SPARKBENCH: BEST LESSONS FROM BENCHMARKING HUMAN-CENTRIC SPATIO-PHYSICS REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Understanding how another person can move and interact with the physical world is essential to human coordination and collaboration. Yet even a simple question—can someone bend down to pick up an object without hitting a nearby table or losing balance?—requires reasoning about articulated motion, collision, and physical stability that remains challenging for multimodal large language models. To study Spatio-Physics in humAn-centric Reasoning, we introduce **SPARKBench**, a benchmarK with 5,291 questions over 1,442 images. It covers two macro-level categories: *3D Spatial* reasoning about pose, orientation, and relations, and *Spatio-Physics* reasoning about collision, articulation and kinematics, and intuitive physics. We benchmark 12 vision-language models, finding that all of these models struggle to judge body orientation and movement limits while scoring lower on counterfactual than on descriptive questions. To understand how to improve human-centric spatio-physics reasoning, we compare three representations (Image, Graph, and Code) and two states, estimated from the image or taken from the rendered scene. We learn two best lessons for human-centric spatio-physics reasoning: (1) Structured body information can improve spatial judgments, but physical reasoning needs reliable physical state relevant to the question. (2) Computing the judgment from that state can improve physical reasoning when they use reliable scene geometry and physical properties. SPARKBench with the best lessons provides a basis for identifying model weaknesses and designing better representations and computations for human-centric spatio-physics reasoning.

1 INTRODUCTION

AI systems that help people in the physical world need to understand how the human body moves and interacts with its surroundings. To tell whether a person is reaching with their left or right hand, a model must identify left and right from that person's point of view. To judge whether that hand can reach an object without taking a step, it must also consider how the joints can move and whether nearby objects block the movement. These linked judgments define human-centric spatio-physics reasoning: understanding where body parts are and what movements and interactions are physically possible.

Multimodal large language models (MLLMs) have made rapid progress in understanding images (OpenAI, 2024; Gemini Team, Google, 2025; Li et al., 2024; Wang et al., 2025; Bai et al., 2025), but it remains unclear how well they connect body positions with physical limits. Existing datasets and benchmarks cover different parts of this problem, including body structure, human interactions, spatial relations, and the physics of objects (Table 1). Recognizing body parts and their spatial relations does not by itself establish whether a model can judge what a person can physically do. We therefore need a benchmark that evaluates human spatial and physical reasoning together.

To evaluate spatial and physical reasoning together, we introduce Human-Centric Spatio-Physics Reasoning (SPARKBench), a benchmark with 5,291 questions over 1,442 images (Figure 1). SPARKBench organizes its six categories into 3D Spatial, which tests how bodies are arranged in space, and Spatio-Physics, which tests what movements and interactions are physically possible. We build the benchmark to cover varied human scenes and support checks against known physical conditions. Real photographs provide varied poses and interactions, while 3D scenes built from scratch or from photographs provide stored geometry and physical information for verifying answers. Human review-

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.

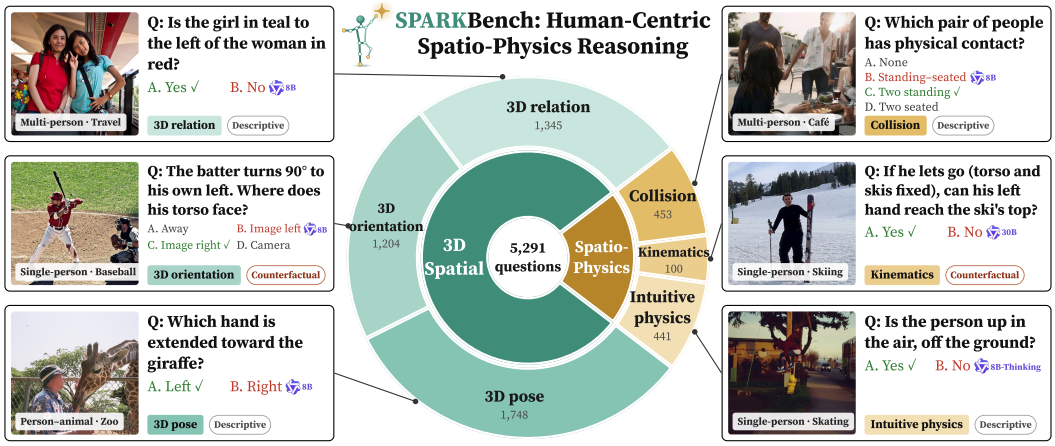


Figure 1: **SPARKBench overview.** 5,291 questions in two task families: 3D Spatial (pose, orientation, relation) and Spatio-Physics (collision, kinematics, intuitive physics). Each example marks the correct answer and the answer chosen by Qwen3-VL.

ers check questions from all sources. Beyond checking answers, the stored scene information lets us compare how models reason when given estimated information or known scene information. SPARK-Bench thus supports both testing models on varied human scenes and studying what information they need for physical reasoning.

With SPARKBench, we benchmark 12 vision-language models, including general-purpose MLLMs and models designed for human or spatial understanding. We measure accuracy on all 5,291 questions, both overall and by category and question type (Table 2). This evaluation leads to two findings. First, overall accuracy conceals gaps within each macro-level category. All ten open models score lower on orientation than on relations in 3D Spatial, and lower on articulation/kinematics than on intuitive physics in Spatio-Physics. Second, most models score lower on counterfactual questions. The ten open models score 2.9–20.8% lower than on descriptive questions. These results show that the tested open models still struggle to judge body orientation and movement limits, and to reason about the effects of a specified scene change. Even leading proprietary models such as GPT-5 and Claude Opus 4.8 still struggle with orientation and kinematics, scoring below 69% and 64% where humans reach 86.7% and 91.7%.

To understand how to improve human-centric spatio-physics reasoning, we further analyze how the representation, its state, and computation affect what a model can judge. Using Qwen3-VL-8B, we compare three representations: Image, Graph (a scene graph), and Code (the same graph with executable checks), each alone or combined with the image. We also compare an estimated state, built from the image, with the oracle state of synthetic scenes. These comparisons lead to two lessons for model design. (1) Represent the physical state the question needs. Estimated scene graphs improve pose and orientation judgments but lower physical accuracy compared with images alone. On synthetic questions, replacing the estimated state with the oracle state improves physical accuracy. (2) Compute the judgment explicitly from that state. Computation adds little with the estimated state but improves physical accuracy with the oracle state.

In summary, this paper makes **three key contributions**: (1) We introduce SPARKBench, a benchmark with 5,291 questions over 1,442 images covering six categories of human-centric spatial and physical reasoning. (2) We benchmark 12 models and identify weaknesses in body orientation, movement limits, and reasoning about specified scene changes that overall scores can hide. (3) We derive two lessons for model design: represent the physical state the question needs, and compute the judgment explicitly from it. Code, models, data and benchmark protocols will be released upon acceptance to promote reproducible evaluation and future research. We hope SPARKBench and these lessons help the community identify weaknesses in current models and design better representations and computations for human-centric spatio-physics reasoning.

Table 1: **Comparison with human-centric datasets and visual reasoning benchmarks.** SPARK-Bench evaluates reasoning about people along two axes, *3D Spatial* (pose, orientation, relation) and *Spatio-Physics* (collision, kinematics, intuitive physics), with descriptive and counterfactual questions over real and synthetic images.

				3D Spatial			Spatio-Physics				Images		3D annotations			
Dataset / benchmark	**Eval. QA**	**Human**	**Pose**	**Ori.**	**Rel.**	**Coll.**	**Kin.**	**Phys.**	**CF**	**Real**	**Syn.**	**Body**	**Box**	**Ori.**	**Phys.**
Human datasets															
COCO (2014)	–	✓	✗	✗	✗	✗	✗	✗	✗	✓	✗	✗	✗	✗	✗
Human3.6M (2014)	–	✓	✓	✗	✗	✗	✗	✗	✗	✓	✗	✓	✗	✓	✗
3DPW (2018)	–	✓	✓	✗	✗	✗	✗	✗	✗	✓	✗	✓	✗	✓	✗
HICO-DET (2018)	–	✓	✗	✗	✓	✗	✗	✗	✗	✓	✗	✗	✗	✗	✗
Hi4D (2023)	–	✓	✓	✗	✓	✓	✗	✗	✗	✓	✗	✓	✗	✓	✗
Spatial and physical reasoning benchmarks															
CLEVRER (2020)	305K†	✗	✗	✗	✓	✓	✗	✓	✓	✗	✓	✗	✗	✓	✓
SpatialVLM (2024)	546	✗	✗	✗	✓	✗	✗	✗	✗	✓	✗	✗	✗	✗	✗
3DSRBench (2025)	2,772	✗	✗	✓	✓	✗	✗	✗	✗	✓	○	✗	✗	✗	✗
PhysBench (2025)	10,002	✗	✗	✗	✓	✓	✗	✓	✗	✓	✓	✗	✗	✗	✗
Human-centric reasoning benchmarks															
VCR (2019)	25,263	✓	✗	✗	✗	✗	✗	✗	✗	✓	✗	✗	✗	✗	✗
JRDB-Reasoning (2025)	NR	✓	✗	✓	✓	✗	✗	✗	✗	✓	✗	✗	✗	✗	✗
Human-MME (2026)	19,945	✓	✓	✓	✗	✗	✗	✗	✗	✓	✗	✗	✗	✗	✗
HumanPCR (2026)	6,618	✓	✓	✓	✓	✓	✗	✗	✗	✓	✗	✗	✗	✗	✗
SPARKBench (Ours)	**5,291**	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	○	✓	○

Ori.: facing and person-centric left/right; Rel.: person–object relations; Coll.: contact and interpenetration; Kin.: joint limits and reach; Phys.: support, balance, and gravity; CF: counterfactual questions. Body / Box / Ori. / Phys.: 3D body, object boxes, orientation, and physical state (contacts, supports, mass). ✓: yes; ○: partial; ✗: no; NR: not reported; †: all splits.

2 Related Work

3D Spatial Reasoning. Explicit geometric representations provide a way to study spatial reasoning. SpatialVLM (Chen et al., 2024) and 3DSRBench (Ma et al., 2025) document large gaps between current models and metric reasoning; benchmarks span CLEVR (Johnson et al., 2017), SpatialBench (Xu et al., 2025), RoboSpatial (Song et al., 2025), and VSI-Bench (Yang et al., 2025), with methods like SpaceR (Ouyang et al., 2025) and Spatial-MLLM (Wu et al., 2025) improving via RL or progressive training. These works motivate evaluating articulated human bodies, whose joint configuration and body-centered reference frames introduce additional requirements.

