HUMANITY'S SIXTH SENSE: BENCHMARKING INTUITIVE VISUAL REASONING IN MULTIMODAL MODELS

Anonymous authors
Paper under double-blind review

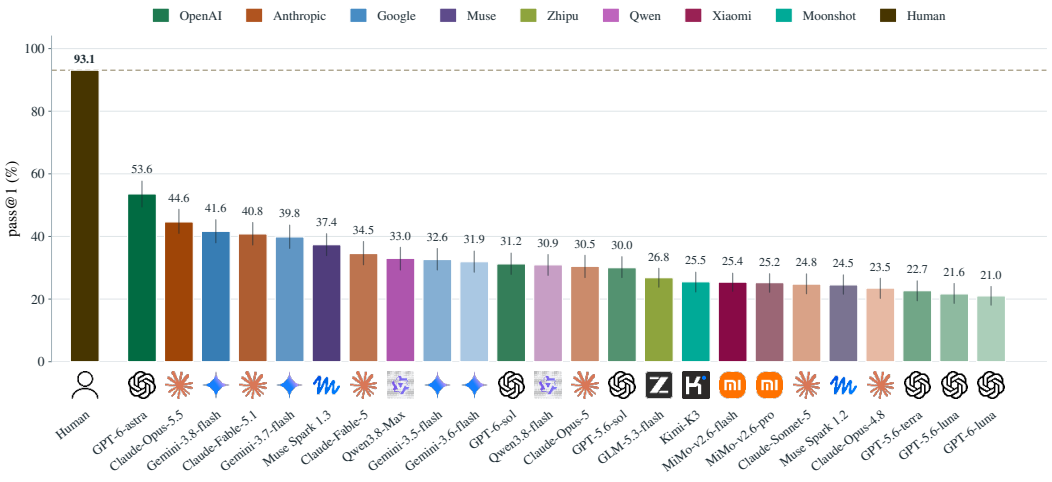


Figure 1: **HSS leaderboard.** Performance on HSS for 24 multimodal models. Every model scores far below people: the best, GPT-6-astra, reaches 53.6% where humans reach 93.1%.

ABSTRACT

Humans perceive far more in a scene than what is explicitly depicted: a single glance captures past causes and future trajectories; a quick peek determines if a vehicle can fit between two parked cars; a few seconds of video reveals who holds authority in a room; and a fleeting clip highlights subtle abstract patterns like unwritten rules or hidden labels. This capacity reflects a form of humanity's *sixth sense*: an intuitive reasoning mechanism that recovers implicit information beyond raw sensory perception. Crucially, this rapid, zero-shot visual intuition underpins everyday navigation and social interaction, making it a vital capability for Multimodal Large Language Models (MLLMs) deployed alongside people. Existing visual benchmarks, however, target either deliberate expert-level analysis in academic and mathematical domains or low-level perception, leaving the intuitive reasoning that people perform largely untested. To bridge this gap, we introduce **Humanity's Sixth Sense (HSS)**, a benchmark for intuitive visual reasoning. HSS spans diverse image and video inputs, organizes items under a structured taxonomy, and pairs each with human-written prompts probing the implicit temporal, spatial, social, and abstract structure that people infer at a glance. State-of-the-art models fall well short of human performance on these tasks: participants reach 93.1% accuracy, while the strongest model, GPT-6-astra, reaches only 53.6% even at maximum reasoning effort. Despite excelling in many complex tasks that require advanced perception and knowledge, current models still struggle significantly on these visual tasks tasks intuitive for humans. We further explore agentic frameworks that apply dynamic visual manipulation to HSS, which narrows but does not close the gap. HSS establishes intuitive visual reasoning as a measurable axis of progress and directs attention to a capability that scaling on current benchmarks has so far left behind.

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.

1 INTRODUCTION

A single glance at a visual scene reveals far more than what is explicitly rendered in pixels. Within a fraction of a second, we intuitively register who holds authority in a room, sense the subtle tension behind a polite smile, or notice a toddler edging toward the stairs—long before anything goes wrong. We call this capacity humanity's *sixth sense*: a rapid, pre-verbal form of visual inference that transforms circumstantial evidence into actionable conclusions. For AI agents deployed in homes, vehicles, and workplaces, this skill is not a luxury, but a core necessity: an agent must *perceive the world through a human lens* so that its predictions and behaviors naturally align with the people around it. This capacity relies on *visual intuition*, a concept deeply rooted in cognitive science. Literature characterizes intuition as an experience-based process that yields spontaneous judgments without conscious derivation (Kotler et al., 2025), functioning as skilled pattern recognition in environments with learnable regularities (Kahneman & Klein, 2009). In human vision, this is remarkably concrete: people categorize natural scenes within roughly 150ms (Thorpe et al., 1996) while effortlessly extracting implied physical dynamics, human motives, and social structures (Spelke & Kinzler, 2007; Battaglia et al., 2013; Baker et al., 2017). We synthesize these insights to define an inference as *intuitive* when it is fast, automatic, grounded in everyday experience rather than step-by-step derivation, and accessible even when its underlying rationale is difficult to articulate.

In text-only domains, benchmarks like SimpleBench (Philip & Hemang, 2024) have shown that frontier models lag significantly behind humans on commonsense spatio-temporal and social reasoning. In vision, the closest counterpart is BabyVision (Chen et al., 2026), which highlights model failures on basic visual tasks that young children solve effortlessly. However, while BabyVision tests whether models perceive what is *explicitly present*, HSS evaluates whether they infer what is *implicit*. Visual scenes carry vast unstated context such as revealing interpersonal relationships, latent intentions, physical affordances, and imminent consequences without ever stating them. Existing visual benchmarks fail to probe this dimension: they either treat images as prompts for specialized academic problem-solving (e.g., MMMU (Yue et al., 2024), HLE (Phan et al., 2025)) or isolate narrow perceptual capabilities like visual searching and object tracking (e.g., V* (Wu & Xie, 2024), Video-MME (Fu et al., 2024a), Video-MME-v2 (Fu et al., 2026)). Meanwhile, existing probes of physical and social intelligence (Zellers et al., 2019; Riochet et al., 2022; Kamath et al., 2023; Zadeh et al., 2019; Bear et al., 2021; Jin et al., 2024) remain fragmented across isolated domains, rely heavily on synthetic environments, or depend on multiple-choice formats that allow models to succeed through process-of-elimination heuristics rather than true visual inference.

To address this gap, we introduce **Humanity's Sixth Sense (HSS)**, a benchmark specifically targeting the domain of intuitive visual reasoning that operates between raw perception and deliberate, step-by-step logic. Concretely, an HSS task requires interpreting visual evidence alongside contextual, physical, social, and commonsense knowledge to deduce unstated implications, latent states, and likely outcomes. For instance, viewers must infer an agent's intended goal, recognize physical affordances, reconstruct prior events, or anticipate imminent consequences purely from visual cues and everyday experience. HSS comprises open-ended tasks spanning four core domains: *temporal & causal dynamics*, *physical & spatial logic*, *social understanding*, and *abstract contextual inference* (Figure 2), categorized into 11 subdomains that reflect human scene interpretation (Section 2.2). To avoid multiple-choice artifacts, models generate free-form responses evaluated against human-annotated atomic rubrics via an LLM-based judging pipeline.

Our experiments reveal a substantial capability gap (Figure 1). Human participants achieve 93.1% pass rate on HSS, whereas the strongest model evaluated, GPT-6-astra, attains only 53.6% even at maximum reasoning effort. Moreover, standard scaling strategies do not close this gap: increased test-time deliberation, higher reasoning effort, and agentic tool use each yield only modest improvements, and performance remains far from saturation. These results indicate that intuitive visual reasoning constitutes a distinct and largely unexplored dimension of visual intelligence, one that does not emerge simply from scaling test-time compute.

Our contributions are summarized as follows:

1. **A structured taxonomy of intuitive visual reasoning.** We formalize intuitive visual reasoning across four core domains and eleven subdomains, establishing a principled cognitive framework for evaluating implicit visual inference.

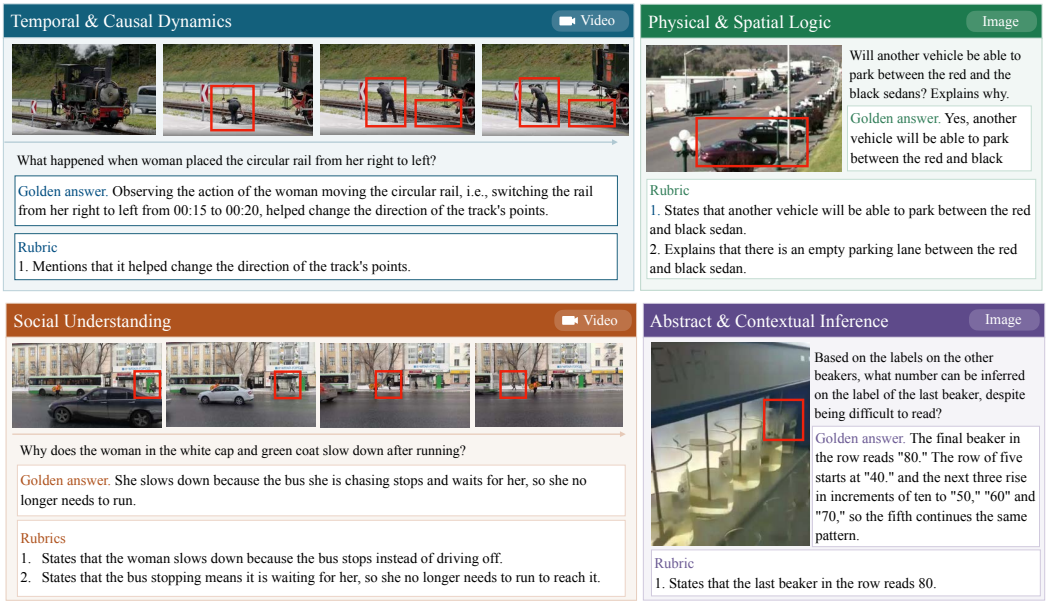


Figure 2: **HSS task examples across the four domains.** Each panel shows the input media (a single image or sampled video keyframes), human-written question, reference answer, and evaluation rubric. *Temporal & causal dynamics*: identifying the physical outcome of a manual track-switching action in video. *Physical & spatial logic*: judging whether a vehicle can fit between two parked sedans without explicit scale markers. *Social understanding*: inferring why a person stops running after a bus in a video clip. *Abstract & contextual inference*: deducing an occluded beaker label from an arithmetic progression of neighboring legible labels. Appendix **??** shows two further tasks for each domain.

2. **A rigorous, human-verified benchmark.** We introduce HSS, a benchmark specifically curated to assess implicit visual reasoning. To address the inherent subjectivity of human intuition, all instances are authored by human annotators and pass through a multi-stage independent review pipeline to guarantee strict human consensus and high data precision.
3. **Comprehensive empirical evaluation.** Evaluating state-of-the-art MLLMs under direct-inference reveals that current MLLMs achieve barely half of human performance (53.6% vs. 93.1%), establishing HSS as a formidable open challenge.
4. **Agentic paradigms and ablation analyses.** We examine the effects of agentic tool-use workflows and reasoning effort, demonstrating that intuitive visual reasoning represents a capability gap that remains resistant to both agentic augmentation and standard compute scaling.

2 The HSS Benchmark

2.1 Task Format

Each benchmark instance is a tuple (m, q, g, R): a visual media m (a static image or a video clip), a natural-language question q, a human-authored reference answer g, and a set of atomic rubrics $R = \{r_1, \ldots, r_k\}$, each independently checkable against g and **mandatory** for an answer to count as correct. Given m and q, a model returns a free-form answer a, and a judge marks every criterion $r_i \in R$ as met or unmet given q, a, g, and R; an instance counts as solved only when all k are met (Section C.4).

2.2 HSS Taxonomy Design

Humanity's intuitive visual reasoning involves the capacity to infer unobserved dynamics, physical constraints, mental states, and contextual rules from visual observation. The division we adopt follows the one cognitive science draws when it studies these inferences separately — core knowledge

of objects and space (Spelke & Kinzler, 2007), intuitive physics (Battaglia et al., 2013), and mentalizing (Baker et al., 2017) — with a fourth domain for the abstract regularities a scene implies. To render this latent capacity measurable, we formalize it into four core domains encompassing eleven subdomains (Table 4):

1. **Temporal & Causal Dynamics (Time):** Reconstructs the unobserved timeline connecting visual states, evaluating inferences about past events (*Retrodiction*), underlying physical drivers (*Mechanistic Causality*), and future trajectories (*Extrapolation*).
2. **Physical & Spatial Logic (Space):** Reasons over unshown content (*Hidden & Invisible*), functional interactions (*Affordance & Feasibility*), novel perspectives (*Spatial "Alien Viewpoint"*), and physical accessibility under spatial constraints (*Spatial Reachability*).
3. **Social Understanding (Mind):** Formalizes the invisible mental and normative substrate of human interaction, capturing intent and belief attribution (*Theory of Mind*) as well as implicit social expectations (*Social Role, Norm & Power Dynamics*).
4. **Abstract & Contextual Inference (Context):** Derives non-perceptual implications that a scene implies but never explicitly depicts, assessing unseen state transitions (*Change & Consequence*) and structural regularities (*Patterns & Pareidolia*).

