SIMVERSE: BENCHMARKING INVERSE MENTAL SIMULATION VIA INTERACTIVE GAMES

Anonymous authors
Paper under double-blind review

ABSTRACT

Inverse mental simulation, the process of reasoning backward from observed physical and spatial outcomes to their causal antecedents, requires compositional inference over continuous physical dynamics and discrete spatial topology. Existing benchmarks predominantly target forward prediction, and none is dedicated to the inverse, effect-to-cause direction in multimodal large language models. We introduce SIMVERSE, a benchmark for inverse mental simulation in multimodal large language models. SIMVERSE comprises 2,486 simulator-verified instances across four interactive game environments, organized into a 2×2 taxonomy by operation order-sensitivity (order-invariant or order-sensitive) and state-transition type (discrete or continuous). Every model output is executed in a domain-specific physics or geometry simulator, so any valid antecedent earns credit and partial credit reflects physical or geometric progress. We evaluate seven models against an IQ-screened human baseline ($n = 50$). The strongest model (Gemini-3.1-Pro) achieves an overall score of 52.7%, falling 22.8 percentage points (95% CI: 19.3–26.4) below the human reference (75.5%), with marked variation across task families. On dice face inference, models identify up to 92.3% of symbols, yet none exceeds 32% once orientation is also required, exposing a bottleneck in cumulative state tracking. A concise direct-answer prompt and removal of the rendered image each lower success for all seven models and raise it on no task, while per-task model rankings remain largely stable. For both humans and models, order-sensitive discrete tasks are markedly harder than order-invariant continuous ones, identifying operation order-sensitivity as a central obstacle for inverse reasoning.

1 INTRODUCTION

Human spatial and physical competence is commonly explained by mental simulation (Shepard & Metzler, 1971; Battaglia et al., 2013; Tenenbaum et al., 2011; Lake et al., 2017): people mentally rotate objects (Cooper, 1975), anticipate mechanical outcomes (Hegarty, 2004), and infer occluded configurations from physical priors (Spelke, 1990; Baillargeon, 2008). Such simulation supports not only forward prediction but also backward inference from an observed outcome to what must have produced it (Sanborn et al., 2013; Hamrick et al., 2016; Hamrick, 2019). We use the term *inverse mental simulation* operationally: the backward reconstruction of *simulator-valid antecedents*, i.e. configurations or action sequences that a simulator verifies to produce the observed outcome.

The inverse direction deserves evaluation in its own right, for reasons established well outside machine learning. First, practical inverse problems are typically *ill-posed*: a solution may fail to be unique or stable, and recovery requires prior structure or regularization (Engl et al., 1996; Stuart, 2010). Second, internal-model theories of motor control treat forward models (predicting the consequences of actions) and inverse models (producing actions that achieve desired outcomes) as distinct computational objects with distinct learning problems (Jordan & Rumelhart, 1992; Wolpert & Kawato, 1998; Kawato, 1999). Third, *determinacy and computational cost come apart*: inverting a map can be far costlier than evaluating it whether or not the antecedent is unique. Forward kinematics is a closed-form composition of transforms, whereas inverse kinematics requires constraint satisfaction over a continuous manifold (Buss, 2009; Aristidou et al., 2018).

Frontier multimodal large language models perform strongly across visual and linguistic tasks (OpenAI, 2025; Gemini Team, 2023), and their physical and spatial reasoning has been probed mainly via

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.

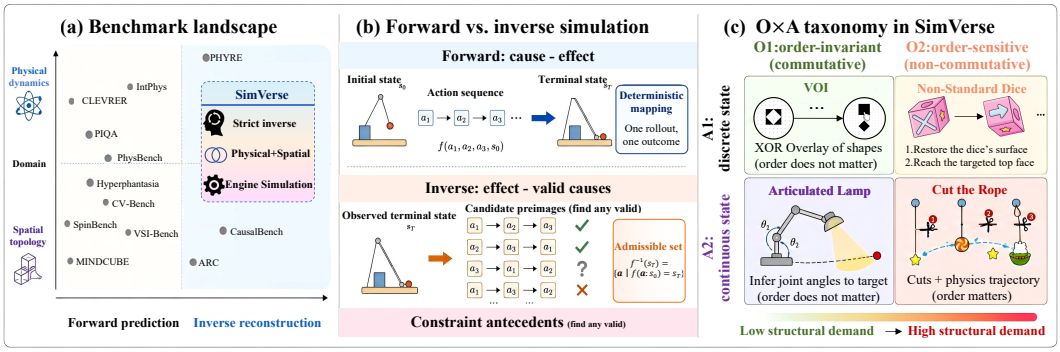

Figure 1: **(a)** Most physical and spatial reasoning benchmarks cluster in the forward-prediction region. SIMVERSE spans both domains under a strict inverse formulation. **(b)** Forward simulation maps an initial state and action sequence to an outcome. Inverse mental simulation seeks any antecedent in the admissible set $f^{-1}(s_T)$, sized by the observation model and constraints. SIMVERSE verifies candidates by simulator execution. **(c)** The four environments in the O×A taxonomy: operation order-sensitivity (O1/O2) and state-transition type (A1: discrete, A2: continuous).

forward-structured tasks: predicting trajectories, resolving spatial relations, and answering physical commonsense questions (Yi et al., 2020; Liu et al., 2023; Chow et al., 2025). Such formats are exposed to distributional shortcuts via surface-level features and training priors (Geirhos et al., 2020; Tong et al., 2024b), so apparent competence can reflect forward-associative recall rather than simulation (Tong et al., 2024b; Li et al., 2026). Crucially, the direction of inference is not behaviorally neutral for current language models: they fail to reverse strictly bijective facts (the reversal curse, Berglund et al., 2024), and their accuracy degrades when math word problems are posed backward, recovering a missing premise from a given answer (Deb et al., 2023). Mathematical invertibility does not imply learned invertibility. Meanwhile, whether an inverse problem has a unique answer is fixed by its observation model and constraints, not by inference direction. An informative inverse benchmark must therefore make the observation model *explicit, parameterized and verified per instance*, and grade answers by execution rather than against a single reference. As Figure 1(a) shows, most existing benchmarks cluster in the forward-prediction region and none offers this combination.

We thus ask: *given an observed outcome and an explicit observation model, can a frontier multimodal model reconstruct an antecedent that a simulator verifies, and where along the inverse problem's structure does reconstruction break down?* We introduce SIMVERSE to answer it. To our knowledge, it is the first benchmark to evaluate multimodal models on simulator-verified inverse reconstruction across both physical and spatial domains. It comprises 2,486 simulator-verified instances, procedurally generated (VOI, Dice) or drawn from parameterized level sets (Lamp, Rope), across four interactive game environments, organized by a principled 2×2 taxonomy: operation order-sensitivity (O1/O2) and state-transition type (A1/A2). Every output runs in a domain-specific simulator, so any goal-satisfying one earns credit, string-matching bias is eliminated, and partial credit is grounded in physical or geometric progress. Explicit difficulty parameters and arbitrary symbolic assignments are built to guard against memorization and co-occurrence shortcuts. An IQ-screened human reference ($n = 50$) is scored by the same pipeline.

Evaluating seven frontier models (Google DeepMind, 2026; OpenAI, 2026; Qwen Team, 2026a;b; Bai et al., 2025; Alibaba Cloud, n.d.; GLM-5V-Turbo Team, 2026; StepFun, n.d.) yields a statistically robust gap: the best model (Gemini-3.1-Pro, Overall 52.7%) trails humans (75.5%) by 22.8 pp [19.3, 26.4], and 27 of 28 human–model family comparisons exclude zero. No model dominates all families. Models recover *what* but lose track of *how*: on Dice Face inference, Pattern Accuracy reaches 92.3% but Pattern+Orientation Accuracy is at most 31.6%, a cumulative state-tracking bottleneck milder in humans than in most models. Both populations pay more for the operation axis than for the state-transition axis (O1+A2 vs. O2+A1). Removing the image or the prompt's auxiliary guidance lowers success for every model, significantly in 26 and 17 of 35 model–task cells, while per-task rankings stay largely intact. Most failures are executable near misses, not formatting errors.

Our contributions are threefold. **(i)** A formal, execution-verifiable formulation of inverse mental simulation that separates state space, action space, observation model and admissible solution set, with

a two-axis difficulty decomposition that yields testable predictions. **(ii)** The SIMVERSE benchmark: 2,486 solvability-verified instances with documented generation parameters, execution-based grading across four heterogeneous engines, a diagnostic PA/POA decomposition, and an interface-matched human reference. **(iii)** An evaluation of seven frontier models with confidence intervals, prompt and modality ablations, and a failure decomposition, revealing a robust human–model gap, family-specific competence, and an operation-dominant difficulty structure shared with humans.

2 RELATED WORK

Physical and spatial reasoning benchmarks. Intuitive-physics benchmarks score plausibility judgments on rendered events (Riochet et al., 2022; Bordes et al., 2025) or answers to object-collision questions (Yi et al., 2020). Physical commonsense is probed with text or multimodal questions (Bisk et al., 2020; Chow et al., 2025) and, for text-to-image models, generated scenes (Meng et al., 2024). Spatial benchmarks probe relation perception, localization and viewpoint reasoning (Liu et al., 2023; Tong et al., 2024a; Yang et al., 2025; Ramakrishnan et al., 2025). Broad suites (Yue et al., 2024; Lu et al., 2024; Yu et al., 2024) cover diverse visual domains, and Hyperphantasia (Sepehri et al., 2025) evaluates mental visualization by completing or extrapolating visual patterns. The closest precedent is PHYRE (Bakhtin et al., 2019), a simulator-graded suite of goal-directed 2D physics puzzles where an agent places one or two balls to reach a goal. It targets agents allowed many single-action attempts per puzzle in one physical domain. SIMVERSE instead evaluates single-shot reconstruction by multimodal models across physical and spatial domains, includes order-sensitive multi-step tasks, makes the observation model explicit and reports a human reference.

Inverse problems and internal models. Ill-posed inverse problems are handled by regularization or Bayesian priors (Engl et al., 1996; Stuart, 2010). Motor control pairs forward and inverse internal models (Jordan & Rumelhart, 1992; Wolpert & Kawato, 1998; Kawato, 1999). Inverse dynamics (Pathak et al., 2017), inverse kinematics (Buss, 2009; Aristidou et al., 2018) and inverse graphics (Kulkarni et al., 2015; Ravi et al., 2020; Kuang et al., 2023) recover actions, joint configurations or scene parameters. These works develop solvers. None evaluates multimodal models.

Inverse and counterfactual reasoning in language models. Language models show direction-dependent behavior (Berglund et al., 2024; Deb et al., 2023). CausalBench (Wang, 2024) and CRASS (Frohberg & Binder, 2022) evaluate causal and counterfactual language reasoning, and ARC (Chollet, 2019) requires inducing transformations from input–output pairs, but none provides physical grounding, simulator-based grading, or a controlled decomposition of inverse difficulty. Concurrent spatial benchmarks (SpinBench (Zhang et al., 2026), VSI-Bench (Yang et al., 2025), MINDCUBE (Wang et al., 2026), TangramPuzzle (Liu et al., 2026)) target richer spatial or compositional structure, predominantly in the forward paradigm (Table 1).

3 INVERSE MENTAL SIMULATION: FORMAL FRAMEWORK

3.1 DEFINITIONS

Let $\mathcal{S}$ be the *environment state space* (object configurations with geometric, topological and dynamic properties) and $\mathcal{A}$ the *primitive action space*. A *forward simulation* is a deterministic transition $f : \mathcal{A}^n \times \mathcal{S} \to \mathcal{S}$ mapping an ordered action sequence $\mathbf{a} = (a_1, \ldots, a_n)$ and an initial state s_0 to a terminal state $s_T = f(\mathbf{a}, s_0)$. An *observation model* specifies what the solver sees, e.g. the rendered target and available pieces, or roll directions and the symbol touching the ground per roll. *Inverse mental simulation* asks for any element of the *admissible inverse solution set*

$$f^{-1}(s_T) = \{\mathbf{a} \in \mathcal{A}^n \mid f(\mathbf{a}, s_0) \in \mathcal{G}(s_T),\ \mathbf{a} \text{ satisfies } \mathcal{C}\},$$

where $\mathcal{C}$ collects domain constraints and the goal set $\mathcal{G}(s_T) \subseteq \mathcal{S}$ contains the terminal states counted as producing the observed or specified outcome s_T: the singleton $\{s_T\}$ when an exact match is required (VOI) and a goal region when success is a goal predicate (Dice Path: target face and orientation, Lamp: beam center within the light radius, Rope: candy in the mouth). In configuration-inference tasks the roles swap: actions are observed and the unknown antecedent is (part of) s_0.

Table 1: Comparison of SIMVERSE with representative benchmarks, per the original papers. Direction: primary task formulation. Sim.-Graded: outputs are scored by simulator execution. Inverse: requires inferring causes, actions or configurations from observed or desired effects (Partial: some question types). Order-Sens.: solutions are ordered multi-step action sequences. Continuous: outputs include continuous values. Multi-Obj.: several objects jointly determine the answer (Partial: some items). Human BL: reports human performance. ‡Evaluates text-to-image generation.