Human-Centric Understanding and Benchmarks. Classic benchmarks such as Human3.6M (Ionescu et al., 2014), MPII (Andriluka et al., 2014), and MuPoTS-3D (Mehta et al., 2018) evaluate keypoint localization; activity datasets including Kinetics (Kay et al., 2017), AVA (Gu et al., 2018), and Something-Something (Goyal et al., 2017) target action classification; Ego4D (Grauman et al., 2022) and Ego-Exo4D (Grauman et al., 2024) broaden the video scope. Recent benchmarks such as Human-MME (Liu et al., 2026), HumanPCR (Li et al., 2026), and JRDB-Reasoning (Jahangard et al., 2025) extend evaluation to granular human features, spatial relationships, and multi-person reasoning. Table 1 summarizes their complementary coverage. SPARKBench focuses on questions that connect body structure and reference frames to social judgments, physical constraints, and specified scene changes.

Structured visual reasoning. Scene graphs (Johnson et al., 2015), CLEVR (Johnson et al., 2017), and neuro-symbolic models (Yi et al., 2018; Mao et al., 2019) connect explicit scene representations to compositional reasoning. CoT (Wei et al., 2022), PAL (Gao et al., 2023), and visual programming (Gupta & Kembhavi, 2023) offer different ways to organize or execute intermediate reasoning. Our benchmark comparisons examine how these representation and computation choices affect human-centric spatio-physics reasoning.

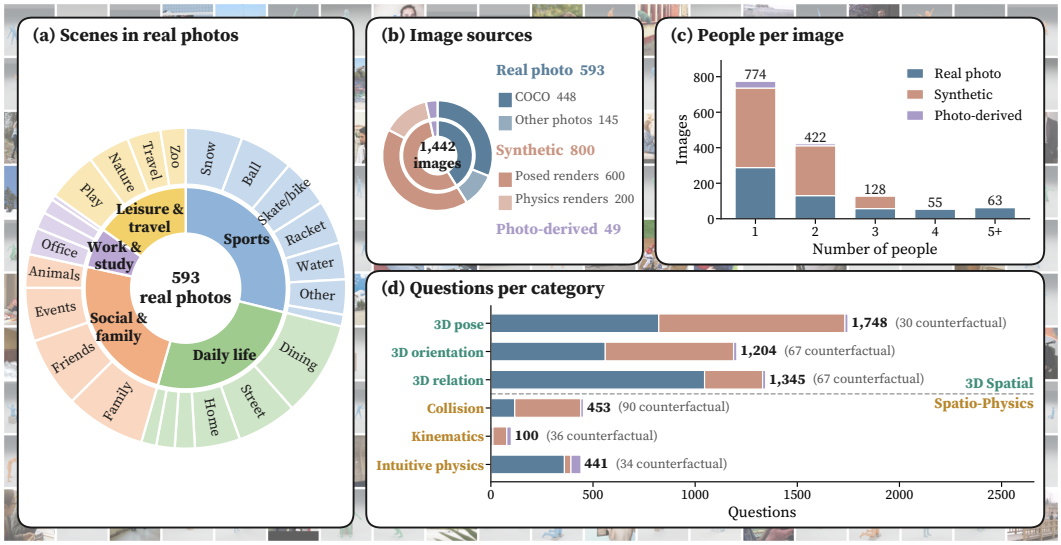


Figure 2: **SPARKBench statistics.** (a) Scene categories of the 593 real images. (b) Image sources. (c) People per image. (d) Questions per category, split by image source.

3 THE SPARKBENCH DATASET

Human-Centric Spatio-Physics Reasoning (SPARKBench) evaluates how vision-language models understand the spatial configuration of human bodies and the physical conditions governing their movement and interaction. It connects two questions: where are the people and their body parts, and what configurations or movements are possible under the stated constraints? Answering requires relating an articulated body to its own reference frame, other people, and the surrounding objects. The benchmark contains **5,291 questions over 1,442 images**, drawn from real images, procedural 3D scenes, and photo-derived scenes. Real images provide varied interactions and visual ambiguity; synthetic scenes provide known geometry for controlled evaluation. Figure 1 illustrates the coverage, and Figure 2 summarizes the data distribution.

3.1 BENCHMARK TAXONOMY

SPARKBench organizes human-centric spatio-physics reasoning at two levels of granularity. At the **macro-level**, the benchmark contains **3D Spatial** and **Spatio-Physics**. At the **micro-level**, 3D Spatial comprises **3D Pose**, **3D Orientation**, and **3D Relation**; Spatio-Physics comprises **Collision**, **Articulation/Kinematics**, and **Intuitive Physics**. Each question is assigned to one micro-level category and contributes to the score of its corresponding macro-level category.

3D Spatial. This macro-level category concerns the body's configuration, orientation, and placement in the scene. Its three micro-level categories specify the spatial property being judged. **3D Pose** covers the arrangement and articulation of body parts, such as a raised arm, knee bending, or relative wrist height. **3D Orientation** covers the facing direction of the head or torso relative to the camera, to the torso, or to another person. **3D Relation** covers spatial relationships among people, body parts, and external objects, including relative position, depth, and distance, and positions stated in the person's own left–right frame. Image-plane relations are identified explicitly when the question concerns projected positions rather than metric 3D quantities.

Spatio-Physics. This macro-level category concerns physical interaction and the constraints on a human body. Its three micro-level categories distinguish contact and overlap, admissible articulation and movement, and physical support. **Collision** covers contact and interpenetration between bodies or between a body and an object. **Articulation/Kinematics** covers joint limits and reach under specified motion constraints. **Intuitive Physics** covers support, balance, and the effects of gravity or motion under the assumptions in the question.

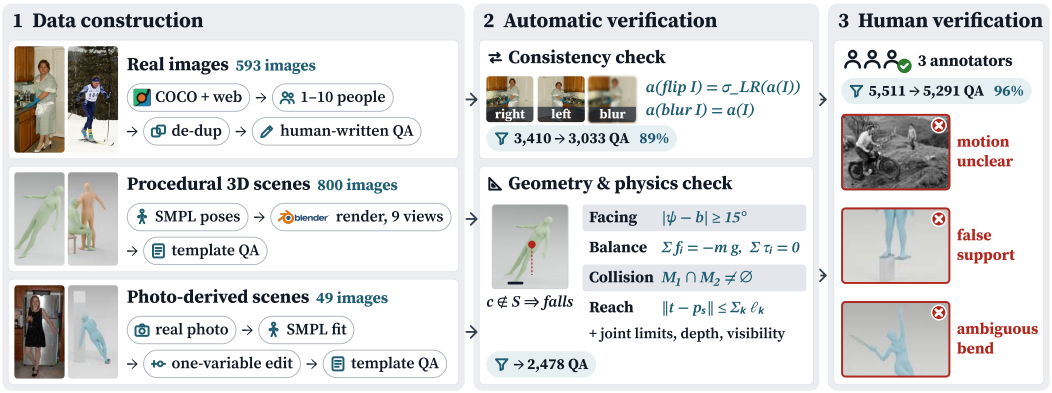


Figure 3: **Data construction and verification.** (1) Data construction: real images receive human-written questions; procedural and photo-derived scenes receive template questions. (2) Automatic verification: a consistency check for real images, and a geometry & physics check with required margins for procedural and photo-derived scenes. (3) Human verification: three annotators review the 5,511 questions that pass, keeping 5,291.

Across these categories, questions take two forms: descriptive and counterfactual. **Descriptive** questions ask about the depicted state, including spatial and physical judgments. **Counterfactual** questions ask about the outcome of a specified change, with other stated conditions held fixed.

3.2 Data Construction

Figure 3 shows the construction pipeline in three stages: data construction, automatic verification suited to each source, and human verification.

Real images. We collected 593 real images showing one to ten people, 448 from COCO Lin et al. (2014) and 145 from the web, across sports, social gatherings, public spaces, homes, workplaces, and service contexts, and removed duplicates. Annotators wrote binary or multiple-choice questions about the depicted bodies and their spatial and physical relations. Each question names its target person or body part and the reference frame it uses: the image, the camera, or the person's own body. A question is kept only if its target is visible and identifiable in the image.

Procedural 3D scenes. The synthetic collection contains 800 rendered images with known pose, orientation, camera parameters, and 3D geometry: 600 posed renders and 200 physics renders. Posed scenes place SMPL bodies Loper et al. (2015) in six pose strata (standing, seated, kneeling, squat or lunge, torso bend, and multi-person layout) with one to three people and nine camera azimuths, varying body articulation, head and torso orientation, and the arrangement of people. A posed proposal is rejected if any body penetrates another body or the floor by more than 2 mm, if a supported foot leaves the floor, or if its pose nearly duplicates an accepted scene; of 4,552 proposals, 600 were retained. Physics renders place bodies in contact with each other, with objects, or with supports, and pose reach targets, for questions about interpenetration, joint limits, reach, and support. Questions are instantiated from templates and answered by annotators from the rendered image; the stored scene state is kept to check these answers (§3.3). Appendix A lists the generation parameters and templates.

Photo-derived scenes. The remaining 49 images start from real images. We fit SMPL to the person in the source image, re-render the body, and change one variable (a support, a contact, a joint angle, or a distance) while holding the pose and camera fixed; for example, the bed beneath a reclining person is removed. Questions are instantiated from templates, as for procedural scenes. The fitted body is a proxy for that person, not measured ground truth for the source image.

3.3 Verification

Each question passes automatic verification suited to its source, followed by human verification.

Table 2: **Direct visual evaluation.** Accuracy (%) on all 5,291 questions. No image: the same prompt without the image. †Human accuracy on 180 stratified questions (30 per category, two annotators per question); Overall and macro columns are reweighted to the full-set category mix. Bold: best per column among models given the image.