Rather than treating visual intuition as an unanalyzable or monolithic capacity, the taxonomy described above provides a principled foundation that fulfills a dual purpose: first, as a generative blueprint that enforces rigorous, multi-faceted coverage during benchmark curation; and second, as a fine-grained diagnostic framework for model evaluation. By projecting performance across these eleven cognitive dimensions rather than relying solely on aggregate scores, the taxonomy disentangles distinct competency profiles and exposes the specific latent reasoning gaps of state-of-the-art multimodal models.

2.3 Task Design Principles

Because intuitive visual reasoning inherently risks subjective interpretation, task curation in HSS is strictly governed by three core design principles:

1. **Exclusion of Pure Perception:** Tasks must require latent inference beyond what is explicitly depicted in the visual media. We reject tasks that rely purely on fine-grained visual recognition, regardless of visual complexity. If a high-resolution crop of the target region makes the answer immediately obvious, the task evaluates visual acuity rather than intuitive reasoning and is discarded.
2. **Exclusion of Specialized Expertise:** Tasks must target intuitive reasoning accessible to any attentive adult. We strictly prohibit items requiring symbolic calculation, specialized academic knowledge, or niche professional training, ensuring the benchmark purely isolates commonsense visual intuition.
3. **Strict Human Consensus and Unambiguity:** Both questions and target answers must command unambiguous human consensus. Prompts are carefully structured to prevent subjective or divergent interpretations, and a task is admitted into the benchmark only if independent reviewers reach unanimous agreement on the ground truth.

2.4 Benchmark Construction Pipeline

Each task in HSS is authored by a trained human annotator through a three-step workflow: (1) selecting a source image or video clip from a curated media pool; (2) assigning the item to one of our four core domains and eleven subdomains; and (3) composing a natural-language prompt q, a gold reference answer g, and a structured rubric R decomposing g into fine-grained, independently checkable criteria. Authored tasks then enter a rigorous quality-control pipeline consisting of three independent review rounds. At each stage, a reviewer renders one of three verdicts: **Pass:** Advances the task to the subsequent review stage. **Edit:** Returns the task to the authoring phase for revision with specific feedback, after which it must re-enter the review pipeline from the beginning. **Drop:** Permanently removes the item if the prompt is ambiguous, the visual media is uninformative, or the target inference fails to command unanimous human consensus.

Only tasks that achieve unanimous approval across all three independent rounds are admitted into the benchmark, guaranteeing that every released rubric has been verified by at least three reviewers beyond its author. The review pipeline is highly lossy, and this attrition is the primary driver of dataset quality: out of 3,466 initially authored tasks, 1,164 survive the first review stage, 559 pass the second, and 522 pass the third, yielding an overall end-to-end acceptance rate of 15.1%. The most significant filter occurs at the first review gate, where approximately two-thirds of submitted tasks are filtered out, primarily for relying on pure visual perception or exhibiting subjective prompts and target answers. Appendix B (Figure 8) details the complete flowchart and attrition mechanics across each review stage. The final released benchmark comprises these 522 strictly vetted tasks across 4 domains and 11 subdomains (288 image-based and 234 video-based instances), carrying 753 fine-grained rubric criteria in total (Table 3).

3 Experiments

We evaluate 24 representative MLLMs on HSS. For models supporting native video inputs, we feed raw video files directly. Otherwise, we apply uniform frame sampling capped at 500 frames; for models with stricter input limits, we default to their maximum supported frame capacity. Static images are evaluated at native resolution, downscaling only when required by API or payload constraints. Model outputs are graded by Claude-Opus-5, serving as an automated LLM judge. Because the HSS data curation process retains only mandatory criteria and excludes non-essential ones, a response is marked as successful if and only if it satisfies all of the corresponding rubric criteria. Further details regarding the experimental setup, prompts, and evaluation protocol are in Appendix C.

3.1 Benchmark Results

Table 1 reports 24 models from eight vendors on HSS. Several key findings stand out:

HSS is far from saturated despite near-perfect human performance. The top-performing model, GPT-6-astra (maximum reasoning effort), achieves only a 53.6% pass rate, compared to 93.1% for human annotators. Most models perform below 40%, with a median pass rate of 30.7%. In addition, across all 24 models, the average reasoning token usage is 4,118 per task. Even on tasks intuitive to humans, models generate substantial reasoning chains before answering. While extended chain-of-thought reflects current reasoning alignment techniques, models frequently overthink without arriving at the correct answer for HSS.

Results breakdown across media type and domains. Video tasks prove more challenging than static images for 23 of the 24 models, with an average drop of 7.7 percentage points. This gap highlights ongoing struggles with long temporal horizons and keyframe localization. Social understanding is the weakest domain across vendors; it is the lowest-performing domain for 20 of the 24 models, averaging 23.9% accuracy compared to 33.5% across the remaining three domains.

Visual inputs are indispensable, preventing pure language-prior guessing. The blind control experiment confirms that HSS cannot be solved through language priors alone: stripping visual media drops the top model's score from 53.6% to 6.6%. The residue is guessing rather than knowledge: only 4 of 522 tasks pass blind on all three attempts, and the passes concentrate on two-way questions, where the model commits to one side and invents supporting detail (Appendix C.6).

Human performance. Twenty annotators answered the tasks under the same free-form conditions as the models and reached 93.1%, ahead of the best model by nearly 40 points. Each answer also carries a self-reported confidence. Annotators were unsure on only 63 of 522 tasks (12.1%) and scored 94.1% on the 459 they were sure about — hesitation tracks prompt defects rather than difficulty, which is what the admission criteria are meant to catch. Even the unsure answers scored 85.7%: the competence HSS measures stays available to people when their confidence is not.

3.2 Model Failure Analysis

To investigate why models fail, we categorize 8,573 failed first attempts from the main evaluation using an eleven-cause taxonomy grouped into three families. **Perception**: errors related to properties explicitly present in the visual signal (e.g., scale, geometry, facing direction, and spatial alignment). Extending beyond low-level object detection (e.g., recognizing a person or car), this involves evalu-

Table 1: **Main results.** Pass@1 (%) on HSS, over the 522 released tasks and three attempts each. A task counts as correct only when every rubric criterion is met; ± is the half-width of a 95% bootstrap interval over tasks. The columns re-cut the same tasks by media type and by domain, and *think-tok* / *ans-tok* are mean thinking and answer tokens per task. The best model value is **bold** and the second best is underlined, and the *blind* control re-runs the strongest model with the media removed.

Vendor	Model (Reasoning Effort)	pass@1	image	video	temporal	physical	social	abstract	think-tok	ans-tok
	Human baseline	**93.1**	94.1	91.9	97.0	94.9	89.1	89.2	—	—
OpenAI	GPT-6-astra (max)	**53.6** ±4.1	**55.6**	**51.1**	**58.2**	**53.6**	**46.1**	**55.6**	2,896	40
	GPT-6-sol (max)	31.2 ±3.4	31.8	30.5	38.1	28.6	22.4	36.3	3,298	30
	GPT-5.6-sol (max)	30.0 ±3.3	28.5	31.9	38.6	26.3	27.3	28.1	3,630	39
	GPT-5.6-terra (max)	22.7 ±3.2	24.1	20.9	28.4	21.0	19.1	21.9	5,794	45
	GPT-5.6-luna (max)	21.6 ±3.2	23.5	19.4	26.9	24.6	13.3	18.6	5,146	42
	GPT-6-luna (max)	21.0 ±3.0	23.3	18.2	26.9	24.1	9.4	20.6	3,993	39
Anthropic	Claude-Opus-5.5 (xhigh)	<u>44.6</u> ±3.8	<u>51.3</u>	<u>36.5</u>	<u>44.0</u>	<u>44.9</u>	34.8	**55.6**	1,496	294
	Claude-Fable-5.1 (xhigh)	40.8 ±3.6	44.7	36.0	43.3	41.1	34.2	44.1	3,478	360
	Claude-Fable-5 (max)	34.5 ±3.7	37.2	31.3	38.1	37.1	30.0	30.4	3,880	463
	Claude-Opus-5 (max)	30.5 ±3.5	32.5	27.9	31.3	35.0	22.4	30.1	2,888	445
	Claude-Sonnet-5 (max)	24.8 ±3.2	29.5	18.9	29.1	27.5	14.2	25.8	3,209	408
	Claude-Opus-4.8 (max)	23.5 ±3.2	27.2	18.9	27.1	26.5	14.2	23.5	2,416	408
Google	Gemini-3.8-flash (high)	41.6 ±3.7	47.2	34.8	43.3	37.5	<u>40.0</u>	<u>48.4</u>	10,407	158
	Gemini-3.7-flash (high)	39.8 ±3.7	45.4	33.0	40.5	40.9	35.8	41.5	1,858	118
	Gemini-3.5-flash (high)	32.6 ±3.4	35.1	29.6	35.6	31.6	31.8	31.4	2,456	184
	Gemini-3.6-flash (high)	31.9 ±3.4	32.5	31.2	37.1	29.7	32.4	28.4	3,166	178
Muse	Muse Spark 1.3 (max)	37.4 ±3.5	44.8	28.2	39.6	38.1	25.8	45.8	6,278	281
	Muse Spark 1.2 (xhigh)	24.5 ±3.1	27.5	20.8	23.6	24.2	21.2	29.7	3,640	412
Qwen	Qwen3.8-Max (max)	33.0 ±3.6	38.3	26.2	38.1	35.2	21.0	34.7	7,423	217
	Qwen3.8-flash (max)	30.9 ±3.3	37.2	22.9	37.9	29.2	18.6	37.1	7,260	200
Zhipu	GLM-5.3-flash (max)	26.8 ±3.0	30.2	22.6	31.3	26.3	18.2	31.0	4,484	247
Moonshot	Kimi-K3 (max)	25.5 ±3.1	29.1	21.2	33.1	27.4	15.5	23.2	4,949	98
Xiaomi	MiMo-v2.6-flash (high)	25.4 ±2.9	29.2	20.6	30.1	29.6	12.1	26.7	2,294	182
	MiMo-v2.6-pro (high)	25.2 ±2.9	30.4	18.6	28.3	29.5	14.5	25.7	2,499	192
Control	GPT-6-astra (max) [blind]	6.6 ±1.5	2.4	11.7	9.5	5.5	4.8	6.5	1,046	41

ating visible visual evidence for intuitive reasoning. **Latent Inference**: errors related to information not directly rendered in pixels (e.g., unobserved past events, hidden quantities behind obstacles, or underlying motives), requiring the model to bridge the gap between what is seen and what remains unobserved or implied. **Reasoning**: errors related to logical deduction, where the model fails to synthesize correct explicit and latent visual facts to draw valid conclusions.

As shown in Figure 3, failures stem predominantly from visual perception (53%) and latent inference (41%), whereas abstract reasoning accounts for only 5%. Overall, the two largest single failure modes are missing a decisive visual cue (21%) and misidentifying an object, person, or role (20%). Furthermore, failure patterns are highly systematic across architectures rather than model-specific: on tasks where five or more models fail, a median of 80% fail due to the exact same cause. Subdomains also exhibit distinct failure signatures: *affordance* and *spatial reachability* fail primarily on size and fit (62% and 56%), *extrapolation* on facing direction (42%), *hidden & invisible* on hidden quantities (44%), and *mechanistic causality* and *theory of mind* on missed cues (47% and 39%). *Alien viewpoint* is the only subdomain with a notable reasoning component, led by line-of-sight errors (23%). Extended failure analyses and case studies are provided in Appendices D and E.

3.3 Agentic Harnesses

To test whether active tool use aids performance on HSS, we evaluate base models inside vendor-provided agentic harnesses—including Claude Code and Codex—across 388 tasks. As shown in Table 2, the agentic setup improves accuracy across all configurations, though performance gains and compute overhead vary widely. Claude Code provides substantial accuracy increases (+16.5 to +20.5 points) at high compute costs (median 57–72 steps; > 42k thinking tokens per task), whereas Codex yields more modest improvements (+4.9 to +7.5 points) with minimal overhead (14–18

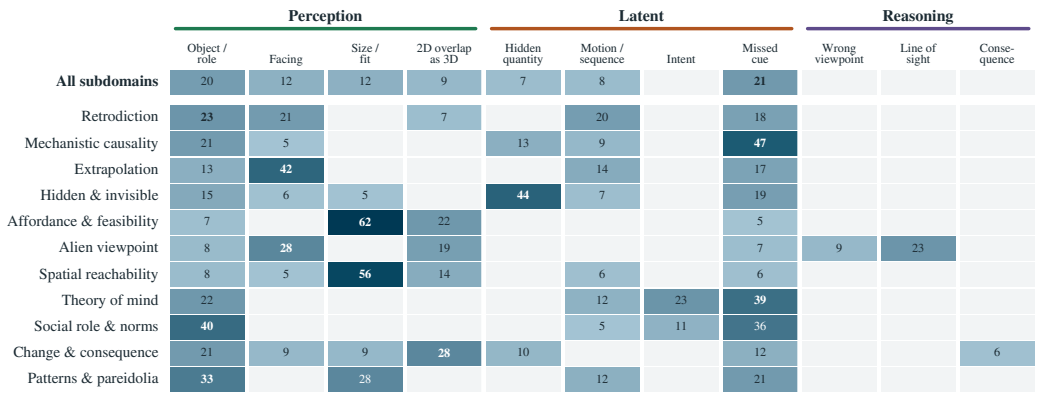


Figure 3: **What models get wrong, by subdomain.** Share (%) of each subdomain's failures attributed to each cause, over 8,573 labelled failures from 24 models on 522 tasks.

steps). However, even the top-performing setup (Codex with GPT-6-astra) reaches only 59.3%, demonstrating that HSS remains far from saturated even when models are equipped with agentic capabilities. Re-labeling agentic failures using our error taxonomy reveals distinct patterns in where agent loops succeed and fail:

- **Active re-inspection repairs perception, not spatial depth:** Total failures decrease from 944 to 759 (Figure 4), driven by a 19–37% reduction in perception and latent errors (e.g., recovering missed cues or misidentified objects). However, 2D-to-3D projection errors persist, demonstrating that zooming into a 2D image cannot recover lost 3D spatial depth.
- **Gains concentrate in evidence-rich subdomains:** Subdomains where visual evidence is explicitly present in-frame benefit most, such as *patterns & pareidolia* (+30.6 points) and *retrodiction* (+18.9 points; Table 8). Tasks whose answer turns on depth or on an unobserved outcome gain least (e.g., *spatial reachability* at +4.8 points).
- **Agentic failure dynamics:** Agent loops introduce unique failure modes, occasionally overanalyzing inputs and talking themselves out of correct initial readings. Additionally, failed agent runs take significantly more steps than successful ones (median 81 vs. 67 steps).