Benchmark	Direction	Domain	Sim.-Graded	Inverse	Order-Sens.	Continuous	Multi-Obj.	Human BL
PHYRE (Bakhtin et al., 2019)	Inverse	Physical	✓	✓	×	✓	✓	×
IntPhys (Riochet et al., 2022)	Forward	Physical	×	×	×	×	×	✓
CLEVRER (Yi et al., 2020)	Forward	Physical	×	Partial	×	×	✓	×
PhysBench (Chow et al., 2025)	Forward	Both	×	×	×	×	✓	✓
PhyBench‡ (Meng et al., 2024)	Forward	Physical	×	×	×	×	Partial	×
PIQA (Bisk et al., 2020)	Forward	Physical	×	×	×	×	×	✓
VSR (Liu et al., 2023)	Forward	Spatial	×	×	×	×	✓	✓
CV-Bench (Tong et al., 2024a)	Forward	Spatial	×	×	×	×	✓	×
ARC (Chollet, 2019)	Forward	Abstract	×	Partial	×	×	✓	×
CausalBench (Wang, 2024)	Mixed	Abstract	×	Partial	×	×	×	×
SpinBench (Zhang et al., 2026)	Forward	Spatial	×	Partial	×	×	✓	✓
VSI-Bench (Yang et al., 2025)	Forward	Spatial	×	Partial	×	✓	✓	✓
MINDCUBE (Wang et al., 2026)	Forward	Spatial	×	×	×	×	✓	✓
IntPhys 2 (Bordes et al., 2025)	Forward	Physical	×	×	×	×	✓	✓
Hyperphantasia (Sepehri et al., 2025)	Forward	Spatial	×	×	×	×	×	✓
SIMVERSE (ours)	**Inverse**	**Both**	✓	✓	✓	✓	✓	✓

Three consequences shape the benchmark. **(1) Identifiability is a property of the observation model.** With complete observations of a bijective f the antecedent is unique. With partial observations it may not be, in either direction. SIMVERSE therefore exposes the observation model through generation parameters (e.g. roll-sequence length determines how many dice faces are ever observed) instead of leaving it implicit. **(2) Determinacy is not cost.** Finding an antecedent can be costly whether or not it is unique. Articulated Lamp instantiates the forward- versus inverse-kinematics cost contrast directly. **(3) Grading by membership.** An output receives full credit when executing it in the simulator satisfies the goal, i.e. when it lies in the admissible set, so correct alternatives to the reference solution are not penalized (in Dice Face, answering never-observed faces with '?' makes the admissible answer unique, so membership reduces to comparison with the simulator-derived assignment). Simulator-grounded partial credit measures proximity to the goal. Per-task specifications of $\mathcal{S}$, $\mathcal{A}$, the observation model and $f^{-1}(s_T)$ are given in Appendix A.

3.2 TWO DIFFICULTY AXES

Operation dimension (O): order-sensitivity. A task is *order-invariant* (**O1**) if f is invariant to every permutation σ of the action sequence. Recovering the *set* of actions then suffices. It is *order-sensitive* (**O2**) if some σ gives $f(a_{\sigma(1)}, \ldots, a_{\sigma(n)}, s_0) \neq f(a_1, \ldots, a_n, s_0)$. The solver must then recover both set and order, which requires tracking cumulative state across every step.

Transition dimension (A): state-transition type. The A axis characterizes how the world state changes in the simulation the solver must run mentally. A task is *discrete* (**A1**) if the state advances by discrete transitions among finitely many configurations (a shape toggling grid cells, a die rolling by $90°$), so every intermediate state can be tracked symbolically and exactly. It is *continuous* (**A2**) if the state evolves continuously under actions (an arm sweeping through its joint angles, a candy swinging and falling under gravity), so the solver must propagate continuous quantities through continuous geometry or dynamics and judge magnitudes along the way. The axis concerns the simulated process, not the answer format: Lamp answers are quantized to a fixed angle step and Rope answers are discrete command sequences, yet in both the simulated state changes continuously. Action and answer encodings are given per task in Appendix A. Since they are not independently crossed with the state-transition type, the taxonomy keeps two axes.

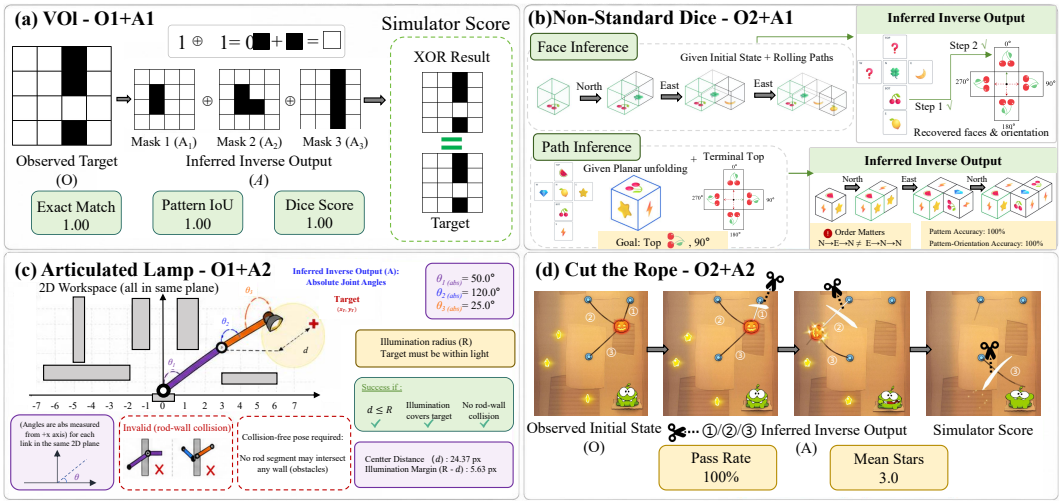

Figure 2: Representative SIMVERSE instances. **(a)** VOI (O1+A1, discrete): infer shape placements whose XOR overlay matches the target. **(b)** Non-Standard Dice (O2+A1, discrete): infer the initial face map with orientations from a roll sequence (Face), or a roll sequence reaching a target face and orientation (Path). **(c)** Articulated Lamp (O1+A2, continuous): infer joint angles that bring the beam onto the target without collisions. **(d)** Cut the Rope (O2+A2, continuous): infer an ordered command sequence that delivers the candy to the mouth while collecting stars.

3.3 PREDICTED DIFFICULTY ORDERING

For a reasoner whose constraints are working-memory capacity and continuous-quantity estimation, as in human mental simulation (Baddeley, 2007; Hegarty, 2004), the framework predicts

$$\text{O1+A1} \prec \{\text{O2+A1}, \text{O1+A2}\}^{\dagger} \prec \text{O2+A2}, \tag{1}$$

where $X \prec Y$ means that Y is expected to be harder than X, and $\dagger$ leaves the order of the two singly-demanding families open. It predicts axis-specific failures: *state-tracking drift* under O2 (early steps correct, errors accumulating) and *numerical imprecision* under A2 (correct approach, misjudged magnitudes), mirroring working memory (Ecker et al., 2010) and spatial imagery (Kosslyn et al., 2006). SIMVERSE tests these predictions as functional signatures in both populations.

4 SIMVERSE: BENCHMARK DESIGN

SIMVERSE follows three principles: **(P1) inversion under an explicit observation model**, **(P2) execution-based grading**, and **(P3) structured difficulty** along O×A. Each prompt contains the task description, a structured state encoding with explicit coordinates, and a JSON output schema, accompanied by a rendered scene image (Figure 2).

Game 1: Visual Overlay Interference (VOI, O1+A1). Binary-mask shapes are placed on a discrete canvas and combined by pixel-wise XOR ($1 \oplus 1 = 0$). The solver infers the placement of each shape so that the overlay reproduces a target silhouette. XOR is commutative and associative, hence O1. *Metrics:* **Match Rate** (exact overlay match, primary), **Pattern IoU**, and **Dice Score**.

Game 2: Non-Standard Dice (O2+A1). A die carries six arbitrary symbols. Rolls compose in the non-Abelian octahedral rotation group (24 elements), simulated with exact 3×3 integer rotation matrices. **Face Inference:** given N roll directions and the bottom-face imprint (symbol and rotation) after each roll, infer the initial face-to-symbol map and each symbol's orientation ($0°/90°/180°/270°$). **Path Inference:** given the initial face map, output a roll sequence bringing a target symbol to a specified face and orientation. *Metrics:* **Pattern Accuracy (PA)** (symbols correct) and **Pattern+Orientation Accuracy (POA)** (symbol and orientation correct). POA is the primary metric. PA−POA isolates orientation tracking from symbol assignment on the same instance. Face **Pass Rate** (all faces fully correct) is reported as the strictest end-to-end reference.

Table 2: Main results (text+image, original prompt) with 95% CIs. Overall: family macro. Gap: human − model on Overall. All values in %/pp.

Model	Overall	Human−model gap
Human (n=50)	*75.5 [72.5, 78.6]*	—
Gemini-3.1-Pro	**52.69** [50.8, 54.6]	+22.8 [+19.3, +26.4]
Qwen3.5-397B	42.53 [40.8, 44.3]	+33.0 [+29.4, +36.4]
Qwen3.6-Plus	37.37 [35.3, 39.4]	+38.1 [+34.5, +41.8]
GPT-5.4	27.37 [25.4, 29.3]	+48.1 [+44.6, +51.7]
GLM-5V-Turbo	20.33 [18.5, 22.1]	+55.2 [+51.7, +58.6]
Qwen3-VL-Plus	19.20 [17.6, 20.8]	+56.3 [+52.9, +59.8]
Step-1O-Turbo-Vision	11.14 [9.8, 12.5]	+64.4 [+61.1, +67.6]

Game 3: Articulated Lamp Positioning (O1+A2). A planar multi-joint arm with a lamp head must illuminate a 2D target: the solver infers joint angles within limits, a continuous inverse-kinematics problem, and a forward-kinematics solver evaluates the pose. The order in which joints are set is irrelevant (O1). *Metrics:* **Success Rate (SR)** (target within the light radius and no link–obstacle collision, primary) and **Center Distance (CD)** (Euclidean error, px).

Game 4: Cut the Rope (O2+A2). Inspired by ZeptoLab (2010), a candy hangs from ropes attached to fixed anchors. The solver outputs an ordered list of gameplay commands (e.g. cutting a given rope, popping a bubble), each optionally triggered by a continuous condition such as candy position or rest duration, to deliver the candy to the mouth while collecting stars. Dynamics are simulated in Box2D (Catto, n.d.) at a fixed $\Delta t = 1/60\,\mathrm{s}$ with 10 velocity and 10 position iterations from a fully specified initial state, so a given command sequence always yields the same trajectory. *Metrics:* **Pass Rate** (primary) and **Mean Stars** (out of 3).

Instance generation and verification. SIMVERSE has 2,486 instances (VOI 600, Dice Face 502, Dice Path 502, Lamp 610, Rope 272). VOI and Dice instances are procedurally generated from machine-readable configurations with explicit difficulty controls: grid size and required/distractor shape counts (seven VOI tiers), roll-sequence length (2–10 moves, five Dice Face tiers) and target rotation offset (Dice Path). Validity filters reject degenerate candidates. Lamp and Rope levels are drawn from parameterized level sets, each shipped with a reachable reference solution (Appendix C). Every instance is simulator-verified to admit a valid solution. A schema-level check enforces a single authoritative answer field. The dice orientation convention is documented in Appendix B.

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

Models and protocol. We evaluate Gemini-3.1-Pro, GPT-5.4, Qwen3.5-397B, Qwen3.6-Plus, Qwen3-VL-Plus, GLM-5V-Turbo and Step-1O-Turbo-Vision via official APIs. All models receive identical instructions, state encodings, output schemas and instances from shared templates (with only minimal API-specific formatting). Inference is single-shot at temperature 0 with chain-of-thought prompting (Wei et al., 2022) and platform-default extended reasoning where exposed, without simulator feedback. Malformed JSON is re-requested, still-unparseable responses count as zero-credit parse failures, and no retry depends on correctness. Parsing and scoring are identical for all systems: we evaluate deployed end-to-end systems under a standardized protocol (Appendix D).

Human reference. Fifty graduate STEM participants at or above the 80th percentile on the 36-item Raven's APM (mean IQ 122.3 ± 7.6) each completed a stratified random sample of 50 instances through the same text-and-image interface, without simulator feedback. Responses were scored by the same simulators (Appendix E). This is a cognitively bounded reference rather than a ceiling.

Scores and statistics. The single aggregate score, **Overall**, is the family macro: the mean of VOI Match Rate, Dice (mean of Face and Path POA), Lamp SR and Rope Pass Rate (Appendix F). Model

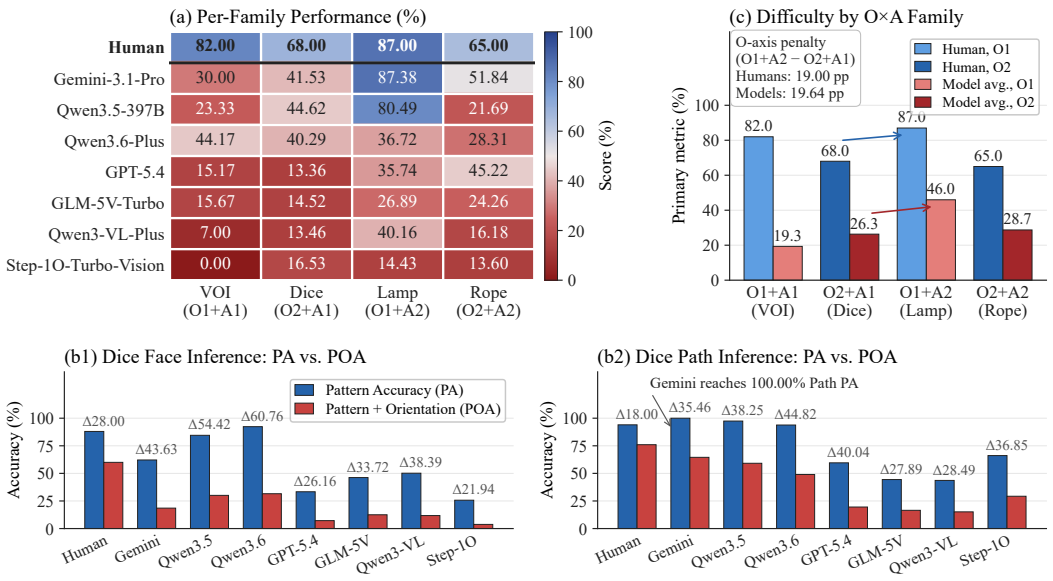

Figure 3: Main results (simulator-graded). **(a)** Per-family primary metrics (VOI Match Rate, mean Dice Face/Path POA, Lamp SR, Rope Pass Rate). Humans (n=50) exceed every model on VOI, Dice and Rope. Only Gemini-3.1-Pro matches the human Lamp SR. **(b1, b2)** PA versus POA on Dice Face and Path, with PA−POA gaps (Δ). Although PA exceeds 92% on Face (Qwen3.6-Plus) and reaches 100% on Path (Gemini-3.1-Pro), POA falls by 22–61 pp (Face) and 28–45 pp (Path). Human gaps are 28 pp (Face) and 18 pp (Path, see Finding 1). **(c)** Difficulty by O×A family. The O1+A2 → O2+A1 drop is 19.0 pp for humans and 19.6 pp for the model mean.

intervals are 95% task-stratified bootstrap intervals (10,000 replicates). Ablation differences use paired bootstrap per model×task cell on binary per-instance success (Dice Face: Pass Rate, Dice Path: POA), with exact McNemar tests, and Wilson intervals for binomial rates. Human intervals are participant-clustered, reported for ICC $\in \{0.05, 0.10, 0.20\}$ (reference 0.10).