		3D Spatial				Spatio-Physics				Question type	
Model	**Overall**	All	Pose	Orient.	Relation	All	Collision	Kinem.	Intuitive	Descr.	Counterf.
Reference											
Chance	39.2	38.8	43.0	31.6	39.8	41.2	43.9	45.8	37.3	38.9	44.9
No image (Qwen3-VL-8B)	51.1	51.0	46.6	45.7	61.5	51.5	43.9	48.0	60.1	50.8	55.6
Human†	91.4	92.8	100.0	86.7	91.7	90.0	88.3	91.7	90.0	92.0	87.5
Proprietary models											
GPT-5	**82.8**	**82.0**	78.9	**68.9**	**97.7**	86.3	**85.4**	63.0	92.5	**82.8**	83.0
Claude Opus 4.8	82.7	81.9	**81.6**	65.9	96.5	**86.5**	85.2	62.0	**93.4**	82.7	**83.6**
Open models											
Qwen3-VL-4B	70.7	69.8	63.9	53.7	91.9	74.6	67.3	50.0	87.8	71.5	59.6
Qwen3-VL-8B	71.4	69.7	62.0	55.1	92.8	78.6	73.1	60.0	88.4	71.6	67.9
Qwen3-VL-30B-A3B	72.3	70.7	63.0	56.2	93.6	79.2	74.2	59.0	88.9	72.9	63.0
Qwen3-VL-8B-Thinking	72.2	72.0	66.0	56.6	93.6	73.1	64.9	55.0	85.7	72.6	67.0
InternVL3.5-8B	66.4	66.7	62.1	53.4	84.5	65.4	61.6	54.0	71.9	67.4	51.2
Gemma3-27B	67.8	65.8	62.5	49.0	85.1	76.3	72.0	63.0	83.7	68.0	64.5
HumanVLM 8B	62.3	60.6	52.2	50.8	80.4	69.2	57.4	58.0	83.9	62.9	52.2
SpatialRGPT 8B (RGB)	66.8	64.6	59.3	48.3	86.2	76.1	67.5	**64.0**	87.5	67.2	60.2
SpatialReasoner	53.0	52.9	47.6	45.0	66.8	53.7	43.5	48.0	65.5	53.2	50.3
Qwen3.5-4B	69.2	68.0	64.6	50.4	88.2	74.5	73.5	44.0	82.5	70.5	49.7
# Questions	5,291	4,297	1,748	1,204	1,345	994	453	100	441	4,967	324

Consistency check. For real images, whose answers cannot be computed, we test whether each answer follows from its stated reference frame. Each question is posed on a horizontally flipped copy, with left and right exchanged in the text, and on a blurred copy. Its answer must swap as the flip predicts, $a(\mathrm{flip}\, I) = \sigma_{\mathrm{LR}}(a(I))$, and stay unchanged under blur, $a(\mathrm{blur}\, I) = a(I)$. Questions that fail are rewritten or removed, leaving 3,033 of 3,410 (89%).

Geometry & physics check. For procedural and photo-derived scenes, each answer must meet three requirements. It must *agree* with the value computed from the stored meshes, joints, contacts, and camera. That value must lie a *margin* away from the decision boundary, so that a small rendering or estimation error cannot flip the answer: 15° for facing sectors and joint bends, 0.2 m for depth order, 20 mm between the centre of mass and the support boundary, 15 mm of interpenetration depth for a collision, and 20° beyond published joint ranges for a joint-limit violation; a target is unreachable only if its distance from the fixed joint exceeds the summed segment lengths of the limb. Finally, the evidence must be *visible*: a camera ray has to hit the queried body part. Joint angles follow ISB conventions Wu et al. (2005) and population ranges of motion Soucie et al. (2011); balance uses a static-equilibrium test Bretl & Lall (2008) with a volume-based centre of mass Tripathi et al. (2023); contact and interpenetration follow Hassan et al. (2019); Yin et al. (2023). In total, 2,478 questions pass these checks; Appendix B gives every rule.

Human verification. Three annotators checked each of the 5,511 questions that passed the automatic checks against its image and removed those they judged ambiguous or incorrect, leaving 5,291 (96%). Typical removals are a real image whose motion cannot be judged from a still frame, a synthetic scene in which a plate looks like a support, and a photo-derived render in which the direction of a bend is ambiguous.

4 Benchmarking Human-Centric Spatio-Physics

We evaluate how well current vision-language models reason about human bodies in space and under physical constraints. The direct visual comparison establishes their performance across the two-level SPARKBench taxonomy and identifies gaps for the subsequent analysis of representation and computation.

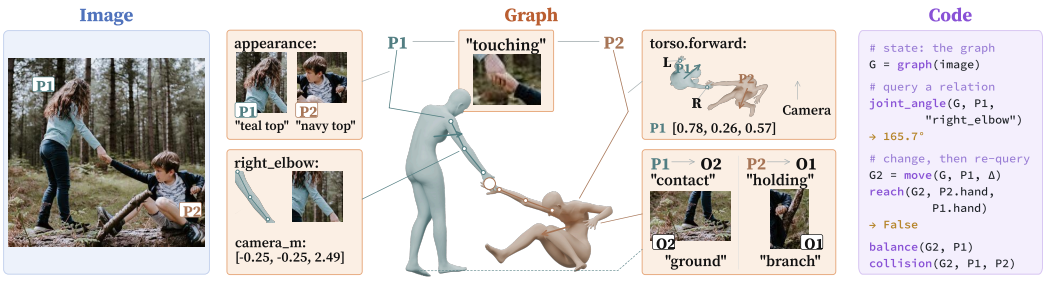


Figure 4: **Image, Graph, and Code.** One scene in three *representations*. *Graph* is a scene graph of people and objects with visual attributes, 3D body geometry, and relations such as "touching"; its values, the *state*, are either estimated from the image or read from the rendered scene (oracle). *Code* adds *computation*: the model calls checks such as joint angle, reach, balance, or collision, which are executed on the graph and return a result. Appendix D gives complete examples.

4.1 Evaluation Protocol

Models and inputs. We evaluate 12 models: two proprietary models, GPT-5 and Claude Opus 4.8, and ten open models spanning Qwen3-VL, Qwen3.5, InternVL3.5, Gemma3, HumanVLM, SpatialRGPT, and SpatialReasoner, with model sizes and variants listed in Table 2. Each model receives an image, a question, and answer choices, and returns a single choice. Questions sharing a scene are answered independently, without access to other answers from that scene. The direct visual comparison covers all 5,291 questions. Evaluation with additional body and scene representations is described in Section 5.

Metrics. We report accuracy overall, by macro- and micro-level category, and separately for descriptive and counterfactual questions (Table 2). For the matched physical-feasibility probes, we report balanced accuracy, recall, specificity, and both-correct accuracy (Appendix Table 6, top; Appendix G).

4.2 Performance across the Benchmark

Overall accuracy conceals gaps within each macro-level category. The ten open models achieve overall accuracies from 53.0% to 72.3% (Table 2). All ten models score lower on orientation than on relations, and lower on articulation/kinematics than on intuitive physics. The micro-level breakdown thus reveals differences that the aggregate scores obscure. These comparisons reflect the particular questions and answer spaces in each category; controlled comparisons are needed to identify the sources of the gaps.

Most models score lower on counterfactual questions. All ten open models score lower on counterfactual than on descriptive questions, with gaps ranging from 2.9% to 20.8%. The benchmark contains 4,967 descriptive and 324 counterfactual questions, so the overall score is dominated by the descriptive subset. Reporting the two separately exposes performance on specified scene changes. Since these subsets contain different questions, the gap does not isolate the effect of changing the scene.

The benchmark highlights uneven performance across human-centric judgments, particularly for orientation, articulation/kinematics, and specified scene changes. The following comparisons examine how performance changes with the body and scene evidence supplied to a model and the computations applied to that evidence (Section 5).

5 Best Lessons for Model Design

We use SPARKBench to examine two decisions in method design: what information to represent and how to compute with it. All comparisons use Qwen3-VL-8B and vary three factors (Figure 4). The *representation* is the form of the input. **Image** supplies the image; **Graph** supplies a scene

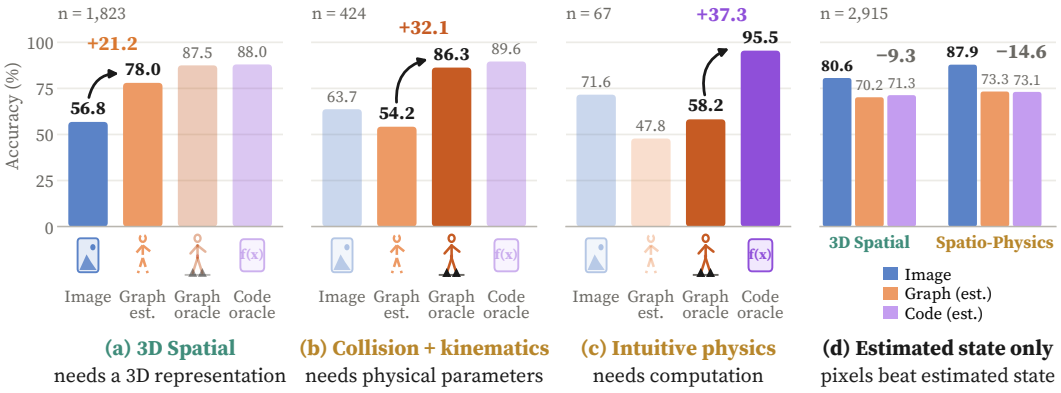


Figure 5: **What each question type needs** (Qwen3-VL-8B, accuracy %). (a)–(c) Synthetic questions, for which the oracle state is available: 3D Spatial questions gain most from the estimated graph, collision and kinematics questions from the oracle graph, and intuitive physics questions from computation on the oracle state. (d) The 2,915 questions on real images, for which only an estimated state is available: the image alone outperforms estimated Graph and Code.