Appendix F reports all four harness/base pairs and provides four annotated trajectories; Appendix D details the complete failure analysis for agentic experiments.

Table 2: **Agentic harnesses against single-pass inference**. *Raw* is the base model called once, *Agent* the same model inside the harness; every Δ is agentic minus single-pass on those tasks, recut by media type and domain as in Table 1. *think-tok* is mean thinking tokens per task inside the harness and *steps* the median number of agent steps.

Harness	Base model	Overall pass@1 raw	agent	Δ	Δ by media image	video	Δ by domain temporal	physical	social	abstract	Cost think-tok	steps
Claude Code	Claude-Opus-5	30.8	51.3	**+20.5**	**+24.3**	+15.8	**+26.3**	**+15.1**	+18.2	**+25.6**	43,764	72
	Claude-Fable-5.1	41.2	57.7	+16.5	+15.4	**+17.9**	+16.8	+14.9	**+22.1**	+13.2	42,237	57
Codex	GPT-6-astra	**54.4**	**59.3**	+4.9	+4.5	+5.5	+6.7	−1.4	+9.5	+9.6	2,040	14
	GPT-5.6-sol	28.4	35.8	+7.5	+8.0	+6.8	+3.4	+6.5	+6.9	+15.5	6,541	18

3.4 Ablation Studies

Effects of Reasoning Effort. We evaluate the impact of reasoning effort on model performance across HSS subdomains (Figure 5). While increased effort generally improves overall performance, accuracy in certain subdomains degrades at higher reasoning levels. For example, in the retrodiction subdomain, GPT-6-astra exhibits a 14**-point drop** when scaling effort from *high* to *xhigh*. Consequently, peak accuracy is frequently achieved at intermediate effort levels rather than the maximum setting. These findings suggest that explicit deliberation is not the primary driver of HSS performance, which aligns with human cognitive observations where similar competencies operate intuitively prior to deliberate reasoning. Extended ablation results are provided in Appendix G.

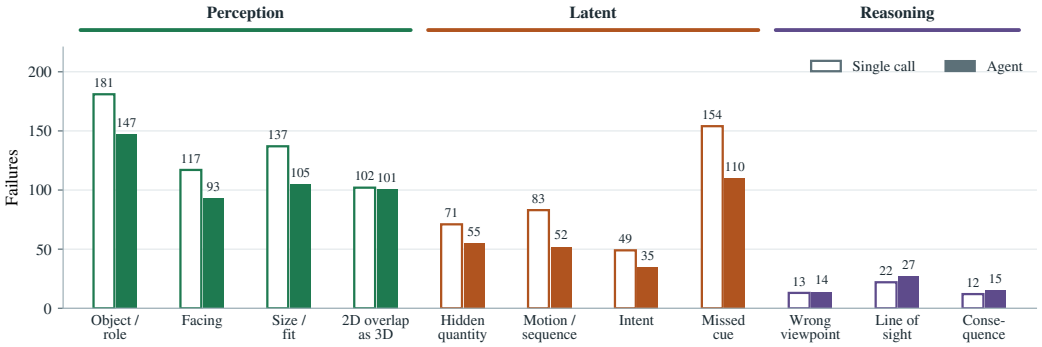


Figure 4: **What the agentic scaffold fixes.** Pooled over the four harness/base pairs of Table 9, first attempts. The agent cuts failures from 944 to 759, mostly on causes a closer look can fix: object/role (181 → 147), facing (117 → 93), size/fit (137 → 105), and missed-cue (154 → 110) errors. Depth errors (a 2D overlap read as 3D alignment) do not fall (102 → 101), and the small reasoning causes grow.

Effects of Video Frame Budget. Sweeping the frame budget from 10 to 500 across 234 video tasks shows that performance scales consistently with denser visual sampling. Across four evaluated models, every model achieves its lowest score at 10 frames, with GPT-6-astra gaining +17.9 points when frame budget increases from 10 to 500. Extended ablations and per-model results are provided in Appendix H.

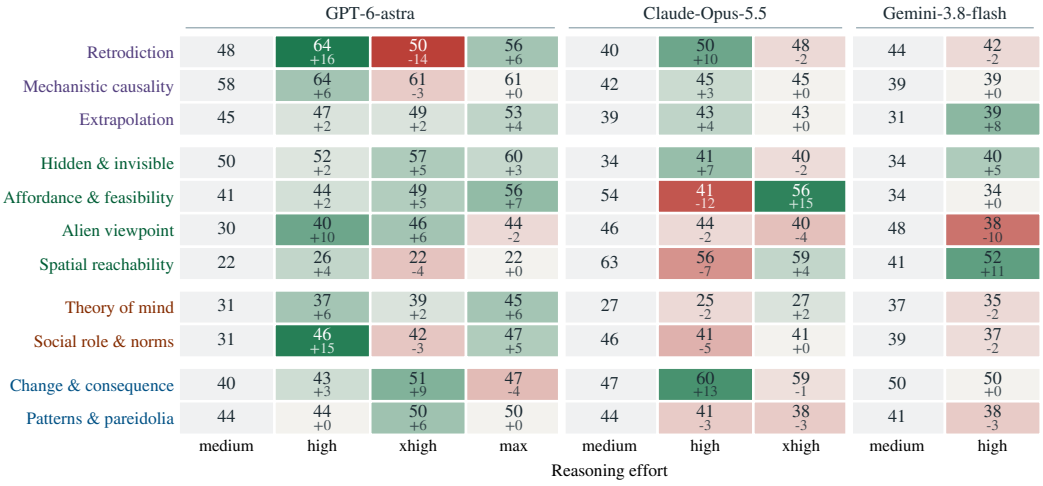


Figure 5: **Performance by subdomains under different reasoning effort** We increase the reasoning effort level from left to right. Green is a gain, red a loss.

Effects of different judge models. To check that the results do not depend on the choice of LLM judge, we re-graded five models with three judges from three vendors: Claude-Opus-5 (the primary judge), GPT-5.6-terra and Gemini-3.6-flash, across 2,620 items and 3,780 criteria. Rankings are preserved: judge-to-judge scores correlate at $r \geq 0.98$ (Figure 6a), and while GPT-5.6-terra is stricter by 2.4 points on average, it is stricter uniformly. Over 95% of items and criteria (Figure 6b), with $\kappa \geq 0.92$ on both units. Finally, there is no self-preference (Panickssery et al., 2024): Claude-Opus-5 grading itself shifts its own score by -0.44 points. Per-model breakdowns are in Appendix I.

4 Related Work

Most current multi-modal benchmarks fall into two families. One wraps problem-solving with visual inputs, such as MMMU, MathVista, Humanity's Last Exam (Yue et al., 2024; Lu et al., 2024; Phan et al., 2025; Zou et al., 2025), where the image supplies explicit givens that could in principle be

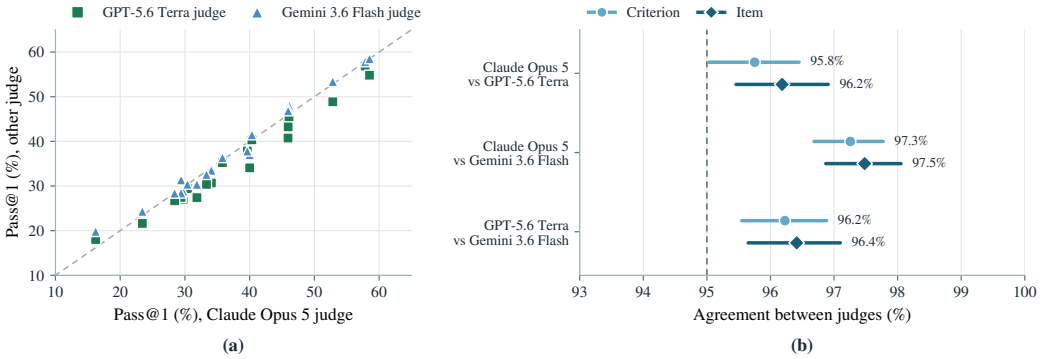


Figure 6: **Judge alignment across vendors.** (a) Pass@1 scores under alternative judges closely track the primary judge (Claude-Opus-5) along the identity line. (b) Pairwise agreement exceeds 95% at both the criterion ($n = 3{,}780$) and the item ($n = 2{,}620$) level for all three judge pairs.

transcribed into text. The other targets perceptual acuity, clutter and content tracking, such as V^*, MMBench, Video-MME, BLINK, Perception Test and MVBench (Wu & Xie, 2024; Liu et al., 2024; Fu et al., 2024a;b; Pătrăucean et al., 2023; Li et al., 2024), a line that runs back to VQA v2 (Goyal et al., 2017). HSS is orthogonal to both: its answers are neither transcribable nor findable by looking harder, because what the task asks for was never in the frame.

HSS is also related to a body of visual abduction benchmarks and to spatial, physical, social and abstract reasoning suites (Zellers et al., 2019; Hessel et al., 2022; Chen et al., 2024; Riochet et al., 2022; Xiao et al., 2021; Zadeh et al., 2019; Sap et al., 2019; Bear et al., 2021; Yi et al., 2020; Jin et al., 2024; Yang et al., 2025; Zhang et al., 2019; Kamath et al., 2023). Each covers part of this territory, typically one competence at a time and often on synthetic scenes; HSS spans physical, spatial, social and abstract inference over natural media under a single admission standard (Appendix A). Our agentic condition follows work showing that active inspection — cropping, zooming, re-sampling, sketching — can lift visual reasoning (Roberts et al., 2025; Guo et al., 2025; Hu et al., 2024), and our grading follows LLM-as-judge practice (Zheng et al., 2023) with an explicit audit of self-preference (Panickssery et al., 2024) (Section 3.4).

5 Limitations

We note three major limitations. **Visual channel only.** HSS evaluates what a model infers from what it can see. Video items are supplied without their audio track and without a transcript, so items whose resolution would depend on speech or off-screen sound are outside the benchmark's scope by construction. **Judge dependence.** Scores are mediated by an LLM judge, which is known to carry biases including self-preference (Panickssery et al., 2024). We report cross-judge agreement and an oracle condition in which human reference answers are scored correct, which bound the effect. **Contamination.** Media sourced from the public web may appear in training data. The questions are newly authored against that media rather than collected with it, which is what the benchmark tests, but prior exposure to an individual image or clip cannot be fully excluded.

6 Conclusion

Humans deduce far more from a visual scene than what is explicitly rendered in pixels: a capability measured directly by Humanity's Sixth Sense (HSS). To benchmark visual intuition objectively, HSS operationalizes it through a structured taxonomy and a rigorous task construction pipeline. Across our benchmarks, frontier models exhibit a stark performance gap compared to the human baseline, a discrepancy that persists despite test-time compute scaling and integration into agentic harnesses. Because 94% of model errors stem from visual perception and latent inference rather than deliberate reasoning, HSS highlights intuitive visual reasoning as a fundamental, measurable axis of multimodal intelligence where current scaling paradigms have so far yielded minimal progress.

AI USE STATEMENT

AI assistants were used strictly to polish draft and assist in developing evaluation and analysis code. All benchmark tasks, including media selection, questions, reference answers, and grading rubrics were authored and reviewed entirely by human annotators. LLMs appear in this work solely as model baselines of benchmark study and automated evaluators.

ETHICS STATEMENT

Media licensing. All media is drawn from three repositories that expose per-item, machine-readable license metadata: Wikimedia Commons, Openverse and the Internet Archive. In addition, only four licenses are admitted: Public Domain, CC0, CC-BY and CC-BY-SA. All four permit redistribution. Any item that is non-commercial, no-derivatives, or whose rights cannot be verified from source metadata is rejected at retrieval time. Creator attribution is carried per item and released with the benchmark, and the source files are mirrored so the frozen benchmark is not exposed to link rot. Furthermore, audio tracks are stripped from all video assets to ensure the benchmark evaluates purely visual reasoning, matching our task design.

People in the media. Many tasks depict identifiable people in public settings. They are used for research evaluation only, and a keyword screen is applied to video candidates to avoid material primarily featuring minors. Theory of mind tasks ask models to infer emotion, intent and belief from appearance and behaviour. Such inferences are culturally situated and can encode stereotypes. The multi-reviewer consensus requirement of Section 2.3 removes tasks on which readings diverge, but agreement among a demographically narrow annotator pool is not evidence of a universal inference. We therefore read theory of mind results as measurements of agreement with annotator judgement, not as ground truth about the people depicted.

Annotation. Annotators were recruited and compensated through a professional data-annotation provider at the provider's standard contracted rates.

