# LEARNING GROUNDED REASONING FROM HUMAN GAZE

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Human visual reasoning is driven by systematic exploration and extraction of task-relevant visual evidence. Human gaze provides a window into this process, making gaze a useful source of guidance for visually-grounded reasoning in vision-language models (VLMs). However, human gaze also includes noise and redundancies that may transfer poorly to VLMs. In this work, we investigate how human gaze data can be effectively leveraged to improve visual reasoning in VLMs, while maintaining alignment with human behavior. We collect 7,202 human gaze trials from a diverse and challenging set of visual reasoning tasks, and train grounded reasoning models with different forms of gaze guidance. Across evaluations in chart understanding, visual search, and counting, we find that selective use of human gaze consistently outperforms supervision that incorporates the full gaze sequence, achieving greater human alignment while maintaining performance competitive with leading grounded-reasoning baselines. These gains persist through reinforcement learning, while gaze-guided models remain more spatially aligned with human behavior. Behavioral analyses further show that gaze supervision induces persistent differences in exploration and verification. Our results suggest that the value of human gaze lies in transferring its exploration structure without requiring complete imitation of human behavior.

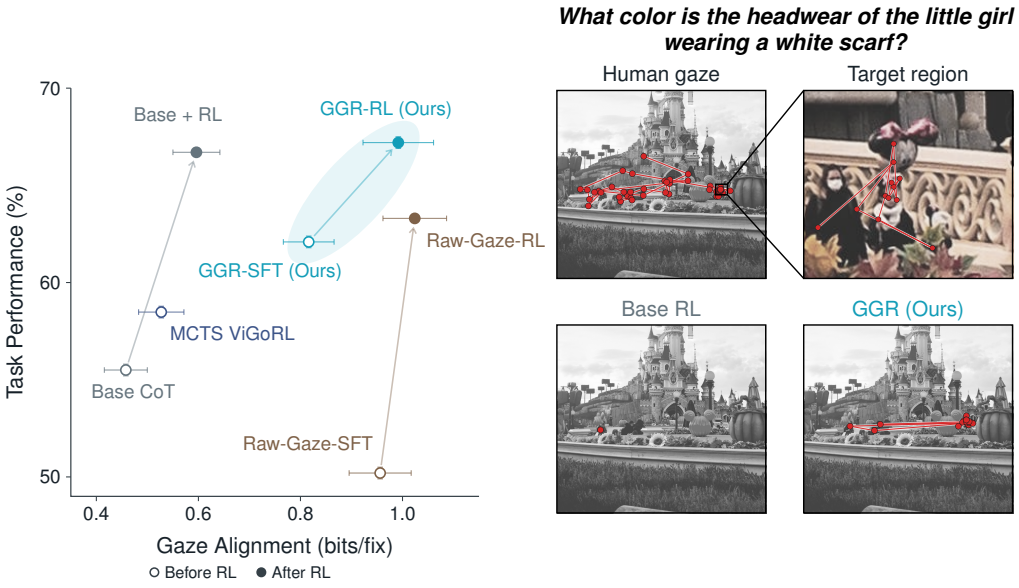


Figure 1: **GGR obtains the highest combination of task performance and human gaze alignment.** Left: mean accuracy across six benchmarks versus human-fixation log-likelihood relative to uniform (± SEM). Open circles (○) denote pre-RL models and filled circles (●) post-RL models. Right: human gaze and model grounding traces from Base RL and GGR on a held-out example.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

# 1 INTRODUCTION

Visual reasoning is fundamentally a process of *selective exploration*: a reasoner must decide where to look, extract the relevant evidence, and use that evidence to determine where to look next. Humans make this process directly observable through their gaze. When comparing two series in a chart, for example, a person may look from the legend to one curve, trace it to a relevant point, inspect the corresponding axis value, and then repeat the process for the other curve. When searching for a person in a crowded scene, gaze moves among candidate people as identifying attributes are checked and distractors are rejected. Human fixations therefore provide a naturally grounded record of task-directed visual exploration.

This signal is particularly appealing for training vision-language models (VLMs). Although chain-of-thought reasoning allows VLMs to decompose a problem into textual steps, those steps need not remain faithfully connected to the image. Models may confuse which features belong to which objects (Campbell et al., 2024; Budny et al., 2026), or hallucinate visual evidence, producing an internally coherent trace whose intermediate claims are unsupported by the image (Xu et al., 2026; Ma et al., 2026). Grounded-reasoning methods address such failures by associating reasoning steps with explicit image locations, making the cited evidence available for inspection (Haputhanthri et al., 2026). ViGoRL uses tree search to discover sequences of spatially grounded reasoning steps (Sarch et al., 2026b), while Ground-R1 uses reinforcement learning (RL) to improve model-generated evidence selection (Cao et al., 2025). These approaches can discover useful strategies, but depend on the model to propose the evidence-gathering behaviors from which it learns.