5.2 Main Results

Table 2 and Figure 3 summarize the results. Per-game tables with secondary metrics are in Appendix G and all intervals in Appendix H.

Finding 1: Models recover *what*, not *how*: a pervasive PA→POA collapse. On Dice Face, Qwen3.6-Plus identifies 92.3% of symbols (PA) but only 31.6% of symbol–orientation pairs (POA), a 60.8 pp [58.3, 63.3] drop. Across models the gap spans 22–61 pp on Face and 28–45 pp on Path, where Gemini-3.1-Pro reaches 100% PA but 64.5% POA. PA reflects symbol assignment, while POA also requires carrying the cumulative rotation across all rolls, so the gap isolates the state-tracking drift predicted for O2. Since Dice Path supplies the full face map, its gap cannot stem from unobservable faces. Humans show the same pattern (gaps: Face 28.0 pp, Path 18.0 pp). On Path, every model's gap exceeds the human gap significantly (+9.9 to +26.8 pp). On Face, five models exceed it significantly (+5.7 to +32.8 pp), including all four with PA $>$ 50% (Qwen3.6-Plus, Qwen3.5-397B, Gemini-3.1-Pro, Qwen3-VL-Plus). For the two models with the lowest Face PA, GPT-5.4 (33.4%) and Step-1O-Turbo-Vision (25.7%), the excess is indistinguishable from zero or negative (-1.8 [-6.9, $+3.1$], -6.1 [-10.8, -1.2]), a floor effect: little pattern accuracy remains for orientation errors to erode (Appendix H.1). End-to-end Face Pass Rate is at most 6.0% for every model (five score exactly 0, Wilson upper bound 0.8%) vs. 18.0% for humans.

Finding 2: A substantial, statistically robust human–model gap. Gemini-3.1-Pro leads with an Overall score of 52.7% [50.8, 54.6], 22.8 pp [19.3, 26.4] below the human reference of 75.5% [72.5, 78.6]. The human interval does not overlap the best model's at any of the three ICC settings. Of the 28 human–model family comparisons, 27 exclude zero. The single exception is Gemini-3.1-Pro on

Lamp (-0.4 [-4.4, +3.7]). Even there, Gemini's mean Center Distance (21.73 px) is about $2.6\times$ the human value (8.50 px), and Qwen3.5-397B approaches but stays significantly below humans (80.49%, +6.5 [+2.0, +11.0]). On Rope, Gemini-3.1-Pro (51.8%) leads GPT-5.4 (45.2%, about $1.15\times$) and remains 13.2 pp [5.9, 20.5] below humans.

Finding 3: Model competence is family-specific. No model dominates. Gemini-3.1-Pro leads on Lamp SR, Rope and Dice Path POA (64.5%) but trails Qwen3.6-Plus on VOI (30.0% vs. 44.2%). Qwen3.6-Plus has the best VOI Match Rate and Face PA/POA (92.3%/31.6%) yet only 36.7% on Lamp and 28.3% on Rope. GPT-5.4 ranks fourth overall but second on Rope. The cross-model spread exceeds 30 pp in every family (73.0 pp on Lamp), whereas the human cross-family range is 22 pp: model competence for inverse reconstruction is heterogeneous and not predictable from a single aggregate. Binary success also understates graded competence: on VOI, Qwen3.6-Plus has 0.786 IoU at 44.2% exact match, and Step-1O-Turbo-Vision 0.303 at 0.0% (Appendix H.2).

Finding 4: The operation axis dominates difficulty in both populations. Humans score 82.0%, 68.0%, 87.0% and 65.0% on O1+A1, O2+A1, O1+A2 and O2+A2. The model means are 19.3%, 26.3%, 46.0% and 28.7%. Both populations find O1+A2 much easier than O2+A1, with nearly identical drops (humans 19.0 pp, models 19.6 pp): the operation axis imposes a larger cost than the state-transition axis in this contrast, resolving the order left open by † in Eq. (1). For humans both O2 families occupy the hard end and O2+A2 is lowest, as predicted. The two departures from Eq. (1) are diagnosable: both populations score O1+A2 above O1+A1 (for humans within overlapping intervals), and models also invert O1+A1 and O2+A1, traceable to a metric asymmetry: VOI requires an exact pixel-level match, whereas POA credits each correct face. For models the two O2 families lie close together (O2+A2 28.7% slightly above O2+A1 26.3%) and far below O1+A2.

5.3 Robustness to Prompt Scaffolding and Input Modality

Prompt sensitivity. A concise direct-answer prompt keeps the task rules, structured state, action constraints and output schema but removes auxiliary guidance, checking reminders and redundant explanation (Appendix D.1). It lowers success for all seven models and raises it on no task: 17 of 35 model$\times$task cells drop significantly, and every model drops significantly on at least one task. The largest drop is Gemini-3.1-Pro on Lamp (-17.0 [-21.6, -12.5]). Per-task model rankings are largely preserved (Spearman $\rho = 0.883$–0.991, undefined on Dice Face, where all models score zero Pass Rate under the concise prompt, Appendix I).

Modality ablation. The text-only condition removes the rendered image and supplies a deterministic, task-specific serialization of the same simulator state. Schema checks verify completeness, while leakage tests exclude answers and reference solutions (Appendix D.2). Removing the image lowers success for all seven models and raises it on no task: 26 of 35 cells drop significantly, and every model drops on at least three tasks. Per-task rankings remain largely preserved ($\rho = 0.927$–0.964). The largest drops occur on Lamp and VOI (up to -26.5 pp, Qwen3.6-Plus on VOI). Invalid outputs rise from 1.4% to 4.7% (macro), mostly on Rope (Qwen3-VL-Plus 0.0%$\rightarrow$52.6%, Step-1O-Turbo-Vision 8.5%$\rightarrow$41.9%). With information held constant, current models extract spatial structure more effectively from renderings than equivalent symbolic encodings: part of the difficulty lies in visual–spatial integration, and action grounding degrades without rendering (Appendix J).

5.4 Failure Decomposition

Each output is assigned to one category: parse failure, parsable but simulator-invalid, executable with zero credit, executable near miss (positive simulator-grounded partial credit), or success (Table 3). In the primary condition, parse failures and invalid actions total 1.4% (cell macro) and near misses 44.7%: when models fail, they almost always produce executable antecedents that are physically or combinatorially wrong (Appendix K).

Table 3: Outcome categories (macro over model×task cells, humans over tasks).

Condition	Parse	Invalid	Zero	Near	Succ.
Original	0.1	1.4	27.6	44.7	26.3
Concise	0.3	1.0	28.9	48.7	21.1
Text-only	0.5	4.7	26.6	51.7	16.5
Human	*0.0*	*0.8*	*3.8*	*29.8*	*65.6*

Task-level signatures (pooled over conditions, Appendix K.1) bear out the O2 prediction. On Dice Path the dominant failure is a correct pattern with wrong rotation (43.3% of outputs). On Dice Face outputs dominated by symbol errors are most frequent (39.8%), but 27.5% get every symbol right and an orientation wrong, the instance-level counterpart of the PA→POA collapse. VOI failures are mostly shape deviations ($0 <$ IoU < 0.6, 60.7%). On Lamp the most frequent failure misses the target and hits an obstacle (27.6%). On Rope, simulation timeouts dominate (23.0% with and 19.5% without stars, cf. trajectory anchoring, Appendix L). In the primary condition, Step-1O-Turbo-Vision emits simulator-invalid joint-angle vectors on 193/610 (31.6%) Lamp instances (next highest: Qwen3.6-Plus, 33/610 = 5.4%).

6 Limitations and Scope

Scope of the term. *Inverse mental simulation* names a task class. It presupposes no particular theory of internal world models. **No matched forward control.** Without a strictly matched forward counterpart, we do not claim that inverse reconstruction is universally harder than forward prediction or that the observed bottlenecks would be weaker in forward form. Shortcut reduction is a design motivation, not a measured result. **PA→POA mechanism.** PA and POA differ only in the orientation requirement, so we attribute the gap to cumulative state tracking. The per-step course of the drift is not measured directly. The released generation parameters directly support matched forward/inverse controls, controlled distribution shifts, identifiability-controlled PA/POA subsets and roll-length stratification. **Decoding.** Results characterize single-shot deterministic decoding. Self-consistency and best-of-k sampling are untested. **Deployed systems.** Models are accessed through commercial APIs whose internal reasoning implementations are not controllable. The robustness analyses show that per-task rankings largely survive changes to our scaffolding, but we do not isolate architectural differences. **Axis contrasts.** Each O×A family is instantiated by one game with its own primary metric, so axis contrasts are partly confounded with task and metric design (e.g. exact matching on VOI). **Human reference.** With $n = 50$ the human intervals are wider. We report them under three ICC assumptions. Human–model comparisons are unpaired (Appendix F). **Other.** Because malformed responses are re-requested and only final responses are stored, the parse-failure rate is a lower bound on raw formatting errors. Environments are two-dimensional, and the released stratified subset of model outputs supports partial offline verification, not full reconstruction of every API run.

7 Conclusion

SimVerse establishes inverse mental simulation (reconstructing simulator-valid antecedents from observed outcomes) as a measurable evaluation regime grounded in inverse-problem and internal-model research. It formalizes inverse difficulty along operation order-sensitivity and state-transition type, instantiates the full 2×2 taxonomy with 2,486 simulator-verified instances with explicit, parameterized observation models, and grades every output by execution. The results are clear: the best of seven frontier models (Overall 52.7%) falls 22.8 pp [19.3, 26.4] below an IQ-screened human reference. Competence is family-specific. Models recover *what* but not *how*, a cumulative state-tracking collapse that exceeds the human one for most models, and both populations pay more for the operation axis than for the state-transition axis (O1+A2 vs. O2+A1). The model ranking is robust: under a concise prompt and without the image, every model loses success, none gains on any task, and per-task rankings stay largely intact. Cumulative state tracking is therefore the capability to target next, e.g. through forward–backward cycle consistency or curricula over sequential operations, and the released generators and simulators support extensions to 3D physics, multi-agent inverse planning and matched forward/inverse controls.

Ethics Statement

The human reference involved adult volunteers (graduate STEM students) screened with Raven's Advanced Progressive Matrices. Recruitment, screening and procedure are described in Appendix E. The benchmark consists of synthetic game instances, procedurally generated (VOI, Dice) or drawn from parameterized level sets (Lamp, Rope).

REPRODUCIBILITY STATEMENT

The code is provided as supplementary material: the instance generators (VOI, Dice) and level sets with reference solutions (Lamp, Rope), the four simulators, the machine-readable generation configurations with random seeds (Appendix C), the evaluation harness and prompt templates (Appendix D), and the statistics scripts (10,000 bootstrap replicates, seed 20260803). The dataset of all 2,486 instances is available at `https://huggingface.co/datasets/SimVer-ano/simverse2026`. A documented, stratified subset of model outputs covering all seven models, all five task variants, and both successes and failures supports qualitative inspection and offline re-scoring on that subset, i.e. partial offline verification rather than full reconstruction of every API run (Appendix M).

AI USE STATEMENT

Large language models are the objects of evaluation in this work (Section 5). No LLM output is used in scoring. We used generative AI tools in four supporting roles. (i) An LLM-based pre-filter flagged potentially degenerate or ambiguous configurations among procedurally generated candidates, with final acceptance decided solely by deterministic simulator verification. (ii) AI tools assisted in writing and editing code, including the analysis scripts that compute all reported statistics and the plotting code for Figure 3. (iii) They helped refine prompt templates on dry-run outputs from LLMs not evaluated here. (iv) They helped draft, edit and check prose (including text describing the reported statistics), obtain feedback on its presentation, and search and summarize related literature. Generative AI was not used to generate data, formulate hypotheses, develop the O×A framework, design experiments, translate, or analyze data qualitatively. The paper contains no proofs. We take full responsibility for the content of this paper, including anything produced with the aid of generative AI.

REFERENCES

Alibaba Cloud. Qwen3-VL-Plus model documentation. Alibaba Cloud Model Studio documentation, n.d. URL `https://help.aliyun.com/zh/model-studio/qwen3-vl-plus`. Accessed September 26, 2026.

Andreas Aristidou, Joan Lasenby, Yiorgos Chrysanthou, and Ariel Shamir. Inverse kinematics techniques in computer graphics: A survey. *Computer Graphics Forum*, 37(6):35–58, 2018. doi: 10.1111/cgf.13310.

Alan Baddeley. *Working memory, thought, and action*. Oxford University Press, 2007. ISBN 9780198528012. doi: 10.1093/acprof:oso/9780198528012.001.0001.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie 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, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025. doi: 10.48550/arXiv.2511.21631.