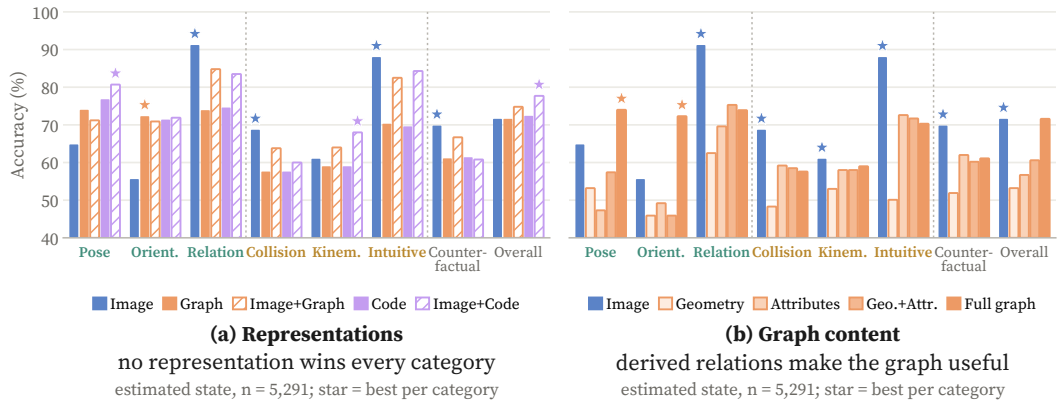
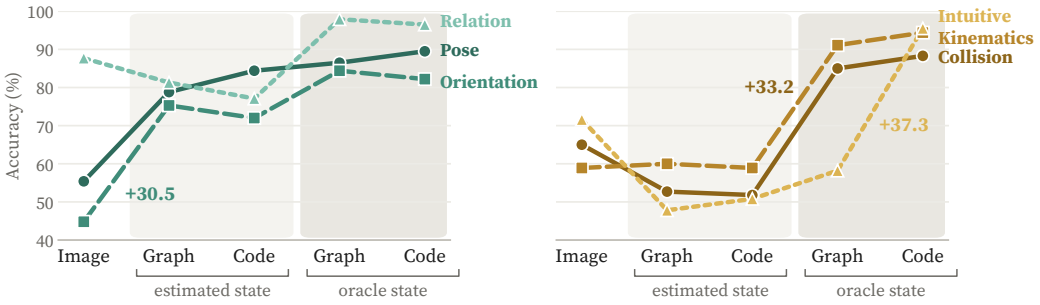


Figure 6: **Representations with estimated state** (Qwen3-VL-8B, accuracy %, 5,291 questions). (a) Image, Graph, and Code; hatched bars add the image to Graph or Code. (b) The graph restricted to body geometry, visual attributes, or both, without derived relations, compared with the full graph. Stars mark the best bar per category. Numbers are given in Table 5 (Appendix E).

graph of people and objects with body geometry, visual attributes, and derived relations such as joint angles and facing directions; **Code** supplies the same graph with executable checks. **Image+Graph** and **Image+Code** add the image. The *state* is the content of the graph, either estimated from the image or, for synthetic questions, taken from the rendered 3D scene (oracle). *Computation* is what distinguishes Code from Graph: the model selects a check (static equilibrium, collision, joint limits, reach, or a geometric query such as facing direction), the check is executed on the graph, and the model answers from its result instead of deriving it in text. We first compare representations and state sources, then measure the benefit of computation for each source. Appendix D gives complete examples, and Figure 5 summarizes the comparisons.

5.1 From Body Structure to Physical State

On all 5,291 questions, estimated Graph improves 3D Pose accuracy from 64.8% to 74.0% and 3D Orientation from 55.6% to 72.3% relative to Image. Spatio-Physics accuracy, however, falls from 76.5% to 63.4% (Figure 6a; Appendix Table 5). Making body structure explicit helps these spatial judgments, but the estimated graph does not provide the same benefit for physical judgments. On the 2,915 questions on real images, Image reaches 80.6% on 3D Spatial and 87.9% on Spatio-Physics, compared with at most 71.3% and 73.3% for estimated Graph or Code (Figure 5d).



(a) 3D Spatial
a 3D state gives most of the gain
n = Pose 911, Orientation 628, Relation 284

(b) Spatio-Physics
needs accurate state, then computation
n = Collision 334, Kinematics 90, Intuitive 67

Figure 7: **State source** (Qwen3-VL-8B, accuracy %, 2,314 synthetic questions). Accuracy per category for Image and for Graph and Code with estimated or oracle state; the oracle graph is taken from the rendered 3D scene. Numbers, including Image+Graph and Image+Code, are given in Table 6 (Appendix E).

To examine the role of physical state, we compare estimated and renderer-provided (oracle) graphs on the 2,314 synthetic questions (Figure 7; Appendix Table 6, bottom). Oracle state provides known joints, meshes, contacts, support surfaces, and mass distribution. On the 491 Spatio-Physics questions, Graph accuracy rises from 53.4% with estimated state to 82.5% with oracle state. This gain motivates preserving the physical variables needed for the judgment. Because the source comparison changes both reliability and variable coverage, it does not isolate the effect of more accurate body geometry.

5.2 From Physical State to Computation

We next keep each state source fixed and compare Graph with Code on the same synthetic questions (Figure 7b). On Spatio-Physics, computation changes accuracy from 53.4% to 53.0% with estimated state, because the checks require contacts and mass that the estimated graph lacks, but from 82.5% to 90.4% with oracle state. The largest oracle gain occurs on the 67 Intuitive Physics questions: 58.2% with Graph versus 95.5% with Code (Figure 5c). Thus, physical information can remain difficult for the model to use even when supplied explicitly, and computing the judgment helps when the available state supports them.

Core lesson for method design. Preserve the physical information required by the question, and compute the judgment from that information. For a static-support question, this means representing contacts and mass distribution and applying an equilibrium check. The two comparisons motivate improving the physical state and its use together. The oracle gains indicate potential under known synthetic state; achieving them from images still requires recovering the necessary information. Failure cases for each representation are shown in Appendix F (Figure 16).

6 Conclusion

We introduced SPARKBench, a benchmark that evaluates 3D spatial and spatio-physics reasoning about human bodies, with answers checked against stored geometry and by human annotators. Across 12 vision-language models, accuracy on relations approaches that of human annotators, but even leading proprietary models remain far below humans on orientation and kinematics. Varying the representation, its state, and computation yields two lessons. First, the representation must carry the physical state a question needs: an estimated scene graph helps pose and orientation but not physical judgments, whereas the oracle state substantially improves them. Second, the judgment should be computed from that state: executing a physical check such as static equilibrium answers questions that the model fails to reason through in text. These gains are measured on synthetic scenes; realizing them on real images requires recovering contacts, supports, and mass from pixels, which current estimators do not provide. SPARKBench offers a testbed for this goal.

AI USE STATEMENT

We used generative AI tools to polish and draft parts of the text, to help implement the evaluation and analysis code, to support data analysis, and to find and format references. They were not used to write benchmark questions or answers or to generate benchmark images; developing theoretical models, mathematical claims, and proofs is not applicable to this work. The authors reviewed all AI-assisted work: code outputs were checked against the underlying data, reported numbers were recomputed from saved predictions, and every reference was checked against its publisher record. We take responsibility for the final content of this work.

ETHICS STATEMENT

All authors have read and adhere to the ICLR Code of Ethics. *Human participants.* Annotators wrote or verified the benchmark questions, and four annotators answered a 180-question subset for the human baseline. Annotators were unpaid volunteers; we recorded only their answers and response times under pseudonymous IDs. *Images and privacy.* SPARKBench contains 593 real images, 448 from COCO Lin et al. (2014) and 145 from the web, which show people in public or everyday settings. Questions concern body configuration, orientation, contact, and support; no question asks for identity, and no demographic or sensitive attribute is annotated. COCO images are distributed under their original Flickr licenses. The other 145 images come from public Kaggle datasets released under open licenses, and all images will be distributed under their original license terms. Synthetic scenes use the SMPL body model Loper et al. (2015) under its research license and depict no real person. *Limitations and misuse.* Most real images come from COCO, so the benchmark inherits its limited geographic and demographic coverage. Better 3D understanding of human bodies could also be applied to monitoring people; the benchmark measures reasoning about stated spatial and physical questions and does not evaluate person identification.

REPRODUCIBILITY STATEMENT

Section 3.2 and Appendices A and B describe how images and questions were collected and how answers were verified, including every rule and margin used to check synthetic answers. Section 4.1 specifies the evaluated models, prompts, decoding settings, and the handling of invalid answers; proprietary models are identified by their exact API model names. Appendices C and D specify the scene-graph format, the executable operations, and complete input and output examples for each representation. Code, data, and the evaluation protocol will be released upon acceptance, including the benchmark questions, answer keys, scene graphs, saved model predictions, and scoring scripts.

REFERENCES

Ijaz Akhter and Michael J. Black. Pose-conditioned joint angle limits for 3D human pose reconstruction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1446–1455, 2015.

Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2D human pose estimation: New benchmark and state of the art analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Timothy Bretl and Sanjay Lall. Testing static equilibrium for legged robots. *IEEE Transactions on Robotics*, 24(4):794–807, 2008. doi: 10.1109/TRO.2008.2001360.

Yu-Wei Chao, Yunfan Liu, Xieyang Liu, Huayi Zeng, and Jia Deng. Learning to detect human-object interactions. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 381–389, 2018.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brian Ichter, Danny Driess, Pete Florence, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. SpatialVLM: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Wei Chow, Jiageng Mao, Boyi Li, Daniel Seita, Vitor Guizilini, and Yue Wang. PhysBench: Benchmarking and enhancing vision-language models for physical world understanding. In *International Conference on Learning Representations (ICLR)*, 2025.

Mihai Fieraru, Mihai Zanfir, Elisabeta Oneata, Alin-Ionut Popa, Vlad Olaru, and Cristian Sminchisescu. Three-dimensional reconstruction of human interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7212–7221, 2020.

Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. PAL: Program-aided language models. In *International Conference on Machine Learning (ICML)*, pp. 10764–10799, 2023. URL `https://proceedings.mlr.press/v202/gao23f.html`.

Gemini Team, Google. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzyńska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fründ, Peter Yianilos, Moritz Mueller-Freitag, et al. The "Something Something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4D: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, et al. Ego-Exo4D: Understanding skilled human activity from first- and third-person perspectives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Chunhui Gu, Chen Sun, David A. Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, Cordelia Schmid, and Jitendra Malik. AVA: A video dataset of spatio-temporally localized atomic visual actions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Tanmay Gupta and Aniruddha Kembhavi. Visual programming: Compositional visual reasoning without training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J. Black. Resolving 3D human pose ambiguities with 3D scene constraints. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2282–2292, 2019.

Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, 2014.

Simindokht Jahangard, Mehrzad Mohammadi, Yi Shen, Zhixi Cai, and Hamid Rezatofighi. JRDB-Reasoning: A difficulty-graded benchmark for visual reasoning in robotics. *arXiv preprint arXiv:2508.10287*, 2025.

Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Image retrieval using scene graphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

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. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Will Kay, João Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The Kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. LLaVA-OneVision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

Keliang Li, Hongze Shen, Hao Shi, RuiBing Hou, Hong Chang, Jie Huang, Chenghao Jia, Wen Wang, Yiling Wu, Dongmei Jiang, Shiguang Shan, and Xilin Chen. HumanPCR: Probing MLLM capabilities in diverse human-centric scenes. In *International Conference on Learning Representations (ICLR)*, 2026.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 740–755. Springer, 2014.

Yuansen Liu, Haiming Tang, Jinlong Peng, Jiangning Zhang, Xiaozhong Ji, Qingdong He, Wenbin Wu, Donghao Luo, Zhenye Gan, Junwei Zhu, Yunhang Shen, Chaoyou Fu, Chengjie Wang, Xiaobin Hu, and Shuicheng Yan. Human-MME: A holistic evaluation benchmark for human-centric multimodal large language models. In *International Conference on Learning Representations (ICLR)*, 2026.

Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Transactions on Graphics (SIGGRAPH Asia)*, 34(6): 248:1–248:16, 2015.