Human behavior offers a complementary source of guidance. Progress in language-model reasoning has benefited from human-written solutions and step-level annotations (Cobbe et al., 2021; Lightman et al., 2023). Beyond answer correctness, the behaviors demonstrated in chain-of-thought—such as verification and backtracking—can shape subsequent improvement through RL (Gandhi et al., 2025). For visual reasoning, gaze records regions that people consult while exploring an image, comparing evidence, and revisiting observations. Could gaze recordings provide similarly rich supervision for grounded reasoning?

Gaze reflects task goals and the coordination of visual processing during problem solving and chart comprehension (Yarbus, 1967; Just & Carpenter, 1976; Grant & Spivey, 2003; Carpenter & Shah, 1998). We investigate which forms of gaze supervision support both strong visual reasoning and closer alignment with human evidence selection. Evaluating both outcomes is necessary to understand what transfers: task accuracy alone does not reveal whether models draw on evidence that humans consult, while gaze similarity alone does not establish that the transferred behavior supports effective reasoning. Human fixations also reflect perceptual and memory constraints that VLMs do not share (Hayhoe & Ballard, 2005; Ballard et al., 1997), so reproducing them in full may not serve both aims. We therefore examine how different uses of gaze—including supervision from every fixation or a selected subset—affect the balance between task performance and human alignment.

We introduce GGR (Gaze-Guided Reasoner), which uses human fixations as candidate locations for constructing grounded reasoning supervision. We collect gaze during challenging chart understanding, counting, and visual search tasks. Across six benchmarks, selective gaze supervision improves performance after SFT while increasing alignment with held-out human gaze relative to the base model and a grounded SFT baseline. Full-fixation supervision produces higher gaze alignment but lower task performance, showing that closer imitation does not necessarily improve performance. After RL, selectively trained models retain greater human-gaze alignment than the strongest baselines while maintaining competitive task performance, along with persistent changes in expressed exploration and evaluation.

Our contributions are as follows:

- **A method for selective gaze supervision.** We introduce GGR, which constructs grounded reasoning traces from selected human fixation locations without requiring human-written explanations or task-specific annotations.
- **A multi-domain gaze dataset.** We release 7,202 eye-tracking trials spanning chart understanding, counting, and visual search.
- **An analysis of transfer from human gaze.** We compare forms of gaze supervision through task performance, spatial alignment with human fixations, and expressed reasoning behaviors, and examine which effects persist through subsequent RL.

## 2 RELATED WORK

**Grounded reasoning in VLMs.** Difficulties in binding object features in multi-object scenes have motivated recent methods that train VLMs to associate their reasoning steps with explicit image locations (Campbell et al., 2024; Budny et al., 2026). Point-based supervision can improve object-centric processing (Deitke et al., 2025; Haputhanthri et al., 2026), and points or bounding boxes can be incorporated directly into reasoning traces to improve performance across broader grounded reasoning tasks. Such supervision is straightforward when the task specifies what to identify, but more complex multi-step tasks require discovering a sequence of relevant regions. ViGoRL constructs grounded traces through tree search before applying RL (Sarch et al., 2026b), while related methods use answer-based rewards to learn where to gather visual evidence, including which regions to crop or zoom into (Cao et al., 2025; Wang et al., 2025). These approaches support flexible grounding, although the strategies available for training depend on model-generated proposals. We use human gaze to guide grounded reasoning trace proposal, comparing against ViGoRL to assess the relative benefit of human- vs. model-guided traces.

**Eye movements and visual reasoning.** Eye movements record how people gather visual information in response to task demands. Different questions about the same image elicit different fixation patterns (Yarbus, 1967), and during natural behavior, gaze is directed toward locations relevant to the current reasoning step, often just before the information is needed (Hayhoe & Ballard, 2005; Land & Hayhoe, 2001). Fixation sequences also provide evidence about the coordination of perceptual and cognitive operations (Just & Carpenter, 1976), including the interplay between reading and interpreting chart elements (Carpenter & Shah, 1998) and the shifts in attention associated with problem solving (Grant & Spivey, 2003). Repeated fixations can allow people to consult external information instead of retaining it in memory (Ballard et al., 1995). Ballard et al. (1997) describe such fixations as pointers connecting cognitive operations to locations in the world, emphasizing how eye movements help organize the use of visual evidence over time.