Renée Baillargeon. Innate ideas revisited: For a principle of persistence in infants' physical reasoning. *Perspectives on Psychological Science*, 3(1):2–13, 2008. doi: 10.1111/j.1745-6916.2008.00056.x.

Anton Bakhtin, Laurens van der Maaten, Justin Johnson, Laura Gustafson, and Ross Girshick. PHYRE: A new benchmark for physical reasoning. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL `https://proceedings.neurips.cc/paper/2019/hash/4191ef5f6c1576762869ac49281130c9-Abstract.html`.

Peter W. Battaglia, Jessica B. Hamrick, and Joshua B. Tenenbaum. Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences*, 110(45):18327–18332, 2013. doi: 10.1073/pnas.1306572110.

Lukas Berglund, Meg Tong, Max Kaufmann, Mikita Balesni, Asa Cooper Stickland, Tomasz Korbak, and Owain Evans. The reversal curse: LLMs trained on "A is B" fail to learn "B is A". In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/5178b2f2d7c44aa390c0777dc77b3f0c-Abstract-Conference.html.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. PIQA: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 7432–7439, 2020. doi: 10.1609/aaai.v34i05.6239.

Florian Bordes, Quentin Garrido, Justine T. Kao, Adina Williams, Michael Rabbat, and Emmanuel Dupoux. IntPhys 2: Benchmarking intuitive physics understanding in complex synthetic environments. *arXiv preprint arXiv:2506.09849*, 2025. doi: 10.48550/arXiv.2506.09849.

Samuel R. Buss. Introduction to inverse kinematics with Jacobian transpose, pseudoinverse and damped least squares methods. Unpublished survey, Department of Mathematics, University of California, San Diego, 2009. URL https://mathweb.ucsd.edu/~sbuss/ResearchWeb/ikmethods/iksurvey.pdf. Revised October 7, 2009.

Erin Catto. Box2D: A 2D physics engine for games. Software, n.d. URL https://box2d.org/. Accessed September 26, 2026.

François Chollet. On the measure of intelligence. *arXiv preprint arXiv:1911.01547*, 2019. doi: 10.48550/arXiv.1911.01547.

Wei Chow, Jiageng Mao, Boyi Li, Daniel Seita, Vitor Campagnolo Guizilini, and Yue Wang. PhysBench: Benchmarking and enhancing vision-language models for physical world understanding. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/f38cb4cf9a5eaa92b3cfa481832719c6-Abstract-Conference.html.

Lynn A. Cooper. Mental rotation of random two-dimensional shapes. *Cognitive Psychology*, 7(1):20–43, 1975. doi: 10.1016/0010-0285(75)90003-1.

Aniruddha Deb, Neeva Oza, Sarthak Singla, Dinesh Khandelwal, Dinesh Garg, and Parag Singla. Fill in the blank: Exploring and enhancing LLM capabilities for backward reasoning in math word problems. *arXiv preprint arXiv:2310.01991*, 2023. doi: 10.48550/arXiv.2310.01991.

Ullrich K. H. Ecker, Stephan Lewandowsky, Klaus Oberauer, and Abby E. H. Chee. The components of working memory updating: An experimental decomposition and individual differences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 36(1):170–189, 2010. doi: 10.1037/a0017891.

Heinz Werner Engl, Martin Hanke, and Andreas Neubauer. *Regularization of Inverse Problems*, volume 375 of *Mathematics and Its Applications*. Kluwer Academic Publishers, Dordrecht, 1996. ISBN 9780792341574. URL https://link.springer.com/book/9780792341574.

Jörg Frohberg and Frank Binder. CRASS: A novel data set and benchmark to test counterfactual reasoning of large language models. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pp. 2126–2140. European Language Resources Association, 2022. URL https://aclanthology.org/2022.lrec-1.229/.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020. doi: 10.1038/s42256-020-00257-z.

Gemini Team. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. doi: 10.48550/arXiv.2312.11805.

GLM-5V-Turbo Team. GLM-5V-Turbo: Toward a native foundation model for multimodal agents. *arXiv preprint arXiv:2604.26752*, 2026. doi: 10.48550/arXiv.2604.26752.

Google DeepMind. Gemini 3.1 Pro model card. Model card, 2026. URL https://deepmind.google/models/model-cards/gemini-3-1-pro/.

Jessica B. Hamrick. Analogues of mental simulation and imagination in deep learning. *Current Opinion in Behavioral Sciences*, 29:8–16, 2019. doi: 10.1016/j.cobeha.2018.12.011.

Jessica B. Hamrick, Peter W. Battaglia, Thomas L. Griffiths, and Joshua B. Tenenbaum. Inferring mass in complex scenes by mental simulation. *Cognition*, 157:61–76, 2016. doi: 10.1016/j.cognition.2016.08.012.

Mary Hegarty. Mechanical reasoning by mental simulation. *Trends in Cognitive Sciences*, 8(6):280–285, 2004. doi: 10.1016/j.tics.2004.04.001.

Michael I. Jordan and David E. Rumelhart. Forward models: Supervised learning with a distal teacher. *Cognitive Science*, 16(3):307–354, 1992. doi: 10.1207/s15516709cog1603_1.

Mitsuo Kawato. Internal models for motor control and trajectory planning. *Current Opinion in Neurobiology*, 9(6):718–727, 1999. doi: 10.1016/S0959-4388(99)00028-8.

Stephen M. Kosslyn, William L. Thompson, and Giorgio Ganis. *The case for mental imagery*. Oxford University Press, 2006. ISBN 9780195179088. doi: 10.1093/acprof:oso/9780195179088.001.0001.

Zhengfei Kuang, Yunzhi Zhang, Hong-Xing Yu, Samir Agarwala, Shangzhe Wu, and Jiajun Wu. Stanford-ORB: A real-world 3D object inverse rendering benchmark. In *Advances in Neural Information Processing Systems*, volume 36, pp. 46938–46957, 2023. doi: 10.52202/075280-2033.

Tejas D. Kulkarni, William F. Whitney, Pushmeet Kohli, and Joshua B. Tenenbaum. Deep convolutional inverse graphics network. In *Advances in Neural Information Processing Systems*, volume 28, 2015. URL https://proceedings.neurips.cc/paper/2015/hash/ced556cd9f9c0c8315cfbe0744a3baf0-Abstract.html.

Brenden M. Lake, Tomer D. Ullman, Joshua B. Tenenbaum, and Samuel J. Gershman. Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40:e253, 2017. doi: 10.1017/S0140525X16001837.

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 *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/96810b6d4752abe7bfb91f234c51e9e6-Abstract-Conference.html.

Daixian Liu, Jiayi Kuang, Yinghui Li, Yangning Li, Di Yin, Haoyu Cao, Xing Sun, Ying Shen, Hai-Tao Zheng, Liang Lin, and Philip S. Yu. TangramPuzzle: Evaluating multimodal large language models with compositional spatial reasoning. *arXiv preprint arXiv:2601.16520*, 2026. doi: 10.48550/arXiv.2601.16520.

Fangyu Liu, Guy Emerson, and Nigel Collier. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 11:635–651, 2023. doi: 10.1162/tacl_a_00566.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. MathVista: Evaluating mathematical reasoning of foundation models in visual contexts. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/663bce02a0050c4a11f1eb8a7f1429d3-Abstract-Conference.html.

Fanqing Meng, Wenqi Shao, Lixin Luo, Yahong Wang, Yiran Chen, Quanfeng Lu, Yue Yang, Tianshuo Yang, Kaipeng Zhang, Yu Qiao, and Ping Luo. PhyBench: A physical commonsense benchmark for evaluating text-to-image models. *arXiv preprint arXiv:2406.11802*, 2024. doi: 10.48550/arXiv.2406.11802.

OpenAI. GPT-5 system card. System card, 2025. URL https://openai.com/index/gpt-5-system-card/.

OpenAI. GPT-5.4 Thinking system card. System card, 2026. URL https://openai.com/index/gpt-5-4-thinking-system-card/.

Deepak Pathak, Pulkit Agrawal, Alexei A. Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 2778–2787. PMLR, 2017. URL https://proceedings.mlr.press/v70/pathak17a.html.

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

Qwen Team. Qwen3.6-Plus: Towards real world agents. Official model release, 2026b. URL https://qwen.ai/blog?id=qwen3.6.

Santhosh Kumar Ramakrishnan, Erik Wijmans, Philipp Krähenbühl, and Vladlen Koltun. Does spatial cognition emerge in frontier models? In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/fb92196c6d2afe416528be19ba9963bc-Abstract-Conference.html.

Nikhila Ravi, Jeremy Reizenstein, David Novotny, Taylor Gordon, Wan-Yen Lo, Justin Johnson, and Georgia Gkioxari. Accelerating 3D deep learning with PyTorch3D. *arXiv preprint arXiv:2007.08501*, 2020. doi: 10.48550/arXiv.2007.08501.

Ronan Riochet, Mario Ynocente Castro, Mathieu Bernard, Adam Lerer, Rob Fergus, Véronique Izard, and Emmanuel Dupoux. IntPhys 2019: A benchmark for visual intuitive physics understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):5016–5025, 2022. doi: 10.1109/TPAMI.2021.3083839.

Adam N. Sanborn, Vikash K. Mansinghka, and Thomas L. Griffiths. Reconciling intuitive physics and Newtonian mechanics for colliding objects. *Psychological Review*, 120(2):411–437, 2013. doi: 10.1037/a0031912.

Mohammad Shahab Sepehri, Berk Tinaz, Zalan Fabian, and Mahdi Soltanolkotabi. Hyperphantasia: A benchmark for evaluating the mental visualization capabilities of multimodal LLMs. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-0132. Datasets and Benchmarks Track.

Roger N. Shepard and Jacqueline Metzler. Mental rotation of three-dimensional objects. *Science*, 171(3972):701–703, 1971. doi: 10.1126/science.171.3972.701.

Elizabeth S. Spelke. Principles of object perception. *Cognitive Science*, 14(1):29–56, 1990. doi: 10.1207/s15516709cog1401_3.

StepFun. Step-1o Turbo Vision. StepFun Open Platform documentation, n.d. URL https://platform.stepfun.com/docs/zh/guides/models/vision. Accessed September 26, 2026.

Andrew M. Stuart. Inverse problems: A Bayesian perspective. *Acta Numerica*, 19:451–559, 2010. doi: 10.1017/S0962492910000061.

Joshua B. Tenenbaum, Charles Kemp, Thomas L. Griffiths, and Noah D. Goodman. How to grow a mind: Statistics, structure, and abstraction. *Science*, 331(6022):1279–1285, 2011. doi: 10.1126/science.1192788.

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, Ziteng Wang, Rob Fergus, Yann LeCun, and Saining Xie. Cambrian-1: A fully open, vision-centric exploration of multimodal LLMs. In *Advances in Neural Information Processing Systems*, volume 37, 2024a.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal LLMs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9568–9578, 2024b. doi: 10.1109/CVPR52733.2024.00914.

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 *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/858de961874816f894285ec02e50c7fe-Abstract-Conference.html.

Zeyu Wang. CausalBench: A comprehensive benchmark for evaluating causal reasoning capabilities of large language models. In *Proceedings of the 10th SIGHAN Workshop on Chinese Language Processing (SIGHAN-10)*, pp. 143–151. Association for Computational Linguistics, 2024. URL https://aclanthology.org/2024.sighan-1.17/.

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*, volume 35, pp. 24824–24837, 2022. doi: 10.52202/068431-1800.

Daniel M. Wolpert and Mitsuo Kawato. Multiple paired forward and inverse models for motor control. *Neural Networks*, 11(7–8):1317–1329, 1998. doi: 10.1016/S0893-6080(98)00066-5.

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*, pp. 10632–10643, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Yang_Thinking_in_Space_How_Multimodal_Large_Language_Models_See_Remember_CVPR_2025_paper.html.

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*, 2020. URL https://openreview.net/forum?id=HkxYzANYDB.

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. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 57730–57754. PMLR, 2024. URL https://proceedings.mlr.press/v235/yu24o.html.

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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9556–9567, 2024. doi: 10.1109/CVPR52733.2024.00913.

ZeptoLab. Cut the Rope. Mobile game, 2010. URL https://www.zeptolab.com/games/cut-the-rope. Accessed September 26, 2026.

Yuyou Zhang, Radu Corcodel, Chiori Hori, Anoop Cherian, and Ding Zhao. SpinBench: Perspective and rotation as a lens on spatial reasoning in VLMs. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/724be4472168f31ba1c9ac630f15dec8-Abstract-Conference.html.

A FORMAL SPECIFICATION OF THE FOUR TASK FAMILIES

Section 3 defines inverse mental simulation over an environment state space $\mathcal{S}$, a primitive action space $\mathcal{A}$, a forward transition $f : \mathcal{A}^n \times \mathcal{S} \to \mathcal{S}$, and an observation model that determines what the solver is shown. Given the observation of an outcome s_T (and, where applicable, of s_0), the object of inference is the admissible inverse solution set

$$f^{-1}(s_T) = \{\mathbf{a} \in \mathcal{A}^n \mid f(\mathbf{a}, s_0) \in \mathcal{G}(s_T),\ \mathbf{a} \text{ satisfies } \mathcal{C}\},$$

where $\mathcal{C}$ collects the task constraints (joint limits, obstacle avoidance, and so on) and $\mathcal{G}(s_T) \subseteq \mathcal{S}$ is the goal set of Section 3: the singleton $\{s_T\}$ for VOI and a goal region for Dice Path, Lamp and Rope (Table 4). For the Dice Face task the unknown is an initial configuration rather than an action sequence. The same definition applies with the roles of $\mathbf{a}$ and s_0 exchanged (the roll sequence is observed and the initial face map is inferred). Table 4 instantiates these objects for each task variant.

