AN EXAM FOR ACTIVE OBSERVERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Human vision is a closed loop: gaze is continuously redirected by intermediate hypotheses rather than a single snapshot. Decades of psychophysics and cognitive science have argued that this *active observation* is essential for a wide range of tasks. Whether today's multimodal large language models (MLLMs) exercise active observation is an empirical question that current vision-language benchmarks do not answer. We introduce **ActiveVision**, a benchmark that makes active observation measurable for MLLMs, comprising 17 tasks across 3 categories. Tasks are designed to force repeated visual perception rather than a single static description. Frontier MLLMs collapse on **ActiveVision**: the highest-scoring model we evaluate, GPT-5.6 Sol, solves at most 23.5% of items at any reasoning-effort tier, and even Claude Fable 5, despite its strong results on reasoning and coding leaderboards, solves just 5.9% at its highest tier, far behind three human participants who average 96.1%. Furthermore, much of the gap persists even when models write and run their own vision code. Such code is unreliable on realistic imagery, and catching its failures itself requires the active perception the models lack. Together, these results indicate that current MLLMs lack robust active visual observation, motivating architectures and training objectives that close the perception–reasoning loop.

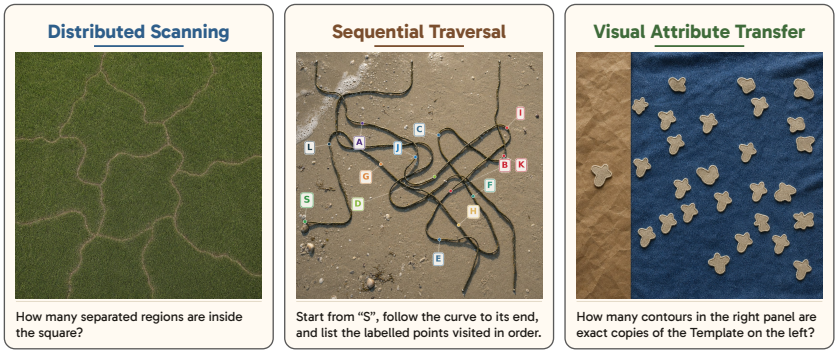


Figure 1: **Seeing is not always a single-glance task.** Many visual problems require an observer to coordinate attention and memory across an image: scanning exhaustively for distributed evidence, following connected structures without losing track, and comparing fine-grained attributes across distant regions. **ActiveVision** turns these abilities into a controlled test of whether MLLMs can keep visual evidence in the reasoning loop.

1 INTRODUCTION

Try the three tasks in Figure 1. Count the separated regions of grassland in the aerial photo on the left. In the middle, trace the rope from the green "S" through all crossings to the far end, reading off each labeled point that you pass. On the right, count every beige fabric patch on the linen whose silhouette matches the one shown on the kraft-paper template. None can be solved at a glance. Each requires several seconds of iterative, hypothesis-driven inspection: fixate, predict, return. The visual system executes this loop effortlessly, and the conscious mind barely notices it. Decades of research in psychophysics, cognitive science, and computer vision have argued that this iterative looking, known as *active observation*, is not an ornament on top of perception but, for a wide range of tasks, an essential part of perception itself (Aloimonos et al., 1988; Bajcsy, 1988; Ballard, 1991; Findlay & Gilchrist, 2003; Hayhoe & Ballard, 2005; O'Regan & Noë, 2001; Yarbus, 1967).

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.

Multimodal large language models (MLLMs) have advanced rapidly, yet it remains unclear whether their benchmark gains reflect progress in this capability. From the GPT-4 family (Achiam et al., 2023) onward, frontier models such as Claude Fable 5 (Anthropic, 2026c), GPT-5.5 (OpenAI, 2026a), and Gemini 3.1 Pro (Google DeepMind, 2026a) have excelled on widely used benchmarks. Performance on multimodal benchmarks such as MMMU-Pro (Yue et al., 2025) and CharXiv (Wang et al., 2024) is nearing saturation, leaving little separation among frontier models.

Yet benchmark saturation does not mean that vision-language understanding is solved. Applications in robotics (Luo et al., 2025), design (Kolodiazhnyi et al., 2026), manufacturing (Jian et al., 2026; Zhang et al., 2026), computer use (Yuan et al., 2026), spatial understanding (Petersson et al., 2025), and scientific discovery (Jiang et al., 2025; Laurent et al., 2024) require models to revisit visual evidence as they reason, a capability current MLLMs do not reliably exhibit. Existing evaluations leave this requirement largely unmeasured. Public leaderboards remain dominated by captioning, single-image visual question answering, and multiple-choice tasks that can often be answered from a static description of the image. Few benchmarks assess whether models can revisit visual evidence to form, test, and refine hypotheses as reasoning unfolds.

Current MLLMs appear to be passive perceivers: the image is encoded once as a fixed sequence of visual tokens, with no explicit perception–action loop. Yet this need not preclude active observation. During autoregressive reasoning, the model may shift attention among visual tokens as hypotheses evolve, letting earlier visual findings guide later steps (Zhang et al., 2025). The question is therefore behavioral rather than architectural:

> *Do MLLMs perform* active observation*, returning to the image to form and test hypotheses as they reason?*

We answer the question with **ActiveVision**, a benchmark that makes iterative visual perception measurable. It consists of 17 tasks spanning three cognitive demands of active vision: **Distributed Scanning** for exhaustive coverage of many local signals, **Sequential Traversal** for ordered stepping along a connected structure, and **Visual Attribute Transfer** for fine-grained comparison across regions. Across these tasks, the underlying configurations resist concise language description, so a model cannot easily summarize an image once in language and answer from that summary alone; it must keep returning to the pixels as its reasoning unfolds.

To create this structural complexity and render it realistically, each task follows a two-stage construction pipeline. First, a procedural scaffold specifies the underlying geometry exactly, including arbitrary positions, shapes, and curves. Second, GPT-image-2 (OpenAI, 2026c) re-renders the scaffold as a photorealistic image while preserving positions, counts, labels, and topology (§3.3). Voronoi regions become aerial fields and rivers, arrow chains become stones linked by footprints, and tangled loops become ropes on driftwood. The resulting images introduce realistic textures and visual complexity absent from sparse, cartoon-like renderings.

Every frontier model we evaluate fails on the overwhelming majority of the 85 released items. We test each model at every reasoning-effort tier exposed by its API. Across all settings, the best result is 23.5%, achieved by GPT-5.6 Sol at *high* effort, which still scores zero on 5 of the 17 tasks. Claude Fable 5, despite its strong results on reasoning and coding leaderboards, solves just 5.9%. Three human participants, by contrast, average 96.1%, about four times the best model's accuracy. Increasing reasoning effort helps only up to a point and leaves most of this gap open.

Tool use narrows but does not close the gap. We evaluate two Codex and three Claude Code agents, which write and run their own vision code and reach only 24.7–68.2% accuracy. Gains concentrate where tasks admit reliable code-based solutions, but traversal remains difficult. On realistic imagery, models often miss tool failures. Tool use therefore shifts the active-observation bottleneck to verification in the agent loop.

In summary, we (i) introduce **ActiveVision**, a controlled benchmark that isolates active observation through 17 diverse tasks spanning three cognitive axes (§3); (ii) systematically evaluate frontier MLLMs across their full reasoning-effort range and, as a tool-use test, autonomous coding agents built on them, finding a large and consistent human–model gap that neither additional reasoning nor tool use closes (§4); and (iii) trace the residual failures to a perceptual bottleneck unresolved by additional reasoning or tooling, positioning active vision as a distinct capability that current models lack but real-world applications require (§5).