REPRODUCIBILITY STATEMENT

We will publicly release our evaluation harness, model registry, and grading prompts. The registry specifies all controlled evaluation parameters, including API endpoints, reasoning effort levels, output token budgets, video processing pathways, and frame caps, which are detailed in Appendix C.1 across all 24 evaluated models and three judges. Where a provider exposes no explicit parameter controls, provider defaults apply. Prompt templates are reproduced verbatim in Appendix C.7. All media assets are content-addressed and locally cached to ensure every model receives byte-identical inputs, with each run logging dataset versioning, task media metadata, and per-criterion judge verdicts. We omit sampling temperature overrides to retain vendor defaults across heterogeneous API endpoints. Finally, all confirmatory metrics in this paper are evaluated on a frozen, versioned release containing 522 tasks (Table 3).

REFERENCES

Chris L. Baker, Julian Jara-Ettinger, Rebecca Saxe, and Joshua B. Tenenbaum. Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour*, 1: 0064, 2017.

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.

Daniel M. Bear, Elias Wang, Damian Mrowca, Felix J. Binder, Hsiao-Yu Fish Tung, R. T. Pramod, Cameron Holdaway, Sirui Tao, Kevin Smith, Fan-Yun Sun, Li Fei-Fei, Nancy Kanwisher, Joshua B. Tenenbaum, Daniel L. K. Yamins, and Judith E. Fan. Physion: Evaluating physical prediction from vision in humans and machines. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2021. arXiv:2106.08261.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brian Ichter, Danny Driess, Pete Florence, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. SpatialVLM: Endowing vision-language models with spatial rea-

soning capabilities. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Liang Chen, Weichu Xie, Yiyan Liang, Hongfeng He, Hans Zhao, Zhibo Yang, Zhiqi Huang, Haoning Wu, Haoyu Lu, Yiping Bao, et al. Babyvision: Visual reasoning beyond language. *arXiv preprint arXiv:2601.06521*, 2026.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Caifeng Shan, Ran He, and Xing Sun. Video-MME: The first-ever comprehensive evaluation benchmark of multi-modal LLMs in video analysis. *arXiv preprint arXiv:2405.21075*, 2024a.

Chaoyou Fu, Haozhi Yuan, Yuhao Dong, Yi-Fan Zhang, Yunhang Shen, Xiaoxing Hu, Xueying Li, Jinsen Su, Chengwu Long, Xiaoyao Xie, et al. Video-mme-v2: Towards the next stage in benchmarks for comprehensive video understanding. *arXiv preprint arXiv:2604.05015*, 2026.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A. Smith, Wei-Chiu Ma, and Ranjay Krishna. BLINK: Multimodal large language models can see but not perceive. In *European Conference on Computer Vision (ECCV)*, 2024b. arXiv:2404.12390.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. arXiv:1612.00837.

Xingang Guo, Utkarsh Tyagi, Advait Gosai, Paula Vergara, Jayeon Park, Ernesto Gabriel Hernández Montoya, Chen Bo Calvin Zhang, Bin Hu, Yunzhong He, Bing Liu, and Sharma Srinivasa Rakshith. Beyond seeing: Evaluating multimodal LLMs on tool-enabled image perception, transformation, and reasoning. *arXiv preprint arXiv:2510.12712*, 2025.

Jack Hessel, Jena D. Hwang, Jae Sung Park, Rowan Zellers, Chandra Bhagavatula, Anna Rohrbach, Kate Saenko, and Yejin Choi. The abduction of Sherlock Holmes: A dataset for visual abductive reasoning. In *European Conference on Computer Vision (ECCV)*, 2022.

Yushi Hu, Weijia Shi, Xingyu Fu, Dan Roth, Mari Ostendorf, Luke Zettlemoyer, Noah A. Smith, and Ranjay Krishna. Visual sketchpad: Sketching as a visual chain of thought for multimodal language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. arXiv:2406.09403.

Chuanyang Jin, Yutong Wu, Jing Cao, Jiannan Xiang, Yen-Ling Kuo, Zhiting Hu, Tomer Ullman, Antonio Torralba, Joshua B. Tenenbaum, and Tianmin Shu. MMToM-QA: Multimodal theory of mind question answering. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024. arXiv:2401.08743.

Daniel Kahneman and Gary Klein. Conditions for intuitive expertise: A failure to disagree. *American Psychologist*, 64(6):515–526, 2009.

Amita Kamath, Jack Hessel, and Kai-Wei Chang. What's "up" with vision-language models? investigating their struggle with spatial reasoning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023. arXiv:2310.19785.

Steven Kotler, Michael Mannino, Karl Friston, György Buzsáki, J. A. Scott Kelso, and Guillaume Dumas. Pathfinding: a neurodynamical account of intuition. *Communications Biology*, 8:1214, 2025. doi:10.1038/s42003-025-08612-9.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, Limin Wang, and Yu Qiao. MVBench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. arXiv:2311.17005.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. MMBench: Is your multi-modal model an all-around player? In *European Conference on Computer Vision (ECCV)*, 2024. arXiv:2307.06281.

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 (ICLR)*, 2024.

Arjun Panickssery, Samuel R. Bowman, and Shi Feng. LLM evaluators recognize and favor their own generations. *arXiv preprint arXiv:2404.13076*, 2024.

Jae Sung Park, Chandra Bhagavatula, Roozbeh Mottaghi, Ali Farhadi, and Yejin Choi. VisualCOMET: Reasoning about the dynamic context of a still image. In *European Conference on Computer Vision (ECCV)*, 2020.

Viorica Pătrăucean, Lucas Smaira, Ankush Gupta, Adrià Recasens, Larisa Markeeva, et al. Perception test: A diagnostic benchmark for multimodal video models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, et al. Humanity's last exam. *arXiv preprint arXiv:2501.14249*, 2025.

Philip and Hemang. SimpleBench: The text benchmark in which unspecialized human performance exceeds that of current frontier models. https://simple-bench.com, 2024. Accessed 2026-09-19.

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.

Jonathan Roberts, Mohammad Reza Taesiri, Ansh Sharma, Akash Gupta, Samuel Roberts, et al. ZeroBench: An impossible visual benchmark for contemporary large multimodal models. *arXiv preprint arXiv:2502.09696*, 2025.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. SocialIQa: Commonsense reasoning about social interactions. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2019. arXiv:1904.09728.

Elizabeth S. Spelke and Katherine D. Kinzler. Core knowledge. *Developmental Science*, 10(1): 89–96, 2007.

Simon Thorpe, Denis Fize, and Catherine Marlot. Speed of processing in the human visual system. *Nature*, 381:520–522, 1996.

Penghao Wu and Saining Xie. V*: Guided visual search as a core mechanism in multimodal LLMs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. arXiv:2312.14135.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. NExT-QA: Next phase of question-answering to explaining temporal actions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

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 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B. Tenenbaum. CLEVRER: Collision events for video representation and reasoning. In *International Conference on Learning Representations (ICLR)*, 2020. arXiv:1910.01442.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, et al. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. arXiv:2311.16502.

Amir Zadeh, Michael Chan, Paul Pu Liang, Edmund Tong, and Louis-Philippe Morency. Social-IQ: A question answering benchmark for artificial social intelligence. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Chi Zhang, Feng Gao, Baoxiong Jia, Yixin Zhu, and Song-Chun Zhu. RAVEN: A dataset for relational and analogical visual reasoning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2023. arXiv:2306.05685.

Chengke Zou, Xingang Guo, Rui Yang, Junyu Zhang, Bin Hu, and Huan Zhang. DynaMath: A dynamic visual benchmark for evaluating mathematical reasoning robustness of vision language models. In *International Conference on Learning Representations (ICLR)*, 2025.

A Extended Related Work

This appendix expands Section 4 with the comparisons that place HSS against individual benchmarks rather than against families of them.

Academic problem-solving in a visual wrapper. MMMU, MathVista and Humanity's Last Exam (Yue et al., 2024; Lu et al., 2024; Phan et al., 2025) test high-level symbolic or domain-specific reasoning over images: a diagram, a plot or a figure supplies givens, and the difficulty lives in the chain of reasoning applied to them. The distinguishing property is that a sufficiently complete verbal description of the image would preserve the task. HSS items do not have this property, and the blind control (Table 1) is the measurement that establishes it: with the media removed, the strongest model drops from 53.6% to 6.6%.

Perception-centred benchmarks. V^*, MMBench and Video-MME (Wu & Xie, 2024; Liu et al., 2024; Fu et al., 2024a) test visual acuity, robustness to clutter, and tracking of content across duration. Here the answer *is* in the frame, and the question is whether the model can find it. HSS inverts the admission criterion (Section 2.3): an item is rejected precisely when a careful enough look would settle it. The two benchmark families are therefore complementary — a model can be excellent at finding what is depicted and still fail here, which is what Section 3.2 observes.

Visual abduction and commonsense. VCR, VisualCOMET and Sherlock (Zellers et al., 2019; Park et al., 2020; Hessel et al., 2022) are the closest ancestors: each asks the model to go beyond the depicted scene to intentions, prior events or likely continuations. They concentrate on social and narrative context and grade plausible completions, often against multiple-choice alternatives. HSS covers physical, spatial, social and abstract inference under one taxonomy, and admits only items whose answer reviewers agree is forced rather than merely plausible, which is what makes open-ended rubric grading possible.

Spatial, physical, and social reasoning. SpatialVLM, IntPhys, Physion and CLEVRER (Chen et al., 2024; Riochet et al., 2022; Bear et al., 2021; Yi et al., 2020) evaluate relation-naming, physical plausibility and collision prediction, largely over synthetic or simulated scenes; NExT-QA, Social-IQ, SocialIQa and MMToM-QA (Xiao et al., 2021; Zadeh et al., 2019; Sap et al., 2019; Jin et al., 2024) evaluate causal, social and mental-state understanding of observed events; VSI-Bench and What's Up (Yang et al., 2025; Kamath et al., 2023) probe spatial memory and relational language; and RAVEN (Zhang et al., 2019) isolates abstract visual analogy. HSS extends both directions to natural media and to inferences about what is *not* observed: an alien viewpoint the camera never occupies, a rope whose tension is visible only in its sag, an event that the clip sets up but never shows.

Tool-augmented inspection. ZeroBench and VisualToolBench (Roberts et al., 2025; Guo et al., 2025) show that letting a model crop, zoom and re-sample improves visual reasoning, and motivate our agentic condition. Section 3.3 uses that lever diagnostically rather than as a method contribution: if HSS failures were failures of looking, active inspection should repair them, and the extent to which it does not (Figure 4) is evidence about what the benchmark measures.

Automated grading and its validity. Open-ended grading at this scale requires an LLM judge (Zheng et al., 2023), which imports known failure modes, self-preference (Panickssery et al., 2024) foremost among them. We audit rather than assume: three judges from three vendors, agreement and κ per pair, and a self-grading contrast (Section 3.4, Appendix I).

B Benchmark Details

This appendix documents the released benchmark: what it contains (Appendix B.1), how each task was built and reviewed (Appendix B.2), and how the taxonomy was arrived at (Appendix B.3).

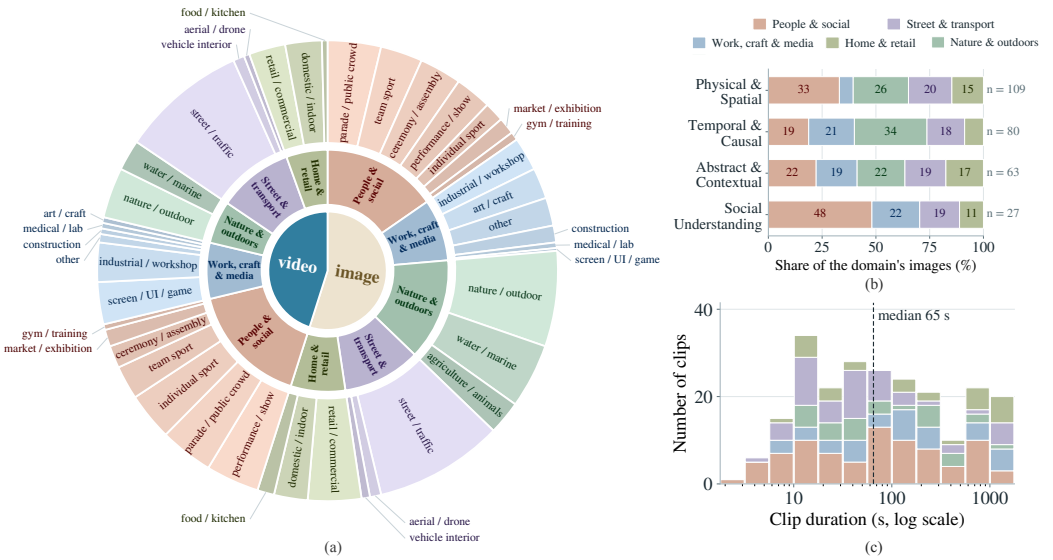


Figure 7: **Media statistics in HSS.** (a) Hierarchical breakdown of image and video assets across broad visual categories (*People & social*, *Work, craft & media*, *Nature & outdoors*, *Street & transport*, and *Home & retail*) and fine-grained sub-scenes. (b) Distribution of visual scene types across the four core reasoning domains, showing diverse environment coverage per domain. (c) Histogram of video clip durations (log scale), spanning a few seconds to 28 minutes; summary statistics are in Table 3.