**Learning from human behavior.** Human behavioral signals have been used to train neural networks toward closer correspondence with human processing (Linsley et al., 2019; Fel et al., 2022; Muttenthaler et al., 2025; Sucholutsky et al., 2025; Binz et al., 2025). Such correspondence is not captured by accuracy alone, since comparable performance can conceal different solution strategies (Geirhos et al., 2020; Kerkouri et al., 2026). Human gaze has been used to supervise attention to task-relevant information. One line of work aggregates fixations into spatial salience maps, supervising where a model attends without retaining the order (Qiao et al., 2018; Saran et al., 2020; Salamatian et al., 2025). A second line retains ordered fixation trajectories to supervise sequential visual processing, as explored in recent radiology models (Li et al., 2026; Lee et al., 2026; Zhu et al., 2026). However, agreement with human attention does not consistently predict task accuracy (Sood et al., 2021), and fixation sequences vary across observers and reflect constraints that VLMs do not share (Wilming et al., 2011; Kerkouri et al., 2026). Our work examines gaze as selective *guidance* for reasoning-trace generation, evaluating task performance and behavioral alignment separately.

## 3 PROBLEM SETUP

**Learning grounded visual reasoning.** Our objective is to learn a vision-language model (VLM) that selects relevant visual evidence and reasons over that evidence to answer a question. We consider a training dataset $\mathcal{D} = \{(I_i, q_i, a_i^*, F_i)\}_{i=1}^N$, where $I_i$ is an image, $q_i$ is a natural-language question, $a_i^*$ is the ground-truth answer, and $F_i = (f_{i,1}, \ldots, f_{i,M_i})$ is a human fixation trace recorded during problem solving. Each $f_{i,j} = (x_{i,j}, y_{i,j})$ denotes a fixation location in image coordinates. Additional details of human data collection can be found in Section 4.3.

We represent a grounded reasoning step as $n_t = \langle s_t, p_t \rangle$, pairing a textual thought $s_t$ with an image location $p_t$ that identifies the visual evidence it refers to. A trajectory $\tau = (n_1, \ldots, n_T, a)$ consists of these steps followed by a predicted answer $a$. Our goal is to design a learning algorithm $\mathcal{A}$ that maps the training dataset $\mathcal{D}$ to a grounded reasoning policy $\pi_\theta = \mathcal{A}(\mathcal{D})$, with the objective of improving answer accuracy on unseen examples. At inference time, $\pi_\theta$ conditions only on the image and question $(I, q)$. Human fixations are available only during training.

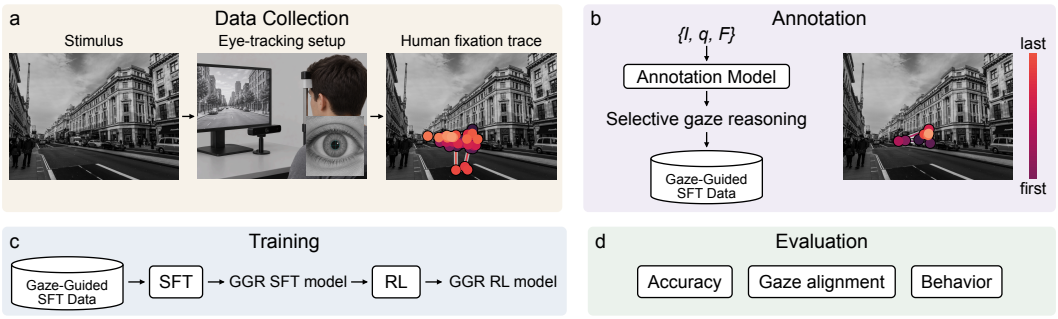


Figure 2: **Learning grounded reasoning with selective gaze guidance. (a)** Eye gaze was recorded while human participants solved visual reasoning problems. **(b)** We give the image, question, and gaze to an annotation model and instruct it to selectively annotate a subset of regions to form a grounded reasoning trace. **(c)** We train on the annotated reasoning traces using SFT followed by RL. **(d)** We evaluate the models on task accuracy, gaze alignment, and chain of thought behavior.