Table 4: Formal specification of the five task variants. $\mathcal{S}$: environment state space. $\mathcal{A}$: primitive action space. "Observed": what the solver is shown (observation model). $f^{-1}(s_T)$: form of the admissible inverse solution set, i.e. what is graded. The A axis classifies how the simulated state changes (discrete transitions vs. continuous evolution), not the answer format.

Task	**State space $\mathcal{S}$**	**Actions $\mathcal{A}$**	**Observed**	**Admissible set $f^{-1}(s_T)$**	**Class**
VOI	binary canvas $\{0,1\}^{N\times M}$	place a shape mask at an integer anchor (optionally rotated). XOR into canvas	canvas size, target silhouette, shape library (may include distractors)	all placement sets whose XOR overlay equals the target exactly	O1+A1
Dice Face	cube orientation (24-element rotation group) $\times$ face labeling $\times$ grid cell	roll N/S/E/W about a bottom edge	roll sequence and the bottom-face imprint (symbol, rotation) after each roll	all initial face maps (symbol and orientation per face) consistent with the imprints. Undetermined faces answered as "?"	O2+A1
Dice Path	as Dice Face	roll N/S/E/W	complete initial face map. Target face constraint	all roll sequences that satisfy the target constraint	O2+A1
Lamp	joint-angle vector $\boldsymbol{\theta} \in \mathbb{R}^d$ (d joints)	set each joint angle within its limits	workspace, base, segment lengths, angle limits, target, light radius, obstacle geometry	all angle vectors whose beam center lies within the light radius of the target without obstacle collision	O1+A2
Rope	continuous 2D rigid-body state (candy, ropes, gameplay objects)	discrete gameplay commands (e.g. cut a rope, pop a bubble), optionally triggered by continuous conditions	map settings and the ordered list of gameplay objects in simulator coordinates	all ordered command programs under which the simulated candy reaches the mouth	O2+A2

State-transition type (A axis). A1 tasks advance the simulated state by discrete transitions among finitely many configurations: in VOI each placement toggles a finite set of grid cells of a binary canvas. In both Dice tasks each roll maps one of the 24 cube orientations to another by a 90° rotation, and the imprint sequence is a sequence of discrete events. A2 tasks require simulating a continuously evolving state: in the Articulated Lamp the arm pose and the beam position vary continuously with the joint angles, and in Cut the Rope the candy follows continuous rigid-body dynamics while ropes are cut and after. The classification concerns the simulated process rather than the answer format: Lamp answers are quantized to a fixed angle step (App. C) and Rope answers are sequences of discrete commands, yet in both the state that must be mentally simulated changes continuously, so

the solver must propagate continuous quantities and judge magnitudes (e.g. where the beam lands, how far the candy swings).

Why there is no third axis. Environment-state type and action-space type are conceptually separable, but in the current benchmark they are not independently crossed: VOI and both Dice tasks have discrete states and discrete actions, the Lamp has a continuous state and (quantized) continuous actions, and Cut the Rope has a continuous state and mixed actions. Introducing a third factorial axis would therefore imply coverage of combinations that SIMVERSE does not instantiate. The A axis is accordingly defined on the evolution of the simulated state, the property hypothesized to drive difficulty, and the per-task specification in Table 4 makes the action and answer encodings explicit.

Identifiability. Whether $f^{-1}(s_T)$ is a singleton is a property of the observation model and of $\mathcal{C}$, not of the direction of inference. SIMVERSE states the observation model of every task explicitly (Table 4) and, for Dice Face, controls it through the roll-sequence length (App. C). In Dice Face, non-identifiability is imposed by construction: faces that never touch the ground during the observed roll sequence leave no imprint and, unless other constraints pin them down, cannot be recovered. The task instructs the solver to report such faces with the sentinel symbol "?" and rotation 0, so that the graded answer is unique for every instance. In the remaining tasks $f^{-1}(s_T)$ is typically non-singleton (many XOR placements, many joint configurations, several command programs), and grading is by membership: any output that the simulator executes to the target outcome is accepted, rather than comparison with a single stored reference.

B GAME SPECIFICATIONS

B.1 GAME 1: VISUAL OVERLAY INTERFERENCE (VOI, O1+A1)

Classification. XOR is commutative ($A \oplus B = B \oplus A$) and associative, so the pixel-wise XOR of any collection of binary masks is independent of the order of application (O1). Placement positions are integer coordinates on an $N \times M$ canvas and the shape set is finite, so each placement changes the canvas by a discrete transition among finitely many states (A1).

State encoding. The canvas dimensions $N \times M$ are specified. The target silhouette is given as the set of filled cells (0-indexed, origin at the top left). Each shape is described by `shape_type`, `width`, `height` and `filled_cells`, a list of (dx, dy) offsets relative to the shape's anchor. The solver outputs a JSON array of placement objects.

Simulation engine. A pixel-wise XOR evaluator initializes an $N \times M$ binary canvas to zero and XORs each placed shape mask, translated to its anchor, into the canvas. The final canvas is compared cell by cell with the target for Match Rate. Pattern IoU and Dice Score are computed over the filled-cell sets.

B.2 GAME 2: NON-STANDARD DICE (O2+A1)

Classification. Rolls are elements of the rotation group of the cube (the octahedral rotation group, 24 elements), which is non-Abelian: rolling east then north differs in general from rolling north then east. The outcome of N rolls therefore depends on their order and not only on their multiset (O2). The state advances by discrete 90° transitions among the 24 cube orientations, so every intermediate state is one of finitely many configurations (A1).

State encoding. For Face Inference the observation is an ordered list of (roll direction, bottom-face imprint) pairs. Directions are drawn from $\{\mathrm{N}, \mathrm{S}, \mathrm{E}, \mathrm{W}\}$ in an explicit top-down world frame, and the die tips about the corresponding bottom edge into the adjacent grid cell. The solver outputs, for each of the six faces (TOP, BOTTOM, FRONT, BACK, LEFT, RIGHT), a pair `{patternId, rotation}`. For Path Inference the complete initial face map is provided in the same format together with the target constraint, and the solver outputs a JSON array of roll directions.

Orientation convention. `rotation` is an integer in $\{0, 90, 180, 270\}$ measured clockwise from the pattern's upright orientation when the face is viewed from outside the cube. Imprints on the

grid are shown top-down (viewed from above, not from underneath). A face that is not uniquely determined by the observations is reported as `patternId` = "?" with `rotation` = 0.

Ground-truth schema. Each released Dice instance carries exactly one authoritative ground-truth field, `answer`, which is the only field read by the grader. A schema-level consistency check run over all released instances enforces this single-field schema, so that the ground truth is unambiguous.

Simulation engine. A rotation-group algebra module represents each roll direction as an exact 3×3 integer rotation matrix and composes the matrices step by step to obtain the orientation after every roll.

B.3 Game 3: Articulated Lamp Positioning (O1+A2)

Classification. The lamp's configuration is a deterministic function of the joint-angle vector $\boldsymbol{\theta} \in \mathbb{R}^d$. The order in which joints are set does not affect the final configuration (O1). The arm pose and the beam position vary continuously with the joint angles, so the simulated state evolves continuously (A2). The task is a planar inverse-kinematics problem: forward kinematics is a closed-form composition of joint transforms, whereas recovering $\boldsymbol{\theta}$ for a target requires constraint satisfaction over a continuous manifold subject to joint limits and obstacles.

State encoding. The kinematic chain is given as an ordered array of link–joint pairs, each with `link_length` (px), `joint_type`, `rotation_axis` and `angle_limits` (degrees). The workspace, base pose, target coordinates, light radius and obstacle geometry are specified in the same frame. The solver outputs a JSON array of joint angles in degrees, restricted to integer multiples of the angle step.

Simulation engine. A forward-kinematics solver composes the joint transforms along the chain to obtain the lamp-head pose and beam direction, and computes the Euclidean distance between the beam center and the target. An instance is solved when this distance is at most the lamp's light radius and no link intersects an obstacle. Center Distance (CD) is reported for all instances whose output yields a valid distance (App. G).

B.4 Game 4: Cut the Rope (O2+A2)

Classification. The candy's motion after each action depends on its position and velocity at the moment the action fires, which are themselves determined by earlier actions and the intervening dynamics. The order of actions therefore changes all subsequent trajectories (O2). Although the solution is a sequence of discrete gameplay commands with optional continuous trigger conditions, the simulated state evolves under continuous Newtonian dynamics (A2).

State encoding and action format. Scene geometry is given in a continuous 2D map frame: ropes (anchor, length, attachment to the candy), the candy, stars and the mouth, together with the other gameplay objects, each with a numeric identifier. The solver outputs an ordered program of discrete gameplay commands—for example cutting a rope or popping a bubble—each optionally guarded by a condition on the continuous state, such as the candy entering a region or remaining at rest for a given duration (e.g. `cut_rope 1 when candy_near 320,260,45`). Several programs may solve a level. Any one is accepted.

Simulation engine and determinism. Scenes are simulated in Box2D (Catto, n.d.) with ropes modeled as joints and the candy as a dynamic circle body. Commands are executed event-driven: the controller checks each command's condition at every step and applies the action when it holds. The world advances at a fixed $\Delta t = 1/60\,\mathrm{s}$ with 10 velocity and 10 position solver iterations per step, from a fully specified initial state, so the same command program produces the same trajectory on every run. This determinism is what makes simulator grading reproducible. A star is collected when the candy's center enters its collection radius. Simulation terminates when the candy comes to rest, leaves the scene, or reaches the episode time limit.

C Instance Generation and Verification

VOI and both Dice tasks are procedurally generated from explicit difficulty parameters with validity filters. The Articulated Lamp and Cut the Rope instances are drawn from fixed level sets, each shipped with a reference solution. Every instance admits at least one solution verified in its simulator. The machine-readable generation configurations, including random seeds where a generator accepts one, are released with the code (App. M).

VOI. 600 instances in seven difficulty tiers. Difficulty increases with the grid size (6–10 cells per side), the number of required shapes (2–4) and distractor shapes (0–3), and the complexity of the shape pool drawn from 11 templates. The overlap and fill ratios of the target are constrained to tier-specific ranges. Candidates whose XOR target is fragmented, or whose overlap or fill ratio falls outside the tier range, are rejected.

Dice Face. 502 instances in five difficulty tiers defined by the length of the roll sequence (2 to 10 rolls, each drawn uniformly from N/S/E/W). Roll-sequence length controls how many faces are observed. Validity filters require a minimum number of distinct visited cells, required faces and observed faces that grows with the number of rolls.

Dice Path. 502 instances. Each target combines a target face with a rotation offset of 90°, 180° or 270°, and only targets reachable within 20 rolls are emitted.

Articulated Lamp. 610 levels in five difficulty tiers, each with one to three obstacles in a fixed 2D workspace and a reachable reference solution. A level is solved when the beam center lies within the light radius of the target and no link intersects an obstacle. Joint angles are quantized to a fixed step.

Cut the Rope. 272 levels drawn from a fixed level set rather than by random sampling, each with a reference command solution that clears the level in the simulator. Physics is simulated in Box2D with a fixed time step ($\Delta t = 1/60\,\mathrm{s}$) from a fully specified initial state. An episode ends when the candy comes to rest, leaves the scene, or reaches the episode time limit.

D Evaluation Protocol

Prompt structure. All prompts are generated from shared task-specific templates with a uniform three-part structure. **(1) Task description:** rules, domain vocabulary, objective and the definition of a valid solution. **(2) State encoding:** the instance state in a structured format with explicit coordinate conventions. **(3) Output schema:** a task-specific JSON schema. The final line of the response must contain the answer object after a fixed marker. In the primary condition the prompt is accompanied by a rendered scene image. Only the minimal formatting differences required by each provider's API interface distinguish the requests sent to different models.

Inference settings. All models are queried through their official APIs at temperature 0, single-shot, with no simulator feedback and no retries conditioned on correctness. All primary results use chain-of-thought prompting (Wei et al., 2022): the model reasons in free text before emitting the structured answer. The reasoning trace is not scored. For models with built-in extended reasoning, the platform's default reasoning mode is used. A response whose answer object cannot be parsed is re-requested by the harness. A response that remains unparseable is scored as a zero-credit parse failure. Only final responses are stored, so the reported parse-failure rate is a lower bound on raw formatting errors. Every reported rate uses all instances of the task as denominator. Parsing and simulator scoring are identical for every system.

Scope. SimVerse evaluates deployed multimodal systems under a standardized protocol. Vendor-side reasoning implementations cannot be inspected or controlled and are treated as part of the evaluated end-to-end system. The comparison does not attempt to isolate architectural differences between the underlying models. Results characterize single-shot deterministic decoding. Sampling-based inference (self-consistency, best-of-k) is not evaluated (Section 6).

D.1 Prompt Variants

The prompt-sensitivity analysis (App. I) compares two single-shot variants, both with text and image. **Original** is the full standardized template used for all primary results: task-specific instructions, validity constraints, solving advice, checking reminders and structured answer requirements. **Concise** is a direct-answer variant that preserves the complete task rules, structured state information, action constraints and required output schema, but removes auxiliary guidance, checking reminders and redundant explanatory text. The amount and organization of instructional scaffolding therefore change while the task information available to the model does not. For example, the original Dice Face template is organized into nine blocks (task, world model, visual legend, input fields, action vocabulary, constraints, solving advice, output schema, final instruction), whereas the concise template consists of the input fields, a compact list of rules and the output schema.

D.2 Text-Only Condition

In the text-only condition the rendered media are removed and replaced by a deterministic, task-specific structured encoding serialized from the same simulator state that produces the rendering. The prompt (original variant), model configuration, output schema, parsing and simulator scoring are unchanged. Before running the ablation, each encoding was verified at the schema level to contain the state needed to solve the task (Table 5). Ground-truth answers, reference solutions (including Dice Path solution roll sequences) and true face maps are excluded from every encoding, and forbidden-field leakage tests cover the exclusion. Because information is held constant, a drop in the text-only condition measures how effectively a model uses a rendering relative to an equivalent symbolic encoding, not a loss of task information.