B.1 Composition and statistics

Four properties of the composition in Table 3 matter for the analysis; Figure 7 shows the same corpus by scene type and clip length. First, no domain dominates the benchmark: the largest, *physical & spatial logic*, holds $176/522$ tasks and the smallest, *abstract & contextual inference*, holds 102, so every domain-level cell is large enough to carry a claim, and the smallest subdomain still holds 27. Second, the media split is near-even (288 images against 234 video clips, 17.6 hours of video in total), but it is deliberately uneven *within* domains, because the domains do not need the same evidence: theory of mind is almost entirely video ($46/51$), since an intention reveals itself over time, whereas the spatial subdomains lean on images, where a single viewpoint is what makes the inference hard. Third, the number of criteria per task varies, and 69.2% of tasks carry exactly one; on those, accuracy and rubric score coincide by construction, which we control for in Section C.4. Fourth, video length varies by an order of magnitude across subdomains, so the frame-budget question of Section C.2 binds very differently in different cells.

B.2 Construction pipeline

Figure 8 diagrams the authoring and review pipeline of Section 2.4 end to end, with the number of tasks surviving each round. We include it because the attrition is itself a description of the benchmark: the three principles of Section 2.3 are cheap to state and expensive to enforce, and the cost shows up as the 85% of authored tasks that do not reach the release.

B.3 Taxonomy design

Table 4 gives the taxonomy in full: each domain with the kind of recovery it demands, and each subdomain with the inferential step that defines it. This section says where the categories came from and how an item is assigned to one.

What the four domains partition. The domains are not subject matter categories — they do not sort scenes into "street", "kitchen", "office". They sort tasks by *what kind of thing is missing* from the depicted moment, which is the variable the benchmark manipulates. Temporal & causal dynam-

Table 3: **HSS dataset composition and statistics.** The released benchmark: 522 tasks over four domains and eleven subdomains, graded against 753 rubric criteria (1.44 per task; 69.2% carry exactly one). Each task is paired with a unique image or video clip, so task counts are also media counts. The video half totals 17.6 hours: median 73 s, interquartile range 20–305 s, longest 28 min. Word counts are means over whitespace-delimited tokens of the prompt, the full task rubric, and the reference answer.

	Tasks				Mean length (words)		
Domain / subdomain	**All**	**Image**	**Video**	**Criteria**	**Prompt**	**Rubric**	**Answer**
Physical & spatial logic	**176**	**114**	**62**	**255**	**37.4**	**27.7**	**73.9**
Hidden & invisible	58	28	30	86	27.8	22.6	69.7
Spatial "alien viewpoint"	50	40	10	76	41.4	32.7	78.9
Affordance & feasibility	41	34	7	61	45.5	30.8	77.8
Spatial reachability	27	12	15	32	38.0	24.8	68.1
Temporal & causal dynamics	**134**	**81**	**53**	**177**	**33.6**	**22.8**	**64.4**
Extrapolation	51	31	20	65	38.5	23.2	63.3
Retrodiction	50	40	10	66	32.7	21.6	73.0
Mechanistic causality	33	10	23	46	27.3	24.2	53.0
Social understanding	**110**	**28**	**82**	**174**	**30.1**	**29.8**	**64.3**
Social role, norm & power	59	23	36	102	29.8	30.8	67.1
Theory of mind	51	5	46	72	30.5	28.5	61.0
Abstract & contextual inference	**102**	**65**	**37**	**147**	**37.9**	**24.6**	**64.0**
Change & consequence	70	47	23	90	38.2	23.4	65.4
Patterns & pareidolia	32	18	14	57	37.2	27.3	61.1
All of HSS	**522**	**288**	**234**	**753**	**35.0**	**26.3**	**67.5**

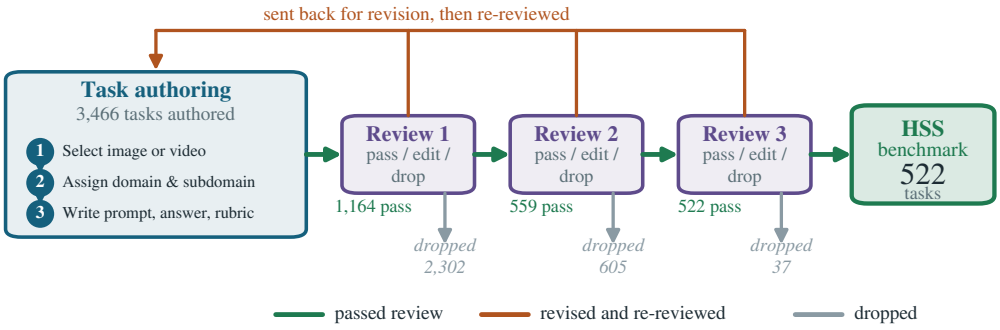


Figure 8: **How an HSS task is built and reviewed**, with the attrition at each stage. A task is authored in three steps, then faces three independent review rounds; each round may pass it on, return it to authoring for revision (after which it re-enters the pipeline from the start), or drop it outright. Counts are net survivors: a task returned for revision and later re-admitted is counted once, at the round it finally passes, so each *dropped* figure is the net loss at that round. Of 3,466 authored tasks, 522 survive all three rounds — a yield of 15.1%. The first gate is by far the steepest: it turns away 2,302 tasks, two in three of everything written, overwhelmingly because the answer was already depicted in the scene, or for want of unambiguous consensus on the ground truth (principles 1 and 3 of Section 2.3).

ics withholds another moment in time: the scene is one slice, and the answer lies behind it, ahead of it, or in the mechanism carrying one to the other. Physical & spatial logic withholds structure: what sits outside the frame, what an object weighs or how it would move, how the layout resolves from a viewpoint the camera never occupied. Social understanding withholds mental and normative state: intent, belief, role, the rule everyone in the scene is following. Abstract & contextual inference withholds the rule itself: the convention, pattern or consequence that makes the arrangement mean something. The split follows the lines along which cognitive science studies these inferences sepa-

rately — core knowledge of objects and space (Spelke & Kinzler, 2007), intuitive physics (Battaglia et al., 2013), mentalizing (Baker et al., 2017) — rather than any partition of image content.

How a task is assigned. Each task carries exactly one subdomain, chosen as its *hardest* inferential step rather than its topic. A photograph of a kitchen whose answer requires knowing that a pan handle turned outward will be caught by a sleeve is *affordance & feasibility*, not "domestic scenes"; a video of a market whose answer requires knowing which trader defers to which is *social role, norm & power*, regardless of the market. Where two steps chain — a physical judgment that only matters because of what someone intends — the reviewer assigns the step the model must get right in order to answer, and the confusable pairs (Section 5) are where that call is least forced. The consequence for reading results is stated in Section 5: per-subdomain numbers are a profile, not a decomposition.

Why eleven. The subdomains are the granularity at which the failure analysis separates: cause associates with subdomain at Cramér's $V = 0.37$ and with domain at $V = 0.42$ (Appendix D), so each subdomain carries a distinct failure signature rather than repeating its parent's. Coarser categories would hide that — affordance and reachability are both "spatial" but fail on size and fit at 62% and 56% while differing in what they ask of the viewer — and finer ones would produce cells too small to carry a claim. Eleven is also what the authoring pipeline can hold: every reviewer applies the same definition table (Table 4) to every task, and the taxonomy doubles as the generative blueprint that keeps coverage even across domains during collection (Table 3).

Table 4: Taxonomy Design for HSS along with descriptions.

Domain	Domain Description	Sub-domain	Sub-domain Description
Temporal & Causal Dynamics	Recovering the scene's state at a moment other than the one depicted, or the causal process connecting those moments. The media is one slice; the answer lies behind it, ahead of it, or in the mechanism driving it.	Retrodiction	Inferring what already happened based on visual traces.
		Mechanistic Causality	Identifying the immediate physical mechanism driving an ongoing event. This is about understanding the "how" and "why" of current motion or state changes, not predicting the future.
		Extrapolation	Predicting the immediate kinematic trajectory or long-term systemic outcome of the current scene.
Physical & Spatial Logic	Recovering unshown structure of the physical scene: what exists outside view, what properties objects hold beyond their appearance, and how the layout resolves from a coordinate the camera never occupied.	Hidden & Invisible	Reasoning about entities or physical properties that are not directly visible in a scene. Deducing occluded, off-screen, or internal objects, non-visual attributes (mass, friction), and invisible forces.
		Affordance & Feasibility	Assessing the physical compatibility between objects or between agents and their environment.
		Spatial "Alien Viewpoint"	Mentally re-projecting the 3D scene to calculate line-of-sight or reachability from a different physical coordinate.
		Spatial Reachability	Evaluating accessible space, paths, and physical volumes an agent or object can extend to within the 3D layout.
Social Understanding	Recovering the unobservable mental and normative layer: what agents believe, want, feel, attend to, and fail to know — and what the situation's unwritten rules permit, forbid, or make awkward.	Theory of Mind	Reasoning about the internal beliefs, knowledge gaps, deceptions, and emotional states of an agent.
		Social Role, Norm & Power Dynamics	Reasoning about unwritten social contracts, etiquette, and cultural appropriateness in a scene.
Abstract & Contextual Inference	Recovering what a scene implies but never shows: the conventions and circumstances that give it meaning, and the outcomes of worlds it doesn't depict. The answer follows from applying a rule or reading a context.	Change & Consequence	Reasoning about a change the scene doesn't contain: given the change, inferring its consequence; or given the consequence, inferring the change required.
		Patterns & Pareidolia	Identifying structure in random data or showing specific hardwired bias to detect familiar patterns—especially faces and human figures—in random noise or objects.

C EXPERIMENTAL SETUP

This appendix fixes every choice that can move a number: which deployment of each model was called and how, how media reach it, how answers are graded and scored, what the blind control shows, and the verbatim prompts.

C.1 MODEL CONFIGURATIONS

Table 5 is the single source of truth for how every model in Table 1 was called. It is generated from the harness's model registry (`configs/models.yaml`), which is released with the evaluation code, so a reader can reproduce any row from the endpoint and the settings given here.

Several conventions apply to every row and are therefore not repeated in the table. *Temperature* and *top-p* are never set: we omit both from the request so that each provider's own default applies, because the alternative — picking a value — silently changes a different thing on every route. *Reasoning effort* is run at the highest level each deployment honours rather than at a common level, since the families do not share a ladder; `default` in the effort column means the route rejects the parameter and the provider's own thinking level applies. The effort listed is the one used for the main table, which is not always the registry default, and the sweep over lower settings is reported separately. *Max output tokens* is the requested ceiling, not the observed usage; thinking tokens count against it, and a response that exhausts it is recorded as a no-answer rather than scored zero, so that an ungenerous budget does not read as a capability failure.

For video, models advertising native ingestion receive the clip as a single content block, and that claim is verified empirically before it is enabled — some deployments accept a video block and silently ignore it (Appendix C.2). Everything else receives uniformly sampled frames at 2 fps, each labelled with its timestamp, capped at the per-model frame limit in the last column: 500 unless a provider enforces something stricter, in which case frames are thinned uniformly with the endpoints retained, so the frames still span the whole clip. Frames are downscaled to a longest side of 768 px and encoded at a fixed JPEG quality; images are sent at native resolution with a longest side of 4096 px and are downscaled only when a payload limit requires it. All media is content-addressed and cached, so every model in the table sees byte-identical input. Requests carry a 1200 s timeout with up to 5 retries under exponential backoff, and failures are recorded per sample rather than silently dropped.

C.2 MEDIA HANDLING

Models differ in how they accept visual input, and this is a confound if left implicit. We handle it explicitly:

Native video. Models that accept a video file receive the clip directly. Whether a model truly ingests video is verified empirically rather than assumed: some deployments accept a video content block and silently ignore it, so we confirm the model describes clip-specific content before enabling this path.

Uniform frame sampling. Otherwise, videos are sampled at a fixed rate of $f = 2$ fps with a hard cap of $N_{\max}$ frames. When $f \cdot T > N_{\max}$ for a clip of duration T, we *lower the rate* to $N_{\max}/T$ rather than truncating, so the frames always span the whole clip:

$$f_{\text{eff}} = \min\left(f,\ \frac{N_{\max}}{T}\right), \qquad N = \min(\lfloor fT \rfloor,\ N_{\max})\,. \tag{1}$$

Each frame is labelled with its timestamp, and the prompt states the frame count, effective rate and duration, so the model can reason about elapsed time rather than mere ordering. We use $N_{\max} = 500$ as a single universal cap for all models, chosen as the most restrictive per-request image limit among the deployments we evaluate; where a provider imposes a stricter limit, frames are thinned *uniformly* (endpoints retained) and the affected samples are flagged so the reduced temporal resolution is visible in the analysis rather than hidden.

Images. Sent at native resolution, downscaled only when a payload limit requires it. All media is content-addressed and cached, so every model sees byte-identical input.

Table 5: **Model configurations for the main experiments.** All 24 models of Table 1, grouped by vendor in the same order, plus the three judges of Section 3.4. *Endpoint* is the identifier sent to the proxy; *API* is the request shape (OpenAI Chat Completions, OpenAI Responses, or Anthropic Messages), which differs across families and is reported because it changes what usage metadata comes back. *Effort* is the reasoning level used for the main table, with `default` meaning the route does not accept the parameter. *Out* is the requested output-token ceiling, which thinking tokens count against. *Native vid.* is whether the clip is sent as a video block rather than as sampled frames, verified empirically. *Frames* is the per-request frame cap that binds for that deployment; sampling is 2 fps for all models, thinned uniformly to fit the cap. A dash is a setting we did not pin: either the provider exposes no such control, or the deployment was reached through the proxy rather than through the committed registry, in which case the provider default applied. The three such rows — Qwen3.8-flash and the two MiMo deployments — have their endpoint and effort recovered from the run records.