## 4 GGR: Learning Reasoning with Selective Gaze Guidance

We seek to obtain a grounded reasoning trajectory $\tau = (\langle s_1, p_1 \rangle, \ldots, \langle s_T, p_T \rangle, a)$, where $p_t$ is the *evidence proposal* and $s_t$ is the *reasoning supervision*. A natural way to incorporate human gaze is to directly use fixation sequences as the *evidence proposal*, and use a strong annotation model to transform these candidates into *reasoning supervision* by determining their semantic content. However, our analysis suggests that this is not the right supervision target. We show in Sections 5.2 and 5.4 that annotating every fixation with verbalized reasoning induces substantially better gaze alignment and more re-examination behavior in the trained model without yielding commensurate gains in task accuracy. At the same time, existing grounded reasoning methods, which use the models themselves to propose locations $p_t$ and reasoning $s_t$, achieve stronger task performance but remain poorly aligned with human gaze (Section 5.2 and Figure 1).

This apparent tradeoff raises the question: can human gaze behaviors be transferred to VLMs in a manner that maximizes both task performance and human alignment? While human fixation traces contain useful task-directed information, they are not themselves ideal reasoning trajectories for a VLM. A fixation may correspond to identifying a target, inspecting an attribute, rejecting a distractor, verifying an intermediate conclusion, or simply revisiting previously observed evidence. Moreover, human scanpaths reflect perceptual and memory constraints that need not apply to VLMs.

This motivates our approach of using gaze not as the evidence proposal targets directly, but as a source of *candidate visual evidence*. Human fixations identify regions that may be useful for solving the task, while an annotation model (a VLM) determines which of these regions should be included in a grounded reasoning trajectory. We detail our approach in the next section.

### 4.1 Constructing Selective Gaze-Guided Reasoning Traces

Our goal during reasoning annotation is to treat $F$ as candidate regions, while allowing an annotation model to selectively annotate a *subset* of regions from the full fixation list $F$. To do so, we provide a strong annotation model (Qwen3.5-122B-A10B) the image $I$, question $q$, and fixation coordinate list $F$, and prompt it to generate a grounded reasoning trace $\tau = (\langle s_t, p_t \rangle, \ldots, n_T, a)$, where $(p_1, \ldots, p_T)$ is a strict subset of $F$ (Figure 2b). The annotation model decides which candidate regions from the fixation list to annotate, while not being constrained to annotate the whole set. Importantly, we do not provide the reference answer $a^*$ while the annotation model constructs the reasoning trace. The model must derive an answer from the image and selected evidence rather than generate a post-hoc rationale for the ground-truth answer. (Zelikman et al., 2022).

**Sampling and verification.** For each trial, we sample 25 candidate trajectories. We apply rejection sampling (Touvron et al., 2023), giving the image, reasoning trace, and answer back to the annotation model to have it flag samples for 1) wrong answers, 2) redundancy in the model's re-

sponse, 3) inconsistency between thinking and final answer, and 4) self-contradictions. We remove any sample from training which triggers any of these flags. The accepted trajectories form our selective gaze-guided SFT dataset. Annotation prompts and details are provided in Appendix A.2.

### 4.2 Training the Grounded Reasoner

Our training pipeline is shown in Figure 2c. We detail each stage below.

**Supervised fine-tuning.** After obtaining the fixation-guided dataset using the process from Section 4.1, we train a compact model with supervised finetuning on the dataset to encourage grounding aligned with selective human gaze behavior. We call the resulting model GGR-SFT.

**Reinforcement Learning.** We then apply on-policy reinforcement learning to GGR-SFT to further refine the grounded model behavior without human gaze. We follow the Vero (Sarch et al., 2026a) recipe, and use GSPO (Zheng et al., 2025) with clip-higher, no KL penalty, and soft overlong penalty. The total reward combines 95% answer accuracy and 5% format compliance (`<think>` tags, boxed answer, and at least one valid coordinate). Our accuracy reward uses text matching for short answer questions, math-verify for numerical answers, and absolute error decreasing exponentially for counting rewards. We call the resulting model GGR-RL.

Full SFT and RL hyperparameters are provided in Appendix A.4.

