CLOSING THE PERCEPTION-ACTION LOOP WITH HUMAN-ALIGNED VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-language models (VLMs) have emerged as model systems for studying human visual cognition. However, their visual capabilities are primarily based on processing high-resolution images of full scenes, which is in sharp contrast to how humans process visual information. Human vision is inherently dynamic, relying heavily on eye movements to bring the high-resolution fovea onto different parts of a scene, so each fixation provides only a partial view. Here, we asked whether VLMs under similar partial observation constraints could efficiently choose where to look and use the information they acquired to solve visual question answering (VQA) and visual search (VS) tasks, and whether their behavior would resemble that of human participants. We found that, unlike humans, VLMs made poor use of successive fixations; their performance plateaued after the first few glimpses, and they frequently chose to answer prematurely. Fine-tuning on human fixation trajectories not only helped the models ground their answers in the available visual information across fixations, but also made their internal representations more predictive of neural responses in the primate brain while giving rise to cross-fixation evidence accumulation reminiscent of signals observed in the primate parietal cortex. Altogether, we show that pretrained VLMs largely lack the ability to engage in active vision, but that targeted fine-tuning on human behavior gives them a human-like perception–action loop and brings their representations into closer alignment with the brain. These properties make them strong candidates for studying the neural computations that underpin active vision in humans and non-human primates.

1 INTRODUCTION

Artificial neural networks trained to recognize objects have been used as models of the primate ventral visual stream for more than a decade, beginning with convolutional neural networks and later extending to vision transformers (Yamins & DiCarlo, 2016; Schrimpf et al., 2020; Gokce & Schrimpf, 2024). Representations across layers in these models replicated various aspects of the neural responses along the ventral visual stream (Yamins et al., 2014; Khaligh-Razavi & Kriegeskorte, 2014). More recently, VLMs have emerged as model systems for studying human visual cognition, reproducing various behavioral signatures, including differences between feature and conjunctive search (Campbell et al., 2024) and interference from task-irrelevant information in selective attention tasks (Wang et al., 2025). Importantly, the majority of the past work has focused on probing the visual capacity of these models in static settings, where they process high-resolution images of full scenes and are asked to respond in one shot.

Vision in humans and other visual animals, by contrast, is inherently dynamic as processing the entire visual field at high resolution would place heavy demands on limited energy and neural resources. Instead, the retina provides high resolution only near the center of gaze, with resolution dropping progressively toward the periphery. As a result, each fixation provides only a partial view of the scene, and eye movements determine which information becomes available next. Therefore, perception depends on an ongoing interaction between seeing and acting, in which the gathered information guides where to look next, and each new fixation shapes what is perceived.

This sequential view of perception naturally connects to VLMs' capacity to process sequences of visual observations (Lin et al., 2024; Bai et al., 2025b; Jiang et al., 2024; Li et al., 2024; Laurençon

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.

et al., 2024). This raises the question of whether they can also serve as models of human active vision and whether these models can use successive foveated observations to decide where to look next and when they have gathered enough information to answer. Here, we investigate these questions in Visual Question Answering (VQA) and Visual Search (VS), examining whether models' sampling behavior resembles that of humans and whether fine-tuning on human behavior improves this correspondence.

Our contributions are:

- We benchmark a range of VLMs on their ability to perform active vision tasks across multiple VQA and VS datasets. Comparing their performance across fixations and exploration patterns with those of humans, we find that models answer prematurely and benefit less from additional fixations, while their scanpaths diverge from those of humans.
- We show that fine-tuning Qwen3.5-9B on human fixation trajectories improves the model's ability to guide its sampling and gather sufficient information before answering. The improvements generalized across tasks, suggesting that the learned active vision behavior extends beyond the fine-tuning task.
- We find that after fine-tuning, VLMs exhibit representation dynamics of *evidence accumulation*, where the model's preference for answering over looking consistently rises as more evidence becomes available. Fine-tuning also improves the alignment between model representations and neural responses in the primate brain during visual search tasks.

2 METHODS

2.1 ACTIVE VISION EVALUATION PIPELINE

Models and Datasets. We evaluate seven models: GPT-5.6-terra, GPT-5.4, GPT-4o, Qwen2.5-VL-72B, Qwen3-VL-32B, Gemini-3-Flash, Qwen3.5-9B (OpenAI, 2026b;a; Hurst et al., 2024; Bai et al., 2025b;a; Google DeepMind, 2025; Qwen Team, 2026), on two visual-cognitive tasks: visual question answering (VQA), where the agent answers a question about an image, and Visual Search (VS), where it locates a target from a cued category. The evaluation uses four open-source datasets with recorded human behavior, including eye-tracking data: two VQA datasets (AiR-D and VQA-MHUG) (Chen et al., 2020; Hudson & Manning, 2019; Sood et al., 2021) and two VS datasets (COCO-Search18 and Natural Design) (Chen et al., 2021; Zhang et al., 2018).

Iterative glimpsing setup. At the initial prompt, the model receives a retinally transformed view of the image centered on the image center. We approximate retinal processing by applying a Gaussian blur whose standard deviation increases with distance from the fixation location. A foveation box of size 4 degrees of visual angle (dva) around the fixation location retains the original image resolution. Outside this box, the degree of blurring increases linearly with eccentricity (see Appendix A for details) (Gupta et al., 2021).

The model is asked to either answer or request another visual sample by specifying fixation coordinates. The model operates in a loop such that after receiving each partial observation, it either provides a final answer (*ANSWER:* followed by an answer to the question or the target coordinates (x, y)) or requests the next fixation (*LOOK:* (x, y)). The trial continues until the agent answers or reaches the limit of 15 steps (Fig. 1).

The model receives a full description of the paradigm and is instructed to answer only when it has acquired enough information. The prompt also specifies the coordinate space: pixel coordinates in the benchmarking analyses or a 10×10 grid in the behavioral and fine-tuning experiments (see example prompts in Appendix B). We use the 10×10 grid in the behavioral experiments to compute a fixation probability distribution over the image. Constraining each output coordinate, x and y, to a digit from 0 to 9 allows us to compute this distribution from the model's logits.

2.2 BEHAVIORAL EVALUATION METRICS

Dynamics of stopping decision and performance. Task performance is quantified using *cumulative correct curves*, which measure the fraction of trials answered correctly by each fixation step

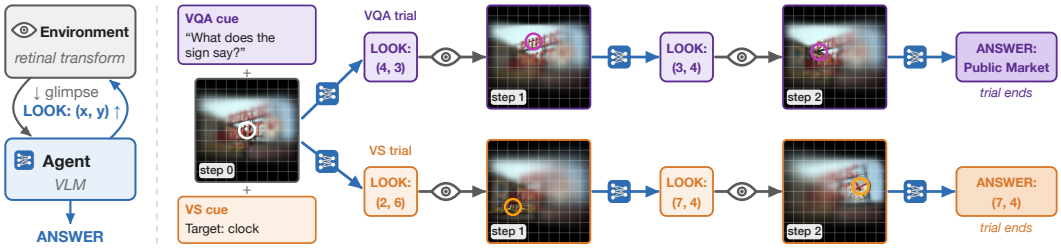


Figure 1: (left) The active-vision environment and the interaction loop. Agents receive glimpses and iteratively output LOOK coordinates or ANSWER. (right) two steps of the process are shown in VQA (purple) and VS (orange) tasks. Real example human and model scanpaths for every dataset and every benchmarked model are shown in Appendix C.

(Zhang et al., 2018; Chen et al., 2021). A VS answer is considered correct if a foveation box with a radius of 2 dva, centered at the reported location, overlaps with the target bounding box. A VQA answer is scored against the human ground truth using normalized string matching (Antol et al., 2015; Goyal et al., 2017). To quantify model–human divergence, we compute *Correct Probability Mismatch* and *Completed Probability Mismatch*: the summed absolute differences between the model and average human cumulative correct and cumulative completed curves, respectively. A trial is *completed* when the model stops requesting glimpses and outputs an *ANSWER* token followed by its answer to the question in VQA or the target location in VS.

Scanpath similarity. We compare model and human fixation sequences on the same trials using methods that capture different aspects of scanpaths. *MultiMatch* (Jarodzka et al., 2010; Dewhurst et al., 2012) measures geometric similarity between scanpaths in terms of shape, saccade direction and length, and fixation position. *Sequence Score* (Yang et al., 2020) measures alignment in the identity and order of visited image locations. Finally, we evaluate how well models predict a human participant's next fixation when conditioned on that participant's fixation history (Kümmerer et al., 2022). This *next-step prediction* method compares the model's probability distribution over possible fixation locations with the participant's actual next fixation. It requires access to model logits and can therefore be applied only to the open-source models in our evaluation. The methods for measuring fixation probability maps and computing different metrics (NSS, AUC, LL, and IG) are included in Appendix G.

2.3 Fine-tuning on human behavior

We fine-tune Qwen3.5-9B on human behavioral data using low-rank adaptation (Hu et al., 2022) (Appendix D). The four datasets differ substantially in their numbers of participants and trials, as well as in difficulty, as reflected in how quickly participants answer (Appendix D.3). To prevent any one dataset from dominating training, we balance their contributions by subsampling trials and participants to match the numbers available in the smallest included dataset.

3 Results

3.1 Do VLMs behave like humans during active vision?

We first ask how VLMs' behavior under the iterative glimpsing protocol (Fig. 1) compares to human behavior on the same trials.

VLMs answer early and benefit less from additional fixations. All models achieve lower accuracy compared to humans after reaching the last fixation on the VS datasets and AiR-D, (cumulative correct curves; Fig. 2, solid lines). GPT-5.6-terra achieves the highest model accuracy on COCO-Search18 and Natural Design. VQA-MHUG dataset is an exception where all models except Qwen2.5-VL-72B match or exceed human accuracy by the final step. Beyond final accuracy, the models also differ from humans in how their performance changes across fixations. Human accuracy generally continues to improve as more fixations are made, whereas model accuracy often

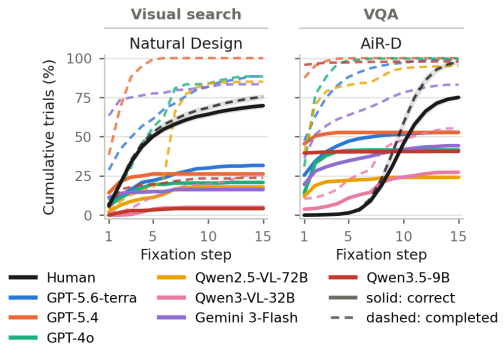


Figure 2: Cumulative *completed* (dashed) and *correct* (solid) curves show the fraction of trials that are answered and correctly answered by the agent by each fixation step. Human curves are the mean over observers with standard error of the mean (SEM) across observers. (other datasets in Appendix H.1).

Agent	Cor PM ↓	Com PM ↓
Natural Design (VS)		
GPT-5.6-terra	**4.22**	2.05
GPT-5.4	4.39	5.45
GPT-4o	5.20	**1.61**
Qwen2.5-VL-72B	5.81	2.42
Qwen3-VL-32B	7.38	6.09
Gemini 3-Flash	5.73	3.38
Qwen3.5-9B	7.41	5.37
AiR-D (VQA)		
GPT-5.6-terra	4.14	7.38
GPT-5.4	4.86	8.95
GPT-4o	4.09	8.02
Qwen2.5-VL-72B	3.84	6.91
Qwen3-VL-32B	**2.87**	**3.27**
Gemini 3-Flash	3.64	5.39
Qwen3.5-9B	4.43	8.76

Table 1: Model–human differences in performance dynamics and stopping behavior. Correct and completed probability mismatch (Cor PM and Com PM): the summed absolute difference between the agent's and the mean human cumulative correct / completed curve across fixation steps (Fig. 2). See Appendix Table A8 for other datasets.

plateaus within the first few fixations. We quantify this divergence using *Correct Probability Mismatch*, the summed absolute difference between the human and model cumulative correct curves (Chen et al., 2021) (Table 1; Cor PM).

The cumulative correct curves alone do not distinguish whether models continue exploring without improving or stop early with incorrect answers. To examine stopping behavior, we therefore consider the *cumulative completed curves*, which measure the fraction of trials on which a final answer has been given by each fixation step (Fig. 2, dashed lines). Many models complete a large fraction of trials within the first few fixations. This difference is particularly clear on the VQA datasets where humans generally make several fixations before answering, whereas models begin answering after the initial central view. Datasets on which humans require more fixations provide a wider window for examining the decision to answer or continue looking. Qwen3-VL-32B shows more gradual completion curves on these datasets and has the lowest *Completed Probability Mismatch* on AiR-D and VQA-MHUG (Fig. 2, dashed lines, Table 1, Completed). Overall, models tend to answer after fewer fixations and benefit less from additional sampling, with lower final accuracy on three of the four datasets.

VLMs successfully predict human fixation locations despite diverging scanpaths. We next compare model and human scanpaths to examine where models look and in what order. Model scanpaths are more similar to human scanpaths than chance when considering the shape and saccade length, but this similarity does not exceed that of a generic center-bias baseline. They also differ from human scanpaths in saccade direction and fixation position, particularly on the more demanding datasets (Table A5, MultiMatch). The models follow different trajectories from humans, with Sequence Scores below those of a center-bias baseline (Table A5, Sequence Score). Together, these results show that VLMs differ from humans in where they look and how their exploration unfolds across fixations. The remaining datasets and the four MultiMatch components are in Appendix F.

However, scanpaths can diverge rapidly following a single different fixation, even when agents use similar strategies to choose where to look. To account for this, we also evaluate next-fixation prediction where we condition the model on a participant's fixation history and measure how well its output distribution predicts that participant's next fixation (Kümmerer et al., 2022). This approach evaluates each fixation as a choice conditioned on previous observations, allowing us to compare model and human choices given the same fixation history. Because this analysis requires access to model logits, we only apply it to two open-source models, Qwen3.5-9B and Qwen3-VL-32B. Both models predict human next-fixation locations significantly better than other baselines, with Qwen3.5-9B

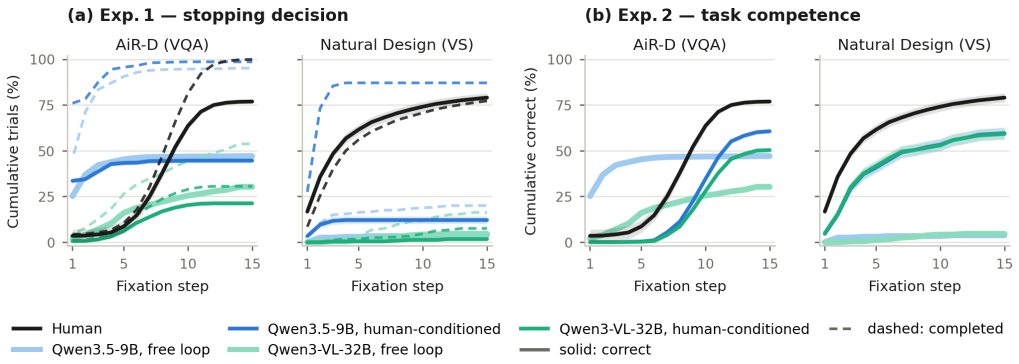


Figure 3: Controlled experiments using human fixation histories. **(a)** Stopping decision: models receive human-selected views and choose whether to answer or continue looking. **(b)** Task competence: models answer after receiving the observations available at the human answering step. Solid lines indicate cumulative correct trials; dashed lines indicate cumulative completed trials. Shading shows ± SEM across trials for models and across observers for humans. Results for the remaining datasets are in Appendix H.2.

performing best (Table A7). This shows that while VLMs' independently generated scanpaths differ from human scanpaths, the models successfully predict human fixation choices when conditioned on human viewing histories.