Figure 2: Overview of all 17 tasks in **ActiveVision**, grouped into three task families that probe distinct dimensions of active observation. For each task, we show one photorealistic instance produced by the rendering pipeline described in §3.3, together with a brief description of the required operation.

2 Related Work

ActiveVision builds on two directions of prior work: the cognitive foundations of active vision and vision-centric benchmarks for MLLMs. We summarize each below; Appendix A expands on both.

Cognitive foundations of active vision. Cognitive science and computer vision have long argued that perception is intrinsically active. *Eye Movements and Vision* (Yarbus, 1967) showed that the same painting elicits different scanpaths under different tasks, and later work showed that problems ill-posed for a passive observer become well-posed once the observer controls its sensors (Aloimonos et al., 1988; Bajcsy, 1988; Ballard, 1991). The sensorimotor account goes further: seeing *is* the mastery of how visual input changes under one's own movements (O'Regan & Noë, 2001; Hayhoe & Ballard, 2005). The shared prediction is that a vision system without iterative redirection fails on problems humans solve routinely; **ActiveVision** tests this prediction for MLLMs.

Vision-centric benchmarks. MMVP (Tong et al., 2024b), MMStar (Chen et al., 2024), and CV-Bench (Tong et al., 2024a) exclude items solvable from text alone, and they are now largely saturated. BLINK (Fu et al., 2024), BlindTest (Rahmanzadehgervi et al., 2024), and BabyVision (Chen et al., 2026) show that models fall short of humans on simple visual tasks, without isolating which capability is missing. Application benchmarks such as OSWorld (Yuan et al., 2026) and BenchCAD (Zhang et al., 2026) bundle perception with coding and reasoning, so a failure cannot be attributed to perception. **ActiveVision** instead isolates one capability, *active observation*, with tasks that resist shortcuts from any single language description and are rendered photorealistically.

3 ActiveVision

ActiveVision collects tasks where a single one-shot glance is unlikely to suffice and the answer benefits from continuously returning to the pixels. §3.1 grounds the design in three elemental operations of human vision, following the definitions of three task families that instantiate them in Figure 2. §3.2 states the design principle every generator follows: discriminative visual state whose

information content exceeds what one language description can losslessly carry. §3.3 introduces each of the 17 tasks and describes the photorealistic rendering pipeline that turns each procedural scaffold into the natural-looking image models actually see.

3.1 Task Families: Grounding the Cognitive Foundations

ActiveVision is built around three *elemental operations*. Specifically, *exhaustive enumeration* beyond the subitizing range (Kaufman et al., 1949; Trick & Pylyshyn, 1994), *curve tracing* along contours (Jolicoeur et al., 1986; Roelfsema et al., 1998), and *fine-grained comparison* under visual working memory limits (Luck & Vogel, 1997; Ballard et al., 1995). Psychophysics and neurophysiology characterize these operations as serial, attention-demanding routines and are carried out *actively* through repeated shifts of gaze (Ullman, 1984; Roelfsema, 2005). We instantiate them below as three task families, each targeting a distinct dimension of active observation.

Distributed Scanning. The image contains many spatially distributed local signals, including dots, strokes, regions, and graph faces. Each must be found and accumulated. Difficulty scales with the signal count and how evenly it covers the canvas. The characteristic failure has two forms. In *partial coverage*, the model counts only five or six of ten items and stops before finishing the scan. In faulty *individuation*, it fails to recognize each signal as a discrete element or distinguish adjacent, similar signals. Items may then be merged, split, or confused with the background before counting begins.

Sequential Traversal. The image encodes a connected structure, such as an arrow chain, a tangled curve through colored regions, or a winding tube. The model must follow it step by step while tracking its current position, direction, and running tally. Difficulty scales with path length, crossing density, and the visual similarity of decoys to the correct next step. The characteristic failure is *gestalt interpolation*. The model guesses the endpoint from the start without traversing intermediate steps.

Visual Attribute Transfer. Fine-grained comparison across regions. The model extracts a visual property from a reference region. This may be length, curvature, thickness, color arrangement, dot pattern, or orientation. It then matches or compares that property against candidates elsewhere. Difficulty scales with the subtlety of the attribute distinction and candidate count. The typical failure is *prior substitution*. Instead of measuring both regions, the model applies a learned linguistic prior.

3.2 Task Design Principle: Forcing Iterative Perception

ActiveVision is built around a single design principle: every task instance carries discriminative visual state whose information content exceeds what a single language description can losslessly carry. An observer that compresses the image once (e.g., *"six red circles in the top-left, three blue squares in the bottom-right"*) and then reasons over that summary will, by construction, lose the information the answer depends on; solving the task requires keeping the image itself in the loop. The principle is realized via three properties as follows.

We realize this principle through three design properties:

- **Arbitrary positions.** Items are placed at continuous, sampled coordinates rather than on a grid or at named anchors. Twenty dots scattered over the canvas carry 190 pairwise spatial relations and twenty real-valued coordinate pairs, far beyond what any single linguistic summary can capture.
- **Arbitrary shapes.** Region boundaries, contours, blob silhouettes, and tile-motif outlines are synthesized fresh for every instance rather than drawn from a fixed library of named shapes. Each is either a closed contour whose radius is modulated by a random number of Fourier harmonics with sampled amplitudes and phases or a periodic spline through jittered ring waypoints. Consequently, the silhouette space is continuous and high-variance, and no two instances repeat a shape. Each boundary is defined by its exact polyline rather than by a named shape such as a triangle.
- **Arbitrary traces.** The routes observers must follow, including paths, arrow chains, tangled loops, and the connecting curve through color zones, are smooth random splines through sampled control points, with dozens of meaningful inflection points that no single description preserves.

Together, these properties make iterative perception the only natural solution. Each instance carries more visual state than a concise description can preserve, so the tasks remain readily tractable for humans who can revisit the image yet consistently difficult for a one-pass observer to solve reliably.

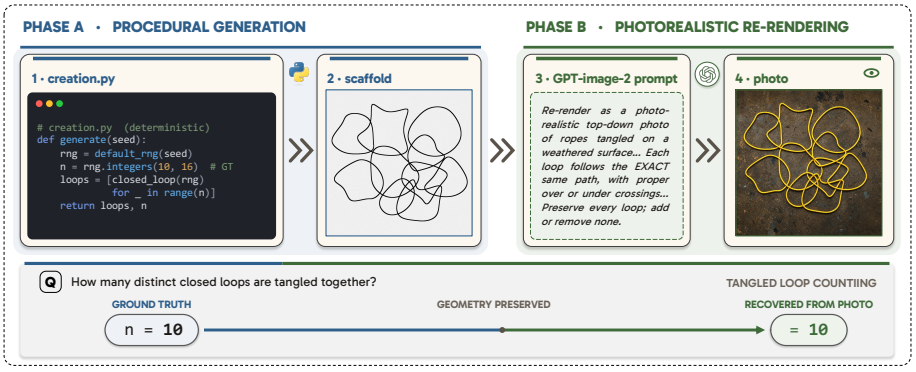


Figure 3: The four-stage generation pipeline, illustrated on *Tangled Loop Counting*. A deterministic Python generator (1) emits the geometric scaffold (2): random smooth closed curves with full ground truth. A task-specific GPT-image-2 prompt (3) re-renders the scaffold into a photorealistic image (4) without altering positions, counts, or topology. The question and ground-truth answer apply unchanged to both renderings; only the photorealistic image is served to the model at evaluation time.