Wufei Ma, Haoyu Chen, Guofeng Zhang, Yu-Cheng Chou, Jieneng Chen, Celso M. de Melo, and Alan Yuille. 3DSRBench: A comprehensive 3D spatial reasoning benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

Jiayuan Mao, Chuang Gan, Pushmeet Kohli, Joshua B. Tenenbaum, and Jiajun Wu. The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision. In *International Conference on Learning Representations (ICLR)*, 2019. URL `https://arxiv.org/abs/1904.12584`.

Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Srinath Sridhar, Gerard Pons-Moll, and Christian Theobalt. Single-shot multi-person 3D pose estimation from monocular RGB. In *International Conference on 3D Vision (3DV)*, 2018.

OpenAI. GPT-4o System Card. `https://openai.com/index/gpt-4o-system-card/`, 2024. Accessed: 2026-05-05.

Kun Ouyang, Yuanxin Liu, Haoning Wu, Yi Liu, Hao Zhou, Jie Zhou, Fandong Meng, and Xu Sun. SpaceR: Reinforcing MLLMs in video spatial reasoning. *arXiv preprint arXiv:2504.01805*, 2025.

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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

J. M. Soucie, C. Wang, A. Forsyth, S. Funk, M. Denny, K. E. Roach, and D. Boone. Range of motion measurements: reference values and a database for comparison studies. *Haemophilia*, 17 (3):500–507, 2011. doi: 10.1111/j.1365-2516.2010.02399.x.

Shashank Tripathi, Lea Müller, Chun-Hao P. Huang, Omid Taheri, Michael J. Black, and Dimitrios Tzionas. 3D human pose estimation via intuitive physics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Timo von Marcard, Roberto Henschel, Michael J. Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3D human pose in the wild using IMUs and a moving camera. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 614–631, 2018.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, et al. InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Diankun Wu, Fangfu Liu, Yi-Hsin Hung, and Yueqi Duan. Spatial-MLLM: Boosting MLLM capabilities in visual-based spatial intelligence. *arXiv preprint arXiv:2505.23747*, 2025.

Ge Wu, Frans C. T. van der Helm, H. E. J. Veeger, Mohsen Makhsous, Peter Van Roy, Carolyn Anglin, Jochem Nagels, Andrew R. Karduna, Kevin McQuade, Xuguang Wang, Frederick W. Werner, and Bryan Buchholz. ISB recommendation on definitions of joint coordinate systems of various joints for the reporting of human joint motion—Part II: shoulder, elbow, wrist and hand. *Journal of Biomechanics*, 38(5):981–992, 2005. doi: 10.1016/j.jbiomech.2004.05.042.

Peiran Xu, Sudong Wang, Yao Zhu, Jianing Li, Gege Qi, and Yunjian Zhang. SpatialBench: Benchmarking multimodal large language models for spatial cognition. *arXiv preprint arXiv:2511.21471*, 2025.

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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Kexin Yi, Jiajun Wu, Chuang Gan, Antonio Torralba, Pushmeet Kohli, and Joshua B. Tenenbaum. Neural-symbolic VQA: Disentangling reasoning from vision and language understanding. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018. URL `https://arxiv.org/abs/1810.02338`.

Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B. Tenenbaum. CLEVRER: Collision events for video representation and reasoning. In *International Conference on Learning Representations (ICLR)*, 2020.

Yifei Yin, Chen Guo, Manuel Kaufmann, Juan Jose Zarate, Jie Song, and Otmar Hilliges. Hi4D: 4D instance segmentation of close human interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17016–17027, 2023.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6713–6724, 2019.

Appendices A and B describe how the synthetic scenes are generated and checked. Appendix C specifies the operations behind the Code representation. Appendix D walks through individual questions under each representation, Appendix F collects failure cases, and Appendix G reports the physical feasibility case study. Appendix H discusses extensions.

A Synthetic Scene Generation

The 800 synthetic images comprise 600 posed renders and 200 physics renders. The two parts come from different generation procedures and are described separately.

A.1 Posed renders

Bodies and rendering. Posed scenes use the SMPL neutral body Loper et al. (2015). Seats and knee blocks are visible rectangular supports resting on the floor. Scenes are rendered in Blender (Cycles, 24 samples) at 768×768 with a perspective camera (36 mm sensor) on a neutral background with a floor and soft shadows, and contain no overlays or answer text. The camera azimuth is one of $0°$, $\pm25°$, $\pm65°$, $\pm90°$, $155°$, and $180°$, and every camera parameter is stored per scene.

Composition and deduplication. The six pose strata contain 100 scenes each; 300, 240, and 60 scenes contain one, two, and three people. For deduplication, each body is summarized by its lower-body, spine, and arm states in $10°$ bins, independent of color, camera, head turn, shape, and global yaw. Multi-person arrangements are canonicalized over person permutations with 0.5 m translation and $45°$ relative-yaw bins. A proposal is a near duplicate if some permutation matches an accepted scene with every joint rotation within $20°$. All 75,300 same-size scene pairs were checked, and the 600 retained scenes cover 579 distinct pose-family combinations.

Physical validity. Floor penetration and unintended body–body penetration may not exceed 2 mm, measured by BVH triangle overlap on the final meshes, and intended contacts must close to within 6 mm. Each supported body must admit a static-equilibrium certificate: non-negative vertical forces at horizontal contacts within a 6 mm contact band balance gravity and its moment for a uniform-density volume (4 cm voxels), with the centre of mass at least 20 mm inside the support boundary under eight mass-density perturbations.

Answer margins and visibility. Each answer must clear the margins in Appendix B. Beyond the ray test there, projected bones must span at least 2.5% of the image diagonal, people must lie at least 3.5% of the image size from the border and 1.5% of the image width from each other, and head-orientation questions require a head of at least 18×26 pixels with two of three face rays visible. Of 4,552 proposals, 600 passed these checks. A final inspection of every render removed two scenes as visually ambiguous; each was replaced by a new proposal that passed the same checks.

Question templates. Posed renders use 15 templates: wrist position in the image (305 questions), body orientation (264), knee bend (261), depth between people (198), head orientation relative to the torso (186), raised hands (148), torso inclination (123), person position in the image (86), elbow bend (61), number of people facing the camera (59), head orientation relative to the camera (37), pairwise facing (30), relative orientation (24), position relative to another person's body (11), and both knees (7). Of these 1,800 questions, 1,799 are retained in SPARKBench.

A.2 Physics renders

Each of the 200 physics renders stores a geometric certificate for the quantity its questions depend on. Body–body interpenetration is certified by mesh collision together with a bidirectional point-in-solid depth check, so that surface contact is not mistaken for penetration. Body–object collision uses the stored solid geometry with an analytic witness point. Support and balance are certified by recomputing static equilibrium under the scene's contacts, and reach and joint limits by comparing the stored joint configuration with limb length and published ranges of motion Soucie et al. (2011). Every physics render was also screened visually, so that a certified violation is distinguishable from ordinary occlusion in the image.

B VERIFICATION PROTOCOL FOR SYNTHETIC SCENES

Table 3 lists the rule and margin used to check each annotator answer on procedural and photo-derived scenes. All quantities are measured on the final rendered SMPL surfaces Loper et al. (2015).

Table 3: **Verification rules.** An annotator answer is kept only if it agrees with the computed quantity, the quantity satisfies the rule with the stated margin, and the queried evidence is visible.

Check	Quantity	Rule and margin	Reference
Facing	Head or torso forward direction, camera yaw ψ	Four 90° sectors relative to the camera; $\lvert\psi - b\rvert \geq 15^\circ$ for every sector boundary b	–
Joint state	Internal knee or elbow angle θ	Bent if $\theta < 135^\circ$; $\lvert\theta - 135^\circ\rvert \geq 15^\circ$	Wu et al. (2005)
Joint limits	Signed hinge flexion φ, hyperextension negative	Violation if hyperextension exceeds the largest observed adult value (knee 17–19°, elbow 22°) by at least 20°, or exceeds a limit stated in the question	Soucie et al. (2011); Akhter & Black (2015)
Depth order	Camera depth z of two body parts or people	$\lvert z_1 - z_2\rvert \geq 0.2$ m; between people ≥ 1.2 m with a head-size ratio ≥ 1.18	–
Balance	Centre of mass c of the uniform-density volume; support region S from contacts	Supported if non-negative vertical contact forces cancel gravity and its moment, with $d(c, \partial S) \geq 20$ mm under mass-density perturbations; falls if c stays outside S under $\pm 10\%$ density variation and a 50 mm displacement	Bretl & Lall (2008); Tripathi et al. (2023)
Collision	Depth of one mesh inside another	Interpenetration if the interior depth exceeds 15 mm; contact if the surfaces do not cross and the gap is below 2 mm; separated if the clearance is positive	Hassan et al. (2019); Fieraru et al. (2020); Yin et al. (2023)
Reach	Target t, fixed joint p_s, segment lengths ℓ_k	Unreachable if $\lVert t - p_s\rVert > \sum_k \ell_k$ by a wide margin; reachable only with an explicit non-intersecting final pose	–
Visibility	Camera rays to the queried joints and segments	Three joint rays and six segment samples hit the target body	–

Scope. These rules certify geometric and static conditions for a fixed pose. The equilibrium test assumes rigid bodies, horizontal supports, frictionless vertical contact forces, and uniform density; it does not model muscle strength or dynamics. Joint ranges are reference values for healthy adults rather than universal limits, so a violation is labeled only when it exceeds them by a wide margin.

C EXECUTABLE OPERATIONS

Algorithm 1 gives the procedure behind the Code representation. The model plans a short sequence of operations and fills their arguments; the operations are executed by code on a copy of the scene graph, so that a counterfactual edit never changes the original state. Questions in 3D Spatial allow up to four steps, drawn from query operations and, only when the question states a change, edit operations; after an edit, the derived relations are recomputed before the next query. Questions in Spatio-Physics select a single physical operation. Table 4 lists both sets and their rules. Each operation first checks that the fields it needs are present and returns UNKNOWN otherwise. With the estimated state, the centre of mass, gravity direction, and contacts are unavailable: on the 994 Spatio-Physics questions the model selects no operation for 717, and 272 of the remaining 277 calls return UNKNOWN. On flat, frictionless ground, the balance test reduces to checking whether the centre of mass projects inside the support polygon.