3.2 Investigating the source of differences between the model and human behavior

Our assessments revealed differences between model and human behavior in accuracy, performance across fixations, stopping behavior, and scanpaths. However, the underlying causes of these differences remain unclear.

We hypothesized that the observed gaps could stem from three different abilities: (1) **stopping decision**, the ability to judge whether enough information has been gathered or further sampling is needed; (2) **task competence**, the ability to answer the question or locate the target given sufficient visual information; and (3) **fixation selection**, the ability to choose informative regions to inspect. Differences in scanpaths could reflect different fixation-selection strategies, difficulty understanding or solving the task, or early stopping that cuts exploration short. To examine these possibilities, we perform three experiments where we hold the observation sequence fixed using human scanpaths and vary the actions available to the model.

Experiment 1 – stopping decision. We target stopping behavior by conditioning the models on human scanpaths which relieves the model from making fixation decisions and specifically focus on when models decide to answer. At each step, the model receives the retinal view corresponding to the participant's fixation and may either answer or request another glimpse. An answer is recorded and ends the trial. If the model requests another glimpse, it receives the view at the participant's next fixation, regardless of the location it requested (following the same protocol as in Section 2.2), but here we compare the *conditional cumulative completed curves* with those from the free rollout of the model (Fig. 3a).

Even with human-selected observations, the models' stopping behavior remains significantly different from that of humans. Qwen3.5-9B answers even earlier than in the free loop, whereas Qwen3-VL-32B answers on fewer trials and none of the VS trials. In the free loop, Qwen3-VL-32B typically answers later than humans, particularly on VS. Under human conditioning, the supplied scanpaths therefore often end while the model is still requesting more glimpses. These results suggest that the models' judgment of *when* to respond is different from humans even when they receive the same information.

Next, we test whether aligning the model's stopping behavior with that of humans is sufficient to improve its overall behavioral similarity to humans. In the targeted **fine-tuning experiment 1**, we train Qwen3.5-9B only on answering at the same step as humans by supervising its *LOOK/ANSWER* tokens at each step to match human decisions. This reduces premature answering, bringing its cumulative completed curves closer to those of humans: completed probability mismatch falls from 7.76 to 1.35 on AiR-D. It also improves performance on both VQA and VS (Fig. 4a, Table 3). By ensuring the model receives high-quality observations matched to humans, experiment 1 results establish that the model lacks the ability to effectively judge *when* it has acquired enough information to answer and should stop sampling in line with the flat cumulative correct curves in Fig. 2, and fine-tuning the model matching only this aspect of its behavior to those of humans leads to significant gains in performance and behavioral alignment. Together, these results identify the ability to decide *when* to stop looking around and answering as one of the weak points of VLMs in performing closed-loop perception-action.

Experiment 2 – task competence. We isolate the task competence ability by matching both fixation selection and stopping decision to that of humans. In this setting, the model receives the full human-matched sequence of retinal views and is forced to answer at the last step. The resulting *conditional cumulative correct curves* show that under these conditions, models answer more accurately, although a gap to human performance remains (Fig. 3b and Fig. A4). This suggests that task difficulty alone does not explain their differences from human behavior.

In **fine-tuning experiment 2**, we teach the model only *what* to answer, not *when* to stop requesting glimpses and answer. We supervise the answer following the *ANSWER* token to match the ground-truth correct answer. This supervision improves performance, though less than the fine-tuning experiment 1, which teaches the model *when* to answer. However, it also eliminates the model's exploration, flattening its cumulative correct and completed curves (Fig. 4a).

Experiment 3 – fixation selection. We assess fixation selection independent of the models' stopping decision by conditioning the models on human fixation histories and requiring them to decide where to look next, without the option to answer (see example prompt in Appendix B). This allows us to evaluate next-fixation predictions at every eligible step across all trials while the earlier analysis included only steps before the model answered (Table A7), leaving many trials and later fixation steps excluded because of early stopping. Both models predict human next-fixation locations better than the baselines, with Qwen3.5-9B showing closer alignment than Qwen3-VL-32B (Table 2). These results suggest that VLMs can identify regions of interest for gathering further task-relevant information. The remaining datasets are in Appendix H.2.

Teaching the model only where to look (**fine-tuning experiment 3**) not only produces more human-like scanpaths but also substantially improves visual search performance (Table 3). The same supervision offers little benefit on VQA where accuracy improves modestly on VQA-MHUG ($68.3\% \rightarrow 73.0\%$) but declines on AiR-D ($45.5\% \rightarrow 41.3\%$) (Fig. 4a).

Together, these behavioral experiments suggest that the models' main difficulty lies in deciding whether to gather more information or stop and answer. Qwen3.5-9B, like most models in the benchmark, answers prematurely, whereas Qwen3-VL-32B tends to answer later than humans. Yet, given human fixation histories, both models can choose next fixation locations similar to where humans would look next and answer relatively accurately when they receive the observations available at the human answering step.

The targeted fine-tuning experiments show that each ability's contribution to behavioral alignment depends on the task: teaching the model *where to look* yields the largest gains in VS (exp3), while teaching it *when to answer* yields the largest gains in VQA (exp1). This is consistent with the spatial demands of VS, which requires the model to locate an object. In VQA, humans typically gather information across several fixations before responding, unlike the base model (mean human fixations-to-answer within the 15-fixation budget: 10.2 on AiR-D versus 7.4 on Natural Design). Teaching the model *when to answer* therefore helps it gather more information before committing to an answer.

Agent	Natural Design (VS) NSS↑	AUC↑	IG↑	LL↑	AiR-D (VQA) NSS↑	AUC↑	IG↑	LL↑
Human-between	1.740 ±0.066	0.828 ±0.008	1.865 ±0.114	1.166 ±0.095	1.908 ±0.018	0.872 ±0.002	0.943 ±0.035	1.445 ±0.029
Qwen3-VL-32B	1.090 ±0.058	0.774 ±0.008	-0.022 ±0.138	-0.721 ±0.173	1.851 ±0.022	0.850 ±0.002	0.095 ±0.046	0.597 ±0.050
Qwen3.5-9B	**1.218 ±0.049**	**0.794 ±0.008**	**1.455 ±0.093**	**0.756 ±0.073**	**1.873 ±0.019**	**0.865 ±0.002**	**0.983 ±0.029**	**1.485 ±0.022**
Centre bias	0.535 ±0.043	0.666 ±0.011	0.000 ±0.000	-0.699 ±0.109	1.218 ±0.021	0.777 ±0.003	0.000 ±0.000	0.502 ±0.031
Uniform (chance)	0.000 ±0.000	0.500 ±0.000	0.699 ±0.109	0.000 ±0.000	0.000 ±0.000	0.500 ±0.000	-0.502 ±0.031	0.000 ±0.000

Table 2: Next-fixation prediction in experiment 3 – fixation selection. Qwen3.5-9B's next fixation probability maps are aligned with those of humans more than center-bias and chance baselines consistently on both VS and VQA as indicated by higher *Normalized scanpath salience (NSS)*, *Area under the curve (AUC)*, *Information gain (IG)* and *Log-likelihood (LL)* scores (Appendix G, Table A9).

Agent	AiR-D (VQA) Acc.↑	Cor PM↓	Com PM↓	Avg. MM↑	SS↑	Natural Design (VS) Acc.↑	Cor PM↓	Com PM↓	Avg. MM↑	SS↑
Human	75.7	–	–	0.883	0.348	69.7	–	–	0.839	0.320
Qwen3.5-9B (base)	45.5	4.184	7.757	0.825	0.197	0.0	7.763	6.147	0.798	0.178
exp 1 (when)	**63.4**	**1.381**	**1.351**	0.861	0.261	13.9	6.263	5.841	0.816	**0.283**
exp 2 (what)	49.8	4.874	9.255	–	–	8.3	6.553	6.631	–	–
exp 3 (where)	41.3	3.691	6.925	**0.863**	**0.284**	**36.1**	**3.001**	**1.510**	**0.830**	0.247

Table 3: Behavioral alignment of Qwen3.5-9B before and after fine-tuning to match human behavior on *when to answer (exp 1)*, *what to answer (exp 2)*, or *where to look (exp 3)*. Accuracy (Acc.), scanpath similarity (MM and SS), and stopping behavior mismatch (Com/Cor PM) (Table A10).

3.3 Behavioral fine-tuning yields human-like active vision and neural alignment

Fine-tuning on full human behavior recovers human-like active vision. We fine-tuned Qwen3.5-9B on complete human behavioral traces, supervising its output at each step to match human responses, simultaneously teaching it where to look, when to answer, and what to answer. After fine-tuning, the model showed higher accuracy, more human-like stopping behavior, and more human-like scanpaths (Fig. 4b, Table 4, Appendix G). This full behavioral supervision improved performance across all datasets, although a significant gap with humans still remained. The behavioral alignment improved with the amount of human behavior used for training (Fig. A7). Moreover, the learned ability to coordinate perception and action transferred across tasks: fine-tuning on human behavior in one task also made the model's behavior more human-like on the other (Fig. 4; Table 4 and Table A12).

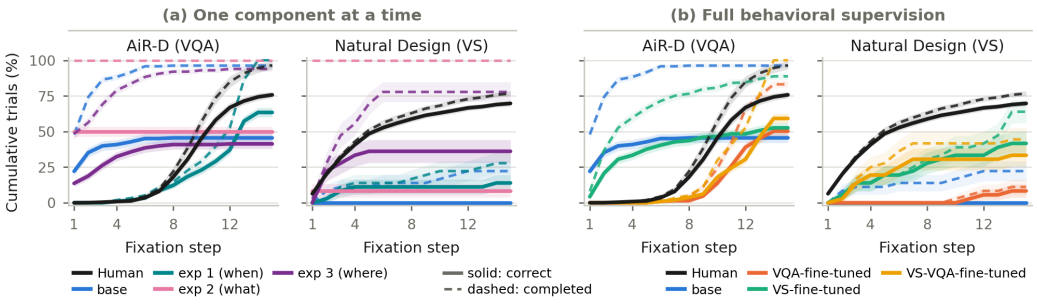


Figure 4: Behavioral fine-tuning. Cumulative correct (solid) and completed (dashed) are shown for the fine-tuned Qwen3.5-9B where it was fine-tuned to match humans on (a) single aspects of behavior: *exp 1–answering step*, *exp 2– answer*, and *exp 3–selected fixations*, or (b) all aspects, either trained only on VS or VQA or both (other datasets: Figures A5- A6).

Agent	AiR-D (VQA)					Natural Design (VS)				
	Acc. ↑	Cor PM ↓	Com PM ↓	Avg. MM ↑	SS ↑	Acc. ↑	Cor PM ↓	Com PM ↓	Avg. MM ↑	SS ↑
Human	75.7	–	–	0.883	0.348	69.7	–	–	0.839	0.320
Qwen3.5-9B (base)	45.5	4.184	7.757	0.825	0.197	0.0	7.763	6.147	0.798	0.178
VQA-fine-tuned	50.2	2.216	1.834	0.885	**0.319**	8.3	7.457	7.897	0.827	0.308
VS-fine-tuned	52.6	3.315	5.008	0.850	0.249	**41.7**	**4.013**	3.814	**0.840**	**0.323**
VS and VQA-fine-tuned	**59.2**	**2.042**	**1.728**	**0.893**	0.310	33.3	4.152	**3.480**	0.840	0.299

Table 4: Behavioral alignment of Qwen3.5-9B before and after supervised fine-tuning with full human behavior, evaluated on held-out validation trials, quantifying alignment in terms of accuracy (Acc., Cor PM), stopping behavior (Com PM), and scanpath similarity (Avg. MM and SS) (other datasets: Table A11).

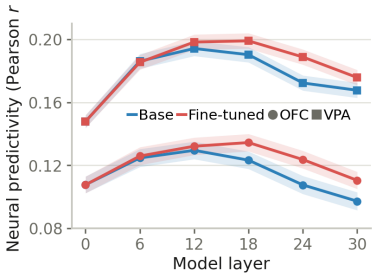


Figure 5: Behavioral fine-tuning increases neural predictivity. Per-neuron Pearson correlations between predicted and recorded responses on held-out trials during free-gaze visual search: 336 OFC neurons (circles) and 881 VPA neurons (squares). Curves compare model layers before (blue) and after (red) human behavioral fine-tuning without neural supervision. Shading: SEM across neurons.

Behavioral fine-tuning increases neural predictivity. Finally, we asked whether behavioral fine-tuning also makes the model's internal representations more brain-like. Using neuronal recordings from macaque orbitofrontal cortex (OFC) and ventral prearcuate area (VPA) during VS (Zhang et al., 2025), we predicted neural responses from model representations before and after fine-tuning (Yamins et al., 2014; Schrimpf et al., 2020) (see Appendix E for the analysis methodology). Fine-tuning increased neural predictivity in both areas without neural supervision, suggesting that learning to coordinate perception and action in visual cognitive tasks also brings model representations into closer alignment with the primate brain (Fig. 5). This gain grows with the amount of human behavior used for training (Appendix I).

Fine-tuning leads to evidence accumulation prior to answering. Our results highlight the importance of deciding when to keep looking and when to answer. Many VLMs answered prematurely and benefited little from additional fixations (benchmark, Section 3.1), and supplying human-selected observations did not resolve their differences from human stopping behavior (behavioral experiments, Section 3.2). Yet, supervising this decision alone substantially improved performance and behavioral alignment, including scanpath similarity (fine-tuning, Section 3.3). Here, we investigate how fine-tuning helps the model choose between gathering more information and answering based on its previous observations.

We first quantify the model's tendency to answer at each step using the *answering margin* $\big(\mathrm{logit}(\textsc{answer}) - \mathrm{logit}(\textsc{look})\big)$, which measures its preference for answering over requesting another glimpse. The base model's margin remains relatively flat throughout the trial before increasing abruptly at the answering step. In contrast, the fine-tuned model's margin starts from a much lower value and rises gradually across fixations leading up to the answer (Fig. 6a).