3.3 Task Instantiation: From Synthetic Scaffolds to Real-world Images

We instantiate this design with 17 task generators grouped by the three families above. Each produces a synthetic scaffold, question, and ground-truth answer from a deterministic seed, making instances reproducible. Answers span broad ranges with flat per-task distributions, preventing the modal answer or task identity from serving as a reliable shortcut. Table 2 in Appendix B lists all 17 tasks.

These tasks are not merely abstract puzzles. Each isolates a perceptual operation that people use in everyday life and across practical settings. For example, *Region Counting* resembles counting countries on a political map or zones in a thermal scan, while *Signal Difference Spotting* mirrors comparing ECGs or reviewing seismic traces. Other tasks capture topological reasoning, path tracing, and attribute matching relevant to wiring inspection, floor-plan analysis, microscopy, manufacturing inspection, and satellite change detection.

We place these operations in realistic settings via a two-stage pipeline (Figure 3). Each task generator first produces a procedural Matplotlib scaffold with full ground truth attached. To make the images noisy and realistic rather than cartoon-clean, we re-render every scaffold with GPT-image-2 using a task-specific prompt that maps the procedural primitives to a real-world setting. Only the rendered image is shown to the model.

The pipeline serves three purposes. First, it removes the cartoon-input confound: the perceptual difficulty seen by models is dominated by the task's discriminative structure rather than by an unfamiliar rendering style. Second, the photorealistic regime is where classical computer-vision baselines degrade sharply, in line with their known brittleness on scanned documents, ultrasound speckle, low-resolution phone snapshots, and similar real-world inputs, so the same images that test MLLMs also bound what a tool-using script can recover. Third, the imagery resembles what downstream applications actually face, making the active-vision diagnosis externally meaningful rather than an artifact of toy renderings.

4 Benchmark Evaluation

We first report the headline accuracy of frontier MLLMs on **ActiveVision** (§4.1), then dive deeper with two ablation studies on the effect of reasoning effort (§4.2) and of agentic tool use (§4.3).

From each of the 17 generators, we sample a fixed evaluation split of $N = 5$ instances, yielding 85 items in total. Every instance passes through the photorealistic rendering pipeline described in §3.3. We use exact-match accuracy as the primary metric. Unless stated otherwise, evaluations use pure CoT with no external tools or code execution. Each item is presented as a single user message containing the question verbatim from the released manifest, followed by the image. No system prompt is used. The model is instructed to wrap its final answer in `<answer>` tags.

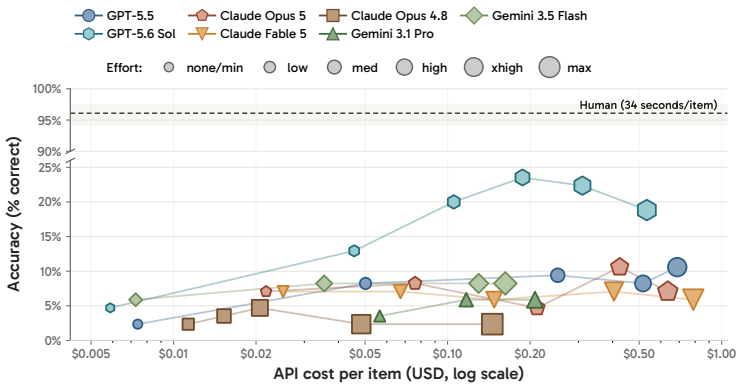


Figure 4: Accuracy versus API cost per item (log scale, public list prices) on the 85-item **ActiveVision** split. Points are complete pure-CoT runs, with marker size encoding the reasoning-effort tier; lines connect runs of the same model. The dashed line and band are the human mean and range (N=3).

We score the last such block by exact match after normalizing case, whitespace, and separators. The human baseline comprises three participants, each of whom completed the full 85-item split unaided through a self-paced web interface and was scored by the same criterion. Figure 4 reports the full reasoning-effort range for GPT-5.5 (OpenAI, 2026a), GPT-5.6 Sol (OpenAI, 2026b), Claude Opus 5 (Anthropic, 2026d), Claude Fable 5 (Anthropic, 2026c), Claude Opus 4.8 (Anthropic, 2026b), Gemini 3.1 Pro (Google DeepMind, 2026a), and Gemini 3.5 Flash (Google DeepMind, 2026b). Table 3 in Appendix C gives per-task scores for these models and Claude Opus 4.7 (Anthropic, 2026a) at their highest reasoning effort.

4.1 Main Results: Every Frontier Model Fails Most Tasks

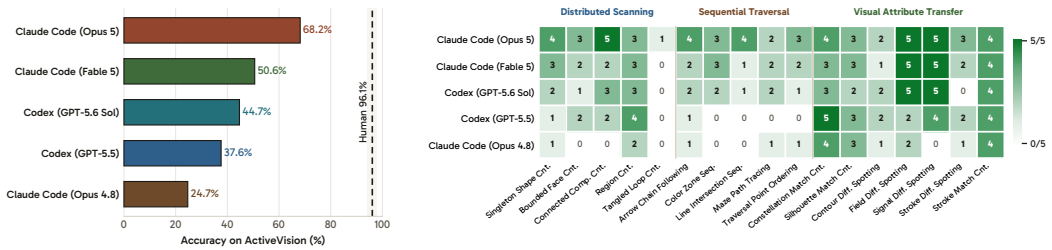
Figure 4 shows that every model, at every reasoning-effort tier, stays far below the three-participant human baseline. The highest-scoring model, GPT-5.6 Sol, reaches only 23.5% at its best tier, *high*, and 18.8% (16 of 85) at its highest tier. Claude Opus 5 and Claude Fable 5, despite their strong results on reasoning and coding leaderboards, solve just 7.1% and 5.9% at their highest tiers. At its highest tier, every model scores **zero** on at least 9 of the 17 tasks (Table 3). The human mean is 96.1%, about four times the best model's accuracy. The models have only weakly overlapping success sets. At their highest tiers, no item is solved by all eight models in Table 3, suggesting that the gap is not specific to a single model. As a shortcut check, a GPT-5.5 question-only control with the image omitted solves only 2 of 85 items, for an accuracy of 2.4%. This result matches the run with both image and question at the same *none* effort, indicating that prompt priors alone do not explain the reported performance.

4.2 Does More Reasoning Help?

The gap is not a deliberation deficit. Figure 4 plots accuracy against API cost per item for every pure-CoT run across reasoning-effort tiers. Every model stays far below the human baseline. Only GPT-5.6 Sol gains clearly from reasoning: it rises from 4.7% at *none* to 23.5% at *high*, then stays at 18.8–22.4% while its per-item cost nearly triples through *xhigh* and *max*. Scaling GPT-5.5's reasoning effort from *none* to *xhigh* increases its per-item cost nearly a hundredfold but improves accuracy only from 2.4% to 10.6%. For the Claude and Gemini models, differences between effort tiers are within sampling noise on 85 items: spending about 30× more per item on Fable 5 or Opus 5 does not make either more accurate. More reasoning helps only up to a point. The models do not run out of reasoning steps; they fail to pull the right visual evidence out of the image. Appendix D shows how: models undercount crowded scenes, lose ordered walks within the first steps, and default to "same" in fine-grained comparison.

4.3 Can Active Vision Be Substituted by Tool Use?

Our main result shows that frontier models lack active observation. We next ask whether computation can substitute for it. Code offers a strong substitution channel: brightness thresholding can segment



(a) Per-agent accuracy: even the strongest agent (Opus 5, 68.2%) stays far below the human band.

