformation that compounds error from both perceptual and reasoning stages. Models without autonomous-driving fine-tuning (`Llama-3.2-11B-Vision-Instruct`, `Mistral-Small-2603`) apply general-purpose directional heuristics that do not respect road geometry, yielding lane-change predictions that are directionally correct but geometrically inconsistent. `Gemma-3-4B-it` and `Qwen3-VL-8B-Instruct` fail at the perceptual stage before any transformation is attempted: lane boundary ambiguity prevents the model from establishing the initial spatial state required as transformation input. Larger models (`GPT-4o-mini`, `Gemini-2.5-Flash`) succeed on straight-road scenarios but fail on curved roads and complex intersections where the turning radius must be inferred from road curvature rather

than arrow cues. The persistence of this failure across parameter scales points to the fundamental challenge of learning continuous kinematic transformations from discrete static image supervision.

E3.5 — LEGO & Assembly Transformation Failure.
Failure to predict the structural result of assembling, rotating, or integrating LEGO bricks or structured mechanical components.

Affected models: `Claude 4.5 Haiku`, `Gemini-2.5-Flash`, `Gemini-3.1-Flash-Preview`, `Gemma-3-4B-it`, `Gemma-4-31B-it`, `Llama-3.2-11B-Vision-Instruct`, `Mistral-Small-2603`, `Kimi K2.5`, `o4-mini`, `GPT-4o-mini`, `GPT-5-mini`, `GLM-4.5V`.

Root-cause analysis. LEGO assembly transformation—predicting the outcome of adding a new brick to an existing structure—requires simultaneously tracking the current structure's geometry, the new brick's orientation, and the stud-connector constraints that determine valid attachment points. Lightweight models (`Gemma-3-4B-it`, `Qwen3-VL-8B-Instruct`) fail to represent the current structure's 3D geometry with sufficient fidelity to reason about which attachment points are accessible, producing structurally impossible assembly predictions. Mid-size models correctly parse the current state but fail when the new piece requires rotation before attachment (a two-operation transformation), indicating a serialisation bottleneck where the second transformation is dropped. Even strong models (`GPT-5-mini`, `Gemini-2.5-Flash`) show elevated error rates on multi-piece additions involving colour-identical bricks, where tracking which brick was added requires maintaining brick identity across isometric views—a capability beyond standard 2D visual grounding.

I.4 E4 — Quantitative Spatial Failure

Definition. A **quantitative spatial failure** occurs when the model is required to produce or reason with a numerical value derived from spatial properties of the scene—counting, distance measurement, size estimation, angle calculation, or area estimation—and produces an incorrect numerical result.

Benchmark examples.

- *"How many bottles of beer are there in the picture in total?"* — model produces an incorrect count.
- *"Can you measure the straight-line distance in meters between the self-car and Entity #1?"* — model provides a wrong metric estimate.
- *"Comparing the total height of the first four floors of the gray house . . . which one is taller?"* — model fails a relative height comparison using image-plane cues.
- *"How many of the yellow taxi cabs could park end-to-end in the space between the front of the department store and the crosswalk?"* — model misestimates the spatial ratio needed.
- *"At approximately what angle is the worker's cap tilted with respect to the ground?"* — model fails angular measurement from a single image.

E4.1 — Counting & Enumeration Failure.
Failure to correctly count discrete objects, structural units, or category instances in the image.

Affected models: `Gemini-2.5-Flash`, `Gemini-3.1-Flash-Preview`, `Gemma-3-4B-it`, `Gemma-4-31B-it`, `Llama-3.2-11B-Vision-Instruct`, `Mistral-Small-2603`, `o4-mini`, `GPT-4o-mini`, `GPT-5-mini`, `Qwen3-VL-8B-Instruct` , `GLM-4.5V`.

Root-cause analysis. Counting requires iterative visual scanning with a held tally—a sequential attentional operation that is fundamentally at odds with the parallel feedforward pass of a standard vision transformer. Lightweight models (`Qwen3-VL-8B-Instruct` , `Gemma-3-4B-it`) exhibit the largest absolute counting errors, attributable to coarse patch tokenisation that merges spatially adjacent, visually similar objects into single tokens. Mid-size models (`Mistral-Small-2603`, `GLM-4.5V`) perform acceptably on small sets but show superlinear error growth with set size, consistent with a fixed attention capacity that is overwhelmed as item count increases beyond approximately seven objects. Larger models (`GPT-4o-mini`, `o4-mini`) maintain lower mean errors but exhibit category-specific failures on objects with complex partial occlusions (e.g., LEGO bricks in dense stacks), where disambiguation requires integrating evidence across multiple spatial locations simultaneously.

E4.2 — Distance & Size Estimation Failure.
Failure to estimate metric distances between objects, object physical sizes, heights, widths, or areas in spatial scenes.

Affected models: all 13 evaluated models.

Root-cause analysis. Metric distance estimation from monocular images requires camera intrinsic calibration or strong learned scene-scale priors—neither of which is reliably available to general-purpose vision–language models trained on uncalibrated internet images. Benchmark questions specify distances in meters (e.g., *"How many meters away is Entity #1 from my car?"*), requiring models to access an implicit mapping from pixel distances to real-world metric units that is only available if the model has internalised strong prior knowledge about the driving scene's scale. Lightweight models (`Gemma-3-4B-it`, `Qwen3-VL-8B-Instruct`) produce physically implausible estimates (e.g., reporting 0.5 m for a vehicle clearly tens of meters away), indicating the complete absence of scene-scale calibration. Mid-size and large models produce more plausible ranges but still exhibit systematic underestimation of large distances and overestimation of small ones, consistent with a regression-to-the-mean artefact from averaged training labels. The universality of this failure underscores the need for depth-supervised or LiDAR-calibrated pre-training data to develop reliable metric spatial grounding.