Vendor	Model	Endpoint	API	Effort	Out	Native vid.	Frames
OpenAI	GPT-6-astra	`openai/gpt-6-astra`	Responses	`max`	128,000	no	500
	GPT-6-sol	`openai/gpt-6-sol`	Responses	`max`	128,000	no	500
	GPT-5.6-sol	`openai/gpt-5.6-sol`	Responses	`max`	128,000	no	500
	GPT-5.6-terra	`openai/gpt-5.6-terra`	Responses	`max`	128,000	no	500
	GPT-5.6-luna	`openai/gpt-5.6-luna`	Responses	`max`	128,000	no	500
	GPT-6-luna	`openai/gpt-6-luna`	Responses	`max`	128,000	no	500
Anthropic	Claude-Opus-5.5	`anthropic/claude-opus-5-5`	Messages	`xhigh`	128,000	no	500
	Claude-Fable-5.1	`anthropic/claude-fable-5-1`	Messages	`xhigh`	128,000	no	500
	Claude-Fable-5	`anthropic/claude-fable-5`	Messages	`max`	128,000	no	500
	Claude-Opus-5	`anthropic/claude-opus-5`	Messages	`max`	128,000	no	500
	Claude-Sonnet-5	`anthropic/claude-sonnet-5`	Messages	`max`	128,000	no	500
	Claude-Opus-4.8	`anthropic/claude-opus-4-8`	Messages	`max`	128,000	no	500
Google	Gemini-3.8-flash	`gemini/gemini-3.8-flash`	Chat	`high`	65,536	yes	500
	Gemini-3.7-flash	`gemini/gemini-3.7-flash`	Chat	`high`	65,536	yes	500
	Gemini-3.6-flash	`gemini/gemini-3.6-flash`	Chat	`high`	16,384	yes	500
	Gemini-3.5-flash	`gemini/gemini-3.5-flash`	Chat	`high`	16,384	yes	500
Muse	Muse Spark 1.3	`meta_ai/muse-spark-1.3`	Responses	`max`	128,000	yes	50
	Muse Spark 1.2	`meta_ai/muse-spark-1.2`	Responses	`xhigh`	128,000	yes	50
Qwen	Qwen3.8-Max	`dashscope/qwen3.8-max`	Chat	`max`	16,384	no	250
	Qwen3.8-flash	`dashscope/qwen3.8-flash`	Chat	`max`	—	—	—
Zhipu	GLM-5.3-flash	`fireworks_ai/glm-5p3-flash`	Chat	`max`	128,000	no	60
Moonshot	Kimi-K3	`bedrock/us.moonshotai.kimi-k3`	Chat	`max`	128,000	no	100
Xiaomi	MiMo-v2.6-flash	`mimo-v2.6-flash`	Chat	`high`	—	—	—
	MiMo-v2.6-pro	`mimo-v2.6-pro`	Chat	`high`	—	—	—
Judges	judge-claude-opus-5	`anthropic/claude-opus-5`	Chat	`high`	8,192	n/a	n/a
	judge-gpt-5.6-terra	`openai/gpt-5.6-terra`	Chat	`high`	8,192	n/a	n/a
	judge-gemini-3.6-flash	`gemini/gemini-3.6-flash`	Chat	`high`	8,192	n/a	n/a

C.3 JUDGING

A single LLM judge grades all runs, following standard practice for free-form evaluation (Zheng et al., 2023). The judge sees the question, the reference answer, the rubric, and the candidate answer, and returns a per-criterion verdict with a one-sentence justification. It does *not* see the media by default: rubrics are written to be checkable against the reference answer, which keeps the judge from re-adjudicating perception. We report the effect of changing judge model in Section 3.4; the effect of showing the judge the media and of changing grading format is left to future work.

Grading protocol. The judge is instructed to: mark a criterion met only if the answer actually asserts what it requires; judge meaning rather than phrasing, crediting synonyms and equivalent directions or units; mark a criterion unmet if the answer states the required content but also asserts a contradicting alternative; ignore extra correct detail; and treat the reference answer as ground truth on disagreement. It returns one verdict per criterion with a short justification citing the candidate answer, in a fixed machine-readable schema. Unparseable output is re-requested a bounded number of times before the sample is recorded as a grading failure rather than a zero. The verbatim system prompt and output schema are in Appendix C.7.

C.4 METRICS

For a model with per-criterion verdicts $v_{ij} \in \{0, 1\}$ over samples $i = 1..n$ with k_i criteria:

$$\text{Accuracy} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\left[\textstyle\sum_j v_{ij} = k_i\right], \qquad \text{Rubric score} = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{k_i}\sum_{j=1}^{k_i} v_{ij}. \tag{2}$$

Accuracy is primary because the benchmark asks whether a model gets the whole answer right; a response that identifies two of three required facts is not a correct answer to the question asked. Rubric score is retained as a diagnostic: it distinguishes systematic near-misses from total misses.

C.5 HARNESS AND VERSIONING

Harness. Models are accessed through a single OpenAI-compatible proxy so that one code path serves all providers. Media is content-addressed and cached; frames are extracted once per (video, sampling configuration) and reused across models, so every model sees byte-identical input.

Determinism and retries. Failures are recorded per sample rather than silently dropped. Transient errors are retried with backoff; parameters a particular deployment rejects are removed and the call retried, and the removal is logged. A response that exhausts its output budget on reasoning without emitting an answer triggers one budget increase before being recorded as no-answer, so that scores are not depressed by an arbitrary budget choice.

Taxonomy labels. Every sample carries a `domain` and a `subdomain` from Table 3. The annotation sheet is inconsistent — it writes the redundant tag about half the time (`Physical & Spatial Logic (space)`), and spelling varies — so labels are normalised on load. Anything unrecognised, or a subdomain used under the wrong domain, is *not* silently absorbed: it surfaces as its own cell with a warning, so labelling errors are visible in the composition table rather than hidden inside a valid-looking cell. In the released 522 tasks exactly one label needs this treatment — a punctuation variant, `Social Role Norm Power`, normalised into *social role, norm & power* and counted there in Table 3.

Dataset versioning. Sample identity is a hash of the question and media reference, so samples are stable under insertion and reordering. Adding samples extends a run rather than restarting it; results for samples no longer in the dataset are excluded from scoring, and any model not yet evaluated on the full current set is reported with reduced coverage rather than imputed zeros. We never report a table in which models have unequal coverage.

C.6 THE BLIND CONTROL

The blind control re-runs the strongest model, GPT-6-astra at maximum effort, on all 522 tasks with the media removed and everything else — prompt, rubric, judge, three attempts — held fixed. It scores 6.6% (Table 1) against 53.6% sighted. The residue is worth looking at directly, because it says what a language prior can still buy on this benchmark.

It buys guesses, not knowledge. Only 4 of the 522 tasks pass blind on all three attempts, while 67 pass at least once: the blind model is not solving a stable subset, it is landing on two-way questions at roughly the rate chance predicts. All four examples below pass on the first attempt and fail on at least one of the other two.

What passes with the media removed

1. A two-way question. "Assuming the person seated in the rowboat was rowing in a constant direction, did they pass in front of or behind the boat on the left?" Blind answer, in full: "They passed behind the boat on the left." Correct, with no stated reason, and it flips on attempts 2 and 3.

2. A yes/no question. "Without standing up, could the musician closest to the left of the director reach out and touch the director with his hand?" Blind answer: "No. The director is too far away for the seated musician to touch him without getting up." The reference says no, so the rubric's single criterion is met by a coin flip dressed in a justification the blind model cannot have.

3. A left/right question. "...from which of his sides did the puck pass through?" Blind answer: "The puck passed on his left side." The sighted model gets this one wrong on all three attempts, which is a fair summary of how little the blind pass means.

4. A question that carries its own answer. "What recent physical process most likely created the muddy channel that the rescue workers are clearing, and what visible evidence supports that conclusion?" Blind answer: "scoured by a recent flash flood or mud-and-debris flow...torn-away banks with exposed roots, uprooted trees, broken logs". The words *muddy channel* and *rescue workers* narrow the process before the image is consulted, and the model then invents evidence that happens to match.

Cases of the first three kinds are a property of the answer space, not of the item: the task still requires the media to be answered *for a reason*, and a rubric that demands the reason as a second criterion removes the free pass. Cases of the fourth kind are a genuine authoring defect — the question over-describes the scene — and the blind control is how we find them. We report the control at the item level in the release so that this subset is auditable rather than merely acknowledged.

C.7 PROMPTS

Every model sees the same two-part input: a media preamble stating what is attached and how, and the annotator's question verbatim. Nothing in the prompt names a domain, hints at an answer format, or mentions the rubric.

Media preamble — image

```
An image is attached.
```

Media preamble — native video

```
A video ({duration}s) is attached.  Reason about motion and event
order across time.
```

Media preamble — sampled frames

```
The video below is provided as {num_frames} frames sampled uniformly at
{fps} frames per second (video duration {duration}s), in chronological
order.  Frame timestamps in seconds:  {timestamps}.
Treat the frames as one continuous video and reason about motion and
event order across them.
```

Timestamps are stated explicitly so a model can reason about elapsed time rather than frame ordering alone.

Blind control system prompt (no media attached)

```
You are a helpful assistant.  Answer the user's question faithfully.
If you do not have enough information to answer, do not guess --- say
so honestly.
```

The licence to decline is deliberate: it makes a refusal a real measurement rather than an artifact of being ordered to guess.

Judge prompt. The judge receives the question, the reference answer, the numbered rubric and the candidate answer, and returns one verdict per criterion in a fixed JSON schema. It does *not* see the media: rubrics are written to be checkable against the reference answer, which keeps the judge from re-adjudicating perception.

Judge system prompt

```
You are a strict, impartial grader for the Humanity's Sixth Sense
(HSS) visual reasoning benchmark.  You are given a question about a
piece of media, a human-written reference answer that is authoritative
and correct, a list of rubric criteria, and a candidate answer from a
model under evaluation.
```

```
Grade the candidate answer against each rubric criterion
independently.
Rules:
1.  A criterion is met only if the candidate answer actually asserts
what the criterion requires.  Do not infer or give credit for
near-misses, hedges, or contradictions.
2.  Wording may differ from the reference; judge meaning, not
phrasing.  Synonyms, equivalent directions (e.g. 'right to left' vs
'leftward'), and equivalent units count.
3.  If the candidate answer states the required content but also
states a contradicting alternative, the criterion is NOT met.
4.  Extra correct detail neither helps nor hurts.  Extra incorrect
detail only matters if it contradicts a criterion.
5.  Treat the reference answer as ground truth when the candidate
disagrees with it.
6.  Judge only the criteria given.  Do not invent criteria.
Return ONLY a JSON object, no prose and no markdown fences, in exactly
this form:
{"criteria":  [{"id":  "<criterion id>", "met":  true|false,
"justification":  "<one short sentence citing the candidate
answer>"}], "overall_comment":  "<one sentence overall assessment>"}
```

D Failure Cause Analysis in Full

Taxonomy and protocol. We label the primary cause of every failed first attempt in the main table: 8,573 failures over 24 models and 522 tasks. The eleven causes fall into three families that separate *where* the answer goes wrong: *perception*, where the model misreads something the scene actually shows; *latent*, where it misses or misreads something the scene only implies; and *reasoning*, where it has the facts right and draws the wrong conclusion. A twelfth label, *no committed answer*, covers non-answers. Labelling is performed by Claude-Opus-5 shown the question, the reference answer and the model's answer — never the media, so it judges the departure from the reference rather than re-adjudicating the scene. Every failure received a label; none errored. Table 6 gives the pooled distribution.

Table 6: **Primary failure causes, pooled over 24 models** (n = 8,573). *Share* is the percentage of all failures; the last two columns split that share by media type. Perception and latent causes together account for 94% of failures.

Family	Cause	Count	Share	Image	Video
Perception	Misidentified an object, person, or role	1,702	19.9	17.7	22.2
	Misread orientation or facing	1,031	12.0	18.2	5.1
	Misjudged size, distance, or fit	1,048	12.2	15.8	8.2
	Read a 2D overlap as 3D alignment	776	9.1	14.6	2.9
Latent	Misread an invisible quantity from its trace	634	7.4	6.9	8.0
	Misread motion or sequence	678	7.9	3.8	12.4
	Misread intent, feeling, or relationship	373	4.4	0.3	8.9
	Missed the decisive cue	1,826	21.3	15.2	28.1
Reasoning	Converted to the wrong viewpoint	120	1.4	1.7	1.1
	Misjudged a line of sight	219	2.6	3.7	1.3
	Mis-traced the consequence or the question's condition	115	1.3	2.1	0.5
—	No committed answer	51	0.6	0.1	1.2

The cause follows the task. Cause is strongly associated with subdomain (Cramér's $V = 0.37$, $\chi^2 = 11{,}564$, dof $= 100$) and with media type ($V = 0.43$), but almost not at all with which model produced the failure ($V = 0.06$, $\chi^2 = 284$, dof $= 230$). The per-task view agrees: on the 503 tasks that five or more models fail, a median 80% of those models fail for the same cause and 90% for the same family. Figure 3 gives the full subdomain × cause matrix.

The media split is the clearest single contrast. Image failures are 66% perception — dominated by misread orientation (18%), misidentification (18%), size and fit (16%) and reading a 2D overlap as 3D alignment (15%). Video failures invert this: 57% are latent, with the missed decisive cue alone at 28%. Video does not make models worse at seeing; it gives them more to miss.