(b) Per-task accuracy (correct out of 5): attribute-transfer tasks largely yield to code, traversal resists the two weakest agents, and *Tangled Loop Counting* resists every agent but Opus 5.

Figure 5: Agentic tool use on the 85-item **ActiveVision** split. (a) Per-agent accuracy against the human reference (N=3); (b) per-task breakdown across the five agents.

Table 1: Accuracy and mean per-item resource use on the 85-item split. Parentheses retain counts and run-level details; dashes indicate quantities that do not apply.

Track	Accuracy (%)	Time/item (min)	API cost/item ($)	Tool calls per item	Output tokens per item	Agent turns per item
Codex (GPT-5.5)	37.6 (32/85)	12.5 (17.8 h total)	2.74 ($232.63 total)	34 (16 code, 18 image)	22.9k (1.95M total)	31 (max 84)
Codex (GPT-5.6 Sol)	44.7 (38/85)	6.7 (9.5 h total)	0.91 ($77.67 total)	21 (13 code, 9 image)	15.4k (1.31M total)	24 (max 104)
Claude Code (Opus 4.8)	24.7 (21/85)	14.7 (20.8 h total)	4.23 ($359.90 total)	53 (27 code, 26 image)	52.2k (4.43M total)	55 (max 115)
Claude Code (Opus 5)	68.2 (58/85)	12.1 (17.1 h total)	3.94 ($335.00 total)	45 (29 code, 16 image)	44.0k (3.74M total)	80 (max 191)
Claude Code (Fable 5)	50.6 (43/85)	13.9 (19.6 h total)	7.63 ($648.63 total)	39 (20 code, 19 image)	44.9k (3.82M total)	86 (max 198)
Human (N=3)	96.1 (94.1–97.6)	0.56 (31–69 min total)	—	0	—	—

regions, while connected-component analysis can return their count, bypassing direct visual inspection. We test this by running five autonomous coding agents per item: Codex with GPT-5.5 and GPT-5.6 Sol, and Claude Code with Opus 4.8, Opus 5, and Fable 5. All operate at *xhigh* effort in fresh sandboxes containing only the image and question.

4.3.1 How Far Does Substitution Carry?

Agentic tools improve performance substantially but do not close the human gap (Figure 5a). The Opus 5 agent solves **58 of 85 items (68.2%)**, Fable 5 **43 (50.6%)**, Codex with GPT-5.6 Sol **38 (44.7%)**, Codex with GPT-5.5 **32 (37.6%)**, and the Opus 4.8 agent **21 (24.7%)**, compared with at most 23.5% for any tool-free run. At the same *xhigh* effort, code lifts Opus 5 from 10.6% to 68.2% and GPT-5.6 Sol from 22.4% to 44.7%. The gains are highly task dependent (Figure 5b). All five agents perform best on **Visual Attribute Transfer** (43–74%), where template cropping and measurement admit clean algorithmic reductions. **Sequential Traversal** is harder: the agents solve 1 to 16 of its 25 items, and only Opus 5 solves more than half. Failure is not limited to traversal; four of the five agents score zero on *Tangled Loop Counting*, and Opus 5 solves one of its five items. Overall, tool use helps most when perception can be reduced to a reliable computation. When no such reduction is available, the gap persists.

4.3.2 Why Does Substitution Stop There?

Figure 6 shows the two recurring failure modes. First, tools are not robust on realistic textures (Figure 6, top). On *Bounded Face Counting*, the agent's color mask dissolves into the cobblestone texture, so nodes and edges are mis-detected and its Euler-formula count comes out wrong (12 vs. a

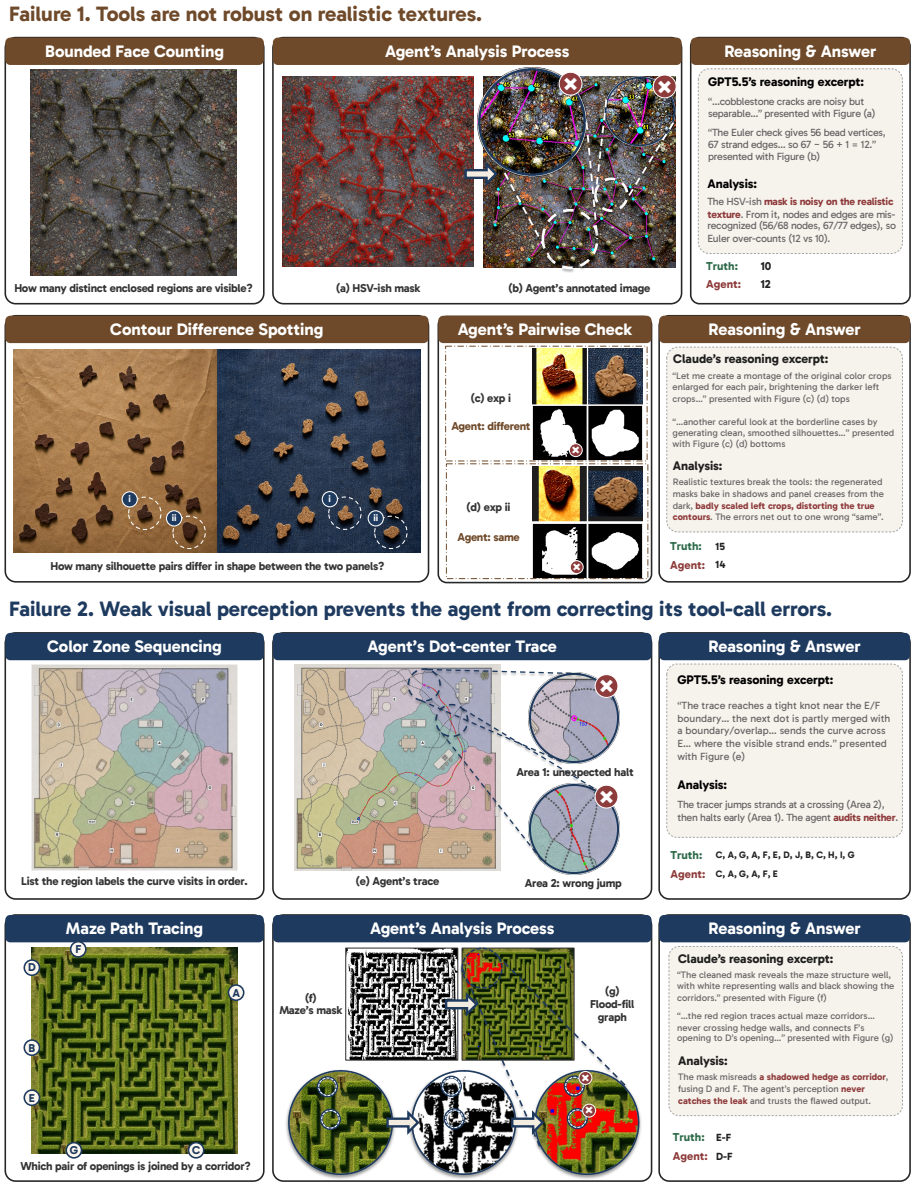


Figure 6: Four examples illustrating two failure modes of agentic tool use, shown through the agents' own intermediate artifacts and verbatim rollout excerpts. For each failure mode, the upper example is from Codex and the lower example is from Claude Code.

truth of 10); on *Contour Difference Spotting*, shadows and paper creases bake into the regenerated silhouettes and a badly scaled crop distorts the true contours, netting out to one wrong "same."