### 4.3 Collecting Human Gaze Traces

We recruit participants to solve visual reasoning problems while we tracked their eye gaze (Figure 2a). Participants completed chart question answering, counting, and visual search trials while participating in the study. Images were scaled so that their longer dimension occupied 1,267 pixels, with questions presented below the image. Real-time synchronization between PsychoPy and the eye tracker allowed participants to zoom into or out of the image around their current gaze location. Participants could also place markers at their gaze location to keep track of counted objects or previously searched regions, reducing working-memory demands during extended trials. We collected 7,202 trials (2,223 visual search, 2,662 counting, 2,317 chart) in total with fixation coordinate traces for each trial. We use 6,504 trials for SFT training and 698 for held-out evaluation trials. Additional details on data collection are in Appendix A.1

**Implementation details.** For supervised finetuning of our models and baselines, we train five epochs and choose the best checkpoint as determined by the validation accuracy. For reinforcement learning, all models train 400 steps, taking the best checkpoint by validation accuracy. For RL, we use 100,000 Vero chart QA examples (Sarch et al., 2026a), 9,667 visual search examples from PixelReasoner and Visual Probe subsets in Vero (Sarch et al., 2026a), and 7,785 counting examples curated from FSC147, REC-8K, PixMo, and RPINE. We sample batches with equal representation from each domain. We provide SFT and RL training configurations in Tables 2 and 3, respectively.

## 5 Experiments

Our experiments aim to understand how human gaze supervision affects visual reasoning performance and gaze alignment in vision-language models. As shown in Figure 2, we seek to understand how each condition and baseline performs on task accuracy and gaze alignment, and how each alters chain of thought behavior. In particular, we investigate the follow questions:

(1) How does GGR compare to baselines on task performance and human gaze alignment?

(2) How does different fixation-guided supervision alter grounded reasoning performance?

(3) How does gaze-guided training alter human gaze alignment on new examples?

(4) How does gaze-guided training shape reasoning behavior?

| Condition | Chart | | Counting | | Search | | Overall | Gaze Alignment [bits/fix] |
|---|---|---|---|---|---|---|---|---|
| | ChartQA Pro | ChartQA | Count BenchQA | CountQA | MME RealWorld Lite | Visual Probe | | |
| *Baselines* | | | | | | | | |
| GPT-5-Nano | 56.0 | 80.1 | 75.4 | 25.7 | 49.8 | 41.5 | 54.8 | – |
| Molmo2-O-7B | 31.4 | 75.2 | 89.4 | 32.1 | 44.4 | 34.8 | 51.2 | – |
| Qwen3-VL-8B-Instruct | $46.0^{\pm 0.2}$ | $82.6^{\pm 0.2}$ | $86.8^{\pm 1.3}$ | $28.4^{\pm 0.5}$ | $44.7^{\pm 0.3}$ | $44.5^{\pm 0.6}$ | $55.5^{\pm 0.3}$ | $0.46^{\pm 0.04}$ |
| *SFT baselines (Qwen3-VL-8B-Instruct)* | | | | | | | | |
| Shuffled | $49.1^{\pm 0.3}$ | $89.2^{\pm 0.1}$ | $82.6^{\pm 2.0}$ | $20.2^{\pm 0.5}$ | $51.5^{\pm 0.2}$ | $42.5^{\pm 1.1}$ | $55.8^{\pm 0.4}$ | $0.01^{\pm 0.00}$ |
| ViGoRL-MCTS | $46.2^{\pm 0.8}$ | $84.8^{\pm 0.4}$ | $90.4^{\pm 0.5}$ | $30.0^{\pm 0.4}$ | $53.6^{\pm 0.4}$ | $\mathbf{45.7}^{\pm 0.6}$ | $58.4^{\pm 0.2}$ | $0.53^{\pm 0.04}$ |
| *Qwen3-VL-8B-Instruct: human-fixation SFT* | | | | | | | | |
| Fix. + Think-Aloud | $31.2^{\pm 0.2}$ | $76.1^{\pm 0.2}$ | $85.0^{\pm 1.2}$ | $24.9^{\pm 0.3}$ | $41.9^{\pm 0.7}$ | $41.9^{\pm 0.7}$ | $50.2^{\pm 0.3}$ | $0.96^{\pm 0.06}$ |
| Fix. All | $50.2^{\pm 0.2}$ | $88.4^{\pm 0.2}$ | $83.3^{\pm 0.5}$ | $22.8^{\pm 0.1}$ | $50.8^{\pm 0.6}$ | $39.8^{\pm 0.8}$ | $55.9^{\pm 0.2}$ | $\mathbf{1.04}^{\pm 0.06}$ |
| GGR-SFT | $\mathbf{58.6}^{\pm 0.1}$ | $\mathbf{91.0}^{\pm 0.2}$ | $\mathbf{93.3}^{\pm 0.4}$ | $\mathbf{32.9}^{\pm 0.4}$ | $\mathbf{55.2}^{\pm 0.8}$ | $\mathbf{46.0}^{\pm 1.0}$ | $\mathbf{62.8}^{\pm 0.2}$ | $0.82^{\pm 0.05}$ |
| *RL baselines (Qwen3-VL-8B-Instruct)* | | | | | | | | |
| Point-Grounded RL | $57.0^{\pm 0.4}$ | $91.4^{\pm 0.0}$ | $\mathbf{96.1}^{\pm 0.4}$ | $44.4^{\pm 0.4}$ | $\mathbf{59.2}^{\pm 0.5}$ | $52.2^{\pm 0.7}$ | $66.7^{\pm 0.2}$ | $0.60^{\pm 0.05}$ |
| Ungrounded | $59.1^{\pm 0.3}$ | $\mathbf{92.3}^{\pm 0.0}$ | $92.7^{\pm 0.4}$ | $39.1^{\pm 0.3}$ | $57.6^{\pm 0.2}$ | $\mathbf{53.1}^{\pm 0.4}$ | $65.7^{\pm 0.1}$ | – |
| *Qwen3-VL-8B-Instruct: SFT → RL* | | | | | | | | |
| Fix. + Think-Aloud | $59.8^{\pm 0.4}$ | $90.9^{\pm 0.3}$ | $89.7^{\pm 0.4}$ | $31.4^{\pm 0.5}$ | $54.4^{\pm 0.2}$ | $\mathbf{53.5}^{\pm 0.5}$ | $63.3^{\pm 0.1}$ | $1.02^{\pm 0.06}$ |
| Fix. All | $58.8^{\pm 0.2}$ | $\mathbf{91.9}^{\pm 0.1}$ | $93.3^{\pm 0.1}$ | $38.1^{\pm 0.5}$ | $58.1^{\pm 0.4}$ | $51.5^{\pm 0.5}$ | $65.3^{\pm 0.1}$ | $\mathbf{1.03}^{\pm 0.07}$ |
| GGR-RL | $\mathbf{61.1}^{\pm 0.3}$ | $\mathbf{92.0}^{\pm 0.1}$ | $94.8^{\pm 0.1}$ | $\mathbf{45.0}^{\pm 0.3}$ | $57.8^{\pm 0.1}$ | $\mathbf{53.1}^{\pm 0.7}$ | $\mathbf{67.3}^{\pm 0.1}$ | $0.99^{\pm 0.07}$ |