To trace this gradual rise in intermediate representations, we next use Jacobian lens (J-Lens) (Gurnee et al., 2026), which allows us to read out the representational shifts that affect the model's output vocabulary (specifically ANSWER vs. LOOK) and compute an internal *answering margin* at each layer. The internal answering margin shows a similar gradual rise in the late layers of the fine-tuned model, whereas the corresponding base-model readouts change relatively little before the answering step (Fig. 6b). Single-trial classification confirmed a ramp-like increase in the answering margin in the late layers (Appendix J.2; Fig. A11; Fig. A13; Table A13). We finally test whether the observed rise depends on the visual information available across fixations or simply tracks the steps within trials. For that, we replace different fractions of the glimpses after the initial center view with uniform grey images. Across conditions that retain some later visual information, removing more

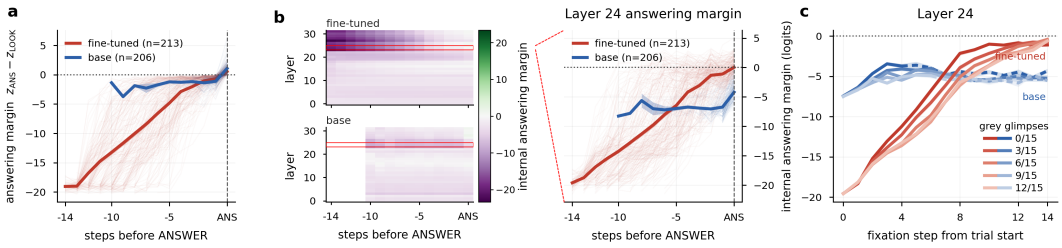


Figure 6: Dynamics of the answering margin. **(a)** Output ANSWER–LOOK logit margin. **(b)** J-Lens margins across layers, with layer 24 enlarged. In (a,b), trajectories align to the answering step; thin lines show individual trials and thick lines show means. **(c)** Fine-tuned layer-24 margins from trial start when different fractions of later glimpses are replaced with grey images, keeping the initial view intact. A dashed line shows a step with fewer than 20 trials reaching that step.

glimpses reduces the mean rate of rise of both the output and internal answering margins (Fig. 6c, Appendix J.3, Fig. A13). We also compare changes in the margin after real and grey glimpses within the same trials. When most glimpses remain real, real glimpses are followed by larger increases in the margin than grey ones, although this difference weakens as grey glimpses become more frequent (Fig. A14). The margin nevertheless rises when all glimpses are grey, even when we remove the initial view, and the model often answers at the latest fixation step possible, 15 steps, which is the maximum number of fixations during fine-tuning (Fig. A12). These results show that visual information shapes the model's increasing preference to answer, but that the rise does not depend entirely on incoming visual evidence.

4 Discussion & Limitations

We asked whether VLMs could perform active vision when they, like humans, see only part of the image at a time. Most models were less accurate than humans, often answered before exploring enough, and looked at different places than humans did. When given human scanpaths, the models approached human accuracy and predicted where humans would look next, but still struggled to decide when to answer. Teaching the model *when to answer* helped most on VQA, while teaching it *where to look* helped most on VS. This points to stopping as a main source of difficulty, in line with Zhang et al. (2026), who found that VLMs often answer rather than abstain when key information is hidden. Most base models also beat humans on VQA-MHUG but not on visual search, which may reflect more exposure to VQA-like tasks during training. Training only on VS also improved VQA performance and behavioral alignment, suggesting that learning to search helps information gathering more broadly. Together with prior work on spatial training (Peng et al., 2023; Chen et al., 2024), this motivates adding tasks such as active visual search to VLM training. Finally, fine-tuning on human behavior alone made the model better at predicting neural responses in the primate brain. Like Centaur (Binz et al., 2025), which fine-tuned a language model on human behavior described in text, our model is trained on human behavior, but it makes decisions from visual observations that it actively selects. This makes it a useful tool for studying active sensing, and an initial step toward a foundation model of active vision.

Limitations. The gradual rise in the answering margin and its sensitivity to visual input are consistent with evidence accumulation across fixations and resemble aspects of decision-related activity in primate LIP (Hanks et al., 2014; Steinemann et al., 2024). However, our margin measures the preference to answer rather than continue looking; it does not measure evidence favoring a particular answer over alternatives. Its rise without visual input also suggests a contribution from processes such as urgency, learned response timing, or a reduced expectation of gaining information from further exploration (Churchland et al., 2008; Hanks et al., 2014; Ditterich, 2006; Cisek et al., 2009; Drugowitsch et al., 2012). Further experiments would need to distinguish these contributions and test whether intermediate representations track evidence supporting specific answers across glimpses.

AI USE STATEMENT

In this work, we used generative AI tools to assist with writing and debugging the code for the experimental pipeline and analyses, and for editing the manuscript text for clarity and grammar. We did not use generative AI to drive the project ideas or design the experiments. AI-assisted text was reviewed and revised by the authors. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We describe the active-vision environment in Section 2 and the retinal transform in Appendix A. The prompts and action spaces are given in Appendix B, and the fine-tuning hyperparameters, supervision masking and dataset composition in Appendix D. The neural predictivity, scanpath similarity and next-fixation analyses are detailed in Appendices E, F and G. We will release the code for the environment, including the retinal-transform implementation, the prompts, the fine-tuning configurations and weights, and the analysis scripts upon publication. All human eye-tracking datasets and the macaque electrophysiology data used in this work are publicly available from their original sources.

REFERENCES

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pp. 2425–2433, 2015.

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

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025b. URL https://arxiv.org/abs/2502.13923.

Marcel Binz, Elif Akata, Matthias Bethge, Franziska Brändle, Fred Callaway, Julian Coda-Forno, Peter Dayan, Can Demircan, Maria K Eckstein, Noémi Éltető, et al. A foundation model to predict and capture human cognition. *Nature*, 644(8078):1002–1009, 2025.

Zoya Bylinskii, Tilke Judd, Aude Oliva, Antonio Torralba, and Frédo Durand. What do different evaluation metrics tell us about saliency models? *IEEE transactions on pattern analysis and machine intelligence*, 41(3):740–757, 2018.

Declan Campbell, Sunayana Rane, Tyler Giallanza, Nicolò De Sabbata, Kia Ghods, Amogh Joshi, Alexander Ku, Steven M Frankland, Thomas L Griffiths, Jonathan D Cohen, et al. Understanding the limits of vision language models through the lens of the binding problem. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pp. 113436–113460, 2024.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14455–14465, 2024.

Shi Chen, Ming Jiang, Jinhui Yang, and Qi Zhao. Air: Attention with reasoning capability. In *European Conference on Computer Vision*, pp. 91–107. Springer, 2020.

Yupei Chen, Zhibo Yang, Seoyoung Ahn, Dimitris Samaras, Minh Hoai, and Gregory Zelinsky. Coco-search18 fixation dataset for predicting goal-directed attention control. *Scientific reports*, 11(1):8776, 2021.

Anne K Churchland, Roozbeh Kiani, and Michael N Shadlen. Decision-making with multiple alternatives. *Nature neuroscience*, 11(6):693–702, 2008.

Paul Cisek, Geneviève Aude Puskas, and Stephany El-Murr. Decisions in changing conditions: the urgency-gating model. *The Journal of Neuroscience*, 29(37):11560–11571, 2009.

Christine A Curcio, Kenneth R Sloan, Robert E Kalina, and Anita E Hendrickson. Human photoreceptor topography. *Journal of comparative neurology*, 292(4):497–523, 1990.

Richard Dewhurst, Marcus Nyström, Halszka Jarodzka, Tom Foulsham, Roger Johansson, and Kenneth Holmqvist. It depends on how you look at it: Scanpath comparison in multiple dimensions with multimatch, a vector-based approach. *Behavior research methods*, 44(4):1079–1100, 2012.

Jochen Ditterich. Evidence for time-variant decision making. *European Journal of Neuroscience*, 24(12):3628–3641, 2006.

Jan Drugowitsch, Rubén Moreno-Bote, Anne K Churchland, Michael N Shadlen, and Alexandre Pouget. The cost of accumulating evidence in perceptual decision making. *The Journal of neuroscience*, 32(11):3612–3628, 2012.

Abdulkadir Gokce and Martin Schrimpf. Scaling laws for task-optimized models of the primate visual ventral stream. *arXiv preprint arXiv:2411.05712*, 2024.

Google DeepMind. Gemini 3 Flash model card. `https://deepmind.google/models/model-cards/gemini-3-flash/`, 2025.

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*, pp. 6904–6913, 2017.

Shashi Kant Gupta, Mengmi Zhang, Chia-Chien Wu, Jeremy Wolfe, and Gabriel Kreiman. Visual search asymmetry: Deep nets and humans share similar inherent biases. *Advances in neural information processing systems*, 34:6946–6959, 2021.

Wes Gurnee, Nicholas Sofroniew, Adam Pearce, Mateusz Piotrowski, Isaac Kauvar, Runjin Chen, Anna Soligo, Paul Bogdan, Euan Ong, Rowan Wang, et al. Verbalizable representations form a global workspace in language models, 2026.

Timothy Hanks, Roozbeh Kiani, and Michael N Shadlen. A neural mechanism of speed-accuracy tradeoff in macaque area lip. *Elife*, 3:e02260, 2014.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL `https://openreview.net/forum?id=nZeVKeeFYf9`.

Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6693–6702. IEEE, 2019.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Halszka Jarodzka, Kenneth Holmqvist, and Marcus Nyström. A vector-based, multidimensional scanpath similarity measure. In *Proceedings of the 2010 symposium on eye-tracking research & applications*, pp. 211–218, 2010.

Dongfu Jiang, Xuan He, Huaye Zeng, Cong Wei, Max Ku, Qian Liu, and Wenhu Chen. Mantis: Interleaved multi-image instruction tuning. *arXiv preprint arXiv:2405.01483*, 2024.

Seyed-Mahdi Khaligh-Razavi and Nikolaus Kriegeskorte. Deep supervised, but not unsupervised, models may explain it cortical representation. *PLoS computational biology*, 10(11):e1003915, 2014.

Matthias Kümmerer, Thomas SA Wallis, and Matthias Bethge. Information-theoretic model comparison unifies saliency metrics. *Proceedings of the National Academy of Sciences*, 112(52): 16054–16059, 2015.

Matthias Kümmerer, Matthias Bethge, and Thomas SA Wallis. Deepgaze iii: Modeling free-viewing human scanpaths with deep learning. *Journal of Vision*, 22(5):7–7, 2022.

Kenneth W Latimer, Jacob L Yates, Miriam LR Meister, Alexander C Huk, and Jonathan W Pillow. Single-trial spike trains in parietal cortex reveal discrete steps during decision-making. *Science*, 349(6244):184–187, 2015.

Hugo Laurençon, Léo Tronchon, and Victor Sanh. Introducing idefics2: A powerful 8b vision-language model for the community. *Hugging Face Blog*, 3, 2024.

Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*, 2024.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pp. 5971–5984, 2024.

OpenAI. Introducing GPT-5.4. `https://openai.com/index/introducing-gpt-5-4/`, 2026a. Released 5 March 2026.

OpenAI. GPT-5.6: Frontier intelligence that scales with your ambition. `https://openai.com/index/gpt-5-6/`, 2026b. Released 9 July 2026; the Terra variant is evaluated here.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023.

Qwen Team. Qwen3.5-9B. `https://huggingface.co/Qwen/Qwen3.5-9B`, 2026. Released 17 February 2026.

Martin Schrimpf, Jonas Kubilius, Michael J Lee, N Apurva Ratan Murty, Robert Ajemian, and James J DiCarlo. Integrative benchmarking to advance neurally mechanistic models of human intelligence. *Neuron*, 108(3):413–423, 2020.

Michael N. Shadlen, Roozbeh Kiani, William T. Newsome, Joshua I. Gold, Daniel M. Wolpert, Ariel Zylberberg, Jochen Ditterich, Victor de Lafuente, Tianming Yang, and Jamie Roitman. Comment on "Single-trial spike trains in parietal cortex reveal discrete steps during decision-making". *Science*, 351(6280):1406, 2016. doi: 10.1126/science.aad3242.

Ekta Sood, Fabian Kögel, Florian Strohm, Prajit Dhar, and Andreas Bulling. Vqa-mhug: A gaze dataset to study multimodal neural attention in visual question answering. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pp. 27–43, 2021.

Natalie Steinemann, Gabriel M Stine, Eric Trautmann, Ariel Zylberberg, Daniel M Wolpert, and Michael N Shadlen. Direct observation of the neural computations underlying a single decision. *Elife*, 12:RP90859, 2024.

Hans Strasburger, Ingo Rentschler, and Martin Jüttner. Peripheral vision and pattern recognition: A review. *Journal of vision*, 11(5):13–13, 2011.

Bingyang Wang, Yijiang Li, Yitong Qiao, Maijunxian Wang, Tianwei Zhao, Yucheng Sun, Binyue Deng, Hokin Deng, Nuno Vasconcelos, and Dezhi Luo. Increasing computation resolves conflicts in vision language models. *arXiv preprint arXiv:2505.18969*, 2025.

Andrew B Watson. A formula for human retinal ganglion cell receptive field density as a function of visual field location. *Journal of vision*, 14(7):15–15, 2014.

Daniel LK Yamins and James J DiCarlo. Using goal-driven deep learning models to understand sensory cortex. *Nature neuroscience*, 19(3):356–365, 2016.

Daniel LK Yamins, Ha Hong, Charles F Cadieu, Ethan A Solomon, Darren Seibert, and James J DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the national academy of sciences*, 111(23):8619–8624, 2014.

Zhibo Yang, Lihan Huang, Yupei Chen, Zijun Wei, Seoyoung Ahn, Gregory Zelinsky, Dimitris Samaras, and Minh Hoai. Predicting goal-directed human attention using inverse reinforcement learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 190–199. IEEE, 2020.

Jie Zhang, Xiaocang Zhu, Huihui Zhou, and Shuo Wang. A large neuronal dataset for natural category-based free-gaze visual search in macaques. *Scientific Data*, 12(1):779, 2025.

Mengmi Zhang, Jiashi Feng, Keng Teck Ma, Joo Hwee Lim, Qi Zhao, and Gabriel Kreiman. Finding any waldo with zero-shot invariant and efficient visual search. *Nature communications*, 9(1):3730, 2018.

Yue Zhang, Zun Wang, Han Lin, Yonatan Bitton, Idan Szpektor, and Mohit Bansal. Seeing isn't knowing: Do vlms know when not to answer spatial questions (and why)? *arXiv preprint arXiv:2605.30557*, 2026.