Second, weak perception cannot catch the tool's errors (Figure 6, bottom). On *Color Zone Sequencing*, the agent's tracer jumps onto a crossing strand and halts early, and the agent audits neither error, submitting the truncated route; on *Maze Path Tracing*, the binarized mask reads a shadowed hedge as corridor and fuses two openings, and the agent trusts the flood-fill's incorrect path. In both cases, a glance at the overlay it had just produced would falsify the answer. In short, a model that cannot see cannot tell when its code is wrong.

4.3.3 How Costly Is Substitution?

Substitution is also expensive and slow. Table 1 reports the per-item cost of the five tracks: each item costs an agent \$0.91–\$7.63 of compute and 7–15 minutes of wall-clock time, spent across dozens of

tool calls and tens of thousands of output tokens, about 12 to 26 times longer than the half minute a human needs, plus compute a human never spends. Even the strongest agent, Opus 5, solves only 68.2% of items at $3.94 and 12 minutes per item; the human participants answer 96.1% with no tools and no code.

5 Discussion

Tool-use gains and remaining bottlenecks. **ActiveVision** evaluates models as pure visual observers. The five agentic tool use runs (§4.3) show why this distinction matters. Classical computer vision primitives such as `findContours`, Canny, and template matching work best on clean inputs. On **ActiveVision**'s photorealistic images, they fragment contours, merge objects, and fail under variations in style. A human can look back at the image and recognize that two masks belong to the same shape or that a dashed route continues beyond where the tracer stopped, while the script may return a confident but incorrect answer. Agentic tools therefore depend on active perception for robustness rather than replacing it. This challenge is greater in real deployments, where medical scans, satellite images, phone snapshots, and scanned documents contain more varied and less predictable noise than fixed pipelines can anticipate.

Scope of the agentic evaluation. The agentic ablation instead measures tool-orchestrated visual problem solving: decomposing a visual question into checkable operations, writing extraction code that works on noisy imagery, and recognizing when an output requires visual verification. In effect, the agent offloads iterative visual reasoning to code and reduces verification to simpler visual checks, a lower bar than solving the task through vision alone. Even so, the strongest agent remains far less accurate and efficient than an unaided human. Because **ActiveVision** is organized around cognitive axes rather than real-world scenarios, fixed pipelines transfer unevenly across tasks. Figure 5b shows this split. The attribute transfer tasks are largely solved, while the traversal tasks remain near zero for the two weakest agents, and every agent but one fails *Tangled Loop Counting*. We therefore report the agentic evaluation as an ablation rather than a standalone track and leave a dedicated benchmark for agentic visual problem solving to future work.

Outlook. **ActiveVision** reveals and quantifies a wide gap between human and machine perception that current vision-language benchmarks do not surface, turning "active vision" from a rhetorical claim into a number that future architectures and training objectives can be held to. The gap matters beyond the benchmark because the three operations it isolates (§3.1) are the documented substance of high-stakes visual work. Exhaustive scanning underlies radiology, cell counting, inventory inspection, and aerial-image search. In chest radiography, about 30% of missed nodules are never fixated at all (Kundel et al., 1978). Sequential tracing is central to connectomics, where proofreaders follow neurites through electron microscopy volumes (Scheffer et al., 2020), as well as to vessel and schematic tracing. Fine-grained comparison underlies latent-print examination, pathology, and industrial quality control (Busey et al., 2011; Krupinski et al., 2006; Drury, 1992). Eye-tracking across these professions shows that expertise largely *is* knowing where and how to look (Gegenfurtner et al., 2011). A model that cannot count, trace, or compare reliably in a controlled image has no path to reliability in these settings; closing the gap is a prerequisite for the broader goal that motivates this work: reliable MLLMs and the multimodal agents built on top of them. We discuss the limitations of this work in Appendix E.

6 Conclusion

We introduce **ActiveVision**, a benchmark that isolates active observation: the ability to keep returning to an image, forming and checking hypotheses as reasoning unfolds. Humans solve nearly all items in about half a minute each, while frontier MLLMs fail most items. More reasoning helps only up to a point, and tools help unevenly: even the strongest agent solves only about two-thirds of the items, at far greater time and cost than a human. Closing this gap is a prerequisite for applications such as robotics, healthcare, manufacturing, and scientific discovery, and **ActiveVision** turns it into a number that future models can be held to.

AI Use Statement

Generative AI tools served only as assistants in this work. We used them to help edit the writing of this paper and to help write and debug analysis and plotting code. In every case, the authors reviewed, verified, and finalized the result, and the authors made all research decisions. Generative models also appear in the benchmark itself: GPT-image-2 re-renders each procedurally generated scaffold into a photorealistic image (§3.3), and the evaluated MLLMs and coding agents are the systems under test. All ground-truth answers come from the deterministic generators, not from a generative model. The authors take full responsibility for the content of this work, including any text, code, or artifacts produced with AI assistance.

Ethics Statement

The human baseline comes from three adult participants who solved the 85 items through a self-paced web interface; the study collected only answers and response times, and no personal data. The benchmark images show synthetic scenes produced from procedural scaffolds and contain no real people or private data.

Reproducibility Statement

Section 3 defines the tasks and the four-stage generation pipeline, and Section 4 gives the evaluation protocol, answer parsing, and scoring. The anonymized supplementary material will contain the 85-item evaluation split and the evaluation and scoring code.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

John Aloimonos, Isaac Weiss, and Amit Bandyopadhyay. Active vision. *International journal of computer vision*, 1(4):333–356, 1988.

Anthropic. Introducing claude opus 4.7. https://www.anthropic.com/news/claude-opus-4-7, 2026a. Accessed: 2026-09-24.

Anthropic. Introducing claude opus 4.8. https://www.anthropic.com/news/claude-opus-4-8, 2026b. Accessed: 2026-09-24.

Anthropic. Introducing claude fable 5 and claude mythos 5. https://www.anthropic.com/news/claude-fable-5-mythos-5, 2026c. Accessed: 2026-09-24.

Anthropic. Introducing claude opus 5. https://www.anthropic.com/news/claude-opus-5, 2026d. Accessed: 2026-09-24.

Ruzena Bajcsy. Active perception. *Proceedings of the IEEE*, 76(8):966–1005, 1988.

Dana H. Ballard. Animate vision. *Artificial Intelligence*, 48(1):57–86, 1991.

Dana H. Ballard, Mary M. Hayhoe, and Jeff B. Pelz. Memory representations in natural tasks. *Journal of Cognitive Neuroscience*, 7(1):66–80, 1995.

Thomas Busey, Chen Yu, Dean Wyatte, John Vanderkolk, Francisco Parada, and Ruj Akavipat. Consistency and variability among latent print examiners as revealed by eye tracking methodologies. *Journal of Forensic Identification*, 61(1):60–91, 2011.

Liang Chen, Weichu Xie, Yiyan Liang, Hongfeng He, Hans Zhao, et al. Babyvision: Visual reasoning beyond language. *arXiv preprint arXiv:2601.06521*, 2026.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*, 2024.

ChildBench Team. Easy for children, hard for ai: The limits of mllms in early learning tasks recognition. https://github.com/Jderder/ChildBench, 2025.

Colin G. Drury. Inspection performance. In Gavriel Salvendy (ed.), *Handbook of Industrial Engineering*, pp. 2282–2314. Wiley, 2nd edition, 1992.

John M. Findlay and Iain D. Gilchrist. *Active Vision: The Psychology of Looking and Seeing*. Oxford University Press, 2003.

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. *arXiv preprint arXiv:2404.12390*, 2024.

Andreas Gegenfurtner, Erno Lehtinen, and Roger Säljö. Expertise differences in the comprehension of visualizations: A meta-analysis of eye-tracking research in professional domains. *Educational Psychology Review*, 23(4):523–552, 2011.