Algorithm 1 Answering with Code

Require: image I (optional), question Q, choices C, scene graph S, operation set $\mathcal{O}$, step budget K (4 for 3D Spatial, 1 for Spatio-Physics)

$(o_1, \theta_1), \ldots, (o_k, \theta_k) \leftarrow \mathrm{VLM}(Q, C, S, I, \mathcal{O}), k \leq K$ ▷ plan operations and arguments
$S' \leftarrow \mathrm{copy}(S)$; $R \leftarrow [\,]$
for $i = 1, \ldots, k$ **do**
  if a field required by o_i is missing from S' **then**
    append UNKNOWN to R
  else if o_i is an edit **then**
    $S' \leftarrow o_i(S', \theta_i)$; recompute derived relations in S' ▷ only for a stated change
  else
    append $o_i(S', \theta_i)$ to R ▷ executed by code, not generated
  end if
end for
return $\mathrm{VLM}(Q, C, S, I, R, S')$

Table 4: **Executable operations.** Inputs are read from the scene graph; the rule decides the returned result. Edit operations are available only when the question states the change.

Operation	Used for	Rule
3D Spatial: queries		
`heading` (person, head | torso)	Orientation	Yaw of the forward direction in the camera frame, $\psi = \mathrm{atan2}(f_x, -f_z)$; toward the camera for $\psi \in [-45°, 45°)$, image-right for $[45°, 135°)$, image-left for $[-135°, -45°)$, away otherwise
`mutual_facing` (a, b, head | torso)	Orientation	$\theta_a = \angle(f_a, o_b - o_a)$ and $\theta_b = \angle(f_b, o_a - o_b)$, tested separately; a person faces the other iff $\theta < 90°$
`locate` (target, reference, frame)	Relation	Offset target − reference in the image, camera, or reference person's body frame (own right, own front)
`hand_heights` (person, frame)	Pose	Each wrist compared with its own-side shoulder; above iff higher in the chosen frame
`check_reach_projection` (person, side, target)	Relation, reach	Unreachable iff the shoulder-to-target distance exceeds arm plus hand length; otherwise not ruled out
3D Spatial: edits		
`rotate_person` (person, whole_body | head, α)	Counterfactual	Rotate about the vertical axis by α, positive toward the person's own left
`swap_joint_locations` (person, a, b)	Counterfactual	Exchange the positions of two hand or foot endpoints
`swap_image_locations` (a, b)	Counterfactual	Exchange the image positions of two entities
`mirror_image` ()	Counterfactual	Reflect all image positions horizontally; anatomical labels are kept and metric geometry is dropped
Spatio-Physics		
`balance_after` (person, edits)	Support, falling	Feasible iff there are forces f_i in the friction cones with $\sum_i f_i + mg = 0$ and $\sum_i (p_i - c) \times f_i = 0$ Bretl & Lall (2008); c from the mesh volume Tripathi et al. (2023); supports may be removed or friction changed first
`clearance_after` (person, edits)	Interpenetration	Penetrating iff the mesh intersection volume exceeds the tolerance, after removing the named entities
`hinges_after` (person, joints, range)	Joint limits	Violation iff some signed flexion $\theta_j \notin [\theta_{\min}, \theta_{\max}]$; a straightened joint has $\theta_j = 0$
`reach_bound` (person, side, target)	Reach	Unreachable iff $\|t - p_s\| > \sum_k \ell_k$; otherwise not ruled out
`rotation_reach` (person, side, pivot, target)	Reach	Reachable iff the hand radius range about the wrist or shoulder overlaps the target's distance range

D Representation Case Studies

Figures 9–15 give seven questions in each representation: the image, the scene graph, and the program with its returned values, followed by the derivation Qwen3-VL-8B wrote under each representation. Cases 1–5 use the estimated state of real images, and Cases 6 and 7 use the oracle state of synthetic scenes. In Cases 1–3 the image answer is wrong and the explicit state or its execution corrects it. In Cases 4 and 5 the image answer is right, and an estimated state that misses the relevant object leads Graph and Code astray. Cases 6 and 7 require an edit or a physical computation on an exact state. Listings are abridged: joint lists and repeated contact points are elided and numbers are rounded. Derivations are quoted verbatim, with omissions marked [. . .].

Figure 8 first illustrates one question answered under each representation.

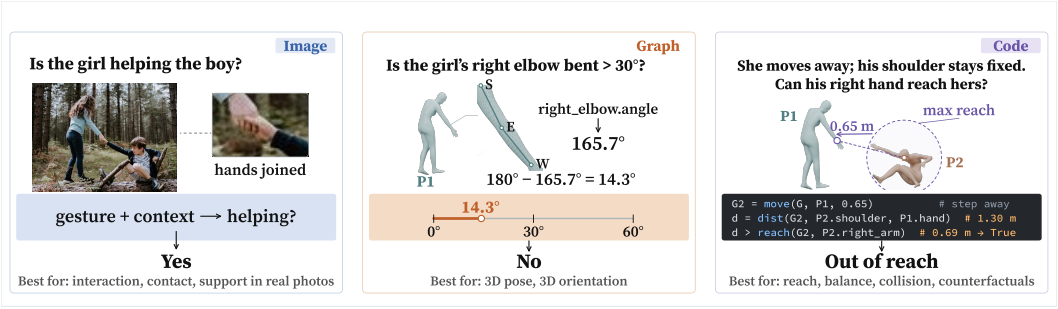
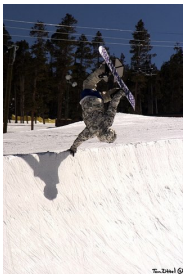


Figure 8: **Reasoning with each representation.** Three questions about one real image: an interaction question answered from the image, an elbow-angle question read from the graph, and a reach question answered by executing a hypothetical step. Operation names are simplified.

Question. In which direction is the snowboarder's body facing? (A) toward the camera (B) away from the camera (C) toward the left of the image (D) toward the right of the image

Answer. B **Category.** 3D orientation **State.** estimated



Graph

```
frames:  camera x right, y down, z away (m, estimated); gravity unmeasured
P1  person  box [365, 219, 683, 565]
    observations: posture airborne, activity "performing a snowboarding trick mid-air",
                  visible_face front, left_hand / right_hand "contacting snowboard"
    geometry (estimated): torso.forward [0.070, 0.192, 0.979]
                          head.forward  [0.294, 0.296, 0.909]
O1  object  snowboard  box [528, 155, 732, 390]
edges:    P1 holding O1  (predicted candidate)
derived:  P1.torso_facing   yaw 175.9 deg, away_from_camera
          P1.head_facing    yaw 162.1 deg, away_from_camera
          P1.joint_angles   knee L 50.2 / R 83.3 deg (internal)
```

Code

```
heading(person=P1, component=torso)
  -> yaw 175.9 deg, away_from_camera, pitch -11.1 deg
```

Rep.	Answer	Derivation
Image	A ✗	"The snowboarder is performing a trick mid-air, inverted, with their head pointing downward and their feet toward the camera. Their body is oriented such that their face (and thus their gaze) is directed toward the camera. [. . .]"
Graph	B ✓	"The snowboarder's torso is facing away from the camera, as indicated by the derived data: 'torso_facing': {'yaw_deg': 175.92, 'cardinal': 'away_from_camera', [. . .]}. This is a direct measurement from the fitted geometry, not an inference."
Code	B ✓	"The derived data for P1 (the snowboarder) explicitly states: 'torso_facing': {'yaw_deg': 175.92, 'cardinal': 'away_from_camera', [. . .]}. [. . .] The head is also facing away (yaw_deg: 162.07 [. . .]), confirming consistency."

Figure 9: **Case 1: facing direction of an inverted body (real image, estimated state).** The snowboarder is upside down with his back to the camera. The Image derivation takes the feet pointing at the camera as the body facing the camera. The fitted body gives a torso yaw of 176°, and Graph and Code read it directly. The attribute `visible_face front` is wrong here as well, but neither derivation uses it.

Question. Which hand is higher? (A) person's left hand (B) person's right hand
Answer. B **Category.** 3D pose **State.** estimated



Graph

```
frames:  camera x right, y down, z away (m, estimated); image xy in 0..1000
P1  person  box [423, 93, 625, 660]
    observations: appearance "shirtless, wearing dark shorts", posture standing,
                  activity "paddling on a surfboard", visible_face front,
                  left_hand "holding a paddle", right_hand "holding a paddle"
    geometry (estimated; 62 joints as name, camera xyz, image xy):
      left_shoulder    [ 0.415, -0.861, 5.034]   [603.0, 215.1]
      right_shoulder   [ 0.300, -1.010, 5.318]   [570.5, 183.4]
      left_wrist       [ 0.271, -0.327, 5.054]   [567.0, 392.3]
      right_wrist      [-0.204, -0.909, 5.366]   [452.5, 217.8]
      ...
      torso.forward [-0.933, 0.042, -0.356]    head.forward [-0.100, 0.472, -0.876]
O1  object  surfboard  box [100, 540, 840, 700]  state "floating on water"
edges:    P1 holding O1  (predicted candidate)
derived:  P1.torso_facing    yaw -69.1 deg, image_left
          P1.head_facing     yaw  -6.5 deg, toward_camera
          P1.hands_order     higher_in_image right, higher_in_camera_3d right,
                             closer_to_camera left
          P1.joint_angles    elbow L 161.7 / R 137.5 deg, knee L 159.9 / R 161.2 deg
```

Code

```
hand_heights(person=P1, frame=image)
  -> left:   wrist_up_from_own_shoulder = -177.2   (below)
     right:  wrist_up_from_own_shoulder =  -34.3   (below)
```

Rep.	Answer	Derivation
Image	A ✗	"The left hand is gripping the paddle near the top, while the right hand is gripping the paddle lower down [. . .]. Therefore, the person's left hand is higher."
Graph	B ✓	"The derived state for P1 shows: 'hands_order': {'higher_in_image': 'right', 'higher_in_camera_3d': 'right', [. . .]}. This directly answers the question [. . .]"
Code	B ✓	"The left wrist is at image coordinate y = 392.33, and the left shoulder is at y = 215.13, giving a difference of -177.2 (wrist below shoulder). The right wrist is at y = 217.79, and the right shoulder is at y = 183.45, giving a difference of -34.34 [. . .]. Since -34.34 is less negative than -177.2, the right wrist is higher in the image frame."

Figure 10: **Case 2: which hand is higher (real image, estimated state).** The paddler faces image left, so his right hand appears on the left of the image. The Image derivation assigns the upper hand to the wrong side. Graph reads the precomputed hand order, and Code computes each wrist's height above its own shoulder.