A RETINAL TRANSFORM

Retinal transformation. The brain does not receive a high-resolution view of the visual field, but it receives the image as it is transformed by the retina, consisting of high-resolution observations at the center of gaze and progressively decaying resolution towards the periphery (Curcio et al., 1990; Strasburger et al., 2011). Here, we implemented a simplified version of the retinal transformation which preserves high-resolution input at the central 4 dva x 4 dva box resembling the fovea size in primates (Watson, 2014), while the image is Gaussian blurred with a standard deviation that increases with increasing eccentricity, approximating the eccentricity-dependent cone density in the primate retina. We did not directly fit the increase in the blurring to the biological cone density, but instead followed a similar approach to the way Gupta et al. (2021) sets the eccentricity-dependent average pooling window size in the intermediate layers of a convolutional neural network.

Concretely, let d be the distance of a pixel from the current fixation and D_{fov} be the diameter of the fovea. The local receptive-field size and the Gaussian blur at that pixel are defined as:

$$\mathrm{RF}(d) = \mathrm{rf}_{\min} + s \cdot \max(0,\ 2d - D_{\text{fov}})\,, \qquad \sigma(d) = \mathrm{RF}(d)/c, \tag{1}$$

Each pixel within the foveal radius is exactly preserved. We set $D_{\text{fov}} = 4°$ of visual angle, $\mathrm{rf}_{\min} = 2\,\text{px}$ and $s = 0.16$. The constant c is the only free parameter which we set to 3 to change the pooling-window diameter into a blur width. We make the blur with an interpolation between nine blur levels, which we compute in advance.

To translate degrees of visual angle to pixels we use the longer side of image in each dataset. COCO-Search18 is $54 \times 35°$ served at $1680 \times 1050\,\text{px}$. Natural Design is $40 \times 32°$ served at $1280 \times 1024\,\text{px}$. The images of the VQA datasets have different sizes. We serve them square at $\max(w, h)$ with aspect-fit padding, and we compute the foveal radius in pixels for each trial. We map the human fixations into the same frame with the same transform before each comparison.

B EXAMPLE PROMPTS AND COMPARING THE ACTION SPACES

We asked the model to output (x, y) coordinates either in continuous pixel space (Section 3.1) or on a discrete 10×10 grid overlaid on the image (Section 3.2 onward). The grid restricts each coordinate to a single digit from 0 to 9, allowing us to compute fixation probability maps from the model's output logits. Here, we examine how these two action spaces affect model performance.

As expected, both models (Qwen3.5-9B and Qwen3-VL-32B) achieve higher accuracy with the coarser action space on COCO-Search18, AiR-D, and VQA-MHUG (Fig. A1). Qwen3.5-9B also shows a small improvement in deciding whether it has enough information to answer or needs to explore further: its cumulative correct curves rise over the first few fixations before plateauing, rather than remaining flat (Fig. A1).

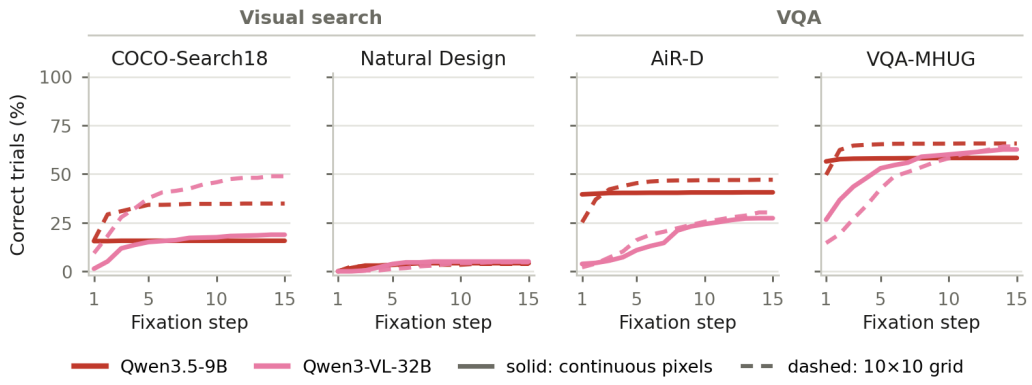


Figure A1: Effect of action space on performance dynamics. Models score higher with a coarser action space (10×10) compared to pixel coordinates. Moreover, Qwen3.5-9B improves more with fixations on a grid action space compared to a pixel space.

System prompt (identical under both protocols).

```
You are a visual reasoning agent.
Your output MUST be exactly one line in one of these formats:
LOOK: (x, y)
ANSWER: (x,y)
Do not output anything else.
```

Protocol 1 — continuous pixel space.

```
You are doing a visual search task. There is an image of size (1680, 1050), and your goal
  is to locate the object from the cue category fork.
You do not receive the image as a clean high-resolution image. Instead, at each step you
  receive a foveated view of the image. In this view, the image is sharpest around one
  fixation location, and it becomes progressively more blurred farther away from that
  location.
Initially, you receive the image foveated around the center location at pixel coordinate
  (x, y)=(840, 525).
If you need more information from another part of the image, you can move the fixation by
  outputting: LOOK: (x, y).
This will give you a new foveated view of the same image, now sharpest around the
  requested location (x, y).
When you know the answer, output: ANSWER: (x,y), specifying the target location.
Coordinates use (x, y) = (column, row), where x increases to the right and y increases
  downward.

The image size is (1680, 1050), so valid coordinates satisfy:
0 <= x < 1680
0 <= y < 1050

Examples:
(0, 0) is the top-left corner
(1679, 0) is the top-right corner
(0, 1049) is the bottom-left corner
(1679, 1049) is the bottom-right corner
(840, 525) is the center
Answer by either outputting LOOK: (x, y) specifying the center of fixation from where to
  receive a foveated view of the image, or by outputting ANSWER: (x,y) specifying the
  target location.

Your output MUST be exactly one line in one of these formats:
LOOK: (x, y)
ANSWER: (x,y)
Do not output anything else.

The image is quite complicated, with a lot of small details, take your time to explore
  the image by requesting views, and then respond. Do not rush it.
x must be an integer in [0, 1679].
y must be an integer in [0, 1049].

Your output MUST be exactly one line in one of these formats:
LOOK: (x, y)
ANSWER: (x,y)
Do not output anything else.
```

Protocol 2 — discrete 10×10 grid space.

```
You are doing a visual search task. There is an image, and your goal is to locate the
  object from the cue category fork.
You do not receive the image as a clean high-resolution image. Instead, at each step you
  receive a foveated view of the image. In this view, the image is sharpest around one
  fixation location, and it becomes progressively more blurred farther away from that
  location.
Initially, you receive the image foveated around the center location at grid index (x,
  y)=(5, 5).
If you need more information from another part of the image, you can move the fixation by
  outputting: LOOK: (x, y).
This will give you a new foveated view of the same image, now sharpest around the
  requested location (x, y).
When you know the answer, output: ANSWER: (x,y), specifying the target location.
Coordinates use (x, y) = (column, row), where x increases to the right and y increases
  downward.
Location layout of (x,y)s from which you can request a new fixation (grid coordinates):
( 0,  0) ( 1,  0) ( 2,  0) ( 3,  0) ( 4,  0) ( 5,  0) ( 6,  0) ( 7,  0) ( 8,  0) ( 9,  0)
( 0,  1) ( 1,  1) ( 2,  1) ( 3,  1) ( 4,  1) ( 5,  1) ( 6,  1) ( 7,  1) ( 8,  1) ( 9,  1)
( 0,  2) ( 1,  2) ( 2,  2) ( 3,  2) ( 4,  2) ( 5,  2) ( 6,  2) ( 7,  2) ( 8,  2) ( 9,  2)
( 0,  3) ( 1,  3) ( 2,  3) ( 3,  3) ( 4,  3) ( 5,  3) ( 6,  3) ( 7,  3) ( 8,  3) ( 9,  3)
( 0,  4) ( 1,  4) ( 2,  4) ( 3,  4) ( 4,  4) ( 5,  4) ( 6,  4) ( 7,  4) ( 8,  4) ( 9,  4)
( 0,  5) ( 1,  5) ( 2,  5) ( 3,  5) ( 4,  5) ( 5,  5) ( 6,  5) ( 7,  5) ( 8,  5) ( 9,  5)
```

```
( 0,  6) ( 1,  6) ( 2,  6) ( 3,  6) ( 4,  6) ( 5,  6) ( 6,  6) ( 7,  6) ( 8,  6) ( 9,  6)
( 0,  7) ( 1,  7) ( 2,  7) ( 3,  7) ( 4,  7) ( 5,  7) ( 6,  7) ( 7,  7) ( 8,  7) ( 9,  7)
( 0,  8) ( 1,  8) ( 2,  8) ( 3,  8) ( 4,  8) ( 5,  8) ( 6,  8) ( 7,  8) ( 8,  8) ( 9,  8)
( 0,  9) ( 1,  9) ( 2,  9) ( 3,  9) ( 4,  9) ( 5,  9) ( 6,  9) ( 7,  9) ( 8,  9) ( 9,  9)
Answer by either outputting LOOK: (x, y) specifying the center of fixation from where to
  receive a foveated view of the image, or by outputting ANSWER: (x,y) specifying the
  target location.

Your output MUST be exactly one line in one of these formats:
LOOK: (x, y)
ANSWER: (x,y)
Do not output anything else.

The image is quite complicated, with a lot of small details, take your time to explore
  the image by requesting views, and then respond. Do not rush it.
x must be an integer in [0, 9].
y must be an integer in [0, 9].

Your output MUST be exactly one line in one of these formats:
LOOK: (x, y)
ANSWER: (x,y)
Do not output anything else.
```

C EXAMPLE SCANPATHS

Figure A2 shows one benchmark trial per dataset, with the scanpath of a human observer and of every model, under the protocol of Section 3.1.



Figure A2: Example scanpaths on one trial of each dataset. Each row is an agent and each column a dataset; An example participant's scanpath and the models' are drawn in yellow. The red box specifies the target in the visual search trials.

D FINE-TUNING DETAILS

D.1 HYPERPARAMETERS

All the fine-tuning runs use the configuration in Table A1. We apply LoRA Hu et al. (2022) to all the linear projections. We set α to two times the rank. The optimizer, the precision, and the parallelism are the same in each run in this paper. Only the training set and the number of epochs are different.

Training and validation sets contained distinct trials. The single-task training sets (VQA only, VS only) are about half the size of the combined set, so we train them for 8 epochs instead of 4 to match the number of optimizer steps (about 1,200 in every run).

Setting	Value
Base model	Qwen3.5-9B
Adaptation	LoRA, all linear layers, $r = 16$, $\alpha = 32$, dropout 0
Objective	token-level cross-entropy on assistant tokens
Optimizer	AdamW, lr 1×10^{-4}, cosine decay, warmup 5%
Precision / attention	bf16, FlashAttention (SDPA)
Parallelism	DeepSpeed ZeRO-3, 4× data-parallel
Per-device batch	1
Gradient accumulation	3 (effective batch 12 sequences)
Max sequence length	16,384 tokens
Max image size	200,704 pixels
Epochs	4 (balanced-all, WHERE/WHAT/WHEN); 8 (VQA-only, VS-only)
Hardware	4× NVIDIA A100, ≈9 h per run (8.6–9.1 h)

Table A1: Fine-tuning hyperparameters. All the runs use these values.

D.2 SUPERVISION MASKING

Tables A2 and A3 show which output tokens contribute to the loss in each fine-tuning experiment. Each single-axis experiment supervises one aspect of the model's behavior, while full behavioral supervision covers all three axes.

In **fine-tuning experiment 1 – stopping decision**, we teach the model only *when* to answer. At each step of each trial, we condition the model on the participant's fixation history. The first token of the model's response specifies whether it requests another glimpse (*LOOK*) or answers the question (*ANSWER*). We apply cross-entropy loss only to this token, matching it to the participant's decision at that step (Table A3). Conditioning on human scanpaths provides enough steps to teach the model whether to continue looking or answer, without directly supervising where it should look or what it should answer.

In **fine-tuning experiment 2 – correct answer**, we teach the model only *what* to answer. To avoid supervising fixation selection or stopping, we let the model explore freely until it generates an answer and ends the trial. We then apply cross-entropy loss only to the answer content following the first *ANSWER* token, using the ground-truth correct answer as the target (Table A3).

In **fine-tuning experiment 3 – fixation selection**, we train the model only on *where* to look. At each step, we condition the model on each participant's fixation history by giving it the human fixation coordinates as part of the prompt. When the model requests a glimpse by outputting *LOOK*, we apply cross-entropy loss only to the fixation coordinates, using the participant's next fixation location as the target (Table A3).

Experiment	Supervised tokens	Supervised ability
WHEN	first token of each assistant turn (`LOOK` vs. `ANSWER`)	stop/continue decision
WHAT	`ANSWER` content tokens only	task competence
WHERE	`LOOK` coordinate tokens only	fixation selection
FULL	every `LOOK` coordinate, the `ANSWER`, all steps	All of the above

Table A2: Loss masking for each supervision axis.

D.3 DATASET COMPOSITION

Each (trial, participant) pair forms a sample in the fine-tuning. The number of unique trials and participants varies across datasets (Table A4). Therefore, to avoid a dataset dominating the fine-tuning when training on both tasks or confounding the comparisons of single-task-fine-tuned models and the maximum number of usable training samples, we took a subset of each dataset to include in

Variant	Output token stream (supervised tokens highlighted)
WHEN	LOOK: (4, 3) LOOK: (4, 5) LOOK: (2, 6) ANSWER: dark
WHAT	LOOK: (4, 3) ANSWER: dark
WHERE	LOOK (4, 3) LOOK (4, 5) LOOK (2, 6) ANSWER: dark
FULL	LOOK : (4, 3) LOOK : (4, 5) LOOK : (2, 6) ANSWER : dark

Table A3: The output tokens that contribute to the loss in each fine-tuning experiment (highlighted), on one example trial. When only teaching the model what to answer, its answer is used to compute the loss at the first step that it is generated, so its scanpath is usually shorter than the other experiments.

the combined training set. To do this, we had to limit the number of participants and unique trials to the smallest number of participants and unique trials per dataset, considering COCO-Search18 and natural design where each participant saw all the unique trials, we could keep 10 participants, and limited by 240 unique trials of natural design, 100 unique trials for the number of trials so that enough is left for evaluation. Then, we excluded all the (trial, participant) samples that had scanpaths longer than 15 fixation steps, resulting in 920 (COCO) and 847 (ND) training samples. However, the experimental setup for AiR-D and VQA-MHUG was different: not all the trials were seen by all the participants (AiR-D: 8-10 and VQA-MHUG: 3 participants per trial), so we sampled more trials than 100, to reach the goal of 1000 (trial, participant) pairs, then applied the same filtering criterion dropping the samples with scanpaths longer than 15 fixation steps, resulting in the inclusion of 994 and 831 training samples from AiR-D and VQA-MHUG in the combined train set (Table A4). Importantly, the training and validation sets were partitioned based on trials, so each trial was either included in the train or validation split.