Google DeepMind. Gemini 3.1 pro: A smarter model for your most complex tasks. https://blog.google/innovation-and-ai/models-and-research/gemini-models/gemini-3-1-pro/, 2026a. Accessed: 2026-09-24.

Google DeepMind. Gemini 3.5: Frontier intelligence with action. https://blog.google/innovation-and-ai/models-and-research/gemini-models/gemini-3-5/, 2026b. Accessed: 2026-09-24.

Mary Hayhoe and Dana Ballard. Eye movements in natural behavior. *Trends in Cognitive Sciences*, 9 (4):188–194, 2005.

Xiangru Jian, Hao Xu, Wei Pang, Xinjian Zhao, Chengyu Tao, Qixin Zhang, Xikun Zhang, Chao Zhang, Guanzhi Deng, Alex Xue, Juan Du, Tianshu Yu, Garth Tarr, Linqi Song, Qiuzhuang Sun, and Dacheng Tao. FORGE: Fine-grained multimodal evaluation for manufacturing scenarios. *arXiv preprint arXiv:2604.07413*, 2026.

Mohan Jiang, Jin Gao, Jiahao Zhan, and Dequan Wang. MAC: A live benchmark for multimodal large language models in scientific understanding. *arXiv preprint arXiv:2508.15802*, 2025.

Pierre Jolicoeur, Shimon Ullman, and Marilynne Mackay. Curve tracing: A possible basic operation in the perception of spatial relations. *Memory & Cognition*, 14(2):129–140, 1986.

Oğuzhan Fatih Kar, Alessio Tonioni, Petra Poklukar, Achin Kulshrestha, Amir Zamir, and Federico Tombari. BRAVE: Broadening the visual encoding of vision-language models. *arXiv preprint arXiv:2404.07204*, 2024.

Siddharth Karamcheti, Suraj Nair, Ashwin Balakrishna, Percy Liang, Thomas Kollar, and Dorsa Sadigh. Prismatic VLMs: Investigating the design space of visually-conditioned language models. *arXiv preprint arXiv:2402.07865*, 2024.

E. L. Kaufman, M. W. Lord, T. W. Reese, and J. Volkmann. The discrimination of visual number. *The American Journal of Psychology*, 62(4):498–525, 1949.

Maksim Kolodiazhnyi, Denis Tarasov, Dmitrii Zhemchuzhnikov, Alexander Nikulin, Ilya Zisman, Anna Vorontsova, Anton Konushin, Vladislav Kurenkov, and Danila Rukhovich. cadrille: Multi-modal CAD reconstruction with reinforcement learning. In *International Conference on Learning Representations (ICLR)*, 2026. arXiv:2505.22914.

Elizabeth A. Krupinski, Allison A. Tillack, Lynne Richter, Jeffrey T. Henderson, Achyut K. Bhattacharyya, Katherine M. Scott, Anna R. Graham, Michael R. Descour, John R. Davis, and Ronald S. Weinstein. Eye-movement study and human performance using telepathology virtual slides: Implications for medical education and differences with experience. *Human Pathology*, 37(12): 1543–1556, 2006.

Harold L. Kundel, Calvin F. Nodine, and Dennis Carmody. Visual scanning, pattern recognition and decision-making in pulmonary nodule detection. *Investigative Radiology*, 13(3):175–181, 1978.

Jon M. Laurent, Joseph D. Janizek, Michael Ruzo, Michaela M. Hinks, Michael J. Hammerling, Siddharth Narayanan, Manvitha Ponnapati, Andrew D. White, and Samuel G. Rodriques. LAB-Bench: Measuring capabilities of language models for biology research. *arXiv preprint arXiv:2407.10362*, 2024.

Steven J. Luck and Edward K. Vogel. The capacity of visual working memory for features and conjunctions. *Nature*, 390(6657):279–281, 1997.

Yulin Luo, Chun-Kai Fan, Menghang Dong, Jiayu Shi, Mengdi Zhao, Bo-Wen Zhang, Cheng Chi, Jiaming Liu, Gaole Dai, Rongyu Zhang, Ruichuan An, Kun Wu, Zhengping Che, Shaoxuan Xie, Guocai Yao, Zhongxia Zhao, Pengwei Wang, Guang Liu, Zhongyuan Wang, Tiejun Huang, and Shanghang Zhang. RoboBench: A comprehensive evaluation benchmark for multimodal large language models as embodied brain. *arXiv preprint arXiv:2510.17801*, 2025.

Brandon McKinzie, Zhe Gan, Jean-Philippe Fauconnier, Sam Dodge, Bowen Zhang, Philipp Dufter, Dhruti Shah, Xianzhi Du, Futang Peng, Floris Weers, et al. Mm1: Methods, analysis & insights from multimodal llm pre-training. *arXiv preprint arXiv:2403.09611*, 2024.

OpenAI. Introducing gpt-5.5. https://openai.com/index/introducing-gpt-5-5/, 2026a. Accessed: 2026-09-24.

OpenAI. GPT-5.6: Frontier intelligence that scales with your ambition. https://openai.com/index/gpt-5-6/, 2026b. Accessed: 2026-09-24.

OpenAI. ChatGPT images 2.0 (GPT-image-2). https://openai.com/index/introducing-chatgpt-images-2-0/, 2026c.

J. Kevin O'Regan and Alva Noë. A sensorimotor account of vision and visual consciousness. *Behavioral and Brain Sciences*, 24(5):939–973, 2001.

Lukas Petersson, Axel Backlund, Axel Wennström, Hanna Petersson, Callum Sharrock, and Arash Dabiri. Blueprint-Bench: Comparing spatial intelligence of LLMs, agents and image models. *arXiv preprint arXiv:2509.25229*, 2025.

Pooyan Rahmanzadehgervi, Logan Bolton, Mohammad Reza Taesiri, and Anh Totti Nguyen. Vision language models are blind. *arXiv preprint arXiv:2407.06581*, 2024.

Pieter R. Roelfsema. Elemental operations in vision. *Trends in Cognitive Sciences*, 9(4):226–233, 2005.

Pieter R. Roelfsema, Victor A. F. Lamme, and Henk Spekreijse. Object-based attention in the primary visual cortex of the macaque monkey. *Nature*, 395(6700):376–381, 1998.

Louis K. Scheffer, C. Shan Xu, Michal Januszewski, et al. A connectome and analysis of the adult Drosophila central brain. *eLife*, 9:e57443, 2020.

Min Shi, Fuxiao Liu, Shihao Wang, Shijia Liao, Subhashree Radhakrishnan, De-An Huang, Hongxu Yin, Karan Sapra, Yaser Yacoob, Humphrey Shi, et al. Eagle: Exploring the design space for multimodal llms with mixture of encoders. *arXiv preprint arXiv:2408.15998*, 2024.

Muzi Tao and Saining Xie. What does a visual formal analysis of the world's 500 most famous paintings tell us about multimodal llms? In *The Second Tiny Papers Track at ICLR 2024*, 2024.

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *arXiv preprint arXiv:2406.16860*, 2024a.

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

Lana M. Trick and Zenon W. Pylyshyn. Why are small and large numbers enumerated differently? A limited-capacity preattentive stage in vision. *Psychological Review*, 101(1):80–102, 1994.

Shimon Ullman. Visual routines. *Cognition*, 18(1–3):97–159, 1984.

An Vo, Khai-Nguyen Nguyen, Mohammad Reza Taesiri, Vy Tuong Dang, Anh Totti Nguyen, and Daeyoung Kim. Vision language models are biased. *arXiv preprint arXiv:2505.23941*, 2025.