E4.3 — Mathematical & Logical Quantification Failure.
Failure on math-embedded spatial tasks including arithmetic, constraint optimisation, combinatorial reasoning, or graph-based calculations.

Affected models: `Claude 4.5 Haiku`, `Gemini-2.5-Flash`, `Gemini-3.1-Flash-Preview`, `Gemma-3-4B-it`, `Llama-3.2-11B-Vision-Instruct`, `Mistral-Small-2603`, `o4-mini`, `GPT-4o-mini`, `GPT-5-mini`, `Qwen3-VL-8B-Instruct` , `GLM-4.5V`.

Root-cause analysis. The tasks in this subcategory (e.g., *"Alfonso the Ostrich buried his head . . . for 98 hours and 56 minutes—when did he start?"*) embed arithmetic inside a narrative or visual context, requiring models to first extract numerical premises from the image and then apply multi-step arithmetic. Lightweight models fail predominantly at extraction: they misread clock faces, misparse tabular data, or misidentify the relevant numerals, so the arithmetic is never correctly posed. Mid-size models extract numbers accurately but commit arithmetic errors in the computation phase, particularly in modular-arithmetic contexts (e.g., clock wrap-around at midnight) that are underrepresented in standard pre-training problem sets. Larger models (`o4-mini`, `GPT-5-mini`) show a distinct pattern: high accuracy on arithmetic given correctly parsed inputs, but residual failures when visual OCR produces noisy numerals, highlighting the brittleness of the visual parsing front-end even in capable reasoning models.

E4.4 — Road & Safety Quantification Failure.
Failure to quantify safety-relevant road-scene properties such as following distance, lane width, or occlusion extent.

Affected models: `Gemini-2.5-Flash`, `Gemini-3.1-Flash-Preview`, `Llama-3.2-11B-Vision-Instruct`, `Mistral-Small-2603`, `Kimi K2.5`, `o4-mini`, `GPT-4o-mini`, `GPT-5-mini`, `Qwen3-VL-8B-Instruct` .

Root-cause analysis. Road-safety quantification tasks ask models to assess whether a spatial gap, speed differential, or visibility margin is sufficient for a safe manoeuvre— a judgement that requires integrating metric estimates with domain-specific safety thresholds. Without autonomous-driving fine-tuning, most evaluated models lack calibrated distance priors for vehicle-following scenarios, causing systematic over- or underestimation of safety margins. `Llama-3.2-11B-Vision-Instruct` and `Mistral-Small-2603` apply generic spatial language (*"close," "far"*) rather than metric reasoning, making their answers inappropriate for numerical safety assessments. `Kimi K2.5` and `Qwen3-VL-8B-Instruct` exhibit binary failures: they correctly identify the proximity relation (near/far) but cannot produce calibrated metric estimates, indicating a qualitative-only spatial representation. Even models with stronger general reasoning (`o4-mini`, `GPT-5-mini`) fail when occlusion is the primary safety cue, reflecting an inability to infer the extent of occluded spatial regions from partial visibility information alone.

I.5 E5 — Multi-Step Reasoning Failure

Definition. A **multi-step reasoning failure** occurs when the model executes one or more spatial or logical reasoning steps correctly but fails on a subsequent step that depends on the output of the previous one. The model demonstrates partial competence—it is not a simple perceptual error —but cannot maintain and propagate intermediate results across a chain of dependent inferences.

Benchmark examples.

- LEGO step-ordering: *"Arrange the following options (A, B, C, D) in the correct order to transition from the current state to the target state."* — model correctly identifies individual transition steps but assembles them in the wrong order.
- LEGO step-sequence: *"Four of these steps correctly fit in the sequence . . . one does NOT belong."* — model identifies most steps but selects the wrong outlier.
- Toy-car collision: *"If the toy car is driving forward, is it enough to remove the stack of books to prevent all collisions? If not, what else should be done?"* — model resolves the first obstacle but fails to identify a second one downstream.
- Visual pattern: *"Divide the following six figures into two categories so that each category has its own common characteristics."* — model identifies one feature dimension but fails to integrate a second.
- Logic puzzle: *"In the addition sum below, a, b, and c stand for different digits. What is the value of $a + b + c$?"* — model produces the correct value for one variable but propagates an error into the others.

E5.1 — Sequential Assembly & Procedural Step Failure.
Failure to correctly order, predict, or validate multi-step LEGO assembly or structured procedural sequences.

Affected models: all 13 evaluated models.

Root-cause analysis. LEGO step-ordering tasks require the model to (1) parse the current and target assembly states, (2) decompose the transformation into atomic brick-addition steps, and (3) order those steps into a valid construction sequence—a three-stage compositional reasoning chain. The universality of failure across all models reveals a fundamental limitation: no evaluated model maintains a persistent symbolic representation of intermediate assembly states across reasoning steps, instead re-parsing each image from scratch at each decision point. Lightweight models (`Gemma-3-4B-it`, `Qwen3-VL-8B-Instruct`) fail at stage (1) due to insufficient 3D scene-parsing capacity, so the ordering problem is never correctly instantiated. Mid-size models succeed at parsing but fail at ordering, often producing permutations that are locally consistent (each adjacent pair is plausible) but globally inconsistent (the final state does not match the target), indicating greedy local reasoning rather than global sequence planning. Larger models (`GPT-5-mini`, `o4-mini`) show above-chance global ordering but still fail on sequences longer than four steps, consistent with a context-window planning horizon that does not scale to longer assembly tasks.