Dataset	Trials	Participants	Samples	Contributing Train Samples
AiR-D	1,422	20	13,875	994
VQA-MHUG	3,986	49	11,958	831
Natural Design	240	16	3,799	847
COCO-Search18	586	10	5,860	920
Total	6,234	95	35,492	3,592

Table A4: Dataset specifications. Number of unique trials per dataset (Trials), participants, and samples, i.e., (trial, participant) pairs in each dataset. *Train Sample* shows the number of (trial, participant) pairs in the subset of each dataset that were included in the training set for the fine-tuning experiments.

D.4 SCALING SUBSETS

We run a set of scaling experiments by fine-tuning the model on subsets of the balanced training dataset (3,592 samples), with $\frac{1}{3}$ full dataset = 1,197 and $\frac{2}{3}$ full dataset = 2,395 (results in Appendix Section I; Fig. A7- A8) . Each smaller dataset is fully included in the larger ones and each set keeps the balance between the datasets. Each subset trains for the same 4 epochs, so the number of optimizer steps increases with the amount of data.

E NEURAL PREDICTIVITY

Recordings. We analyzed recordings from orbitofrontal cortex (OFC) and the ventral prearcuate area (VPA) of lateral prefrontal cortex (LPFC) in two rhesus macaques performing a category-based free-gaze visual search task (Zhang et al., 2025). Each trial began with an example image from the cue category, such as a face on a grey background, while the monkey maintained central fixation. A search array then appeared, containing two targets matching the cued category (face or house) and nine distractors at 11 randomly selected locations among 20 predefined positions. Objects were

sampled from a set of 160 images of monkey faces, houses, flowers, and monkey hands (40 per category). Because our study focuses on single-target visual search, we used only the data up to the first target identification in each trial. Our analysis included 15 sessions, with 22.4 ± 11.2 OFC and 58.7 ± 18.2 VPA units per session, resulting in a total of 336 OFC and 881 VPA units. We summarize the neural activity preceding each fixation as the average firing rate between the previous fixation (i) and the onset of the saccade to the next fixation ($i + 1$).

Predicting neural responses. At each step, we condition the model on the monkey's fixation history and collect its activations. We extract the 4096-dimensional decoder multilayer perceptron (MLP) activations for each of the 180 tokens representing the latest foveated input image preceding the ANSWER/LOOK output token. We average-pool these activations into 10 bins and concatenate them into a single feature vector. We then apply principal component analysis (PCA), retain the top 512 principal components, and fit ridge regression ($\alpha = 10^5$) to map model activations to neural responses. We use five-fold cross-validation, keeping all fixations from a trial in the same fold. Feature standardization, PCA, and regression are fitted only on the training trials and applied to the held-out trials.

We quantify neural predictivity using the Pearson correlation r between predicted and recorded responses on held-out trials. Within each session (15 sessions, 709 ± 142 trials per session), we average predicted and recorded responses separately within conditions before computing the correlations. We define conditions by the cue category and the current and upcoming fixation locations on the 10×10 grid. We retain conditions with at least two fixation transitions (246 ± 44 conditions per session; 4 ± 3 distinct trials per retained condition). We compute a neural predictivity score for each neuron in each session, covering 336 OFC and 881 VPA units in total.

F SCANPATH SIMILARITY

	Natural Design		AiR-D	
Agent	Avg. MM ↑	SS ↑	Avg. MM ↑	SS ↑
Human-between	0.847 ± 0.001	0.380 ± 0.003	0.886 ± 0.000	0.433 ± 0.001
GPT-5.6-terra	**0.811 ± 0.003**	0.258 ± 0.008	**0.837 ± 0.001**	0.242 ± 0.003
GPT-5.4	0.796 ± 0.003	0.193 ± 0.005	0.813 ± 0.002	0.189 ± 0.003
GPT-4o	0.799 ± 0.003	0.241 ± 0.006	0.814 ± 0.001	0.178 ± 0.001
Qwen2.5-VL-72B	0.799 ± 0.003	0.273 ± 0.006	0.825 ± 0.001	0.210 ± 0.002
Qwen3-VL-32B	0.799 ± 0.003	0.226 ± 0.007	0.796 ± 0.002	0.225 ± 0.003
Gemini 3-Flash	0.798 ± 0.003	0.185 ± 0.006	0.754 ± 0.001	0.171 ± 0.002
Qwen3.5-9B	0.800 ± 0.002	0.193 ± 0.006	0.799 ± 0.005	0.152 ± 0.009
Centre bias	0.810 ± 0.001	**0.323 ± 0.006**	0.834 ± 0.001	**0.261 ± 0.002**
Chance	0.736 ± 0.002	0.305 ± 0.006	0.757 ± 0.001	0.240 ± 0.002

	COCO-Search18		VQA-MHUG	
Agent	Avg. MM ↑	SS ↑	Avg. MM ↑	SS ↑
Human-between	0.874 ± 0.002	0.431 ± 0.003	0.880 ± 0.000	0.345 ± 0.002
GPT-5.6-terra	0.816 ± 0.004	0.382 ± 0.007	0.814 ± 0.001	0.206 ± 0.002
GPT-5.4	0.817 ± 0.004	**0.415 ± 0.008**	0.801 ± 0.002	0.204 ± 0.006
GPT-4o	**0.831 ± 0.003**	0.411 ± 0.007	0.802 ± 0.001	0.179 ± 0.001
Qwen2.5-VL-72B	0.811 ± 0.003	0.375 ± 0.007	0.808 ± 0.001	0.198 ± 0.001
Qwen3-VL-32B	0.785 ± 0.003	0.181 ± 0.005	0.793 ± 0.001	0.203 ± 0.002
Gemini 3-Flash	0.820 ± 0.004	0.369 ± 0.009	0.795 ± 0.001	0.180 ± 0.002
Qwen3.5-9B	–	–	0.806 ± 0.003	0.180 ± 0.008
Centre bias	0.794 ± 0.001	0.107 ± 0.001	**0.830 ± 0.000**	0.254 ± 0.001
Chance	0.726 ± 0.001	0.119 ± 0.001	0.754 ± 0.001	**0.260 ± 0.001**

Table A5: Model–human scanpath similarity. Avg. MM averages MultiMatch length, shape, direction, and position scores; SS compares visited locations and their order. ± values are SEM across trials. Human-between averages pairwise participant similarity. Chance and Centre bias sample fixations with replacement from uniform and center-biased distributions, respectively. Individual MultiMatch scores and other datasets appear in Appendix Table A6.

	Natural Design					AiR-D				
Agent	Len ↑	Shp ↑	Dir ↑	Pos ↑	SS ↑	Len ↑	Shp ↑	Dir ↑	Pos ↑	SS ↑
Human-between	0.926 ±0.001	0.932 ±0.001	0.685 ±0.003	0.845 ±0.003	0.380 ±0.003	0.947 ±0.000	0.952 ±0.000	0.746 ±0.001	0.900 ±0.001	0.433 ±0.001
GPT-5.6-terra	0.911 ±0.003	0.925 ±0.002	0.607 ±0.006	0.801 ±0.007	0.258 ±0.008	**0.930 ±0.001**	0.934 ±0.001	0.623 ±0.003	**0.860 ±0.002**	0.242 ±0.003
GPT-5.4	0.897 ±0.004	0.903 ±0.003	0.568 ±0.006	**0.815 ±0.005**	0.193 ±0.005	0.906 ±0.002	0.914 ±0.001	0.581 ±0.004	0.849 ±0.003	0.189 ±0.003
GPT-4o	0.921 ±0.002	0.920 ±0.002	0.584 ±0.006	0.772 ±0.006	0.241 ±0.006	0.923 ±0.001	0.920 ±0.001	0.562 ±0.002	0.851 ±0.002	0.178 ±0.001
Qwen2.5-VL-72B	0.888 ±0.004	0.909 ±0.003	0.599 ±0.007	0.799 ±0.005	0.273 ±0.006	0.922 ±0.001	0.926 ±0.001	0.598 ±0.003	0.856 ±0.002	0.210 ±0.002
Qwen3-VL-32B	**0.937 ±0.002**	**0.953 ±0.001**	0.574 ±0.006	0.734 ±0.007	0.226 ±0.007	0.894 ±0.002	0.924 ±0.001	0.627 ±0.003	0.737 ±0.003	0.225 ±0.003
Gemini 3-Flash	0.903 ±0.003	0.912 ±0.003	0.568 ±0.008	0.808 ±0.007	0.185 ±0.006	0.856 ±0.003	0.899 ±0.002	0.559 ±0.003	0.701 ±0.004	0.171 ±0.002
Qwen3.5-9B	0.921 ±0.001	0.948 ±0.001	0.524 ±0.007	0.806 ±0.004	0.193 ±0.006	0.912 ±0.007	**0.941 ±0.006**	0.507 ±0.014	0.838 ±0.007	0.152 ±0.009
Centre bias	0.907 ±0.002	0.914 ±0.001	0.616 ±0.003	0.804 ±0.003	**0.323 ±0.006**	0.906 ±0.001	0.920 ±0.000	0.679 ±0.001	0.833 ±0.001	**0.261 ±0.002**
Chance	0.765 ±0.004	0.835 ±0.002	**0.623 ±0.003**	0.719 ±0.002	0.305 ±0.006	0.762 ±0.002	0.843 ±0.001	**0.685 ±0.001**	0.736 ±0.001	0.240 ±0.002

	COCO-Search18					VQA-MHUG				
Agent	Len ↑	Shp ↑	Dir ↑	Pos ↑	SS ↑	Len ↑	Shp ↑	Dir ↑	Pos ↑	SS ↑
Human-between	0.897 ±0.002	0.908 ±0.002	0.759 ±0.005	0.932 ±0.002	0.431 ±0.003	0.942 ±0.000	0.947 ±0.000	0.733 ±0.001	0.896 ±0.001	0.345 ±0.002
GPT-5.6-terra	0.870 ±0.003	0.880 ±0.003	0.639 ±0.008	0.875 ±0.004	0.382 ±0.007	0.911 ±0.001	0.919 ±0.001	0.588 ±0.003	**0.839 ±0.002**	0.206 ±0.002
GPT-5.4	0.870 ±0.004	0.878 ±0.003	0.643 ±0.010	0.877 ±0.005	**0.415 ±0.008**	0.887 ±0.003	0.907 ±0.002	0.583 ±0.006	0.826 ±0.004	0.204 ±0.006
GPT-4o	0.877 ±0.003	0.893 ±0.002	0.671 ±0.008	**0.883 ±0.004**	0.411 ±0.007	**0.916 ±0.001**	0.914 ±0.000	0.541 ±0.002	0.837 ±0.001	0.179 ±0.001
Qwen2.5-VL-72B	0.871 ±0.003	0.881 ±0.002	0.621 ±0.008	0.871 ±0.004	0.375 ±0.007	0.909 ±0.001	0.917 ±0.001	0.573 ±0.002	0.833 ±0.002	0.198 ±0.001
Qwen3-VL-32B	0.868 ±0.002	**0.909 ±0.002**	0.590 ±0.009	0.771 ±0.004	0.181 ±0.005	0.909 ±0.001	**0.935 ±0.001**	0.574 ±0.002	0.755 ±0.003	0.203 ±0.002
Gemini 3-Flash	0.861 ±0.003	0.885 ±0.003	**0.672 ±0.009**	0.863 ±0.004	0.369 ±0.009	0.888 ±0.002	0.903 ±0.001	0.566 ±0.003	0.824 ±0.002	0.180 ±0.002
Qwen3.5-9B	–	–	–	–	–	0.903 ±0.004	0.928 ±0.003	0.554 ±0.009	0.838 ±0.004	0.180 ±0.008
Centre bias	**0.907 ±0.001**	0.887 ±0.001	0.518 ±0.003	0.866 ±0.002	0.107 ±0.001	0.904 ±0.000	0.918 ±0.000	0.663 ±0.001	0.833 ±0.001	0.254 ±0.001
Chance	0.806 ±0.002	0.818 ±0.001	0.546 ±0.003	0.734 ±0.002	0.119 ±0.001	0.769 ±0.001	0.842 ±0.000	**0.667 ±0.001**	0.737 ±0.001	**0.260 ±0.001**

Table A6: Four MultiMatch axes. MultiMatch measures the similarity of saccade length (Len), scanpath shape (Shp), saccade direction (Dir), and fixation position (Pos) between two scanpaths and scores each on [0-1]. We saw the average in Table A5 for AiR-D and Natural Design datasets. Here, the individual scores for each axis. The ± values are SEM across trials. Human-between measures the similarity between participants. Chance uses scanpaths whose fixation locations are sampled uniformly. Centre bias uses scanpaths sampled from a centre-bias prior.

G NEXT-FIXATION PREDICTION

To compute the fixation probability map, we constrain the model to output coordinates (x, y), with each coordinate represented by a digit from 0 to 9. For each coordinate token, we extract the logits corresponding to these ten digits and normalize them into a probability distribution. The outer product of the two distributions gives an approximate two-dimensional probability map of the model's next fixation. We evaluate this map against the participant's next fixation using four established saliency metrics (Bylinskii et al., 2018; Kümmerer et al., 2015). (1) *Normalized scanpath salience (NSS)* measures the z-scored map value at the fixated location, indicating how strongly that location stands out in the predicted map. (2) *Area Under the Curve (AUC)* measures how highly the fixated location ranks relative to non-fixated locations. (3) *Log-likelihood (LL)*, measured relative to a uniform baseline, quantifies how much more probability the model assigns to the observed fixation than a uniform map does. (4) *Information Gain (IG)* measures the corresponding improvement over a center-biased baseline.