Zirui Wang, Mengzhou Xia, Luxi He, Howard Chen, Yitao Liu, Richard Zhu, Kaiqu Liang, Xindi Wu, Haotian Liu, Sadhika Malladi, Alexis Chevalier, Sanjeev Arora, and Danqi Chen. CharXiv: Charting gaps in realistic chart understanding in multimodal LLMs. In *Advances in Neural Information Processing Systems (NeurIPS) – Datasets and Benchmarks Track*, volume 37, pp. 113569–113697, 2024. URL https://arxiv.org/abs/2406.18521.

Alfred L. Yarbus. *Eye Movements and Vision*. Plenum Press, New York, 1967.

Hengwei Ye, Yuanting Guan, Yuxuan Ge, Tianying Zhu, Zhenhan Guan, Yijia Zhong, et al. Children's intelligence tests pose challenges for mllms? kidgym: A 2d grid-based reasoning benchmark for mllms. *arXiv preprint arXiv:2603.20209*, 2026.

Mengqi Yuan, Zilong Zhou, Xinzhuang Xiong, Weiming Wu, Jiayang Sun, Jiamin Song, Tianbao Xie, Tao Yu, et al. OSWorld2.0: Benchmarking computer use agents on long-horizon real-world tasks. *arXiv preprint arXiv:2606.29537*, 2026.

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, Yu Su, Wenhu Chen, and Graham Neubig. MMMU-Pro: A more robust multi-discipline multimodal understanding benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15050–15070. Association for Computational Linguistics, 2025. URL https://aclanthology.org/2025.acl-long.736/.

Haozhe Zhang, Kaichen Liu, Miaomiao Chen, Lei Li, Shaojie Yang, Cheng Peng, and Hanjie Chen. BenchCAD: A comprehensive, industry-standard benchmark for programmatic CAD. *arXiv preprint arXiv:2605.10865*, 2026.

Jiarui Zhang, Ollie Liu, Tianyu Yu, Jinyi Hu, and Willie Neiswanger. Euclid: Supercharging multimodal llms with synthetic high-fidelity visual descriptions. *arXiv preprint arXiv:2412.08737*, 2024.

Jiarui Zhang, Mahyar Khayatkhoei, Prateek Chhikara, and Filip Ilievski. MLLMs know where to look: Training-free perception of small visual details with multimodal llms. *arXiv preprint arXiv:2502.17422*, 2025.

A Extended Related Work

ActiveVision draws on three lines of work: a long-running argument from cognitive science and computer vision that perception is intrinsically active, evidence that MLLMs struggle with basic visual perception, and existing benchmarks constructed to stress vision rather than language. We elaborate on each below.

A.1 Cognitive Foundations of Active Vision

Eye Movements and Vision (Yarbus, 1967) showed that the same painting elicits different scanpaths under different task prompts, establishing that gaze is allocated by the task rather than the image alone. Later works (Aloimonos et al., 1988; Bajcsy, 1988) formalized the consequence in computer vision: inverse problems such as shape from shading, structure from motion, and optical flow are ill-posed for a passive observer but become well-posed once the observer can actively control its sensors as part of inference.

Animate vision (Ballard, 1991) reframed the same argument in terms of computational economy, using gaze-targeted local computation rather than global computation. The *sensorimotor account* (O'Regan & Noë, 2001) advanced the strongest version of this position, arguing that seeing *is* the mastery of how visual input changes under one's own movements. Eye-tracking of everyday tasks supports this account beyond the laboratory (Hayhoe & Ballard, 2005; Findlay & Gilchrist, 2003). The shared prediction is sharp: without iterative sensor redirection, a vision system will fail on perceptual problems that humans solve routinely. **ActiveVision** operationalizes that prediction for multimodal large language models.

A.2 Vision-Centric Benchmarks for MLLMs

A first line of work set out to isolate visual perception from linguistic priors. Early diagnoses found frontier models leaning on memorized priors and overlooking queried visual details (Vo et al., 2025; Zhang et al., 2025), and benchmarks such as MMVP (Tong et al., 2024b), MMStar (Chen et al., 2024), and CV-Bench (Tong et al., 2024a) were built to exclude items solvable from text alone and reward reading the image itself, alongside a parallel push on vision-centric model design (McKinzie et al., 2024; Karamcheti et al., 2024; Shi et al., 2024; Kar et al., 2024; Zhang et al., 2024). These benchmarks are now largely saturated, indicating that current MLLMs can often read the image when linguistic shortcuts are controlled, while leaving their remaining perceptual limitations unresolved.

A second line instead targets the human–model gap directly. BLINK (Fu et al., 2024) collects tasks a person answers in the blink of an eye, BlindTest (Rahmanzadehgervi et al., 2024) poses questions so simple that a sighted human never misses them, ArtQA (Tao & Xie, 2024) asks painting questions that require looking, and BabyVision (Chen et al., 2026), ChildBench (ChildBench Team, 2025), and KidGym (Ye et al., 2026) draw on tasks an infant or child can already solve. What these share is a framing by *human-model gap* alone: they establish that models fall short of people without isolating *which* capability is missing; we recognize active vision as a major part of that missing capability, one that deserves to be measured on its own terms.

A third line is application-driven: examples include OSWorld (Yuan et al., 2026) for real-world computer-use tasks and BenchCAD (Zhang et al., 2026) for parametric CAD generation. These usefully track practical deployment value, but each task bundles coding, reasoning, and visual perception together, so a failure cannot be attributed to perception. **ActiveVision** isolates exactly this capability. It targets *active observation*, the iterative looking loop that prior benchmarks do not explicitly name. Its tasks resist shortcuts from any single language description and are rendered photorealistically to resemble real-world inputs encountered in downstream applications.

B TASK LIST

Table 2 lists the 17 tasks of **ActiveVision** with a one-line description of each.

Table 2: The 17 task generators in **ActiveVision**, grouped by cognitive axis. Each row gives the task name and a one-line description of what the model is asked to do.

Distributed Scanning	
Bounded Face Counting	Count bounded faces in a planar graph drawing.
Connected Component Counting	Count connected components in a graph of dots and edges.
Region Counting	Count separated color regions inside a Voronoi-like partition.
Singleton Shape Counting	Count blob silhouettes that appear exactly once among many duplicates.
Tangled Loop Counting	Count distinct closed loops in a tangle of curves.
Sequential Traversal	
Arrow Chain Following	Follow the chain from a marked arrow; report the labeled terminus.
Traversal Point Ordering	List the labeled points encountered along the curve.
Color Zone Sequencing	List, in order, the color regions a curve visits.
Line Intersection Sequencing	List, in order, the labels of the curves crossed by the target curve.
Maze Path Tracing	Identify which pair of border openings is connected by a path.
Visual Attribute Transfer	
Constellation Match Counting	Count copies of a template dot-pattern in a larger field.
Silhouette Match Counting	Count field silhouettes that are exact translated copies of a template.
Stroke Match Counting	Count field strokes that are exact translated copies of a template stroke.
Contour Difference Spotting	Count silhouette pairs whose shape differ between two paired panels.
Field Difference Spotting	List the indexed cells that differ between two paired field-grid panels.
Signal Difference Spotting	List the indexed segments that differ between two paired 1-D-signal panels.
Stroke Difference Spotting	Count stroke pairs whose shape differ between two paired panels.

C PER-TASK RESULTS

Table 3 gives the per-task scores of every model at its highest reasoning effort.