E5.2 — Action & Event Prediction Chain Failure.
Failure to predict subsequent events or outcomes by chaining scene observation to forward simulation across multiple steps.

Affected models: `Gemini-2.5-Flash`, `Gemini-3.1-Flash-Preview`, `Llama-3.2-11B-Vision-Instruct`, `Mistral-Small-2603`, `Kimi K2.5`, `o4-mini`, `GPT-4o-mini`, `GPT-5-mini`, `Qwen3-VL-8B-Instruct` , `GLM-4.5V`.

Root-cause analysis. Event-chain prediction asks models to simulate a causal sequence starting from a perceived scene—a forward-planning task that requires a world model capable of propagating physical or social state across multiple time steps. Most evaluated models produce accurate first-step predictions (e.g., correctly identifying the first obstacle in the toy-car collision task) but fail to continue the simulation to second- or third-order consequences. `Llama-3.2-11B-Vision-Instruct` and `Mistral-Small-2603` systematically terminate prediction after the most salient first event, ignoring subsequent causal chains—a behaviour consistent with an attention mechanism dominated by

high-saliency foreground objects. Instruction-tuned models (`GPT-4o-mini`, `GPT-5-mini`) exhibit language-prior saturation: having produced a plausible first answer, the model treats the question as resolved and does not search for additional downstream consequences, leading to incomplete multi-step answers. `Kimi K2.5` and `Qwen3-VL-8B-Instruct` show a related pattern of premature termination, compounded by limited world-knowledge depth that prevents the model from anticipating non-obvious physical consequences of first-order events.

E5.3 — Driving Multi-Step Inference Failure.
Failure to chain multiple driving-scene observations into a correct multi-step safety, navigation, or motion decision.

Affected models: `Claude 4.5 Haiku`, `Gemini-2.5-Flash`, `Gemini-3.1-Flash-Preview`, `Gemma-3-4B-it`, `Gemma-4-31B-it`, `Llama-3.2-11B-Vision-Instruct`, `Kimi K2.5`, `o4-mini`, `GPT-5-mini`.

Root-cause analysis. Driving multi-step inference requires composing lane-position grounding, distance estimation, and safety-policy reasoning into a single coherent output—a three-step chain where errors at any stage invalidate the final answer. Lightweight models fail early in the chain: imprecise lane-position grounding invalidates all downstream reasoning regardless of safety-policy competence. Mid-size models (`Claude 4.5 Haiku`, `Gemma-4-31B-it`) frequently resolve lane positions correctly but apply safety rules inconsistently, suggesting that their safety-policy knowledge is encoded as surface-level patterns rather than generalisable principles. `Kimi K2.5` and `o4-mini` show a distinct failure mode on multi-vehicle scenes: they correctly assess the ego vehicle's situation but fail to account for the simultaneous manoeuvres of other traffic participants, indicating a single-agent rather than multi-agent reasoning capacity. The combined effect of these compounding error sources produces a total failure rate substantially higher than the sum of individual sub-step error rates, consistent with error multiplication in multi-step inference chains.

E5.4 — Mathematical & Algorithmic Reasoning Chain Failure.
Failure to execute multi-step mathematical or algorithmic reasoning where earlier correct steps are prerequisites for later ones.

Affected models: `Claude 4.5 Haiku`, `Gemini-2.5-Flash`, `Gemini-3.1-Flash-Preview`, `Gemma-3-4B-it`, `Gemma-4-31B-it`, `Llama-3.2-11B-Vision-Instruct`, `Mistral-Small-2603`, `Kimi K2.5`, `o4-mini`, `GPT-4o-mini`, `GPT-5-mini`, `Qwen3-VL-8B-Instruct` , `GLM-4.5V`.

Root-cause analysis. Multi-step mathematical tasks (e.g., sum puzzles with symbolic variable constraints) require models to establish partial solutions as binding intermediate constraints before proceeding to the next variable—a backtracking-aware reasoning process that does not arise naturally in auto-regressive language generation. Lightweight models fail primarily on the visual parsing step: misreading digits or symbols from the rendered puzzle image introduces a corrupted premise that invalidates the entire solution chain. Mid-size models parse the puzzle correctly but fail to propagate constraints across variables, typically solving each variable independently rather than jointly, leading to solutions that satisfy individual constraints but violate global consistency. Larger models (`GPT-5-mini`, `o4-mini`) are competitive with human performance on standard two-variable puzzles but fail on three-variable or modular-arithmetic variants, revealing the boundary of their constraint- propagation capacity. The near-universal nature of this failure across all models highlights a fundamental tension between the sequential, left-to-right generation process of language models and the constraint-satisfaction or backward-chaining inference required by multi-step mathematical reasoning.

I.6 E6 — Language–Spatial Mapping Failure

Definition. A **language–spatial mapping failure** occurs when the model misinterprets the spatial language used in the question itself, independently of any perceptual or reasoning error. The failure lies at the linguistic–spatial interface: spatial prepositions, directional terms, or reference-frame markers are parsed incorrectly, producing a spatially valid but semantically inverted answer.