Table 5: Schema-level content of the text-only encoding.

Task	Exposed in the text-only encoding
VOI	lossless row-band encoding of the target XOR mask (with the shape descriptions)
Dice Face	ordered observed path faces with `patternId`, rotation and flip flags
Dice Path	every cell of the unfolded grid (initial face map) plus the target face
Rope	map settings and the ordered gameplay-object list in simulator map coordinates
Lamp	workspace, base, joint angles, segment lengths, angle limits, target, light radius and complete obstacle geometry

E Human Reference

Participants and screening. The human reference consists of $n = 50$ graduate STEM participants (ages 22–28) screened with the extended 36-item Raven's Advanced Progressive Matrices (mean IQ 122.3 ± 7.6, range 112–136). The screening was administered at least 48 hours before the main session to avoid fatigue confounds. Candidates scoring below the 80th percentile were compensated for their time but excluded. All 50 reported participants passed the screen. The reference is treated as a cognitively bounded reference rather than a theoretical ceiling.

Procedure. Each participant completed a stratified random sample of 50 instances through a web interface identical to the model interface (same text and image), without simulator feedback. A printed instruction sheet described the interface and answer format, and each game was preceded by one unscored practice instance. Instances were randomized within each game block. Game order was fixed across participants. No time limit applied. Sessions were conducted remotely. Human responses are scored by submitting them to the same simulation pipeline used for models.

Uncertainty. Because each participant contributes several responses, human intervals are computed with participant-clustered resampling. The intra-class correlation is itself a modeling choice, so results are reported for $\mathrm{ICC} \in \{0.05, 0.10, 0.20\}$, with $\mathrm{ICC} = 0.10$ as the reference (Table 6). The human overall interval does not overlap the strongest model's interval (Gemini-3.1-Pro, 52.69 $[50.8, 54.6]$) under any of the three settings. Continuous human metrics: Lamp Center Distance 8.50 px (SD 12.0), 95% CI $[7.38, 9.65]$. Rope Mean Stars $2.05/3$ (SD 0.90), 95% CI $[1.95, 2.14]$.

Table 6: Human reference ($n = 50$) with 95% participant-clustered confidence intervals under three intra-class correlation assumptions.

Metric	**Point**	**ICC** $= 0.10$ **(ref.)**	**ICC** $= 0.05$	**ICC** $= 0.20$
VOI Match Rate	82.00	[78.0, 85.5]	[79.0, 84.6]	[76.4, 86.4]
Dice Face Pass Rate	18.00	[14.5, 21.9]	[15.4, 20.9]	[13.6, 23.5]
Dice Face PA	88.00	[84.6, 90.7]	[85.3, 90.2]	[83.2, 91.5]
Dice Face POA	60.00	[55.2, 64.6]	[56.4, 63.5]	[53.5, 66.1]
Dice Path PA	94.00	[91.4, 95.9]	[92.0, 95.4]	[90.3, 96.4]
Dice Path POA	76.00	[71.8, 79.8]	[72.8, 79.1]	[70.0, 81.0]
Lamp Success Rate	87.00	[83.6, 89.9]	[84.4, 89.2]	[82.0, 90.7]
Rope Pass Rate	65.00	[60.3, 69.5]	[61.5, 68.5]	[58.7, 70.9]
Overall (family macro)	75.5	[72.5, 78.6]	[73.3, 77.8]	[71.4, 79.5]

F Statistical Methodology and Aggregate Score

Resampling. All model analyses use the complete benchmark (VOI 600, Dice Face 502, Dice Path 502, Lamp 610, Rope 272, 2,486 instances) for each of the seven models and each of the three conditions (original, concise, text-only). Bootstrap analyses use 10,000 replicates with seed 20260803. *Task-stratified bootstrap:* instances are resampled with replacement within each task, per-task metrics are recomputed, and the aggregate is formed from the resampled tasks. *Paired bootstrap:* for comparisons of two conditions of the same model on the same instances, instances are aligned by identifier and resampled jointly, so that instance difficulty cancels in the difference. PA−POA gaps are paired within instance in the same way. *Wilson intervals* are used for single binomial proportions (Dice Face Pass Rate, App. H). *Exact McNemar tests* (two-sided) complement the paired bootstrap for per-cell binary outcomes in the ablations. *Rank stability* is measured by Spearman's ρ and Kendall's τ_b over the seven models. *Human–model differences* combine the model bootstrap distribution with a normal approximation to the human interval ($\mathrm{se} = (\mathrm{hi} - \mathrm{lo})/3.92$, ICC $= 0.10$). The excess of a model's PA−POA gap over the human gap is computed in the same way.

Aggregate score. **Overall score (family macro)**, the paper's single aggregate, is the mean over the four O×A families of VOI Match Rate, Dice (mean of Face POA and Path POA), Lamp Success Rate and Rope Pass Rate. The ablation cell tests (App. I and J) use binary per-instance task success: VOI Match, Dice Face Pass, Dice Path POA, Lamp Success and Rope Pass.

G PER-GAME RESULTS

Tables 7–11 report all metrics for the original condition (text and image). Models are listed in order of overall score. The best model value in each column is in bold.

Table 7: Game 1: Visual Overlay Interference (VOI, O1+A1).

Model	Match Rate	Pattern IoU	Dice Score
Human	*82.00*	*0.945*	*0.971*
Gemini-3.1-Pro	30.00	0.655	0.753
Qwen3.5-397B	23.33	0.691	0.785
Qwen3.6-Plus	**44.17**	**0.786**	**0.862**
GPT-5.4	15.17	0.632	0.746
GLM-5V-Turbo	15.67	0.545	0.677
Qwen3-VL-Plus	7.00	0.503	0.646
Step-1O-Turbo-Vision	0.00	0.303	0.433

Table 8: Game 2a: Dice Face Inference (O2+A1). POA is the primary Dice metric. Pass Rate (all six faces correct in symbol and orientation) is reported as the strictest end-to-end reference.

Model	Pass Rate	PA	POA
Human	*18.00*	*88.00*	*60.00*
Gemini-3.1-Pro	3.19	62.15	18.52
Qwen3.5-397B	0.00	84.50	30.08
Qwen3.6-Plus	**5.98**	**92.34**	**31.58**
GPT-5.4	0.00	33.35	7.20
GLM-5V-Turbo	0.00	46.22	12.50
Qwen3-VL-Plus	0.00	50.17	11.78
Step-1O-Turbo-Vision	0.00	25.70	3.77

Table 9: Game 2b: Dice Path Inference (O2+A1). PA: the roll sequence brings the target symbol to the target position. POA: additionally with the target orientation.

Model	PA	POA
Human	*94.00*	*76.00*
Gemini-3.1-Pro	**100.00**	**64.54**
Qwen3.5-397B	97.41	59.16
Qwen3.6-Plus	93.82	49.00
GPT-5.4	59.56	19.52
GLM-5V-Turbo	44.42	16.53
Qwen3-VL-Plus	43.63	15.14
Step-1O-Turbo-Vision	66.14	29.28

Remarks. (i) On VOI the continuous metrics separate models that the binary metric does not: Step-1O-Turbo-Vision has a Match Rate of 0.00 but a Pattern IoU of 0.303, i.e. it produces geometrically approximate overlays that fail the pixel-exact constraint. (ii) On the Lamp, Qwen3-VL-Plus has a higher Success Rate than Qwen3.6-Plus and GPT-5.4 (40.16 vs. 36.72 and 35.74) but a larger mean Center Distance (64.48 vs. 60.35 and 62.70). Threshold and distance metrics capture different aspects of continuous precision and are therefore both reported. (iii) Even in failed Rope instances models collect 0.48–1.41 stars on average, indicating partially correct trajectory reasoning.

Table 10: Game 3: Articulated Lamp Positioning (O1+A2). Center Distance (px) is averaged over instances whose output yields a distance (simulator-invalid joint-angle vectors yield none). Success Rate uses all 610 instances as denominator (no distance counts as failure).

Model	**Center Distance** (↓)	**Success Rate**
Human	*8.50*	*87.00*
Gemini-3.1-Pro	**21.73**	**87.38**
Qwen3.5-397B	23.75	80.49
Qwen3.6-Plus	60.35	36.72
GPT-5.4	62.70	35.74
GLM-5V-Turbo	86.87	26.89
Qwen3-VL-Plus	64.48	40.16
Step-1O-Turbo-Vision	159.58	14.43

Table 11: Game 4: Cut the Rope (O2+A2). Outputs that cannot be parsed are counted as 0 stars in Mean Stars and Stars (Fail).

Model	**Pass Rate**	**Mean Stars** (/3)	**Stars (Pass)**	**Stars (Fail)**
Human	*65.00*	*2.05*	*2.55*	*1.12*
Gemini-3.1-Pro	**51.84**	**2.00**	**2.54**	**1.41**
Qwen3.5-397B	21.69	0.83	1.73	0.58
Qwen3.6-Plus	28.31	1.25	2.01	0.95
GPT-5.4	45.22	1.24	1.61	0.94
GLM-5V-Turbo	24.26	1.12	2.12	0.80
Qwen3-VL-Plus	16.18	1.06	2.39	0.81
Step-1O-Turbo-Vision	13.60	0.61	1.41	0.48

H UNCERTAINTY ANALYSIS

Overall score and human–model gaps. Table 2 reports the Overall score with 95% task-stratified bootstrap intervals and the human–model gap on it. All seven gaps exclude zero, ranging from $+22.8$ $[+19.3, +26.4]$ (Gemini-3.1-Pro) to $+64.4$ $[+61.1, +67.6]$ (Step-1O-Turbo-Vision).

Family-level gaps. Of the 28 human–model family cells (Table 12), 27 exclude zero. The single exception is Gemini-3.1-Pro on the Lamp (-0.4 $[-4.4, +3.7]$), the only cell in which a model matches the human reference. Qwen3.5-397B approaches the human Lamp score but remains significantly below it ($+6.5$ $[+2.0, +11.0]$). Outside the Lamp, the narrowest gap is Gemini-3.1-Pro on Rope ($+13.2$ $[+5.9, +20.5]$).

Table 12: Family-level model scores and human$-$model gaps [95% CI]. Dice = mean of Face POA and Path POA. ^{0}: interval contains zero.

	VOI (H 82.00)		**Dice** (H 68.00)		**Lamp** (H 87.00)		**Rope** (H 65.00)	
Model	Score	Gap	Score	Gap	Score	Gap	Score	Gap
Gemini-3.1-Pro	30.00	+52.0 [+46.8, +57.3]	41.53	+26.5 [+22.5, +30.3]	87.38	-0.4 [-4.4, $+3.7$]^0	51.84	+13.2 [+5.9, +20.5]
Qwen3.5-397B	23.33	+58.7 [+53.5, +63.7]	44.62	+23.4 [+19.5, +27.3]	80.49	+6.5 [+2.0, +11.0]	21.69	+43.3 [+36.6, +50.0]
Qwen3.6-Plus	44.17	+37.8 [+32.3, +43.3]	40.29	+27.7 [+23.8, +31.6]	36.72	+50.3 [+45.2, +55.2]	28.31	+36.7 [+29.8, +43.6]
GPT-5.4	15.17	+66.8 [+62.1, +71.6]	13.36	+54.6 [+51.1, +58.2]	35.74	+51.3 [+46.3, +56.2]	45.22	+19.8 [+12.4, +27.1]
GLM-5V-Turbo	15.67	+66.3 [+61.6, +71.1]	14.52	+53.5 [+49.9, +56.9]	26.89	+60.1 [+55.4, +64.7]	24.26	+40.7 [+33.9, +47.4]
Qwen3-VL-Plus	7.00	+75.0 [+70.7, +79.2]	13.46	+54.5 [+51.0, +58.0]	40.16	+46.8 [+41.8, +51.8]	16.18	+48.8 [+42.5, +55.3]
Step-1O-Turbo-Vision	0.00	+82.0 [+78.3, +85.8]	16.53	+51.5 [+47.7, +55.1]	14.43	+72.6 [+68.3, +76.7]	13.60	+51.4 [+45.3, +57.3]

Spread across families. Across models the family scores span 44.2 pp on VOI (0.00–44.17), 31.3 pp on Dice (13.36–44.62), 73.0 pp on the Lamp (14.43–87.38) and 38.2 pp on Rope (13.60–51.84), i.e. more than 30 pp on every family. The human range across families is 22 pp (65.00–87.00). Mean model scores per family are 19.33 (O1+A1), 26.33 (O2+A1), 45.97 (O1+A2) and 28.73 (O2+A2), against human scores of 82.00, 68.00, 87.00 and 65.00. From O1+A2 to O2+A1 the mean model score falls by 19.6 pp (45.97 to 26.33) and the human score by 19.0 pp (87.00 to 68.00).

H.1 PA$-$POA GAPS

Table 13 reports the within-instance PA$-$POA gap for every model with paired 95% intervals, together with its excess over the human gap. On Dice Path, all seven models exhibit a gap significantly larger than the human gap of 18.00 pp (excess $+9.9$ to $+26.8$). On Dice Face, five of seven models exhibit a gap significantly larger than the human gap of 28.00 pp (excess $+5.7$ to $+32.8$): every model whose PA exceeds 50% (Qwen3.6-Plus, Qwen3.5-397B, Gemini-3.1-Pro, Qwen3-VL-Plus) and, narrowly, GLM-5V-Turbo (Face PA 46.22, $+5.7$ $[+0.7, +10.8]$). For the two models with the lowest Face PA, the excess is not distinguishable from zero (GPT-5.4, PA 33.35, -1.8 $[-6.9, +3.1]$) or is negative (Step-1O-Turbo-Vision, PA 25.70, -6.1 $[-10.8, -1.2]$). This is a floor effect: with little pattern accuracy left, there is little for orientation errors to erode, so a small gap reflects low PA rather than intact orientation tracking. The largest gap is Qwen3.6-Plus on Dice Face ($92.34 \rightarrow 31.58$, 60.76 pp).