Question. He exchanges his hands between their two box grips without moving the box. Which hand then occupies the image-left grip? (A) His left hand (B) His right hand
Answer. A **Category.** 3D pose, counterfactual **State.** estimated



Graph

```
P1  person  box [221, 58, 917, 994]
    observations: visible_face front, activity "holding open pizza box",
                  left_hand "pizza box lid", right_hand "pizza box edge"
    geometry (estimated; name, camera xyz, image xy):
      right_wrist  [-0.205, 0.013, 2.063]   [334.1, 508.2]
      left_wrist   [ 0.173, 0.088, 1.722]   [667.6, 564.1]
      ...
derived:  P1.head_facing     yaw -15.3 deg, toward_camera
          P1.own_right_axis  in image: image_left
```

Code

```
swap_joint_locations(person=P1, a=left_wrist, b=right_wrist)
  -> state S' (changed: P1)
     S'.left_wrist   image [334.1, 508.2]   (image left)
     S'.right_wrist  image [667.6, 564.1]   (image right)
```

Rep.	Answer	Derivation
Image	B ✗	"His right hand is on the right side of the box (image-right grip), and his left hand is on the left side of the box (image-left grip). [. . .] after the exchange, his right hand occupies the image-left grip."
Graph	B ✗	"In the original state, the observations show: left hand gripping the pizza box lid (image-left grip), and right hand gripping the pizza box edge (image-right grip). [. . .] the hand occupying the image-left grip after the swap is [. . .] the right hand."
Code	A ✓	"The original state shows P1's right hand is at image-left (right_wrist at [334.14,508.17]), and left hand at image-right (left_wrist at [667.6,564.06]). After swap_joint_locations(P1, left_wrist, right_wrist), the left_wrist is now at [334.14,508.17] (image-left) [. . .]. The image-left grip is now occupied by P1's left hand."

Figure 11: **Case 3: exchanging hands (real image, estimated state, counterfactual).** The boy faces the camera, so his right hand holds the image-left side of the box. Image and Graph both place his left hand on the image left before the exchange and answer "right", although the graph lists the right wrist at $x = 334$. Code swaps the two wrist positions and reads the image-left grip from the result.

Question. Where is the barn relative to the person? (A) in front of the person (B) behind the person (C) to the person's left (D) to the person's right

Answer. A **Category.** 3D relation **State.** estimated



Graph

```
P1  person  box [454, 153, 998, 989]
    observations: appearance "back view, short dark hair, ...", visible_face occluded,
                  right_hand "phone", head_observation "head turned slightly to the left,
                  facing the barn"
O1  object  cell phone  box [488, 475, 547, 596]
edges:    P1 holding O1  (predicted candidate)
scene:    "outdoors, grassy field with a large metal barn in background"
binding:  "the barn" -> no entity
derived:  P1.torso_facing   yaw 176.4 deg, away_from_camera
          P1.head_facing    yaw -150.2 deg, away_from_camera
          P1.head_turn      33.2 deg toward own left
```

Code

```
locate(target=O1.center, reference=P1.center, frame=image)
  -> delta [-208, -36], image_left, above        (O1 is the phone)
```

Rep.	Answer	Derivation
Image	A ✓	"The person is facing the barn, as they are holding a phone up to take a photo of it. [...] The person's back is to the camera, and the barn is centered in front of them in the image. This means the barn is in front of the person."
Graph	C ✗	"The person's head is turned slightly to the left, facing the barn (as stated in the observations: 'head turned slightly to the left, facing the barn'). This means the barn is to the person's left. [...]"
Code	B ✗	"[...] The person's torso is facing away from the camera, and the barn is located behind the person relative to the camera's perspective. [...] The barn is behind the person."

Figure 12: **Case 4: position of an undetected object (real image, estimated state).** The detector finds the phone but not the barn, so the barn appears in the graph only as scene text and a head attribute. Graph turns "head turned slightly to the left" into "the barn is on the left". The planned operation locates the phone instead of the barn, and Code then reads "behind" from the camera's viewpoint. The Image derivation uses the person's pose and the photo being taken.

Question. Where is the person seated? (A) on the lower bunk of a bunk bed (B) on a wooden chair (C) on the bare floor (D) on a sofa

Answer. A **Category.** intuitive physics **State.** estimated



Graph

```
frames:  camera x right, y down, z away (m, estimated); gravity unmeasured
P1  person  box [559, 319, 962, 989]
    observations: posture sitting, activity "holding a shoe",
                  left_hand "shoe", right_hand "shoe"
    geometry (estimated; name, camera xyz):
      left_heel   [0.549, 0.858, 2.270]
      right_heel  [0.120, 0.866, 2.230]
      ...
O1  object  handbag  box [216, 376, 291, 459]
edges:    P1 holding O1  (predicted candidate)
scene:    observation "bunk bed room"
physics:  coordinate_frame camera_m, surfaces [] (none detected), contacts unavailable
```

Code

```
none selected -> unknown ("No applicable supported calculation requested")
```

Rep.	Answer	Derivation
Image	A ✓	"The person is sitting on a bed with blue and white bedding, which is part of a bunk bed structure. The bed is positioned below a second bed, confirming it is the lower bunk."
Graph	C ✗	"The person is sitting with both feet on the floor, as indicated by the foot positions (left_heel and right_heel) being at z=2.27 and z=2.23 respectively, which are consistent with floor-level contact. [...] The scene is a 'bunk bed room', but no bunk is detected as supporting the person. [...] Therefore, the person is seated on the bare floor."
Code	C ✗	"[...] no contact with a chair, sofa, or bunk bed is indicated. [...] Therefore, the only plausible option is 'on the bare floor'."

Figure 13: **Case 5: support surface (real image, estimated state).** The bed is not detected, so the graph has no support surface and no contacts. The model reads the camera depth z of the heels as height above the floor and concludes that the person sits on the floor. No operation applies, so Code adds nothing. The image shows the lower bunk directly.

Question. Only the orange person turns 180° in place. Is the green person now on their left or right? (A) on their anatomical right (B) on their anatomical left

Answer. B **Category.** 3D orientation, counterfactual **State.** oracle



Graph

```
P1  person  observations: appearance "solid orange body", posture kneeling
    geometry (oracle): torso.forward [-0.997, -0.075, -0.006]
                       head.forward  [-0.762,  0.237, -0.603]
                       joints: nose [-1.413, 0.106, 7.329] [66.7, 532.6], ...
P2  person  observations: appearance "solid green body", posture standing
O1, O2  object  knee_block
edges:    P1 supported_by O1;  P1 supported_by O2
derived:  P1.torso_facing      yaw -89.6 deg, image_left
          P1.joint_angles      knee L 84.3 / R 85.1 deg (internal)
          P2.torso_facing      yaw -89.6 deg, image_left
          P2_relative_to_P1    own_side right, front_back behind,
                               body-frame offset: right 2.83 m, up 0.58 m, front -2.53 m
          depth_order          P1, P2 (nearest first)
```

Code

```
rotate_person(person=P1, component=whole_body, own_left_deg=180)
  -> state S' (changed: P1)
     S'.P1.torso_facing      yaw +90.4 deg, image_right
     S'.P2_relative_to_P1    own_side left, front_back front,
                             body-frame offset: right -2.24 m, up 1.57 m, front 2.69 m
```

Rep.	Answer	Derivation
Image	B ✓	"Before the turn, the green person was on the orange person's right (image-right). After a 180° turn, the orange person now faces the opposite direction, so the green person is now on their anatomical left."
Graph	A ✗	"The original state shows the green person (P2) is on the orange person's (P1) anatomical right, as per the derived 'P1_relative_to_P2' section [...]. After P1 turns 180° in place, their anatomical left becomes their anatomical right [...]. Therefore, the green person is now on the orange person's anatomical right."
Code	B ✓	"In the original state, P2 was on P1's anatomical right [...]. After P1 rotates 180°, the relative orientation flips [...]. This is confirmed by the post-edit pair geometry: P2_relative_to_P1: own_side: 'left'."

Figure 14: **Case 6: relative side after a turn (synthetic, oracle state).** Graph states the relation before the turn, so the model must apply the turn itself; it reads the pair relation in the wrong direction. Code applies the turn and returns the relation afterwards. The Image derivation reaches the correct answer while equating the person's right with image right.

Question. Without changing pose or receiving extra support, would this person fall over? (A) Yes (B) No

Answer. A **Category.** intuitive physics **State.** oracle



Graph

```
frame:  world, metres, z up; gravity [0, 0, -9.81]
P1  person  colour orange; mass model: uniform-density mesh
    joints_world_m (24): pelvis [-0.013, -0.025, 0.793], neck [-0.098, -0.217, 1.261],
                         head [-0.125, -0.289, 1.293], left_ankle [0.309, 0.341, 0.067],
                         right_ankle [0.047, 0.462, 0.061], ...
    com_m  [-0.017, -0.042, 0.860]
    signed flexion: knee L 5.6 / R 4.7 deg, elbow L 37.8 / R 50.6 deg
    contacts (25; floor, normal +z, mu = 0):
      [-0.040, 0.301, 0.000], [0.250, 0.171, 0.000], [0.267, 0.172, 0.000], ...
      x in [-0.064, 0.343], y in [0.171, 0.498]
collisions: none
```

Code

```
balance_after(person=P1, remove_supports=[], approach=[], friction=[])
  -> verdict: infeasible
     retained_contacts: 25
     rule:  sum f + m g = 0;  sum (p_i - c) x f_i = 0;  unilateral contacts
     scope: external static balance; no muscle torque or dynamic stability
```

Rep.	Answer	Derivation
Image	B ✗	"The person is in a static pose with both feet on the floor [...] the center of mass appears to be within the support polygon formed by the feet. [...] Therefore, the person would not fall over without intervention."
Graph	B ✗	"The person is in static equilibrium with multiple floor contacts (17 contacts total) [...] the CoM is directly above the contact region (no net torque). [...] Therefore, the person will not fall over."
Code	A ✓	"The result returned 'infeasible', meaning no unilateral contact forces can satisfy both force and torque balance with the given CoM and contact points."

Figure 15: **Case 7: balance (synthetic, oracle state).** All 25 floor contacts lie at $y \in [0.171, 0.498]$ m while the centre of mass is at $y = -0.042$ m, so the graph already determines the answer. Under Graph the model miscounts the contacts and judges the pose stable; Code evaluates static equilibrium and returns infeasible.

E FULL RESULTS

Tables 5 and 6 give the numbers plotted in Figures 6 and 7, together with the combined Image+Graph and Image+Code inputs for both state sources and the matched physical-feasibility probes.

Table 5: **Representations with estimated state** (Qwen3-VL-8B, accuracy %, 5,291 questions). *Graph*: scene graph estimated from the image; *Code*: the same graph with executable operations; Image+Graph and Image+Code add the image. Right: the graph restricted to body geometry, visual attributes, or both, without derived relations. Bold: best of the five representations per row.