Benchmark examples.

- *"If the 'chimpanzee forest' arrow is pointing south, which direction is the other arrow pointing?"* — model confuses the allocentric mapping, reporting the opposite cardinal direction.
- *"The pictures are taken continuously from a first-person perspective. At the moment of the last picture, in which direction is the door relative to you?"* — model conflates "relative to me" with "relative to the camera's world orientation."
- *"When I was taking photo 2, in which direction was the stove relative to me?"* — model correctly identifies the stove but reverses the directional label.

Note on prevalence. E6 was the sparsest failure category: only `Llama-3.2-11B-Vision-Instruct` exhibited an explicit E6 error cluster in the source data, and in other models E6-type failures appear to be subsumed under E2 (spatial grounding), where the language-parsing error is conflated with the grounding error it induces. Three second-level subcategories are defined below for completeness and to support future annotation with finer linguistic granularity.

E6.1 — Directional Preposition Confusion.
Model swaps or misapplies directional terms such as left/right, above/below, or in-front/behind.

Affected models: `Gemini-3.1-Flash-Preview`, `Llama-3.2-11B-Vision-Instruct`.

Root-cause analysis. Directional prepositions are among the most context-sensitive spatial expressions in English: "left of X" is only well-defined once both the reference frame (whose left?) and the reference object are established. `Llama-3.2-11B-Vision-Instruct` exhibits systematic left–right reversals on allocentric descriptions given from a named compass direction (e.g., *"facing south, which is to the west?"*), suggesting that its spatial-language representations conflate multiple reference-frame conventions. `Gemini-3.1-Flash-Preview` fails on relative directional navigation tasks that require updating a spatial frame incrementally across multiple described movements—a discourse-level tracking operation for which its context window representation appears inadequate. Both models show markedly higher accuracy when directional terms are accompanied by visual arrows or diagram markers, confirming that the failure is linguistic rather than perceptual: the models can perceive the relevant spatial relationships but cannot correctly decode the verbal description of them. This failure mode is likely under-reported in existing benchmarks because most spatial VQA datasets provide visual rather than purely verbal directional cues, masking the linguistic-interface deficit.

E6.2 — Reference-Frame Ambiguity Failure.
Failure to resolve whether a spatial description is given in an allocentric (world-centred) or egocentric (observer-centred) frame.

Affected models: `Gemini-3.1-Flash-Preview`, `Llama-3.2-11B-Vision-Instruct`.

Root-cause analysis. Natural spatial language is systematically ambiguous between egocentric and allocentric readings: "the door on the left" may refer to the speaker's left or to an absolute compass orientation, depending on context. Current vision–language models are not explicitly trained to track which reference frame is active, leading to arbitrary frame selection that produces correct answers only when the two frames coincide. `Llama-3.2-11B-Vision-Instruct` defaults strongly to an allocentric frame, causing errors on tasks that explicitly invoke an egocentric perspective (e.g., first-person photo sequences with questions about "what is to my right"). `Gemini-3.1-Flash-Preview` shows the reverse tendency: defaulting to an egocentric frame, producing errors on tasks that require a stable world-centred compass interpretation (e.g., *"facing south, which room is on the west side?"*). Both models perform markedly better when the reference frame is made explicit in the question text (*"from your perspective"* vs. *"from a bird's-eye view"*), confirming that the failure is one of frame disambiguation rather than spatial reasoning per se.

I.7 E7 — Format / Refusal Failure

Definition. A **format or refusal failure** occurs when the model does not produce a valid, answerable response to a spatial question, regardless of its underlying spatial competence. This includes outright refusals, responses that acknowledge the question but decline to commit to an answer (e.g., "I cannot

determine this from the image"), malformed outputs that do not conform to the required response format, and answers that address a different question than the one asked.

Benchmark examples.

- Driving ego-motion: *"How would you describe the driving scene involving our car? Please explain, focusing on our car's turning manoeuvre."* — model produces a generic description that avoids committing to any specific turning direction.
- LEGO rotation necessity: model is asked to respond with only "A" or "B" but produces a paragraph-length justification without a letter selection.
- Narrative collision: *"If the toy car is driving forward . . . E. Removing the stack of books is sufficient; F. Not Sure."* — model selects "F. Not Sure" without attempting a spatial analysis.
- Road-lane classification: model asked to select from four named lane options produces a free-text lane description that cannot be mapped to any of the offered choices.

E7.1 — Driving-Scene Output Failure.
Model refuses or produces an invalid answer for driving-scene tasks including ego-motion description, vehicle turning, lane-change, or road-safety judgements.

Affected models: all 13 evaluated models.

Root-cause analysis. Driving-scene description questions are the dominant source of E7 failures across all models, reflecting a systematic tension between the open-ended free-text nature of the required answer and the models' tendency to hedge when spatial confidence is low. All evaluated models produce evasive or underspecified descriptions (e.g., *"the vehicle appears to be making a turn"*) when the spatial evidence is ambiguous, rather than committing to a specific directional answer as required by the benchmark. Instruction-tuned models (`GPT-5-mini`, `o4-mini`) are additionally susceptible to safety-fine-tuning artefacts: they occasionally refuse to answer driving-scene questions by citing uncertainty about road conditions, even when the benchmark provides sufficient information for a definitive answer. Lightweight models (`Gemma-3-4B-it`, `Qwen3-VL-8B-Instruct`) produce format violations by ignoring the required multiple-choice structure and generating free-text responses, indicating insufficient instruction-following calibration for constrained-output spatial tasks. The near-universal distribution of this failure underscores the need for calibration training that teaches models to commit to best-estimate spatial answers under uncertainty rather than defaulting to refusal or vagueness.