H.2 CONTINUOUS PARTIAL CREDIT

Simulator-graded partial credit provides a graded view of each output (Table 14). Three observations follow. (i) Binary and continuous metrics can diverge sharply: on VOI, Qwen3.6-Plus reaches 0.786 partial credit against a 44.17% Match Rate, and Step-1O-Turbo-Vision 0.303 against 0.00%, so a zero binary score does not indicate unstructured output. (ii) On the Lamp, the Success Rate ranks Gemini-3.1-Pro above Qwen3.5-397B (87.38 vs. 80.49), whereas the continuous score ranks them in the opposite order (0.716 $[0.692, 0.740]$ vs. 0.770 $[0.738, 0.800]$) with overlapping intervals. The threshold and distance-based metrics are not interchangeable at the top of the Lamp ranking, and no single "best Lamp model" is identified. (iii) Dice Face partial credit stays at or below 0.457 for every model while the strict Pass Rate stays at or below 5.98%. Both metrics identify Dice Face as the task on which models are furthest from a complete solution.

Dice Face Pass Rate. Table 15 gives the strict Dice Face Pass Rate with Wilson intervals. Five of seven models score exactly 0 (Wilson upper bound 0.8%), so the metric does not discriminate among

Table 13: PA−POA gaps (pp) with paired 95% bootstrap CIs and excess over the human gap. Rows sorted by gap. ^{0}: interval contains zero.

Subject	PA	POA	PA−POA	95% CI	Excess over human [95% CI]
Dice Face Inference					
Human	88.00	60.00	28.00	[23.9, 32.6]	—
Qwen3.6-Plus	92.34	31.58	60.76	[58.3, 63.3]	+32.8 [+27.7, +37.7]
Qwen3.5-397B	84.50	30.08	54.42	[52.3, 56.5]	+26.4 [+21.6, +31.2]
Gemini-3.1-Pro	62.15	18.52	43.63	[40.6, 46.7]	+15.6 [+10.4, +20.9]
Qwen3-VL-Plus	50.17	11.78	38.39	[35.9, 41.0]	+10.4 [+5.4, +15.4]
GLM-5V-Turbo	46.22	12.50	33.72	[31.1, 36.3]	+5.7 [+0.7, +10.8]
GPT-5.4	33.35	7.20	26.16	[23.6, 28.7]	−1.8 [−6.9, +3.1]^0
Step-1O-Turbo-Vision	25.70	3.77	21.94	[19.9, 24.1]	−6.1 [−10.8, −1.2]
Dice Path Inference					
Human	94.00	76.00	18.00	[14.5, 21.9]	—
Qwen3.6-Plus	93.82	49.00	44.82	[40.4, 49.2]	+26.8 [+21.1, +32.4]
GPT-5.4	59.56	19.52	40.04	[35.9, 44.2]	+22.0 [+16.3, +27.7]
Qwen3.5-397B	97.41	59.16	38.25	[34.1, 42.6]	+20.2 [+14.7, +25.9]
Step-1O-Turbo-Vision	66.14	29.28	36.85	[32.7, 41.2]	+18.9 [+13.3, +24.5]
Gemini-3.1-Pro	100.00	64.54	35.46	[31.3, 39.6]	+17.5 [+11.9, +23.0]
Qwen3-VL-Plus	43.63	15.14	28.49	[24.7, 32.5]	+10.5 [+5.1, +16.0]
GLM-5V-Turbo	44.42	16.53	27.89	[23.9, 31.9]	+9.9 [+4.4, +15.3]

Table 14: Continuous partial credit with 95% bootstrap CIs (original condition). VOI: Pattern IoU. Dice Face: the grader's per-instance overall face accuracy (distinct from POA). Dice Path: target pattern correct. Rope: stars/3. Lamp: $\max(0,\, 1 - d/(2r))$ with d the beam-center distance and r the light radius.

Model	VOI	Dice Face	Dice Path	Rope	Lamp
Gemini-3.1-Pro	0.655 [0.632, 0.678]	0.263 [0.243, 0.283]	1.000 [1.000, 1.000]	0.665 [0.623, 0.708]	0.716 [0.692, 0.740]
Qwen3.5-397B	0.691 [0.669, 0.711]	0.457 [0.435, 0.478]	0.974 [0.958, 0.986]	0.276 [0.234, 0.319]	0.770 [0.738, 0.800]
Qwen3.6-Plus	0.786 [0.769, 0.803]	0.443 [0.420, 0.465]	0.938 [0.916, 0.958]	0.417 [0.368, 0.464]	0.394 [0.358, 0.428]
GPT-5.4	0.632 [0.613, 0.652]	0.160 [0.143, 0.177]	0.596 [0.552, 0.639]	0.414 [0.368, 0.462]	0.339 [0.310, 0.370]
GLM-5V-Turbo	0.545 [0.527, 0.564]	0.232 [0.213, 0.251]	0.444 [0.400, 0.488]	0.373 [0.326, 0.420]	0.259 [0.225, 0.291]
Qwen3-VL-Plus	0.503 [0.487, 0.519]	0.253 [0.237, 0.270]	0.436 [0.394, 0.480]	0.354 [0.309, 0.400]	0.369 [0.336, 0.403]
Step-1O-Turbo-Vision	0.303 [0.288, 0.318]	0.129 [0.115, 0.144]	0.661 [0.620, 0.703]	0.202 [0.162, 0.244]	0.126 [0.103, 0.152]

them. This is why POA is the primary Dice metric and Pass Rate serves as the strictest end-to-end reference. The contrast between humans (18.00%) and the best model (Qwen3.6-Plus, 5.98%) is nonetheless clear.

Table 15: Dice Face strict Pass Rate (%) with Wilson 95% CIs for models. The human interval is participant-clustered (ICC $= 0.10$, App. E).

Subject	**Pass Rate**	**Wilson 95% CI**
Human	18.00	[14.5, 21.9]
Qwen3.6-Plus	5.98	[4.2, 8.4]
Gemini-3.1-Pro	3.19	[2.0, 5.1]
Qwen3.5-397B	0.00	[0.0, 0.8]
GPT-5.4	0.00	[0.0, 0.8]
GLM-5V-Turbo	0.00	[0.0, 0.8]
Qwen3-VL-Plus	0.00	[0.0, 0.8]
Step-1O-Turbo-Vision	0.00	[0.0, 0.8]

I PROMPT SENSITIVITY

The concise prompt (App. D.1) lowers success for all seven models and raises it on no task. At the cell level (Table 17), 17 of 35 model×task differences exclude zero, and in every cell the bootstrap interval and the exact McNemar test ($\alpha = 0.05$) agree. Six of the non-significant cells are floor cells in which both conditions score 0.0%. Every model drops significantly on at least one task. The largest drop is Gemini-3.1-Pro on the Lamp (-17.0 $[-21.6, -12.5]$). Per-task rank correlations between the two conditions range from 0.883 to 0.991 (Table 16, for seven models a single adjacent transposition yields $\rho = 0.964$). On Dice Face the ranking is undefined under the concise prompt because all seven models score exactly zero. Absolute scores are therefore responsive to instructional scaffolding, most visibly for strong models on tasks where they score highest, while per-task between-model structure changes little.

Table 16: Rank stability of the seven models between the original condition and each ablation (per-task success rates). —: undefined (all models tied at zero).

	Original vs. concise		**Text+image vs. text-only**	
Scope	**Spearman ρ**	**Kendall τ_b**	**Spearman ρ**	**Kendall τ_b**
VOI	0.991	0.976	0.927	0.823
Dice Face	—	—	—	—
Dice Path	0.991	0.976	0.964	0.905
Rope	0.883	0.781	0.964	0.905
Lamp	0.893	0.810	0.964	0.905

Table 17: Prompt sensitivity per model and task (success rate, %, original vs. concise, both text+image). Δ: paired bootstrap 95% CI. p: exact McNemar. ✓: interval excludes zero.

Model	Task (n)	Orig.	Concise	Δ [95% CI]	p	Sig.
Gemini-3.1-Pro	VOI (600)	30.0	18.7	$-11.3\,[-15.3, -7.2]$	7.59×10^{-8}	✓
	Dice Face (502)	3.2	0.0	$-3.2\,[-4.8, -1.8]$	3.05×10^{-5}	✓
	Dice Path (502)	64.5	57.8	$-6.8\,[-11.8, -2.0]$	0.0102	✓
	Rope (272)	51.8	43.4	$-8.5\,[-15.4, -1.8]$	0.0206	✓
	Lamp (610)	87.4	70.3	$-17.0\,[-21.6, -12.5]$	1.86×10^{-12}	✓
Qwen3.5-397B	VOI (600)	23.3	16.8	$-6.5\,[-11.5, -1.5]$	0.0142	✓
	Dice Face (502)	0.0	0.0	$+0.0\,[+0.0, +0.0]$	1	
	Dice Path (502)	59.2	48.4	$-10.8\,[-15.7, -6.0]$	2.66×10^{-5}	✓
	Rope (272)	21.7	18.0	$-3.7\,[-11.0, +4.0]$	0.387	
	Lamp (610)	80.5	65.7	$-14.8\,[-19.7, -9.8]$	1.11×10^{-8}	✓
Qwen3.6-Plus	VOI (600)	44.2	31.0	$-13.2\,[-18.2, -8.0]$	7.72×10^{-7}	✓
	Dice Face (502)	6.0	0.0	$-6.0\,[-8.2, -4.0]$	1.86×10^{-9}	✓
	Dice Path (502)	49.0	48.4	$-0.6\,[-4.6, +3.6]$	0.85	
	Rope (272)	28.3	25.7	$-2.6\,[-8.1, +2.9]$	0.443	
	Lamp (610)	36.7	32.3	$-4.4\,[-8.0, -0.8]$	0.0217	✓
GPT-5.4	VOI (600)	15.2	3.0	$-12.2\,[-14.8, -9.7]$	2.12×10^{-22}	✓
	Dice Face (502)	0.0	0.0	$+0.0\,[+0.0, +0.0]$	1	
	Dice Path (502)	19.5	16.5	$-3.0\,[-7.2, +1.2]$	0.188	
	Rope (272)	45.2	43.4	$-1.8\,[-5.9, +2.2]$	0.473	
	Lamp (610)	35.7	30.7	$-5.1\,[-9.8, -0.2]$	0.0433	✓
GLM-5V-Turbo	VOI (600)	15.7	13.3	$-2.3\,[-4.8, +0.2]$	0.0925	
	Dice Face (502)	0.0	0.0	$+0.0\,[+0.0, +0.0]$	1	
	Dice Path (502)	16.5	15.7	$-0.8\,[-5.6, +4.2]$	0.814	
	Rope (272)	24.3	13.6	$-10.7\,[-16.2, -4.8]$	0.000422	✓
	Lamp (610)	26.9	22.3	$-4.6\,[-9.8, +0.8]$	0.0976	
Qwen3-VL-Plus	VOI (600)	7.0	0.0	$-7.0\,[-9.2, -5.0]$	4.55×10^{-13}	✓
	Dice Face (502)	0.0	0.0	$+0.0\,[+0.0, +0.0]$	1	
	Dice Path (502)	15.1	14.5	$-0.6\,[-5.4, +4.2]$	0.87	
	Rope (272)	16.2	16.2	$+0.0\,[-4.0, +4.0]$	1	
	Lamp (610)	40.2	30.2	$-10.0\,[-14.9, -5.2]$	9.6×10^{-5}	✓
Step-1O-Turbo-Vision	VOI (600)	0.0	0.0	$+0.0\,[+0.0, +0.0]$	1	
	Dice Face (502)	0.0	0.0	$+0.0\,[+0.0, +0.0]$	1	
	Dice Path (502)	29.3	25.1	$-4.2\,[-9.6, +1.2]$	0.16	
	Rope (272)	13.6	13.2	$-0.4\,[-5.1, +4.4]$	1	
	Lamp (610)	14.4	5.2	$-9.2\,[-11.5, -6.9]$	2.78×10^{-17}	✓

J MODALITY ABLATION

Removing the rendered image and supplying the schema-complete structured encoding instead (App. D.2) lowers success for all seven models and raises it on no task. Per-task rank correlations range from 0.927 to 0.964 (Table 16). At the cell level, 26 of 35 differences exclude zero (Table 18). Bootstrap and McNemar agree in every cell, and six of the nine non-significant cells are floor cells with both conditions at 0.0%. Every model drops significantly on at least three tasks. The largest drops occur on the Lamp (Qwen3.5-397B -23.6 $[-27.5, -19.7]$, Gemini-3.1-Pro -23.3 $[-28.0, -18.4]$) and on VOI (Qwen3.6-Plus -26.5 $[-31.5, -21.5]$), i.e. on continuous geometry, where numeric fields must be integrated into a spatial configuration, and on pixel-exact composition.

Action grounding. Removing the image also degrades the executability of outputs. The macro share of invalid or non-executable outputs rises from 1.4% (original) to 4.7% (text-only, 1.0% under the concise prompt), concentrated in Cut the Rope, where the per-model invalid share changes as follows (original $\rightarrow$ text-only): Gemini-3.1-Pro 0.0 $\rightarrow$ 19.5%, Qwen3.5-397B 0.0 $\rightarrow$ 22.8%, Qwen3.6-Plus 0.0 $\rightarrow$ 5.9%, GPT-5.4 0.0 $\rightarrow$ 0.0%, GLM-5V-Turbo 0.0 $\rightarrow$ 16.9%, Qwen3-VL-Plus 0.0 $\rightarrow$ 52.6%, Step-1O-Turbo-Vision 8.5 $\rightarrow$ 41.9%.