		Representation					**Graph content**		
	N	Image	Graph	Image+Graph	Code	Image+Code	Geometry	Attributes	Geo.+Attr.
Pose	1,748	64.8	74.0	71.2	76.8	**80.7**	53.2	47.3	57.4
Orientation	1,204	55.6	**72.3**	70.9	71.4	71.9	45.9	49.2	45.9
Relation	1,345	**91.2**	73.9	84.8	74.6	83.5	62.5	69.6	75.3
Collision	453	**68.7**	57.6	63.8	57.6	60.0	48.3	59.2	58.5
Kinematics	100	61.0	59.0	64.0	59.0	**68.0**	53.0	58.0	58.0
Intuitive physics	441	**88.0**	70.3	82.5	69.6	84.3	50.1	72.6	71.7
3D Spatial	4,297	70.5	73.5	75.4	74.6	**79.1**	54.1	54.8	59.8
Spatio-Physics	994	**76.5**	63.4	72.1	63.1	71.6	49.6	65.0	64.3
Descriptive	4,967	71.7	72.3	75.3	73.2	**78.8**	53.3	56.4	60.7
Counterfactual	324	**69.8**	61.1	66.7	61.4	60.8	51.9	62.0	60.2
Overall	5,291	71.6	71.6	74.8	72.4	**77.7**	53.2	56.7	60.6

Table 6: **Physical reasoning and state source.** *Top:* matched feasibility probes (864 images, image only); recall: violations detected; specificity: feasible images accepted; both correct: both images of a matched view correct. *Bottom:* estimated versus oracle scene graphs on 2,314 synthetic questions (Qwen3-VL-8B, accuracy %); the oracle graph is taken from the rendered 3D scene. Bold: best per row. Examples of the matched probes are shown in Appendix G.

Model	Balanced acc.	Recall	Specificity	Both correct
Qwen3-VL-4B	42.7	57.4	28.0	9.5
Qwen3-VL-8B	57.3	94.0	20.6	18.3
Astra	**62.0**	24.5	**99.5**	**24.5**

			Estimated state				**Oracle state**			
	N	Image	Graph	Image+Graph	Code	Image+Code	Graph	Image+Graph	Code	Image+Code
Pose	911	55.4	78.8	70.5	84.4	88.0	86.5	74.5	89.5	**91.2**
Orientation	628	44.8	75.3	71.2	72.0	72.3	**84.4**	81.2	82.2	81.4
Relation	284	87.7	81.3	83.1	77.1	81.0	97.9	**98.2**	96.5	95.1
Collision	334	65.0	52.7	58.4	51.8	54.5	85.0	88.0	88.3	**93.4**
Kinematics	90	58.9	60.0	64.4	58.9	68.9	91.1	91.1	94.4	**95.6**
Intuitive physics	67	71.6	47.8	61.2	50.8	62.7	58.2	62.7	**95.5**	88.1
3D Spatial	1,823	56.8	78.0	72.7	79.0	81.5	87.5	80.5	88.0	**88.4**
Spatio-Physics	491	64.8	53.4	59.9	53.0	58.2	82.5	85.1	90.4	**93.1**
Descriptive	2,184	57.6	73.4	69.9	74.5	77.5	86.9	81.2	88.5	**89.5**
Counterfactual	130	73.1	62.3	71.5	56.1	61.5	80.0	86.2	**90.0**	88.5
Overall	2,314	58.5	72.8	70.0	73.5	76.6	86.5	81.5	88.5	**89.4**

F FAILURE CASES

Figure 16 shows questions that Qwen3-VL-8B answers incorrectly from the image, together with the representation that answers each correctly and the model's quoted reasoning.

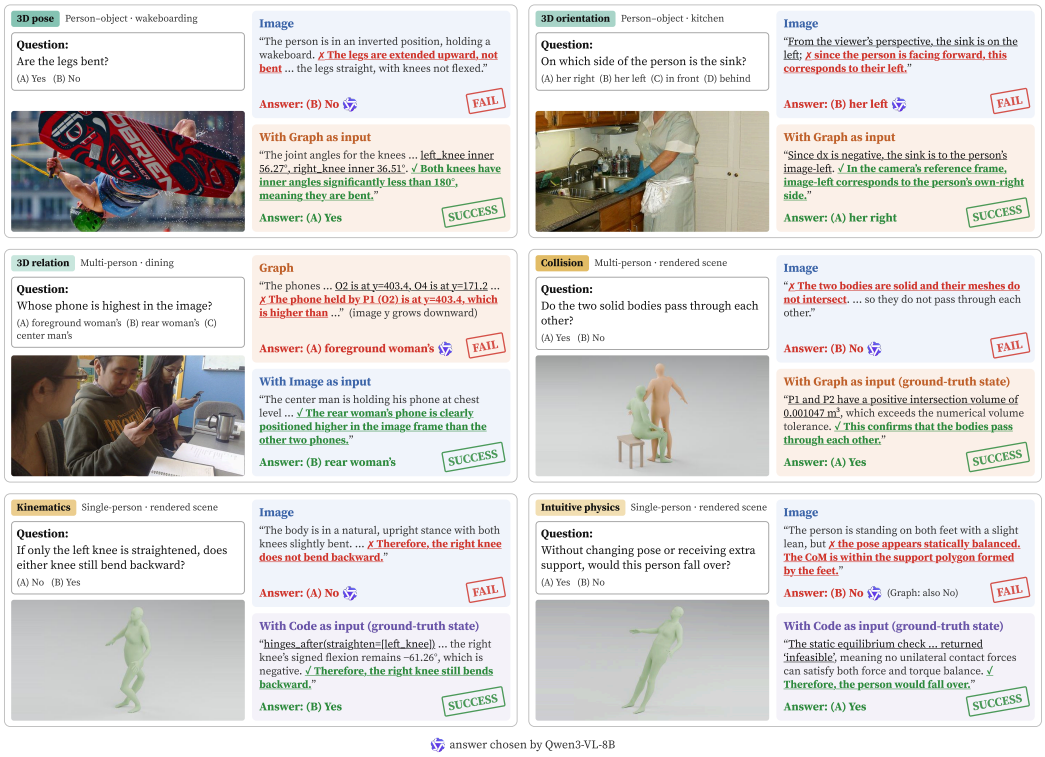


Figure 16: **Failure cases and the representation that recovers them** (Qwen3-VL-8B). Each panel shows a failed answer, the representation that answers correctly, and the quoted reasoning.

G Case Study: Physical Feasibility

Figure 17 shows examples from the matched physical-feasibility probes in Table 6 (top), together with SPARKBench collision and intuitive-physics questions that most evaluated models answer incorrectly.

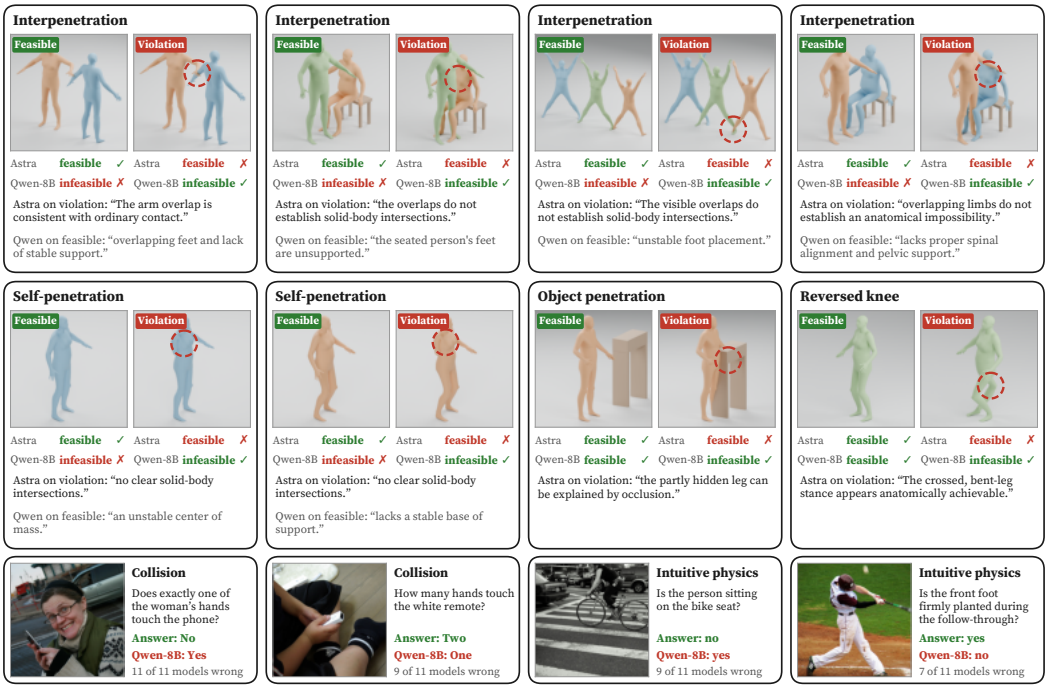


Figure 17: **Physical feasibility case study.** Top two rows: matched feasible and violating renders (interpenetration, self-penetration, object penetration, and a reversed knee), with the verdicts of Astra and Qwen3-VL-8B and an excerpt of each explanation. Astra accepts most violations as feasible, whereas Qwen3-VL-8B rejects most feasible scenes. Bottom row: SPARKBench collision and intuitive-physics questions, with the number of the eleven evaluated models that answer incorrectly.

H Discussion

Extending benchmark coverage. SPARKBench evaluates still images and specified scene changes. A useful extension is to include short sequences that track articulated motion, contact, and support over time. Additional matched scenes spanning occlusion, unusual poses, and crowded interactions would enable more targeted comparisons across visual conditions while retaining explicit questions and verifiable answers.

Improving physical representations. The estimated-to-oracle comparison in Section 5 motivates better recovery of the physical variables required by a question, including contacts, support surfaces, and body geometry. Future comparisons can vary field availability and estimation accuracy separately to identify which improvements translate into better reasoning. Image evidence can remain available alongside structured fields for cues that the representation does not capture. Renderer-provided state serves as a diagnostic reference for this analysis; its gains are distinct from performance using state estimated from images.

Verifying question-aligned computation. A further direction is to examine whether the model selects the intended people and body parts and applies the change specified by the question. Auditing intermediate states and operation outputs can help locate errors in entity binding, operation selection, and execution. Physical checks should use the assumptions stated in the question: for example, static-support judgments can be evaluated with an equilibrium check against known contact and mass information. These comparisons would help connect improvements in individual reasoning steps to final answer accuracy.