E7.2 — LEGO & Assembly Output Failure.
Model refuses or produces an invalid answer for LEGO rotation, adjacency, or assembly prediction tasks.

Affected models: `Claude 4.5 Haiku`, `Gemini-2.5-Flash`, `Gemini-3.1-Flash-Preview`, `Gemma-3-4B-it`, `Gemma-4-31B-it`, `Llama-3.2-11B-Vision-Instruct`, `Mistral-Small-2603`, `Kimi K2.5`, `o4-mini`, `GPT-4o-mini`, `GPT-5-mini`, `Qwen3-VL-8B-Instruct` , `GLM-4.5V`.

Root-cause analysis. LEGO assembly questions require models to respond with a single letter (A–E) from a constrained option set—a format that is simple in principle but frequently violated in practice. Lightweight models (`Gemma-3-4B-it`, `Qwen3-VL-8B-Instruct`) produce extended justifications rather than single-letter answers, reflecting a training regime that rewards verbose explanations over concise constrained outputs. Mid-size models (`Mistral-Small-2603`, `Gemini-3.1-Flash-Preview`) produce correctly formatted but null-valued outputs (e.g., outputting only `None` or `N/A`) when their internal spatial confidence falls below a threshold, effectively converting spatial uncertainty into a format failure. RLHF-trained models (`GPT-4o-mini`, `o4-mini`) occasionally produce a letter *followed by* a multi-sentence justification that contradicts the chosen letter, creating an internally inconsistent output that cannot be reliably scored. The breadth of this failure across all models suggests that constrained-output compliance for multi-image spatial tasks requires dedicated format-compliance training beyond what standard instruction-tuning provides.

E7.3 — Narrative & Abstract Output Failure.
Model refuses or produces an invalid answer for tasks involving narrative interpretation, lyrical reasoning, or abstract visual classification.

Affected models: `Claude 4.5 Haiku`, `Gemini-2.5-Flash`, `Gemini-3.1-Flash-Preview`, `Llama-3.2-11B-Vision-Instruct`, `Mistral-Small-2603`, `Kimi K2.5`, `o4-mini`, `GPT-4o-mini`, `GPT-5-mini`, `Qwen3-VL-8B-Instruct` , `GLM-4.5V`.

Root-cause analysis. Narrative-spatial tasks (*"what measures should be taken to avoid a collision?"*) require the model to commit to a complete action plan rather than a single categorical answer, creating an opportunity for under-specification that most models exploit. Models trained with strong safety fine-tuning (`Claude 4.5 Haiku`, `GPT-4o-mini`) show elevated refusal rates on toy-collision scenarios, possibly because collision-avoidance phrasing activates content-policy classifiers even though the task is purely spatial and benign. `Llama-3.2-11B-Vision-Instruct` and `Mistral-Small-2603` produce answers that enumerate all visible objects as potential collision hazards rather than identifying only those in the vehicle's path, generating technically valid but informationally vacuous responses that fail on precision. `Qwen3-VL-8B-Instruct` and `Kimi K2.5` select "Not Sure" (option F in several benchmark templates) at rates substantially above chance, indicating a calibration bias toward epistemic hedging for open-ended spatial tasks. The distribution of this failure suggests that output-format compliance for open-ended spatial action plans requires training on graded-confidence response templates that reward precise, committed answers over non-committal hedging.

E7.4 — 3D Spatial Reasoning Output Failure.
Model refuses or produces an invalid answer for 3D adjacency, cube rotation, or geometric transformation tasks requiring a specific letter or category selection.

Affected models: `Claude 4.5 Haiku`, `Gemma-4-31B-it`, `GLM-4.5V`.

Root-cause analysis. This subcategory is the most narrowly distributed E7 failure, appearing in only three models—all of which show a specific inability to commit to a single-letter response on three-image LEGO assembly tasks involving rotation necessity judgements. `Claude 4.5 Haiku` and `Gemma-4-31B-it` both produce truncated outputs in which the required letter (A or B) is omitted at the end of an otherwise complete reasoning chain, suggesting a generation-length or early-stopping artefact specific to multi-image prompts. `GLM-4.5V`'s failures in this subcategory cluster on 3D adjacency questions where the answer is a binary adjacency judgement—a task type that appears to conflict with its decoding preferences, possibly because adjacency classifications were absent or rare in its fine-tuning data. All three models produce the correct answer in a verbose prose form but fail to extract and present it in the required constrained format, indicating that the failure is one of output post-processing rather than spatial reasoning capacity.

J Hard-Case Construction and Evaluation

Hard-Case Construction. To test whether agentic reasoning can improve performance beyond direct VLM inference, we construct a *hard-case subset* from instances that remain unsolved by all evaluated frontier VLMs. For each task category, we sample 25 such instances, yielding a balanced set of 175 cases that represents regions where current VLMs fail. This subset is used exclusively to assess whether additional reasoning, decomposition, and tool use can recover errors that persist under direct model inference.