What the agent loop repairs. Table 7 pairs each agentic run against its own single-call run on the same tasks, pooling the four harness/base pairs of Table 9. This analysis uses first attempts only and is restricted to the released 522 tasks, one of which the agentic subset predates; it therefore runs over 387 tasks where Table 2 runs over the 388 the harnesses were given. Total failures fall from 944 to 759.

The pattern is sharp. Every perception and latent cause falls by 19–37%, with one exception: reading a 2D overlap as 3D alignment does not move (102 → 101), and only 24% of single-call failures of that kind are repaired — the lowest fix rate of any common perception cause. Cropping and zooming recovers a missed cue or a misidentified object; it does not recover depth, because the extra looking returns more of the same projection. Meanwhile all three reasoning causes *grow* in absolute terms (line of sight +23%, consequence +25%, wrong viewpoint +8%) off small bases: as the perceptual failures are cleared away, what remains is increasingly a failure of inference rather than of sight.

Table 7: **How the agent loop shifts the failure distribution**, four harness/base pairs pooled on 387 paired tasks, first attempts. *Fixed* is the percentage of single-call failures with that cause that the harness repairs.

Family	Cause	Single	Agentic	Δ %	Fixed %
Perception	Misidentified an object, person, or role	181	147	−19	35
Perception	Misread orientation or facing	117	93	−21	40
Perception	Misjudged size, distance, or fit	137	105	−23	39
Perception	Read a 2D overlap as 3D alignment	102	101	−1	24
Latent	Misread an invisible quantity from its trace	71	55	−23	37
Latent	Misread motion or sequence	83	52	−37	33
Latent	Misread intent, feeling, or relationship	49	35	−29	31
Latent	Missed the decisive cue	154	110	−29	34
Reasoning	Converted to the wrong viewpoint	13	14	+8	31
Reasoning	Misjudged a line of sight	22	27	+23	9
Reasoning	Mis-traced the consequence or the question's condition	12	15	+25	25
—	No committed answer	3	5	+67	33

Table 8: **Where the agent loop pays, by subdomain.** Pass@1 (%) for the four harness/base pairs pooled on the 387 paired tasks of Table 7, first attempts. n counts pair × task observations, so each subdomain appears four times. Re-inspection helps most where the evidence is in the frame and least where the answer turns on depth or on an outcome the media never shows.

Subdomain	n	Single	Agentic	Δ
Patterns & pareidolia	68	33.8	64.7	**+30.9**
Retrodiction	152	34.9	55.3	+20.4
Theory of mind	136	32.4	48.5	+16.2
Social role, norm & power	168	32.1	47.6	+15.5
Affordance & feasibility	140	43.6	55.0	+11.4
Change & consequence	224	40.6	49.1	+8.5
Hidden & invisible	180	40.6	48.9	+8.3
Mechanistic causality	84	47.6	56.0	+8.3
Spatial "alien viewpoint"	152	42.1	50.0	+7.9
Extrapolation	160	45.0	52.5	+7.5
Spatial reachability	84	34.5	39.3	+4.8

Limitations of this analysis. Three, stated plainly. The labels are produced by an LLM and have not been verified by a human annotator, so the taxonomy shares should be read as a consistent automated reading of 8,573 answers rather than as adjudicated ground truth. The labeller, Claude-Opus-5, is also one of the 24 evaluated models; the near-zero cause-by-model association (V =

0.06) is evidence against a self-favouring labelling bias, but it is not a controlled test of one. And the agentic comparison uses first attempts only, so its single-call accuracies differ from Table 2 by a few tenths of a point (for instance Claude-Opus-5 is 32.0 → 51.2 here against 30.8 → 51.3 there).

E ONE FAILURE PER SUBDOMAIN

Aggregate scores and the taxonomy of Appendix D both abstract away from what an individual HSS task actually asks. This section grounds them: one representative failure from each of the eleven subdomains, shown with the media the models were served.

Each case is a task that *all three* of GPT-6-astra, Gemini-3.8-flash and Claude-Opus-5.5 missed on every one of their three attempts. Selecting on unanimous, repeated failure is deliberate — it removes both sampling noise and any single model's idiosyncrasy, so what remains is a property of the task rather than of a grader or a seed. There are 79 such tasks in the release, between 3 and 13 per subdomain, and we take one from each (Figures 9–14). The model answer quoted in each panel is the judge's condensed reading of what the models actually said; where the three differed in wording they agreed in substance, and we quote the shared error.

Two things are visible across the set that the numbers alone do not convey. The first is how little of the difficulty is perceptual. In every panel the relevant evidence is plainly in frame — a taut chain, a flag blowing toward the lens, a gap on a shelf, skate marks on ice — and the models describe that evidence correctly while drawing the wrong conclusion from it. The second is how ordinary the correct answers are. Nothing here needs expertise: an adult recognises that a bolted cage does not spin with the shaft inside it, that an adult and a child cannot comfortably high-five, and that a mark written on an inspected panel most likely means it passed. That combination — visible evidence, unremarkable inference, consistent failure — is what the benchmark is built to isolate.

F AGENTIC HARNESSES IN FULL

Table 9 gives all four harness/base pairs behind Table 2, paired on the same 388 tasks. All four improve significantly. The size of the gain tracks the harness rather than the base model: Claude Code lifts both of its models by 16–21 points at 12–14× the base model's thinking tokens, Codex lifts both of its by 5–8 at 1–2×. Across the four pairs the token multiplier and the accuracy gain rank together perfectly ($\rho = 1.0$).

Table 9: **All agentic harness/base pairs**, paired on 388 tasks. *Blank* counts tasks where the harness returned no answer, scored as zeros; p is two-sided McNemar.

Harness	Base model	Raw	Agent	Δ	Win	Loss	p	Think	Blank
Claude Code	Claude-Opus-5	30.8	51.3	+20.5	112	30	$< 10^{-4}$	43,764	0
Claude Code	Claude-Fable-5.1	41.2	57.7	+16.5	96	30	$< 10^{-4}$	42,237	1
Codex	GPT-5.6-sol	28.4	35.8	+7.5	63	33	0.003	6,541	4
Codex	GPT-6-astra	54.4	59.3	+4.9	53	28	0.007	2,040	0

Trajectories. Agent steps are logged per task. For the harnesses that search hardest, failures run *longer* than successes: Claude Code takes a median 81 steps on tasks it gets wrong against 67 on tasks it gets right around Opus-5 (rank-sum $p = 0.01$) and 65 against 54 around Fable-5.1 ($p = 0.04$). The loop does not know when it has seen enough. Codex, which searches least, shows no such gap. What the extra looking buys, and what it costs, is easiest to see in a whole run: Appendix F.1 walks through four.

F.1 FOUR TRAJECTORIES

Every harness logs its steps, so a run can be read end to end. Below are four runs, condensed to the steps that change what the agent knows: two the loop repaired and two it broke, one pair per harness. Step numbers are `step_id` values from the logged trajectory and include the tool observations the agent reads back, which is why counts differ between harnesses for comparable amounts of work.

Hidden & invisible



Prompt
Which way is the wind blowing, from the striped flag?

All three models answered
From right to left across the frame.

Reference
Toward the camera.

Spatial “alien viewpoint”



Prompt
If the child in the gold jacket turned 90° right, would they face the woman?

All three models answered
They would face her left side.

Reference
Not at all; they would have to turn around.

Figure 9: **One representative failure per subdomain (1 of 6):** the two subdomains of physical & spatial logic that turn on what lies outside the frame: hidden & invisible, and the alien viewpoint. Every task shown was missed by GPT-6-astra, Gemini-3.8-flash and Claude-Opus-5.5 on all three attempts; the model line is the judge’s condensed reading of what all three said. All eleven cases are image tasks, so each card shows exactly the media the models were served.

Affordance & feasibility



Prompt
Would two more white books fit on this shelf?

All three models answered
No, there is too little space.

Reference
Yes, in the gaps already there.

Spatial reachability



Prompt
If the blonde woman at centre stage extended her right arm straight out at shoulder height, would her hand reach the dark-haired woman on the left?

All three models answered
Yes, her hand would make contact; no step needed.

Reference
No; there is a clear gap, so she would have to step closer.

Figure 10: **One representative failure per subdomain (2 of 6):** the two that turn on fit and reach: affordance & feasibility, and spatial reachability. Selection and panel layout as in Figure 9.

Extrapolation



Prompt
The green truck reverses straight back. Who is in most danger?

All three models answered
A man in a red vest on the left.

Reference
The man in green hi-vis, directly behind.

Retrodiction



Prompt
From which side of the frame did the traveller with the two dogs enter?

All three models answered
The lower-left side.

Reference
The bottom edge, tracing the path back.

Figure 11: **One representative failure per subdomain (3 of 6):** the forward and backward temporal subdomains: extrapolation and retrodiction. Selection and panel layout as in Figure 9.

Mechanistic causality



Prompt
Are these chains bearing the plough's drag weight?

All three models answered
No, they hang slack.

Reference
Yes, they are taut and loaded.

Theory of mind



Prompt
Who in the room seems to be paying the least attention to the meeting?

All three models answered
The man in the black polo against the back wall, or the man in glasses at front left.

Reference
The blonde-haired woman behind the man in the blue shirt, looking down rather than at the discussion.

Figure 12: **One representative failure per subdomain (4 of 6):** mechanistic causality and theory of mind. Selection and panel layout as in Figure 9.

Social role, norm & power



Prompt
Are the people around the aircraft visitors, or crew preparing it?

All three models answered
Visitors examining it.

Reference
Crew, from the open access panels.

Change & consequence



Prompt
If the shaft spins, does the cage around it spin too?

All three models answered
Yes, it turns with the shaft.

Reference
No; it is bolted to the engine casing.

Figure 13: **One representative failure per subdomain (5 of 6):** social role, norm & power, and change & consequence. Selection and panel layout as in Figure 9.

Patterns & pareidolia



Prompt
Whose direction of gaze breaks the pattern set by the rest?

All three models answered
The person in front with a pom-pom hat.

Reference
The man in the black beanie, centre, looking right.

Figure 14: **One representative failure per subdomain (6 of 6):** patterns & pareidolia. Selection and panel layout as in Figure 9.

Extended-thinking blocks are redacted in the vendor logs, so every quotation below comes from an agent's own tool descriptions or from the answer file it wrote.

Repaired: "Which team likely won this game of rugby? Why?"

Claude Code + Claude-Opus-5, social role & power, video (30 s), 121 steps. Single call: **fail** (all three attempts). Agent: **pass**.

Reference. The team in white won: after the whistle several white players raise both arms above their heads, while the black-and-red side stands with arms down and shoulders slumped.

Single call. "The home side in the dark red-and-black jerseys almost certainly won" — reading a kneeling white player as "the classic posture of a beaten player".

Agent. "The away team — the side playing in all white — almost certainly won."

Trace. (3) probes duration and resolution; (4–8) extracts one frame per second and tiles 36 of them into a contact sheet; (10–14) opens individual frames hunting the celebration; (16–18) crops a 240×110 region around two players and upscales it to 1920 px; (20) rebuilds the contact sheet at 2 fps because 1 fps was too coarse; (24–32) zooms progressively from 900×450 down to 420×220 on the first 3.5 seconds after the whistle; (47) writes a preliminary answer already naming the white team, then keeps working for 70 more steps; (50) re-samples the first 1.6 s at 8 fps; (63, 86, 106) three attempts to read the stadium scoreboard, all defeated by LED overexposure; (90) runs scene-cut detection, gets an empty list, and so establishes that the clip is one continuous shot and really is the final whistle; (99–104) re-extracts with a contrast boost to trace the goal line and shows the red-and-black side standing in its own in-goal; (113) "I have enough evidence to write a much stronger answer now"; (120) edits its own answer to soften "the scoreboard reads LOU" to "appears to read" and deletes a hallucinated advertiser name.

Why it worked. The cue was in the frame but brief. Re-sampling the same three seconds at 1, 2 and 8 fps and zooming five-fold on the celebrating cluster turned a glance into a measurement, and the pitch-line and scene-cut checks supplied corroboration the single call never sought.

Broken: "If the lightbulb hanging above the bald-headed man with red glasses were to fall, where would it end up relative to his body?"

Claude Code + Claude-Opus-5, change & consequence, image, 36 steps. Single call: **pass** (all three attempts). Agent: **fail**.

Reference. Behind him. The bulb is not lighting his face — no shadow under the chin, only a sheen on the top of his head — so it hangs behind his head from the camera's viewpoint.

Single call. "The fixture hangs from the ceiling well behind him…it would land just behind his right shoulder rather than squarely on it."

Agent. "It would land on his head — specifically on the front-left of his shaved head…It would *not* land behind him."

Trace. (3–4) reads the image and its dimensions; (8) writes a preliminary answer; (10–19) crops the bulb at 300%, the head at 400% and the glasses at 500% and reads each back; (24) dumps raw grayscale pixel rows across the support pole at five heights to test whether the camera is rolled; (26) does the same across the bulb and computes an intensity-weighted centroid at $x \approx 449$, then scans rows $y = 150$–240 for the left edge of the head; (28) measures the skull's apex and right edge; (30) "I have precise measurements now"; (31) draws a red plumb line at $x = 449$ down the whole image and crops it for inspection; (34) "The overlay confirms the plumb line"; (35) writes the final answer — "I can just read off what that line passes through" — and adds a section titled *One caveat about depth*: "If it hangs, say, half a metre behind him, it would fall past the back of his head", which it then overrules on the grounds that his forehead "is clearly lit from above-front-left by this light".