Table 3: Per-task exact-match accuracy on **ActiveVision**. Model entries report correct responses out of five at each model's highest reasoning effort. Human entries report the mean across three participants on the same scale (e.g., 14/15 is reported as 4.7/5). The Overall row reports scores out of 85; individual human accuracies are 97.6%, 96.5%, and 94.1%.

Task	GPT		Claude				Gemini		Human
	5.5 *xhigh*	5.6 Sol *max*	Opus 4.7 *max*	Opus 4.8 *max*	Opus 5 *max*	Fable 5 *max*	3.1 Pro *high*	3.5 Flash *high*	avg over $N=3$
Distributed Scanning									
Bounded Face Counting	0/5	0/5	1/5	1/5	1/5	1/5	0/5	0/5	5/5
Connected Component Counting	0/5	0/5	0/5	0/5	0/5	0/5	1/5	1/5	4.7/5
Region Counting	2/5	1/5	0/5	0/5	2/5	0/5	2/5	1/5	5/5
Singleton Shape Counting	1/5	1/5	0/5	0/5	0/5	1/5	0/5	1/5	4/5
Tangled Loop Counting	0/5	0/5	0/5	0/5	0/5	0/5	0/5	0/5	5/5
Sequential Traversal									
Arrow Chain Following	1/5	0/5	0/5	0/5	2/5	1/5	1/5	2/5	5/5
Traversal Point Ordering	0/5	2/5	0/5	0/5	0/5	0/5	0/5	0/5	5/5
Color Zone Sequencing	0/5	0/5	0/5	0/5	0/5	0/5	0/5	0/5	5/5
Line Intersection Sequencing	0/5	1/5	0/5	0/5	0/5	0/5	0/5	0/5	5/5
Maze Path Tracing	0/5	1/5	1/5	0/5	0/5	1/5	0/5	0/5	5/5
Visual Attribute Transfer									
Constellation Match Counting	3/5	5/5	0/5	0/5	1/5	0/5	0/5	0/5	4.3/5
Silhouette Match Counting	1/5	3/5	0/5	0/5	0/5	0/5	0/5	1/5	5/5
Stroke Match Counting	1/5	2/5	0/5	0/5	0/5	0/5	1/5	1/5	4.7/5
Contour Difference Spotting	0/5	0/5	1/5	0/5	0/5	0/5	0/5	0/5	5/5
Field Difference Spotting	0/5	0/5	0/5	0/5	0/5	0/5	0/5	0/5	4.3/5
Signal Difference Spotting	0/5	0/5	0/5	1/5	0/5	0/5	0/5	0/5	4.7/5
Stroke Difference Spotting	0/5	0/5	1/5	0/5	0/5	1/5	0/5	0/5	5/5
Overall	**9/85** (10.6%)	**16/85** (18.8%)	**4/85** (4.7%)	**2/85** (2.4%)	**6/85** (7.1%)	**5/85** (5.9%)	**5/85** (5.9%)	**7/85** (8.2%)	**81.7/85** (96.1%)
Avg output tokens / time	22.5k	26.4k	9.0k	5.5k	25.1k	15.4k	16.8k	17.5k	33.6 s

D ADDITIONAL ANALYSIS: WHAT DO MODELS DO WHEN THEY FAIL?

Pooling the 29 pure-CoT runs of the evaluation sweep, we examine the response behavior of each task family: what the models count, how far they trace, and when they declare a difference.

Counting falls behind as scenes grow crowded. Figure 7 plots predicted count against ground-truth count on the eight counting tasks, with a least-squares fit per model. All seven fitted slopes are below one (0.44–0.91), so the fits drop further below the perfect-counting diagonal as the true count grows. The models are conservative counters: the more there is to count, the larger the share left uncounted, as if the model glimpses the image rather than scanning it exhaustively.

Tracing loses its way at the very start. Figure 8 shows, for the three ordered-walk tasks, the share of walks whose first k steps are all correct. Survival collapses within the first one or two moves, and only 10 of the 435 walks are completed exactly, all by GPT-5.6 Sol. The models cannot reliably trace along a structure in the image; failure is not gradual drift late in a long walk but the loss of the walk's frame at its very beginning.

"Same" becomes the safe default in visual comparison. Table 4 reports error rates for every run on the two panel-naming difference tasks: the share of real differences missed and the share of identical panels falsely flagged. Miss rates are high across the board while false alarms stay low, and the highlighted runs answer "none" on nearly every item, missing all real differences. Unable to resolve the fine-grained comparison, the models fall back on "same" because it is the safe answer, a response bias rather than an act of perception.

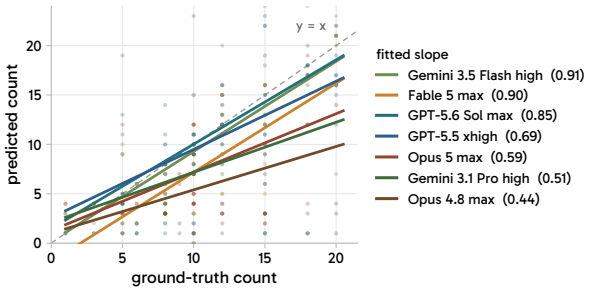


Figure 7: Predicted versus ground-truth counts across eight counting tasks at each model's highest reasoning effort. Solid lines are least-squares fits (slopes in the legend); the gray dashed line is the identity $y = x$. All slopes are below one, so undercounting grows with the true count.

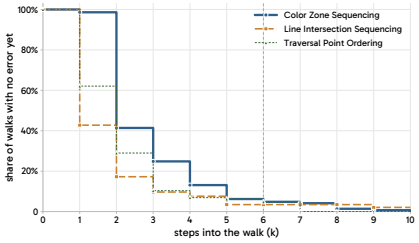


Figure 8: Prefix survival on three ordered-walk tasks (435 walks over 29 pure-CoT runs): the share of walks whose first k steps are all correct. Dashed line: shortest walk length (k=6). Only 10 walks, all from GPT-5.6 Sol, are completed exactly.

Table 4: Miss rate / false-alarm rate (%) on the two panel-naming difference tasks, per model and reasoning-effort tier. A miss is a real difference left unlisted; a false alarm is an identical cell or segment listed as different. For Gemini 3.5 Flash the *none* column is its *minimal* tier. Highlighted runs answer "none" on nearly every item, missing all real differences.

Model	*none*	*low*	*med*	*high*	*xhigh*	*max*
GPT-5.5	73 / 21	96 / 0	100 / 0	100 / 0	100 / 0	—
GPT-5.6 Sol	63 / 10	88 / 1	94 / 1	85 / 5	92 / 0	92 / 0
Claude Opus 4.8	—	—	—	—	—	81 / 17
Claude Opus 5	—	87 / 6	85 / 10	83 / 14	81 / 13	87 / 13
Claude Fable 5	—	96 / 12	88 / 12	85 / 9	77 / 6	88 / 6
Gemini 3.1 Pro	—	96 / 6	96 / 3	83 / 13	—	—
Gemini 3.5 Flash	73 / 13	94 / 5	87 / 9	90 / 12	—	—

E LIMITATIONS

Our images are synthetic even with photorealistic re-rendering: GPT-image-2 produces realistic-looking outputs from a controlled prompt rather than samples of the natural-image distribution. We accept this in exchange for exact ground truth and controlled task structure. External validity therefore rests on the elemental visual operations the tasks measure, including scanning, tracing, and comparing, rather than on the renderings themselves (§3.1). Furthermore, our tasks are built so that no short language description carries the answer; as models get better at describing images, this property can erode, and future versions of the benchmark would need to tighten the tasks to restore it.