Agent-Based Evaluation Setup. We evaluate two agentic configurations: (i) **SpatioXolver**, a multi-agent system that decomposes spatial reasoning across specialized agents, and (ii) a **Claude Code**-based agent that uses sequential reasoning with programmatic tools. Both receive the same inputs and operate under the same evaluation constraints. The goal is not to compare agent architectures, but to test whether agentic scaffolding can solve cases that frontier VLMs fail to solve through direct inference.

Evaluation Protocol. Each agent generates a final answer for every instance in the hard-case subset. We evaluate responses using the same LLM-based judging framework described in Section 3.2. We report the number and proportion of correctly solved instances overall and within each category. Because all instances were selected from the failure set of the evaluated VLMs, any successful solution directly indicates recovery beyond raw VLM performance.

Comparative Analysis Framework. We compare agent performance with the direct-inference results of the frontier VLMs on the same hard cases. This analysis asks a focused question: *can agentic scaffolding recover spatial reasoning failures that remain unsolved by the underlying VLMs?* We report category-level success rates to determine whether such gains occur consistently across spatial reasoning dimensions or are concentrated in particular types of tasks.

K DISCUSSION

K.1 BROADER IMPACTS

Advances in spatial reasoning for multimodal large language models (MLLMs) have meaningful societal implications, spanning both beneficial applications and potential risks.

Positive Impacts. Improved spatial reasoning is a key enabler for safe and effective embodied AI. Models with stronger spatial understanding can advance applications in autonomous driving, robotic manipulation, augmented reality, and geospatial analytics. In addition, such capabilities can support assistive technologies for visually impaired users by enabling more accurate, context-aware scene descriptions and navigation assistance. By providing a rigorous and comprehensive benchmark, FRONTIERSPATIAL facilitates the development of AI systems that are more reliable, interpretable, and structurally grounded.

Risks and Negative Impacts. At the same time, enhanced spatial intelligence introduces dual-use risks. Capabilities such as fine-grained spatial tracking, multi-agent temporal reasoning, and cross-view inference could be repurposed for large-scale surveillance, drone tracking, or military applications. Moreover, our analysis highlights important fairness and safety concerns: current models exhibit geographic and domain biases, often underperforming on scenes or layouts that deviate from dominant training distributions. In safety-critical settings (e.g., navigation or medical imaging), such systematic failures could lead to uneven performance and potentially harmful outcomes for underrepresented populations or environments.

Mitigation and Responsible Use. While this work focuses on evaluation rather than deployment, FRONTIERSPATIAL itself contributes to risk mitigation. By systematically exposing failure modes, including reference-frame confusion, geographic bias, and spatial misestimation, the benchmark provides actionable diagnostics for developers to identify and address weaknesses prior to real-world deployment. We encourage future work to incorporate bias-aware training, uncertainty estimation, and domain-diverse data to improve robustness and fairness in spatial reasoning systems.

K.2 LIMITATIONS

LLM Judge Variance and Calibration. The automated evaluation pipeline relies on a vision-language model (e.g., *Gemini-3.1-Flash-Image-Preview*) as a judge. Although temperature is fixed at 0.1 to reduce stochasticity, automated evaluation can introduce systematic or instance-level judgment errors. We mitigate this through a dedicated human validation study (Section F), where the LLM judge achieves perfect agreement with human majority annotations ($\kappa = 1.0$), exceeding measured inter-human agreement ($\kappa = 0.81$). Thus, while the judge is not a substitute for human evaluation in principle, these results support its use as a consistent proxy for large-scale benchmarking.

Scope of Visual Representation. The benchmark evaluates spatial reasoning primarily through static monocular images, including 2D projections of 3D scenes and temporally ordered frames. It does not directly evaluate embodied or interactive spatial intelligence in dynamic environments. This is an intentional scope choice: isolating visual reasoning from navigation policies, motor control, and environment-specific action loops enables controlled and comparable evaluation of the underlying reasoning capabilities.

Reliance on Existing Data. FRONTIERSPATIAL introduces no new images or questions; it instead combines, curates, deduplicates, and contextualizes existing datasets under a unified taxonomy and evaluation framework. This limits our ability to introduce entirely new visual phenomena beyond those represented in the source datasets. However, new data collection is outside the intended contribution of this work: the goal is to provide a systematic measurement framework that makes fragmented existing resources jointly usable, comparable, and diagnostically informative.

Source Dataset Biases. Because the benchmark is constructed from existing datasets, it inherits biases in their visual domains, annotation schemes, language, task formulations, and sampling procedures.Through our four-stage curation pipeline, we have systematically sought to mitigate these biases through taxonomy-based coverage, quality filtering, de-duplication, and balanced task sampling. Nevertheless, some biases may persist due to limitations inherited from the source datasets, which cannot be fully addressed through post-hoc curation.

Coverage of Spatial Intelligence. Although the taxonomy spans seven broad categories and 65 fine-grained tasks with both single and multi-image reasoning, spatial intelligence is substantially broader than any finite framework. Abilities such as very large long-horizon physical interaction, active exploration, continuous 3D scene reconstruction, and sensorimotor reasoning remain outside the current scope. The taxonomy should therefore be viewed as a structured operationalization of major visual spatial reasoning dimensions, rather than an exhaustive definition of spatial intelligence.