Why it broke. The tools answered the wrong question precisely. A 2D overlap was converted into a 3D plumb line to sub-pixel accuracy, and the arithmetic gave the agent enough confidence to overrule the depth objection it had itself raised — and to misread the very lighting evidence that settles the task. This is the failure mode of Table 7 in miniature: reading a 2D overlap as 3D alignment is the one cause the agent loop does not reduce.

Repaired: "…he approaches a couple. What does he notice that causes him to approach them?"

Codex + GPT-6-astra, theory of mind, video (20 min at 320×240), 21 steps. Single call: **fail**. Agent: **pass**.

Reference. He notices the woman drop her handkerchief, and approaches to retrieve it and return it.

Single call. "He notices the woman peering at him through her opera glasses."

Agent. "He notices that the woman. . . has dropped her handkerchief over the edge of their theater box."

Trace. (5) probes the file; (6) builds one 5×6 contact sheet covering the first 300 s at one frame per 10 s, with the timestamp burned into each tile; (7) reads the sheet and locates the two-chair act near $t \approx 205$ s; (8–9) a re-extraction fails and the image read errors out; (10) debugs its own `ffmpeg` invocation, moving the duration flag and switching to an absolute burned-in clock; (12–14) zooms $t = 215$–223 at 2 fps and takes one 960×720 close-up at $t = 250.8$ s; (15–16) "He notices the woman drop her handkerchief" and writes the answer; (17–21) a verification pass over $t = 216$–246 at 2 fps confirms it.

Why it worked. Timestamped contact sheets localised a twenty-minute film to a ten-second window, and a three-fold upscale of a 320×240 source resolved an object a few pixels wide — the detail the single call replaced with a plausible story.

Broken: "For what reason does the woman with darker sleeves and a dark apron take objects from the other workers?"

Codex + GPT-6-astra, theory of mind, video (148 s), 27 steps. Single call: **pass**. Agent: **fail** (rubric 0.5: the job title is right, the evidence is not).

Reference. A quality inspection: she takes each piece, examines it closely, and *returns it* to the worker.

Single call. "She takes them briefly, holds them up and turns them to examine them, then hands them back and moves to another worker."

Agent. "She is acting as an inspector or supervisor, checking that the pieces have been made correctly" — describing her as taking a piece and moving on, never noting the hand-back.

Trace. (6–8) overview sheet at one frame per 10 s; (9–11) re-samples $t = 8$–63 at one frame per 3 s plus a full-resolution frame; (12–15) two more sheets covering the hand-off window and the tail of the clip; (16) writes a first answer — "inspecting and gathering their output" — and in the same command builds a 40-tile sheet of $t = 25$–45 at 2 fps cropped to her work area; (19–22) crops the hand-off at 1 fps; (23–25) the densest look of the run, $t = 52$–55 at 4 fps upscaled two-fold; (26) revises the answer, changing "inspecting and gathering their output" to "checking that the pieces have been made correctly".

Why it broke. It sampled the right seconds at the right rate and still narrated only half the exchange. Its one revision tightened the job title rather than re-examining the action, so the return — the evidence the rubric asks for — never entered the answer. More looking does not help when the agent has already decided what it is looking at.

G Reasoning Effort in Full

Table 10 gives per-subdomain accuracy at every reasoning level, the numbers behind Figure 5. At the domain level, moving from the lowest to the highest setting gives GPT-6-astra $+8.2$ (temporal), $+12.5$ (physical), $+15.5$ (social) and $+7.8$ (abstract); Claude-Opus-5.5 $+6.7$, $+9.1$, $+0.9$ and $+4.9$; and Gemini-3.8-flash $+0.7$, $+5.7$, $+3.6$ and -1.0. Seven of the 33 model $\times$ subdomain cells end no better than they started, and three of Gemini-3.8-flash's eleven end lower. In only five of the twelve model $\times$ domain curves is the best score at the highest available setting.

H Video Frame Budget in Full

Table 11 gives the headline sweep and Table 12 repeats it with a 95% bootstrap interval on every cell. The sweep runs over the 234 video tasks of the release; frames are sampled at 2 fps and thinned uniformly to the cap, endpoints retained, so every budget spans the whole clip and the comparison is one of temporal *resolution* rather than coverage. The 500-frame column is the configuration used everywhere else in the paper, scored here on the first attempt alone, which is why it sits a few tenths from the three-attempt video column of Table 1.

Three things are worth separating. First, the lever is real: every model improves from 10 to 500 frames, which is not true of reasoning effort (Appendix G). Second, it saturates. The $10 \rightarrow 100$ step is worth $+11.5$, $+6.0$, $+2.1$ and $+9.8$ points to the four models; the $100 \rightarrow 500$ step, a five-fold further increase in tokens, is worth $+6.4$, $+2.1$, $+3.8$ and -1.3. Third, at this sample size most of it is not resolved: the per-cell interval is about ± 6 points and only GPT-6-astra's end-to-end gain has non-overlapping intervals at its endpoints. We therefore read the table as ruling out a large monotone gain from temporal resolution beyond roughly 100 frames, and as unable to resolve a small one.

Table 10: **Accuracy by subdomain and reasoning effort** (%), all 522 tasks. Columns are each deployment's levels in order — five for GPT-6-astra (`low–max`), four for Claude-Opus-5.5, three for Gemini-3.8-flash — and Δ is last minus first.

Subdomain	Reasoning level, lowest → highest					Δ
GPT-6-astra						
Hidden & Invisible	48	50	52	57	60	+12
Affordance & Feasibility	39	41	44	49	56	+17
Spatial "Alien Viewpoint"	28	30	40	46	44	+16
Spatial Reachability	22	22	26	22	22	+0
Extrapolation	41	45	47	49	53	+12
Retrodiction	50	48	64	50	56	+6
Mechanistic Causality	55	58	64	61	61	+6
Theory of Mind	29	31	37	39	45	+16
Social Role, Norm & Power Dynamics	32	31	46	42	47	+15
Change & Consequence	39	40	43	51	47	+9
Patterns & Pareidolia	44	44	44	50	50	+6
Claude-Opus-5.5						
Hidden & Invisible	24	34	41	40		+16
Affordance & Feasibility	41	54	41	56		+15
Spatial "Alien Viewpoint"	38	46	44	40		+2
Spatial Reachability	59	63	56	59		+0
Extrapolation	35	39	43	43		+8
Retrodiction	40	40	50	48		+8
Mechanistic Causality	42	42	45	45		+3
Theory of Mind	25	27	25	27		+2
Social Role, Norm & Power Dynamics	41	46	41	41		+0
Change & Consequence	51	47	60	59		+7
Patterns & Pareidolia	38	44	41	38		+0
Gemini-3.8-flash						
Hidden & Invisible	33	34	40			+7
Affordance & Feasibility	32	34	34			+2
Spatial "Alien Viewpoint"	34	48	38			+4
Spatial Reachability	41	41	52			+11
Extrapolation	33	31	39			+6
Retrodiction	44	44	42			−2
Mechanistic Causality	42	39	39			−3
Theory of Mind	33	37	35			+2
Social Role, Norm & Power Dynamics	32	39	37			+5
Change & Consequence	49	50	50			+1
Patterns & Pareidolia	44	41	38			−6

Table 11: **Video frame budget.** Pass@1 (%) on the 234 video tasks at each per-request frame cap, sampled at 2 fps and thinned uniformly to fit. Δ is 500 minus 10 frames; *best* is the budget at which each model peaks.

Model	10	50	100	200	500	Δ	best
GPT-6-astra	33.8	41.9	45.3	47.0	**51.7**	**+17.9**	500
Claude-Opus-5.5	30.3	32.1	32.5	**37.2**	36.3	+6.0	200
GPT-6-sol	23.1	27.8	29.1	26.1	**31.2**	+8.1	500
Claude-Opus-5	18.8	19.7	**28.6**	27.4	27.4	+8.5	100

The finding that matters for the benchmark is what the budget does *not* fix. At the full 500-frame budget the video–image gap of Table 1 is still open: 51.7 against 55.6 for GPT-6-astra, and 36.3 against 51.3 for Claude-Opus-5.5. If the video deficit were a sampling artifact it should close as the budget grows, and it does not.

Table 12: **Video frame budget, with intervals.** Pass@1 (%) on the 234 video tasks at each per-request frame cap, with 95% bootstrap intervals over tasks.

Model	10 frames	50	100	200	500
GPT-6-astra	33.8 [28,40]	41.9 [36,48]	45.3 [39,51]	47.0 [40,53]	51.7 [45,58]
Claude-Opus-5.5	30.3 [25,36]	32.1 [26,38]	32.5 [26,38]	37.2 [31,44]	36.3 [30,42]
GPT-6-sol	23.1 [18,29]	27.8 [22,33]	29.1 [23,35]	26.1 [21,32]	31.2 [25,37]
Claude-Opus-5	18.8 [14,24]	19.7 [15,25]	28.6 [23,35]	27.4 [22,33]	27.4 [22,33]

I JUDGE ALIGNMENT IN FULL

This appendix expands the summary in Section 3.4. Five models spanning the leaderboard — GPT-6-astra, Gemini-3.8-flash, Claude-Fable-5, Claude-Opus-5 and GLM-5.3-flash — were re-graded end to end by three judges from three vendors, on identical rubrics, reference answers and grading prompts: 2,620 model–task pairs (five models over a 524-task snapshot taken two tasks before the final release, one attempt each) carrying 3,780 individual criteria. Per-model scores under each judge are in Table 13.

Table 13: **Cross-judge alignment.** Pass@1 (%) for the same five models under three judges from three vendors, single attempt per task. Bold marks the case where a judge grades a model from its own vendor. The rightmost column is the largest disagreement between any two judges about a model; the largest anywhere in the table (3.6 points, Claude-Fable-5) is inside the ±4-point sampling interval of Table 1. Every pairwise rank correlation is $\rho = 1.0$: the three judges disagree about levels, never about order. Agreement statistics for the same runs are plotted in Figure 6.

	Judge			
Model	Claude-Opus-5	GPT-5.6-terra	Gemini-3.6-flash	Spread
GPT-6-astra (max)	53.8	50.8	**54.2**	3.4
Gemini-3.8-flash (high)	42.7	40.8	**43.5**	2.7
Claude-Fable-5 (max)	**33.8**	30.2	33.0	3.6
Claude-Opus-5 (max)	31.5	30.0	**31.7**	1.7
GLM-5.3-flash (max)	26.9	25.2	**27.1**	1.9
Mean over the five	37.7	35.4	37.9	

The judges agree, and where they differ they differ by a constant. Pairwise agreement is 95.8–97.2% at the criterion level and 96.2–97.5% at the item level, with Cohen's κ between 0.914 and 0.946 on both units (Figure 6b); all three judges return the same verdict on 95.2% of items. The disagreement that remains is almost entirely a level shift rather than a reordering. GPT-5.6-terra is the strict judge, scoring every one of the five models on average 2.4 points below the mean of the other two, while Claude-Opus-5 and Gemini-3.6-flash sit within 0.3 points of each other on average. Spearman rank correlation between judges is exactly 1.0 for all three pairs, so no conclusion in this paper that rests on model ordering changes under a different grader.

Self-preference is not detectable at this scale. LLM judges are known to favour their own family's outputs (Panickssery et al., 2024), which matters here because our primary judge shares a vendor with two evaluated models. Each judge has an overall leniency — the mean gap between the score it gives a model and the mean the other two give the same model — of +1.1 (Claude-Opus-5), −2.4 (GPT-5.6-terra) and +1.3 points (Gemini-3.6-flash). Subtracting it isolates whatever is specific to a same-vendor pair, and what remains is −0.80 for GPT-5.6-terra on GPT-6-astra, +0.38 for Gemini-3.6-flash on Gemini-3.8-flash, +1.09 for Claude-Opus-5 on Claude-Fable-5, and −0.44 points for the literal self-grading case, Claude-Opus-5 judging Claude-Opus-5. All four are smaller than the sampling interval on a single cell, and two have the wrong sign for a self-preference account. We read this as no evidence of vendor bias at the resolution HSS currently supports, not as proof of its absence.

Disagreement concentrates where the inference is about people. The 4.8% of items on which the three judges are not unanimous are not spread evenly: they are 7.3% of social understanding items and 6.4% of temporal & causal items, against 3.3% of physical and 2.7% of abstract ones, and 6.0% of multi-criterion items against 4.4% of single-criterion ones. Media type makes no difference (5.0% video, 4.8% image). This is worth stating plainly, because social understanding is also the domain models are worst at, and a sceptical reader should ask whether that deficit is a grading artifact rather than a capability gap. It is not: social understanding is the weakest of the four domains under *every* judge, and the judge-to-judge spread within it (1.7 points) is several times smaller than the 8 to 11 points separating it from the other three domains (Table 14). Judges argue more often about social items, and still agree that models fail them.

Table 14: **Per-domain accuracy under each judge**, pooled over the five re-graded models. The ordering of the domains is identical under all three judges, and social understanding is last under each.

Domain	Claude-Opus-5	GPT-5.6-terra	Gemini-3.6-flash	Spread
Temporal & causal dynamics	41.5	37.3	40.9	4.2
Physical & spatial logic	38.5	36.5	38.8	2.3
Abstract & contextual inference	38.2	37.4	39.0	1.6
Social understanding	30.2	28.8	30.6	1.7

What this does not establish. All three judges are LLMs, and shared training-data biases would be invisible to a comparison among them. The check that would close this gap is agreement against human graders on a stratified subset, which we report alongside the human baseline.