	Natural Design				AiR-D			
Agent	NSS ↑	AUC ↑	IG ↑	LL ↑	NSS ↑	AUC ↑	IG ↑	LL ↑
Human-between	2.056 ± 0.087	0.898 ± 0.008	1.107 ± 0.126	1.838 ± 0.074	2.078 ± 0.046	0.901 ± 0.004	0.486 ± 0.060	1.875 ± 0.048
Qwen3-VL-32B	1.029 ± 0.056	0.754 ± 0.009	-0.117 ± 0.140	-0.774 ± 0.168	**1.841 ± 0.023**	0.848 ± 0.002	**0.114 ± 0.045**	0.721 ± 0.049
Qwen3.5-9B	**1.324 ± 0.085**	**0.842 ± 0.010**	**0.418 ± 0.089**	**1.149 ± 0.105**	1.823 ± 0.045	**0.873 ± 0.005**	0.108 ± 0.062	**1.496 ± 0.051**
Centre bias	0.978 ± 0.096	0.817 ± 0.013	0.000 ± 0.000	0.731 ± 0.131	1.757 ± 0.068	0.861 ± 0.005	0.000 ± 0.000	1.389 ± 0.063
Chance	0.000 ± 0.000	0.500 ± 0.000	-0.731 ± 0.131	0.000 ± 0.000	0.000 ± 0.000	0.500 ± 0.000	-1.389 ± 0.063	0.000 ± 0.000

	COCO-Search18				VQA-MHUG			
Agent	NSS ↑	AUC ↑	IG ↑	LL ↑	NSS ↑	AUC ↑	IG ↑	LL ↑
Human-between	2.536 ± 0.066	0.914 ± 0.005	2.071 ± 0.123	2.073 ± 0.091	2.075 ± 0.056	0.871 ± 0.005	0.291 ± 0.136	0.945 ± 0.137
Qwen3-VL-32B	**1.551 ± 0.063**	**0.791 ± 0.008**	0.208 ± 0.138	-0.162 ± 0.154	1.732 ± 0.017	0.835 ± 0.002	-0.130 ± 0.030	0.493 ± 0.034
Qwen3.5-9B	1.106 ± 0.053	0.781 ± 0.008	**0.682 ± 0.078**	**0.685 ± 0.077**	**1.846 ± 0.049**	**0.866 ± 0.005**	**0.831 ± 0.065**	**1.484 ± 0.054**
Centre bias	0.512 ± 0.043	0.742 ± 0.008	0.000 ± 0.000	0.002 ± 0.075	1.015 ± 0.049	0.803 ± 0.006	0.000 ± 0.000	0.653 ± 0.063
Chance	0.000 ± 0.000	0.500 ± 0.000	-0.002 ± 0.075	0.000 ± 0.000	0.000 ± 0.000	0.500 ± 0.000	-0.653 ± 0.063	0.000 ± 0.000

Table A7: Next-fixation prediction. Qwen3.5-9B's fixation probability maps are well-aligned with those of humans compared to baselines as indicated by higher NSS, AUC, IG and LL scores. ± values indicate SEM across trials. Scores include only steps at which the model chooses to continue looking. Because of early stopping, Qwen3.5-9B can only be scored on 39.5% (1,021/2,586) of human fixation steps on AiR-D and 17.7% (194/1,096) on Natural Design. Qwen3-VL-32B continues looking more often, covering 90.5% and 94.4%, respectively.

H Results for the Remaining Datasets

H.1 Benchmark

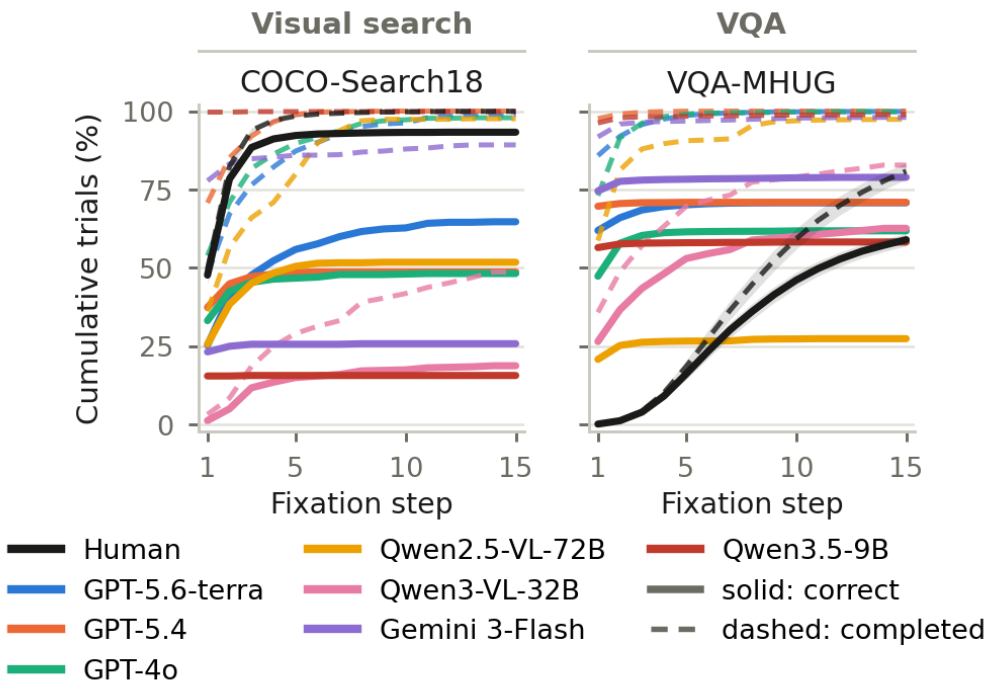


Figure A3: Cumulative *completed* (dashed) and *correct* (solid) curves on the two datasets that Figure 2 does not show. The format is that of Figure 2: the solid lines are the fraction of trials answered correctly by each fixation step and the dashed lines the fraction answered at all. The shaded bands show the ±SEM across trials for the models and across participants for the humans.

	COCO-Search18		VQA-MHUG	
Agent	Cor PM ↓	Com PM ↓	Cor PM ↓	Com PM ↓
GPT-5.6-terra	**4.80**	0.94	5.61	8.44
GPT-5.4	6.16	**0.29**	5.80	8.73
GPT-4o	6.35	0.80	4.24	8.32
Qwen2.5-VL-72B	6.04	1.46	**2.81**	7.42
Qwen3-VL-32B	11.06	9.15	3.23	**4.39**
Gemini 3-Flash	9.46	1.80	6.91	8.31
Qwen3.5-9B	10.94	0.80	3.89	8.53

Table A8: Model–human differences in performance dynamics and stopping behavior, on the two datasets that Table 1 does not show. The format and the metrics are those of Table 1: correct and completed probability mismatch against the mean human cumulative curves (Fig. A3), summed over the fixation steps. Lower is more human-like.

H.2 CONTROLLED EXPERIMENTS

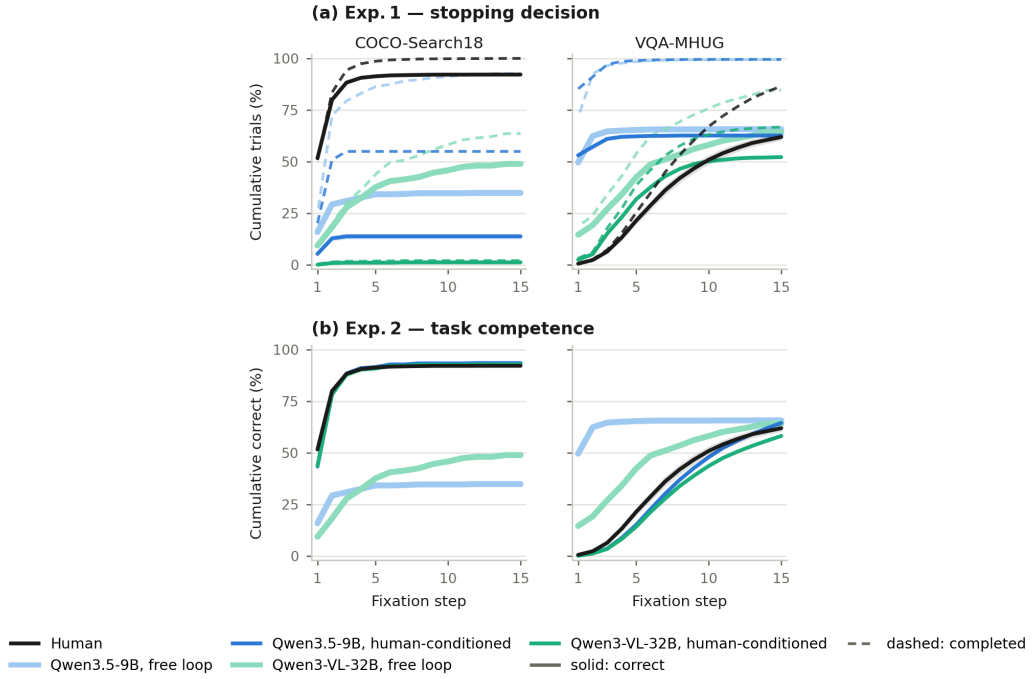


Figure A4: Controlled experiment 1 (stopping decision) and 2 (task competence), on the two datasets that the main figure does not show. The format is that of Figure 3. In experiment 1, we condition the model on a scanpath randomly drawn from the pool of participants who visited a trial, and at each step ask the model to output either an answer or where to look next, assessing its ability to decide *when* to keep looking for more information versus stop sampling to answer. In experiment 2, the model receives the full human sequence of observations and is asked to answer. Thus, in experiment 2, the model does not decide *when* to answer, so there are no cumulative completed curves, only conditional cumulative correct curves specifying its accuracy across fixations. The shaded bands are ±SEM across trials for the models, and ±SEM across observers for the human reference.

	COCO-S18				VQA-MHUG			
Agent	NSS↑	AUC↑	IG↑	LL↑	NSS↑	AUC↑	IG↑	LL↑
Human-between	2.675 ± 0.053	0.915 ± 0.004	2.499 ± 0.094	2.121 ± 0.075	–	–	–	–
Qwen3-VL-32B	**1.410 ± 0.060**	**0.789 ± 0.008**	-0.095 ± 0.139	-0.473 ± 0.160	1.636 ± 0.014	0.825 ± 0.001	-0.427 ± 0.029	0.058 ± 0.034
Qwen3.5-9B	1.150 ± 0.047	0.774 ± 0.007	**0.975 ± 0.067**	**0.597 ± 0.069**	**1.642 ± 0.012**	**0.837 ± 0.001**	**0.705 ± 0.018**	**1.190 ± 0.015**
Centre bias	0.379 ± 0.024	0.702 ± 0.007	0.000 ± 0.000	-0.378 ± 0.059	1.146 ± 0.014	0.778 ± 0.002	0.000 ± 0.000	0.486 ± 0.020
Uniform (chance)	0.000 ± 0.000	0.500 ± 0.000	0.378 ± 0.059	0.000 ± 0.000	0.000 ± 0.000	0.500 ± 0.000	-0.486 ± 0.020	0.000 ± 0.000

Table A9: Next-fixation prediction in experiment 3 – fixation selection for the remaining two datasets. The format is the same as Table 2. Qwen3.5-9B's next fixation probability maps are aligned with those of humans more than center-bias and chance baselines consistently on both VS and VQA.

Agent	VQA-MHUG					COCO-Search18				
	Acc.↑	Cor PM↓	Com PM↓	Avg. MM↑	SS↑	Acc.↑	Cor PM↓	Com PM↓	Avg. MM↑	SS↑
Human	60.4	–	–	0.877	0.329	93.9	–	–	0.869	0.459
Qwen3.5-9B (base)	68.3	5.025	8.154	0.816	0.228	42.0	7.586	1.180	0.819	0.411
exp 1 (when)	**83.4**	**1.889**	**2.517**	**0.849**	**0.284**	68.2	3.813	2.351	0.815	0.431
exp 2 (what)	70.2	5.527	8.607	–	–	60.2	4.598	0.740	–	–
exp 3 (where)	73.0	5.629	8.095	0.826	0.269	**83.0**	**1.439**	**0.735**	**0.864**	**0.541**

Table A10: Behavioral alignment of Qwen3.5-9B before and after fine-tuning to match human behavior on when to answer (exp 1), what to answer (exp 2), or where to look (exp 3). Same as Table 3 for the remaining two datasets.

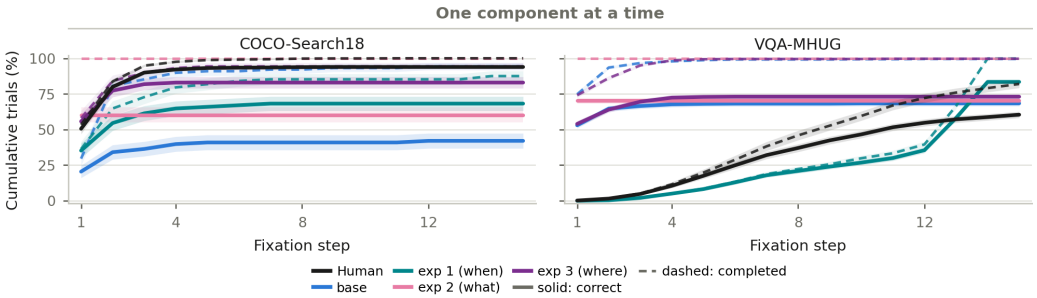


Figure A5: Single-aspect behavioral fine-tuning. Cumulative correct (solid) and completed (dashed) are shown for the fine-tuned Qwen3.5-9B where it was fine-tuned to match humans on: *exp 1– answering step*, *exp 2– answer*, and *exp 3–selected fixations*. The format is that of Figure 4a.

H.3 Fine-tuning Experiments

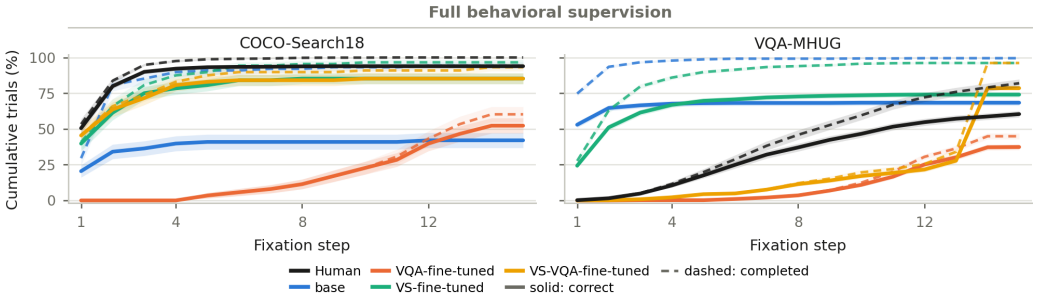


Figure A6: Behavioral fine-tuning with the full behavior supervised either trained only on VS or VQA or both. The format is that of Figure 4b.