Agent-Based Hard-Case Analysis. The agent-based evaluation is intentionally limited to a small hard-case subset and is designed to test whether agentic scaffolding can recover failures of direct VLM inference. It is not a comprehensive comparison of agent architectures, prompting strategies, or tool-use policies. The results should therefore be interpreted as evidence that additional reasoning and tool use can recover some otherwise unresolved cases, rather than as a definitive ranking of agentic approaches.

L PROMPTS USED IN FRONTIERSPATIAL

FRONTIERSPATIAL: Prompt for Answering Model

Answer the question using clear visual and logical reasoning.
Question:
`{question}`
Context (Some possible answers):
`{choices}`
Instruction: Give only the final answer in one sentence.

FRONTIERSPATIAL: Prompt for Error Taxonomy

You are an expert in semantic classification and taxonomy design.

Map each cluster (with one example) to an **Error taxonomy category**.

Constraints:

- Minimum total taxonomy categories: 30
- Maximum total taxonomy categories: 50
- Reuse existing categories whenever possible
- Only create new categories when necessary
- Keep names concise, distinct, and generalizable

Input:
`Clusters:{clusters}`

Output (STRICT JSON):

```
{
    "cluster_id": "taxonomy"
}
```

FRONTIERSPATIAL: Promt For Error Category Generation (E1–E7)

You are an expert error taxonomy classifier for spatial reasoning models.

You will be given a set of clustered model failures. Your task is to assign each cluster to exactly **ONE primary error type** from a fixed taxonomy.

Return ONLY valid JSON. Do NOT include explanation, markdown, or extra text.

========== FIXED ERROR TAXONOMY ==========

`{error_taxonomies}`

========== INPUT ==========

Cluster ID: `{cluster_id}`

Cluster Examples:

`{list_of_failure_samples}`

========== TASK ==========

1. Carefully analyze the cluster behavior.
2. Identify the **SINGLE** best matching error type.
3. Do NOT assign multiple labels.
4. Prefer the dominant failure mode in the cluster.

========== OUTPUT FORMAT ==========

Return **STRICT JSON only** (no trailing commas, no Python types, use `null` not `None`):

```
{
  "cluster_id": "{cluster_id}",
  "error_type": "",
  "error_name": "string label of selected type"
}
```

FRONTIERSPATIAL: Prompt for Judging Model

You are an expert evaluation agent for spatial reasoning tasks. Your job is to rigorously evaluate a model's response.
Return ONLY a valid JSON object. Do NOT include any explanation, markdown, or extra text.
========== INPUT ==========
Question:
`{question}`
Ground Truth Answer:
`{ground_truth}`
Model Response:
`{model_output}`
========== INSTRUCTIONS ==========

1. **Accuracy Score (0/1):**
   - 1 if the final answer is correct.
   - 0 otherwise.
   - For OE: use semantic equivalence.
2. **Refusal Score (0/1):**
   - 1 if the model refuses (e.g., "I cannot answer", "insufficient information").
   - 0 otherwise.
3. **Off Topic Score (0/1):**
   - 1 if response is irrelevant.
   - 0 if relevant.
4. **Extract Reasoning:**
   - Extract and summarize the reasoning for the score.
5. **Spatial Elements Score (0/1):**
   - 1 if key spatial objects/entities are correctly identified.
   - 0 otherwise.
6. **Spatial Relations Score (0/1):**
   - 1 if spatial relations (left/right/above/below/etc.) are correct.
   - 0 otherwise.
7. **Spatial Reasoning Score (0/1):**
   - 1 if logical spatial reasoning is correct.
   - 0 otherwise.

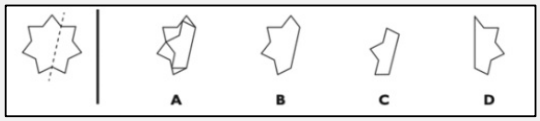
========== OUTPUT FORMAT ==========
Return **STRICT JSON only** (no trailing commas, no Python types, use `null` not `None`):

```
{
  "accuracy_score": 0,
  "refusal_score": 0,
  "off_topic_score": 0,
  "spatial_elements_score": 0,
  "spatial_relations_score": 0,
  "spatial_reasoning_score": 0,
  "reasoning": ""
}
```

M REPRESENTATIVE BENCHMARK SAMPLES ACROSS ALL TASKS

Here (including supplementary PDF) we include at least one example (with responses from different models) from each of 65 tasks.

Spatial & Geometric Reasoning : 2D Reasoning : 2D Geometry



Question: Which figure that shows how the left-hand figure will look when folded along the dotted line? Possible Choices: A,B,C,D

Gold Answer: A

anthropic_claude_haiku_4.5: Analysis of Paper Folding Looking at the lefthand figure a sixpointed star when folded along the vertical dotted line The right half of the star will fold over onto the left half The resulting silhouette will show the stars profile from the side The folded shape will appear as an irregular quadrilateralpolygon outline Comparing the options Option A Shows a shape that matches the profile of the folded star Options B C D Show different orientations that dont match the fold result The answer is A

google_gemini_2.5_flash: When the lefthand figure is folded along the dotted line the right half of the star will overlap the left half resulting in a shape that resembles option B The dotted line is not a line of symmetry so the folded shape will not be a perfect half of the original star Option A shows the two halves overlapping but not perfectly aligned as if folded Options C and D are incorrect as they do not represent the result of folding the given figure along the dotted line The final answer is boxedB

google_gemini_3.1_flash_lite_preview: The correct answer is A because folding the star along the dotted line causes the overlapping segments of the stars points to create the complex outline shown in that figure