Interpretation. Because the text-only encoding is verified to be information-complete at the schema level, these drops do not reflect missing task information. They show that current models extract spatial structure more effectively from a rendering than from an equivalent symbolic encoding, so part of the difficulty measured by SIMVERSE lies in visual–spatial integration and action grounding, in addition to inverse reconstruction itself. The multimodal setting therefore remains the primary benchmark condition, and the text-only condition serves as a diagnostic that localizes where a given system's difficulty arises.

Table 18: Modality ablation per model and task (success rate, %, text+image vs. text-only, both with the original prompt). Δ: paired bootstrap 95% CI. p: exact McNemar. ✓: interval excludes zero.

Model	Task (n)	Text+img	Text-only	Δ [95% CI]	p	Sig.
Gemini-3.1-Pro	VOI (600)	30.0	11.8	$-18.2\,[-22.7, -13.7]$	5.28×10^{-15}	✓
	Dice Face (502)	3.2	0.0	$-3.2\,[-4.8, -1.8]$	3.05×10^{-5}	✓
	Dice Path (502)	64.5	48.2	$-16.3\,[-20.7, -12.0]$	1.09×10^{-12}	✓
	Rope (272)	51.8	35.3	$-16.5\,[-22.8, -10.3]$	5.2×10^{-7}	✓
	Lamp (610)	87.4	64.1	$-23.3\,[-28.0, -18.4]$	2.65×10^{-19}	✓
Qwen3.5-397B	VOI (600)	23.3	13.2	$-10.2\,[-15.0, -5.3]$	4.52×10^{-5}	✓
	Dice Face (502)	0.0	0.0	$+0.0\,[+0.0, +0.0]$	1	
	Dice Path (502)	59.2	48.4	$-10.8\,[-15.7, -5.8]$	4.65×10^{-5}	✓
	Rope (272)	21.7	17.3	$-4.4\,[-9.9, +1.1]$	0.155	
	Lamp (610)	80.5	56.9	$-23.6\,[-27.5, -19.7]$	2.58×10^{-29}	✓
Qwen3.6-Plus	VOI (600)	44.2	17.7	$-26.5\,[-31.5, -21.5]$	6.13×10^{-22}	✓
	Dice Face (502)	6.0	0.0	$-6.0\,[-8.2, -4.0]$	1.86×10^{-9}	✓
	Dice Path (502)	49.0	44.0	$-5.0\,[-11.4, +1.4]$	0.148	
	Rope (272)	28.3	19.9	$-8.5\,[-14.7, -2.2]$	0.0117	✓
	Lamp (610)	36.7	19.8	$-16.9\,[-20.5, -13.6]$	2.81×10^{-20}	✓
GPT-5.4	VOI (600)	15.2	0.0	$-15.2\,[-18.0, -12.3]$	8.08×10^{-28}	✓
	Dice Face (502)	0.0	0.0	$+0.0\,[+0.0, +0.0]$	1	
	Dice Path (502)	19.5	10.0	$-9.6\,[-14.1, -4.8]$	9.83×10^{-5}	✓
	Rope (272)	45.2	32.0	$-13.2\,[-19.5, -7.0]$	5.57×10^{-5}	✓
	Lamp (610)	35.7	22.1	$-13.6\,[-17.9, -9.3]$	2.04×10^{-9}	✓
GLM-5V-Turbo	VOI (600)	15.7	9.3	$-6.3\,[-9.8, -2.8]$	0.000534	✓
	Dice Face (502)	0.0	0.0	$+0.0\,[+0.0, +0.0]$	1	
	Dice Path (502)	16.5	6.2	$-10.4\,[-14.5, -6.2]$	1.18×10^{-6}	✓
	Rope (272)	24.3	14.3	$-9.9\,[-14.3, -5.5]$	4.19×10^{-5}	✓
	Lamp (610)	26.9	17.0	$-9.8\,[-14.6, -5.1]$	0.000106	✓
Qwen3-VL-Plus	VOI (600)	7.0	0.0	$-7.0\,[-9.2, -5.0]$	4.55×10^{-13}	✓
	Dice Face (502)	0.0	0.0	$+0.0\,[+0.0, +0.0]$	1	
	Dice Path (502)	15.1	4.0	$-11.2\,[-14.9, -7.4]$	7.32×10^{-9}	✓
	Rope (272)	16.2	13.2	$-2.9\,[-7.7, +1.8]$	0.291	
	Lamp (610)	40.2	24.4	$-15.7\,[-19.5, -12.1]$	2.63×10^{-16}	✓
Step-1O-Turbo-Vision	VOI (600)	0.0	0.0	$+0.0\,[+0.0, +0.0]$	1	
	Dice Face (502)	0.0	0.0	$+0.0\,[+0.0, +0.0]$	1	
	Dice Path (502)	29.3	19.3	$-10.0\,[-15.5, -4.4]$	0.000727	✓
	Rope (272)	13.6	5.1	$-8.5\,[-12.9, -4.0]$	0.000191	✓
	Lamp (610)	14.4	5.2	$-9.2\,[-11.5, -6.9]$	2.78×10^{-17}	✓

K FAILURE DECOMPOSITION

Categories. Every output is assigned to exactly one of five categories, checked in this order: *success* (the simulator reports the task solved), *parse failure* (no answer object conforming to the schema), *invalid* (parsed, but containing simulator-invalid or non-executable actions), *near miss* (executable, unsuccessful, with positive simulator partial credit as defined in Table 14), and *zero* (executable, unsuccessful, zero partial credit). Near misses are thus defined by simulator-based partial credit, not by textual similarity. Table 19 reports the macro average over the 35 model×task cells for each condition, and Table 20 the average over the seven models for each task.

Caveat on parse failures. Malformed responses are re-requested by the harness, and a response that remains unparseable is counted as a zero-credit parse failure. Because only final responses are stored (App. D), the parse-failure column is a lower bound on raw formatting-error rates and should not be read as a model's intrinsic parse-failure rate.

Table 19: Failure decomposition by condition (%, macro over the 35 model×task cells, human row: macro over tasks).

Condition	Parse fail	Invalid	Zero	Near miss	Success
Original (text+image)	0.1	1.4	27.6	44.7	26.3
Concise (text+image)	0.3	1.0	28.9	48.7	21.1
Original (text-only)	0.5	4.7	26.6	51.7	16.5
Human	0.0	0.8	3.8	29.8	65.6

Table 20: Failure decomposition by task and condition (%, mean over the seven models).

Task	Condition	Parse fail	Invalid	Zero	Near miss	Success
VOI	Original	0.0	0.0	0.9	79.8	19.3
	Concise	0.0	0.0	3.2	85.0	11.8
	Text-only	0.0	0.0	4.9	87.6	7.4
Dice Face	Original	0.3	0.0	27.8	70.5	1.3
	Concise	0.0	0.0	25.8	74.2	0.0
	Text-only	0.0	0.0	30.3	69.7	0.0
Dice Path	Original	0.0	0.0	27.9	36.0	36.2
	Concise	0.0	0.0	30.7	36.9	32.4
	Text-only	0.0	0.0	17.2	57.1	25.7
Rope	Original	0.0	1.2	38.0	32.0	28.7
	Concise	1.5	4.7	27.1	41.9	24.8
	Text-only	2.3	22.8	20.8	34.5	19.6
Lamp	Original	0.0	5.6	43.2	5.2	46.0
	Concise	0.0	0.4	57.6	5.3	36.7
	Text-only	0.0	0.6	60.0	9.5	30.0

Reading. In the primary condition, parse failures and invalid actions together account for 1.4% of outputs (macro), while near misses account for 44.7%: model outputs are overwhelmingly well-formed and executable but physically or combinatorially wrong, so the measured gaps are not formatting artifacts. Humans succeed far more often (65.6%) and produce few zero-credit outputs (3.8%). The decomposition also differs qualitatively by task. VOI failures are almost entirely near misses (79.8%), whereas Lamp failures are mostly zero-credit (43.2%), because the Lamp partial credit vanishes once the beam center is more than two light radii from the target. The Lamp is also the only task with a notable invalid share in the primary condition (5.6%), which is driven by Step-1O-Turbo-Vision (below).

K.1 Task-Specific Failure Semantics

Table 21 refines failures with task-specific semantics, pooled over the three conditions and seven models (so n is 21 times the number of instances). Percentages are of all outputs of the task.

Table 21: Task-specific failure semantics, pooled over three conditions and seven models (% of all outputs of the task).

Task	Outcome	%
VOI (n=12,600)	success	12.9
	shape deviation ($0 < \text{IoU} < 0.6$)	60.7
	shape close ($0.6 \leq \text{IoU} < 0.9$)	22.8
	no overlap ($\text{IoU} = 0$)	3.0
	near exact ($\text{IoU} \geq 0.9$, not matched)	0.7
Dice Face (n=10,542)	success	0.4
	pattern partly right, more pattern errors	39.8
	pattern fully right, orientation wrong	27.5
	pattern all wrong	20.8
	pattern partly right, more orientation errors	11.4
	parse error / no answer	0.1
Dice Path (n=10,542)	success	31.4
	pattern right, rotation wrong	43.3
	pattern wrong	25.3
Rope (n=5,712)	success	24.4
	simulation timeout (stars > 0)	23.0
	simulation timeout (0 stars)	19.5
	candy dropped / lost (stars > 0)	13.1
	invalid action output	9.6
	candy dropped / lost (0 stars)	9.2
	parse error	1.3
Lamp (n=12,810)	success	37.5
	missed target and hit obstacle	27.6
	distance far ($> 5\times$ radius)	16.5
	distance short ($\leq 5\times$ radius)	15.1
	invalid joint angles	2.2
	on target but hit obstacle	1.1

The dominant failure on Dice Path is a correct pattern with the wrong rotation (43.3% of all outputs), and on Dice Face 27.5% of all outputs identify every symbol correctly but get at least one orientation wrong. These instance-level categories isolate cumulative orientation tracking independently of the aggregate PA/POA metrics. On the Lamp the most frequent failure combines a missed target with an obstacle collision (27.6%), and on Rope most failures are timeouts or lost candy, frequently after some stars have been collected.

Model-specific observations. Step-1O-Turbo-Vision produces simulator-invalid joint-angle vectors on 193/610 = 31.6% of Lamp instances in the primary condition, compared with 33/610 for Qwen3.6-Plus, 15/610 for GLM-5V-Turbo and none for the other models. This failure mode accounts for much of its low Lamp score. The concentration of text-only invalid outputs in Cut the Rope is reported in App. J.

L Qualitative Failure Modes

Inspection of reasoning traces across all games reveals four recurring failure modes, aligned with the axis-specific predictions of Section 3. The traces are qualitative evidence. The quantitative counterparts are given in App. K.

FM-1 (State-tracking drift, O2). Models simulate the early steps of an operation sequence correctly but accumulate error in their state representation over successive steps. In the dice tasks, traces

typically show correct orientation tracking for the first few rolls and progressive divergence thereafter, particularly when the sequence contains direction reversals. Quantitatively, this mode corresponds to the "pattern right, rotation wrong" outcome that dominates Dice Path (43.3% of all outputs) and to PA−POA gaps of 22–61 pp (Face) and 28–45 pp (Path).

FM-2 (Numerical imprecision, A2). Models identify a qualitatively correct geometric approach but misjudge continuous parameter magnitudes. On the Lamp, weaker models often produce joint-angle vectors with a plausible beam direction but mean Center Distances of 60–160 px (60.35 to 159.58), against 8.50 px for humans. Even the most accurate model (Gemini-3.1-Pro, 21.73 px) remains well above the human value.

FM-3 (Trajectory anchoring, O2+A2). Specific to Cut the Rope: models preferentially act on the rope geometrically nearest the mouth, implicitly assuming a direct-drop trajectory and ignoring the constraining effect of the remaining ropes on subsequent candy motion. The resulting programs execute but leave the candy swinging, at rest or falling off the scene, which matches the prevalence of simulation timeouts (23.0% of all Rope outputs with and 19.5% without stars collected, pooled over the three conditions, Table 21) and of lost candy.

FM-4 (Discrete–continuous specialization). Model strengths are family-specific rather than uniform. Qwen3.6-Plus attains the highest VOI Match Rate (44.17%) and the highest Dice Face PA (92.34%) but reaches only 36.72% on the Lamp. Gemini-3.1-Pro attains the highest Lamp Success Rate (87.38%) and Rope Pass Rate (51.84%) but only 30.00% on VOI. Across models the family scores span more than 30 pp on every family and 73.0 pp on the Lamp, whereas the human range across families is 22 pp. No single aggregate therefore predicts a model's profile. One plausible explanation is heterogeneous, family-specific inductive biases acquired in training rather than a single transferable inverse-reasoning competence.

M Released Artifacts

The release comprises six parts. (i) The benchmark data for all five task variants (2,486 instances), each with a single authoritative ground-truth field and passing the schema-level consistency check (App. B). (ii) The four simulators (pixel-wise XOR evaluator, rotation-group algebra module, forward-kinematics solver, Box2D environment) and the parsing and scoring code. (iii) The machine-readable generation configurations for all five task variants, including seeds where a generator accepts one. (iv) The prompt templates for the original, concise and text-only conditions and the evaluation harness. (v) The statistics scripts (bootstrap seed 20260803, 10,000 replicates) that produce every interval in this paper. (vi) A documented, stratified subset of model outputs covering all seven models, all five task variants, and both successful and unsuccessful cases. The benchmark data (i) are available at `https://huggingface.co/datasets/SimVer-ano/simverse2026`. The code (ii)–(v) is provided as supplementary material. The output subset supports qualitative inspection and offline re-execution of simulator scoring on the released instances. It provides partial offline verification. It is not a complete reconstruction of all reported API runs, whose raw responses and execution records exceed the practical storage and maintenance constraints of the repository.