Agent	VQA-MHUG Acc.↑	Cor PM↓	Com PM↓	Avg. MM↑	SS↑	COCO-Search18 Acc.↑	Cor PM↓	Com PM↓	Avg. MM↑	SS↑
Human	60.4	–	–	0.877	0.329	93.9	–	–	0.869	0.459
Qwen3.5-9B (base)	68.3	5.025	8.154	0.816	0.228	42.0	7.586	1.180	0.819	0.411
VQA-fine-tuned	37.4	3.289	4.357	0.872	**0.294**	52.3	10.563	11.101	0.807	0.158
VS-fine-tuned	74.0	5.055	6.564	0.834	0.272	**85.2**	1.586	**1.021**	0.850	**0.504**
VS and VQA-fine-tuned	**78.6**	**2.880**	**3.606**	**0.880**	0.280	**85.2**	**1.518**	1.612	**0.861**	0.502

Table A11: Behavioral evaluation of the fine-tuned models on the remaining two datasets. The format follows that of Table 4. Behavioral alignment of Qwen3.5-9B before and after supervised fine-tuning with full human behavior, evaluated on held-out validation trials, quantifying alignment in terms of accuracy (Acc., Cor PM), stopping behavior (Com PM), and scanpath similarity (Avg. MM and SS). Fine-tuning on VQA alone reduced VQA-MHUG accuracy from 68.3% to 37.4% by delaying answers until the fixation limit.

Agent	AiR-D NSS↑	AUC↑	IG↑	LL↑	Natural Design NSS↑	AUC↑	IG↑	LL↑
Human-between	1.927 ± 0.047	0.873 ± 0.005	0.969 ± 0.090	1.470 ± 0.073	1.668 ± 0.152	0.834 ± 0.020	1.606 ± 0.294	1.072 ± 0.327
Qwen3.5-9B (base)	1.833 ± 0.126	0.873 ± 0.012	-0.138 ± 0.130	1.438 ± 0.140	1.230 ± 0.193	**0.851 ± 0.024**	0.452 ± 0.188	**1.172 ± 0.225**
FT: VQA only	1.988 ± 0.054	0.867 ± 0.005	0.873 ± 0.080	1.375 ± 0.076	1.041 ± 0.158	0.742 ± 0.021	0.666 ± 0.264	0.121 ± 0.256
FT: visual search only	1.914 ± 0.059	0.862 ± 0.005	0.672 ± 0.086	1.423 ± 0.074	1.464 ± 0.171	0.793 ± 0.025	1.121 ± 0.322	0.576 ± 0.373
FT: both (balanced)	**2.120 ± 0.049**	**0.888 ± 0.004**	**1.261 ± 0.074**	**1.762 ± 0.054**	**1.535 ± 0.165**	0.807 ± 0.024	**1.454 ± 0.271**	0.920 ± 0.278
Centre bias	1.192 ± 0.054	0.777 ± 0.007	0.000 ± 0.000	0.501 ± 0.074	0.534 ± 0.095	0.690 ± 0.026	0.000 ± 0.000	-0.534 ± 0.277
Uniform (chance)	0.000 ± 0.000	0.500 ± 0.000	-0.501 ± 0.074	0.000 ± 0.000	0.000 ± 0.000	0.500 ± 0.000	0.534 ± 0.277	0.000 ± 0.000

Agent	VQA-MHUG NSS↑	AUC↑	IG↑	LL↑	COCO-Search18 NSS↑	AUC↑	IG↑	LL↑
Human-between	–	–	–	–	2.723 ± 0.140	0.916 ± 0.011	2.620 ± 0.247	2.155 ± 0.185
Qwen3.5-9B (base)	**2.154 ± 0.114**	**0.900 ± 0.009**	0.885 ± 0.150	**1.883 ± 0.108**	1.314 ± 0.162	0.798 ± 0.022	0.843 ± 0.204	0.807 ± 0.224
FT: VQA only	1.852 ± 0.038	0.846 ± 0.003	0.446 ± 0.055	0.990 ± 0.059	1.766 ± 0.173	0.807 ± 0.019	1.030 ± 0.252	0.562 ± 0.308
FT: visual search only	2.112 ± 0.049	0.875 ± 0.004	0.697 ± 0.065	1.585 ± 0.060	2.244 ± 0.182	0.859 ± 0.017	1.880 ± 0.216	1.426 ± 0.239
FT: both (balanced)	1.994 ± 0.034	0.870 ± 0.003	**1.023 ± 0.047**	1.565 ± 0.037	**2.355 ± 0.180**	**0.872 ± 0.014**	**2.135 ± 0.174**	**1.670 ± 0.190**
Centre bias	1.178 ± 0.037	0.784 ± 0.005	0.000 ± 0.000	0.542 ± 0.053	0.343 ± 0.057	0.695 ± 0.019	0.000 ± 0.000	-0.465 ± 0.164
Uniform (chance)	0.000 ± 0.000	0.500 ± 0.000	-0.542 ± 0.053	0.000 ± 0.000	0.000 ± 0.000	0.500 ± 0.000	0.465 ± 0.164	0.000 ± 0.000

Table A12: Next-fixation prediction for the fine-tuned checkpoints. The metric definitions are those of Table A7. The model fine-tuned on both tasks best predicts the next fixation location of humans, although single-task fine-tuned models also predict human next fixations above the centre-bias baseline on the other task, indicating that the effective human-like exploration policy learned on one active visual cognition task can generalize to other tasks.

I Scaling Experiments

Both behavioral alignment and representational alignment improve as we increase the amount of training data (details of dataset composition and training in Appendix D.4). We ran the neural-predictivity analysis of Appendix E, unchanged, on the three models trained on different dataset sizes (Fig. A8).

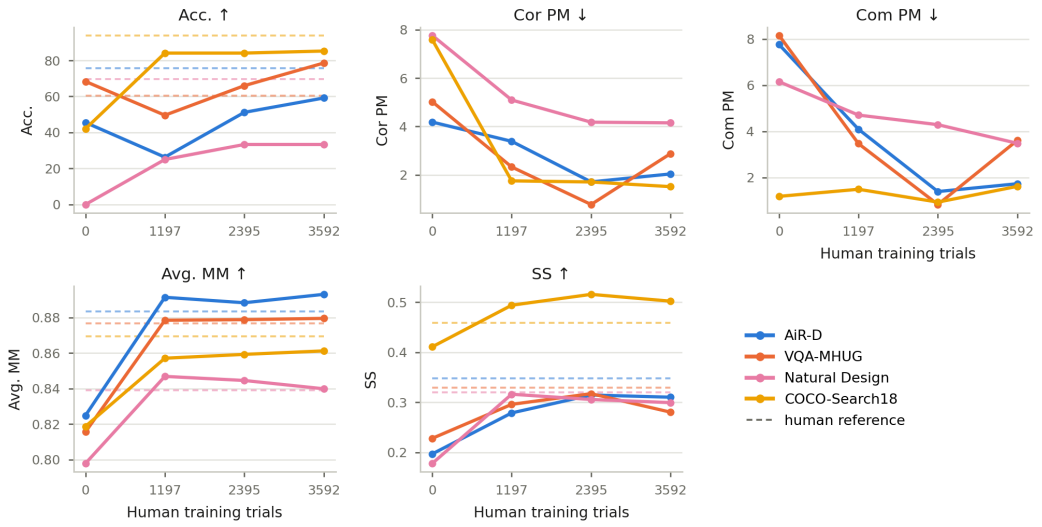


Figure A7: Behavioral alignment against the training set size. Each panel is one metric of Table 4. The initial point with 0 training samples is the base model. The dashed lines are human references: average accuracy (Acc.) or between-participant scanpath similarity (Avg. MM and SS).

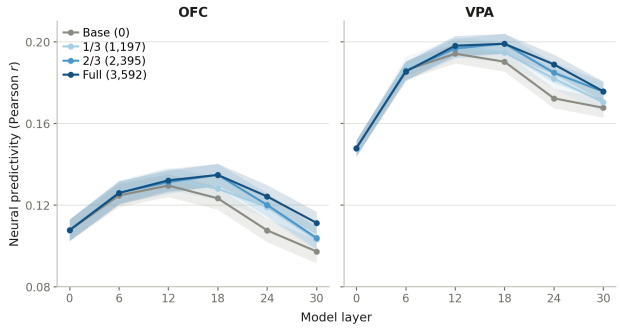


Figure A8: Representational alignment with scaling the training set size. We measure the neural predictivity of OFC and VPA for different checkpoints of the model that are trained with different training set sizes. The analysis, the recordings, and the layers are those of Figure 5.

J Evidence accumulation: controls

J.1 Answering margin across models and datasets

The fine-tuned model shows a gradually rising answering margin across datasets, while the base model's margin remains relatively flat across fixation steps (Figure 6). Late layers (23–31) consistently show the strongest gradual increase and the largest overall rise in the answering margin across all datasets (Fig. A9- A10). These are also the layers that show the largest changes in representational alignment with the brain after fine-tuning (Fig. 5).

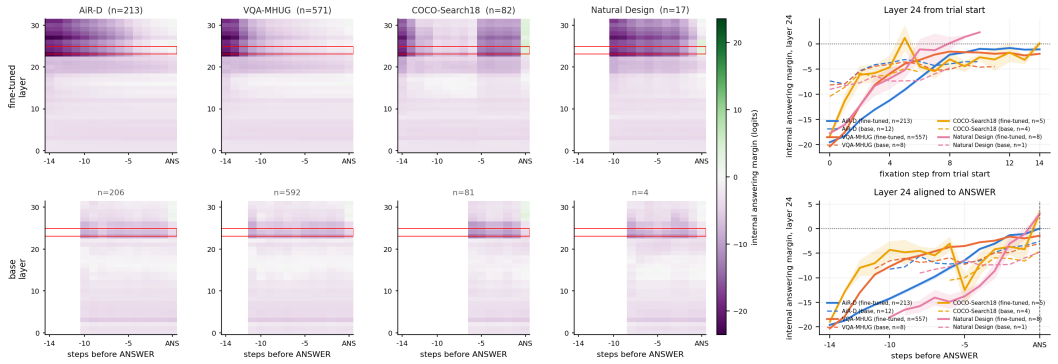


Figure A9: The internal answering margin across layers on all four datasets, for the fine-tuned model (top row) and the base model (bottom row). The format of each matrix is that of Figure 6b, with a red box on layer 24. On the right, the answering margin read out from layer 24 is shown for all the datasets relative to the trial start (top), and the answering step (bottom). We exclude trials in which the model does not answer or answers too early (Appendix J.2).

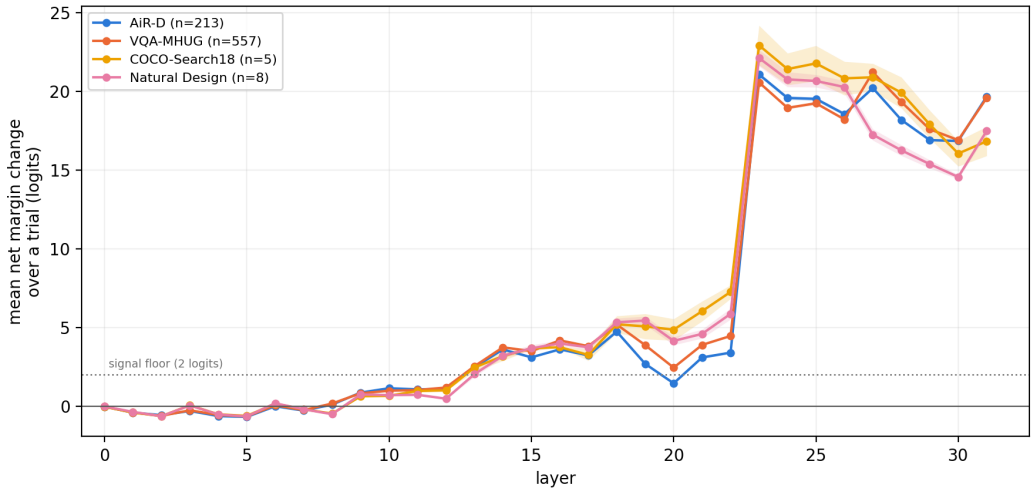


Figure A10: Net rise in answering margin over the layers. Layers 23-31 consistently have the highest net answering margin increase. The shaded bands show ± SEM across trials.

J.2 Ramping versus stepping: single-trial classification

Averaging across trials can produce a ramp even when every individual trial follows a step, if the jumps occur at different fixation steps. This confound is central to the neuroscience debate over whether LIP activity ramps or steps during perceptual decisions (Latimer et al., 2015; Shadlen et al., 2016). We therefore classify single-trial trajectories directly by comparing a linear-ramp model with the best-fitting step model using the Bayesian information criterion (BIC).

We include only trials in which the model answers, providing a complete trajectory from the initial fixation to the answering step. We also exclude trials with fewer than six fixations to ensure enough data points for classification. This excludes many trials for the base Qwen3.5-9B, which answers quickly on most datasets, including AiR-D. Although these short trajectories show little gradual increase before answering, they provide too few observations to reliably distinguish a ramp from a step. For the fine-tuned model, these criteria leave 213 analyzable trials on AiR-D and 557 on VQA-MHUG in the validation sets. In layer 24, a ramp fit the answering margin better than a step in 74%

of trials, suggesting that the observed gradual rise is not simply an artifact of averaging step-like trajectories with different jump points (Figure A11).

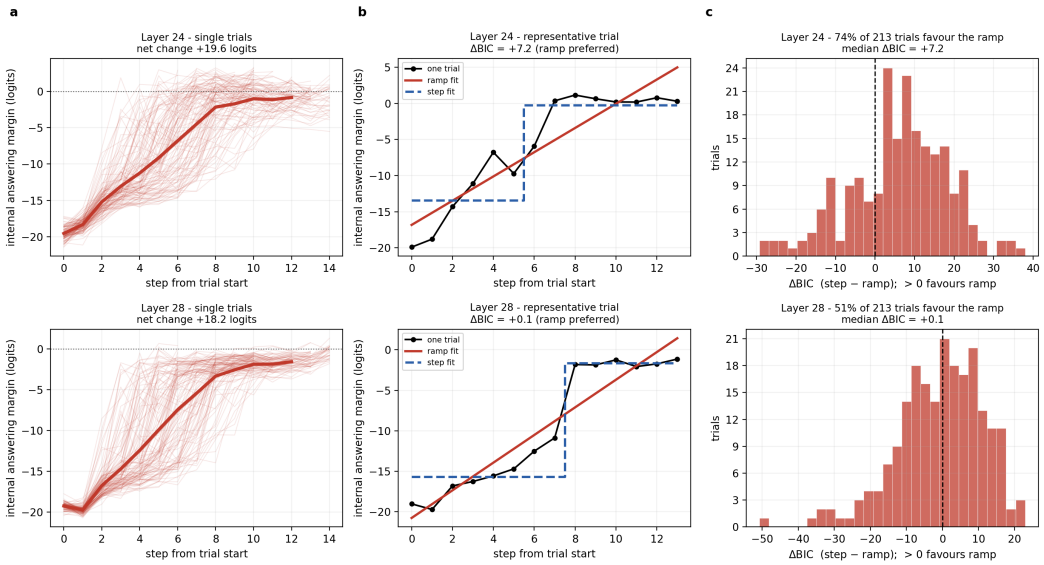


Figure A11: Classification of single-trial answering margin trajectories into ramp or step. (a) Answering margin rises similarly in the two layers 24 and 28 shown here: 19.6 and 18.2, respectively. These two layers are the two with the highest rise in answering margin in the network. (b) Two typical (median-$\Delta BIC(step - ramp)$) ramp trials identified based on the answering margin of layers 24 and 28 are shown. (c) Histogram of $\Delta BIC(step - ramp)$ values over trials.