google_gemma_3_4b_it: Figure B accurately depicts the folded shape of the lefthand figure when folded along the dotted line

google_gemma_4_31b_it: Based on visual reasoning folding the lefthand figure along the dotted line reflects the left side onto the right side resulting in the shape shown in figure B Final Answer B

meta_llama_llama_3.2_11b_vision_instruct: The correct answer is C

mistralai_mistral_small_2603: B

moonshotai_kimi_k2.5: Let me look at the shape again The original figure is an 8pointed star The dotted line goes through two opposite valleys of the star Actually looking at the image the dotted line is positioned such that it passes through two indentations valleys of the star Let me count the points on the star It looks like there are 8 points total The dotted line is vertical Let me trace the left half Starting from the top theres a point Then a valley Then a point Then a valley where the dotted line passes through Then a point Then a valley Then a point Then a valley where the dotted line passes through Wait that doesnt look right Let me look more carefully Actually the dotted line goes through the center and passes through two valleys or indentations Let me think about the shape of the left half When folded the left half will be reflected onto the right side But actually looking at the options these dont look like full stars They look like halfshapes or folded shapes Wait I need to understand what the question is asking It says Which figure shows how the lefthand figure will look when folded along the dotted line When you fold a paper along a line the result is that the paper now has half the area roughly and the shape you see is the outline of the folded paper So if I fold the star along the vertical dotted line the left half will fold over onto the right half The resulting shape will be the

o4_mini: The correct folded shape is shown in figure A

openai_gpt_4o_mini: The answer is C

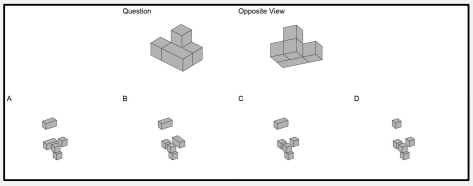
openai_gpt_5_mini: The correct figure is C

qwen_qwen3_vl_8b_instruct: The correct figure is B as it accurately represents the folded shape of the lefthand figure along the dotted line maintaining the correct alignment and symmetry of the original starlike form

z_ai_glm_4.5v: To determine the correct figure visualize folding the left hand shape along the dotted line the resulting shape must match the symmetry and structure of the original when folded Among the options figure A exhibits the appropriate folded appearance

Figure 5: Example from the benchmark and evaluation results. Red denotes incorrect and Green denotes correct answer marked by the judge model.

Spatial & Geometric Reasoning : 3D Reasoning : 3D Assembly



Question: The two images above show a 3D structure from different angles. Which one of the options below could be constructed to appear the same as both given views when observed from the corresponding perspectives without rotation and overlaps? Select the most likely one.
Possible Choices: A,B,C,D

Gold Answer: C

anthropic_claude_haiku_4.5: Comparing the base arrangements in the given views with the options the structure shows one elevated cube with 4 cubes forming an Lshaped base pattern Options A B and C all show this configuration but examining the exact spatial arrangement and how they would appear from both angles Option B matches the configuration most accurately The answer is B

google_gemini_2.5_flash: The final answer is boxedC

google_gemini_3.1_flash_lite_preview: The correct answer is D as it represents the minimal set of cubes required to form the structure shown in both views without any hidden or redundant blocks

google_gemma_3_4b_it: Option B is the most likely answer because it maintains the same relative arrangement of the cubes as seen in the original images when viewed from the corresponding perspectives without rotation or overlap

google_gemma_4_31b_it: Based on the visual analysis of the two perspectives the 3D structure consists of five cubes arranged such that there is a base of three cubes in an Lshape with two additional cubes stacked on top one in the corner and one adjacent which matches the configuration shown in option B Final Answer B

meta_llama_llama_3.2_11b_vision_instruct: The correct answer is C

mistralai_mistral_small_2603: B

moonshotai_kimi_k2.5: 5 cubes or 4 cubes if the rectangular prism counts as 2 cubes Wait in the Opposite View I see A base with what looks like 3 sections A vertical piece at the back Let me think about the Opposite View If the Question view is from one angle the Opposite View is from the opposite angle Let me try to map coordinates Lets say the Question view is looking from the front In the Question image Theres a cube on the left In the center theres a long block 2 cubes long On the right theres a cube On top of the center block theres a cube Actually looking more carefully at the Question image the center piece is oriented horizontally and is 2 cubes long The cube on the left is attached to the left end of this center piece The cube on the right is attached to the right end of this center piece And theres a cube on top of the center piece So if we use coordinates x y z where x is leftright y is frontback and z is updown Lets say the center piece occupies 000 and 100 as a rectangular prism Left cube 1 0 0 Center 0 0 0 and 1 0 0 as one piece or two cubes Right cube 2 0 0 Top cube

o4_mini: The structure shown in option B is the only one that can match both the given front view and the opposite view without rotation or overlaps

openai_gpt_4o_mini: C

openai_gpt_5_mini: D

qwen_qwen3_vl_8b_instruct: D

z_ai_glm_4.5v: To determine the correct option we analyze the 3D structures consistency with both given views The Question and Opposite View images show specific arrangements of cubes and we identify which options 3D structure matches these perspectives without rotation or overlap After evaluating the shapes and cube positions option B is the most likely match

Figure 6: Example from the benchmark and evaluation results. Red denotes incorrect and Green denotes correct answer marked by the judge model.