J.3 Visual evidence manipulation experiment

Experimental setup. In Section 3.3, we asked whether the gradual rise in the fine-tuned model's behavioral and internal answering margins reflects accumulation of visual evidence across steps or whether it is tracking the answering dynamics that the model has learned during fine-tuning. In other words, the ramp-like rise in the answering margin could be due to the prior probability of answering on each step (induced by fine-tuning) rather than the posterior probability of answering given each incoming observation. To investigate this, we varied the amount of visual information available across steps and compared the resulting answering margin trajectories. We randomly replaced a fraction $p \in \{1/5, 2/5, 3/5, 4/5\}$ of the glimpses after the first one with uniform grey images. The initial center-fixated glimpse remained intact to keep the task meaningful.

Evidence accumulation and answering-prior traces in answering margin dynamics. The strength of available visual evidence modulates the answering margin dynamics in layer 24: stronger visual evidence produces steeper ramps (Table A13). However, even when all glimpses after the first are grey, many trials (69%) still show a ramp-like trajectory, and the trial-averaged answering margin rises with the same slope as when all glimpses are informative (grey; Fig. A12).

Together, the modulation by ***visual evidence*** and the persistence of ramps without further informative glimpses suggest that the answering margin reflects two effects: evidence accumulation, corresponding to a *posterior probability of answering* versus looking updated by each incoming glimpse, and an *answering-step prior* learned during fine-tuning. We further test this interpretation by removing all visual information, including the initial glimpse. In this condition, the answering margin moves rapidly to slightly below indifference ($z(\text{LOOK}) = z(\text{ANSWER})$) during the first four fixations (dashed; Fig. A12), consistent with the model recognizing that further exploration will not provide useful information. The margin remains near this level, and the model answers at the latest possible step (15) in 181 of 213 trials.

The rise in the answering margin without incoming visual evidence resembles the urgency signal in LIP, which is independent of evidence, adds to accumulated evidence, and drives the decision variable toward the bound as time passes (Churchland et al., 2008; Hanks et al., 2014; Ditterich, 2006; Cisek et al., 2009; Drugowitsch et al., 2012). Both contributions appear in the same signal: informative glimpses increase the margin (Fig. A14), but the margin also rises when no further visual information arrives.

Moreover, less visual evidence leads to a higher fraction of trials classified as ramp-like rather than step-like in layer 24 (Table A13). This pattern fits a simple account where the *answering-step prior* rises gradually, like a ramp, while the *accumulated evidence* arrives in sudden jumps, like steps. When more visual information is available, a single highly informative glimpse can raise the answering margin sharply within one fixation, so a step function fits that trial better than a ramp. When less information is available, such jumps are rarer, and more trials instead follow the slow, steady rise driven by the answering-step prior.

	Condition	Info.	Answered	Fix.	Slope, output	Slope, output (ramp)	Slope, L24	Slope, L24 (ramp)	Net travel, L24	Ramp fraction, L24
Base	No grey glimpses	1.00	206/213	3.20	$+0.73 \pm 0.05$	$+0.34 \pm 0.03$ (7)	$+1.59 \pm 0.08$	$+1.07 \pm 0.10$ (9)	$+3.2 \pm 0.1$	75% (12)
	1/5 of glimpses after the first grey	0.80	203/213	3.67	$+0.66 \pm 0.04$	$+0.29 \pm 0.03$ (7)	$+1.34 \pm 0.07$	$+0.78 \pm 0.09$ (11)	$+3.1 \pm 0.1$	61% (18)
	2/5 of glimpses after the first grey	0.60	205/213	3.89	$+0.64 \pm 0.04$	$+0.21 \pm 0.06$ (8)	$+1.22 \pm 0.06$	$+0.75 \pm 0.07$ (16)	$+3.0 \pm 0.1$	48% (33)
	3/5 of glimpses after the first grey	0.40	212/213	3.90	$+0.70 \pm 0.04$	$+0.27 \pm 0.07$ (9)	$+1.19 \pm 0.06$	$+0.65 \pm 0.06$ (20)	$+2.9 \pm 0.1$	49% (41)
	4/5 of glimpses after the first grey	0.20	213/213	3.62	$+0.85 \pm 0.04$	$+0.41 \pm 0.05$ (6)	$+1.32 \pm 0.06$	$+0.67 \pm 0.05$ (17)	$+2.9 \pm 0.1$	50% (34)
Fine-tuned	No grey glimpses	1.00	213/213	12.80	$+1.85 \pm 0.03$	$+1.95 \pm 0.04$ (150)	$+1.85 \pm 0.04$	$+1.99 \pm 0.05$ (157)	$+19.6 \pm 0.1$	74% (213)
	1/5 of glimpses after the first grey	0.80	213/213	13.35	$+1.72 \pm 0.03$	$+1.78 \pm 0.03$ (163)	$+1.71 \pm 0.03$	$+1.76 \pm 0.04$ (162)	$+19.5 \pm 0.1$	76% (213)
	2/5 of glimpses after the first grey	0.60	212/213	13.88	$+1.63 \pm 0.02$	$+1.67 \pm 0.02$ (178)	$+1.60 \pm 0.02$	$+1.62 \pm 0.02$ (182)	$+19.4 \pm 0.1$	86% (212)
	3/5 of glimpses after the first grey	0.40	212/213	14.40	$+1.55 \pm 0.01$	$+1.57 \pm 0.01$ (186)	$+1.49 \pm 0.02$	$+1.51 \pm 0.02$ (192)	$+19.2 \pm 0.1$	91% (212)
	4/5 of glimpses after the first grey	0.20	211/213	14.61	$+1.50 \pm 0.01$	$+1.53 \pm 0.01$ (179)	$+1.45 \pm 0.02$	$+1.46 \pm 0.02$ (186)	$+19.1 \pm 0.1$	88% (211)
	All glimpses after the first grey	0.00	209/213	14.45	$+1.53 \pm 0.02$	$+1.60 \pm 0.02$ (138)	$+1.50 \pm 0.02$	$+1.56 \pm 0.02$ (145)	$+18.6 \pm 0.1$	69% (209)
	All glimpses grey (first included)	0.00	202/213	14.23	$+1.11 \pm 0.04$	$+1.90 \pm 0.27$ (14)	$+1.13 \pm 0.05$	$+2.23 \pm 0.23$ (30)	$+17.6 \pm 0.1$	15% (202)

Table A13: Visual evidence manipulation experiment on AiR-D. We randomly grey-out a fraction of glimpses that the model requests. Info. specifies the fraction of intact glimpses carrying information about the scene. Answered shows the fraction of trials where the model answered in each condition. Fix. is the average number of fixations until answering. Results of single-trial ramp vs step classification are shown for the answering margins of the output and layer 24 readouts: average slope of identified ramps (Slope), net increase in the answering margin from the start until the end of the trial (Net travel), and the fraction of trials classified as having ramp-like answering margin curves (Ramp fraction) are shown. n specifies the number of analyzable trials, which are the ones that ended with a model answer within the 15-fixation window and had at least six fixations. The $\pm$ values are SEM over trials.

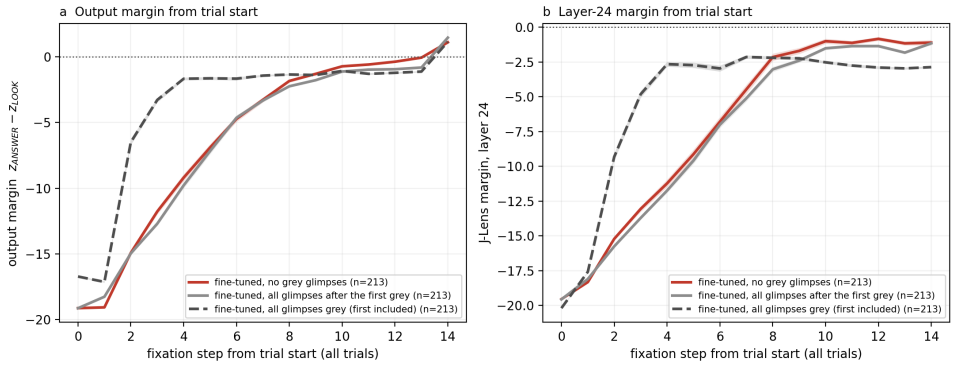


Figure A12: The **(a)** output and **(b)** the layer-24 answering margin (mean $\pm$ SEM over all 213 trials). The three conditions are all informative glimpses (red), no informative glimpses (dashed black), and only the initial center-fixated informative glimpse (grey). The dotted line corresponds to the indifference level. With one informative view, the margin rises along the real trajectory although no further information arrives, and it holds just below indifference until the maximum fixation, the 15th step. With no informative view, it moves to near indifference within four fixations, and then waits.

***Posterior* and *prior* probability of answering across layers.** In layers 21 and 22, the answering margin rises considerably less and more slowly than in layers 23–31 (green vs. purple bands, Fig. A13b,c). Yet, an equal or higher fraction of trials show ramp-like trajectories, and this fraction shows little modulation by visual evidence (green band, Fig. A13a). Together, these observations suggest that the ramps in layers 21–22 primarily reflect an answering prior learned during fine-tuning, whereas those in layers 23–31 reflect a combination of this prior and evidence accumulation. The latter still include a prior component, as ramps persist even when all glimpses after the first are grey (grey curves, Fig. A13).

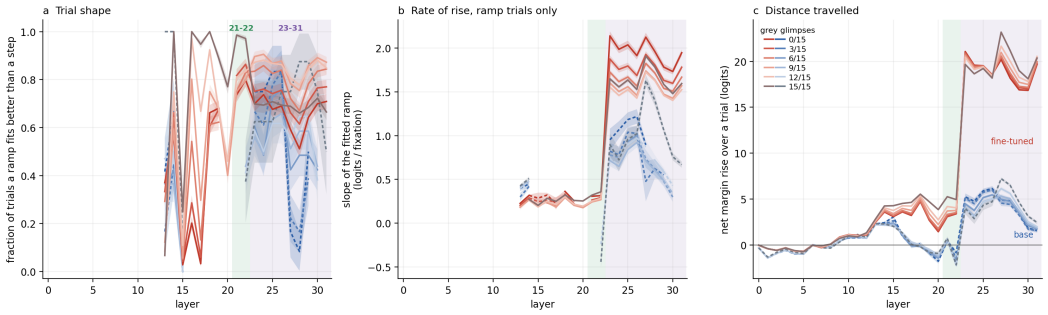


Figure A13: The visual evidence manipulation experiment results across layers. **(a)** Fraction of trials classified as having ramp-like versus step-like answering margin dynamics, **(b)** the average slope and rate of rise of the ramp-like trials, and **(c)** the total traveled distance by the answering margin, its net increase, over the trial are depicted for the fine-tuned (red) and base (blue) model across the six conditions of varying levels of information availability during the trials (darker: higher information level). Shaded bands show SEM across trials. Dashed lines represent model-conditions with fewer than 20 analyzable trials (scanpaths with at least 6 fixations that end with an answer on or before the 15th step).

Sample-by-sample evidence: informative and uninformative glimpses. Informative glimpses generally produce larger increases in the answering margin than grey glimpses, particularly when most glimpses remain informative (Figure A14a-d). The size of these increases also depends on the fixation step, with later steps generally showing smaller increases (Figure A14a-d). As informative glimpses become less frequent, informative and grey glimpses produce more similar changes in the margin. This is consistent with a greater contribution from the *prior* probability of answering when the input stream provides less visual information.

We next examine how the difference in answering margin increases after informative versus grey glimpses varies across layers (Figure A14e). In the fine-tuned model, this difference is negative in layers 21–22 across all conditions: informative glimpses increase the margin less than grey glimpses. In layers 23–31, the difference becomes positive when enough informative glimpses remain and generally increases with the amount of available information. This supports our interpretation that the answering margins in layers 21–22 primarily track an answering prior, whereas those in layers 23–31 also reflect updates from visual evidence, consistent with a posterior probability of answering.

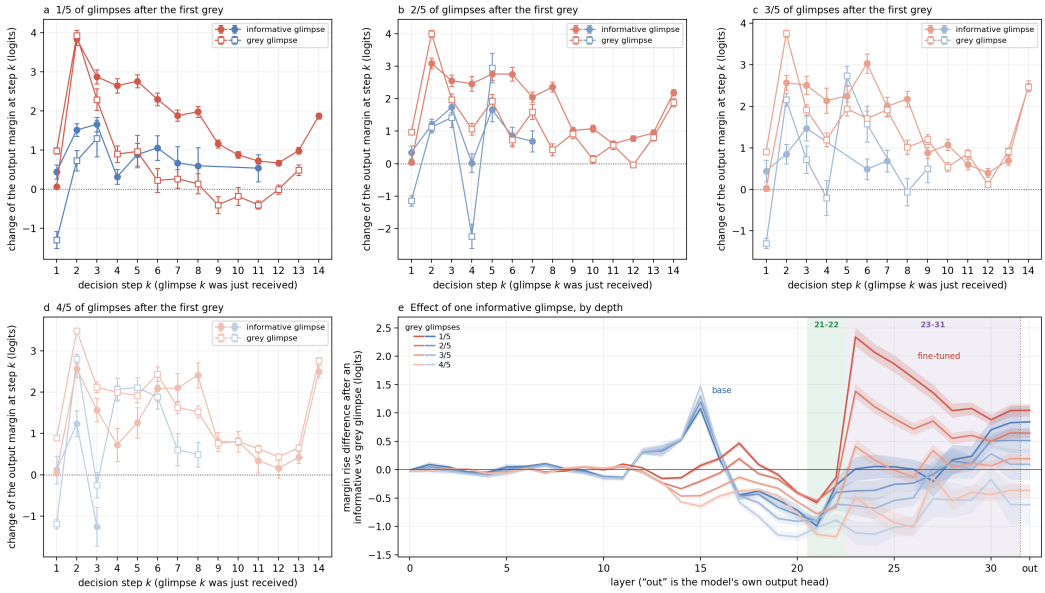


Figure A14: Changes in the answering margin following individual glimpses. **(a–d)** Step-wise changes in the model's output margin after informative or grey glimpses under different levels of visual information. **(e)** Mean margin increase after informative glimpses minus that after grey glimpses, computed within each trial and averaged across trials, for each layer and the output head of the base and fine-tuned models. Error bars and shaded bands indicate SEM across trials.