Table 1: **GGR obtains the best combination of performance and gaze alignment.** Task performance and gaze alignment after SFT (upper block) and RL (lower block). Task performance is reported as Avg@3 ($\pm$ SEM) across three evaluation runs. Gaze alignment is reported as mean $\pm$ SEM across 362 held-out images. Overall is the unweighted mean of the six benchmarks. We bold the best SFT and RL runs separately. If multiple runs have mean scores within 0.5% of the best mean, we bold all of them.

### 5.1 Experimental Setup.

**Gaze-supervision conditions.** We investigate the impact of training on data generated by conditioning a strong model (Qwen3.5-122B-A10B) on human fixations, while enforcing different constraints on the outputs generated. In particular, we generate grounded reasoning data with the following constraints: (1) we require the annotation model to annotate each fixation with an accompanying reasoning step associated with it (*Fix. All*), (2) a strict subset (and no additional points) of the fixations are accompanied by a reasoning step (GGR), (3) directly using the list of fixation coordinates interleaved with transcribed think aloud (*Fix. + Think-Aloud*), and (4) Our Fix. Full condition using shuffled fixation lists across examples (*Shuffled*).

**Non-gaze baselines.** We further compare to the following baselines: (1) Baseline vision-language models Qwen3-VL-8B-Instruct, Molmo2-O-7B, and GPT-5-nano with medium reasoning effort, (2) ViGoRL-MCTS (Sarch et al., 2026b), which uses the Qwen3.5-122B-A10B teacher with tree search to propose and search over grounded reasoning steps that lead to correct answers.

**Evaluations.** We curate a suite of benchmarks to test the ability of our method and baselines on novel problems in chart understanding, counting, and visual search. We evaluate on ChartQA (Masry et al., 2022) and ChartQA-Pro (Masry et al., 2025) for chart understanding, CountBenchQA (Paiss et al., 2023) and CountQA (Tamarapalli et al., 2025) for counting, and MME-RealWorld-Lite (Zhang et al., 2025b) and Visual Probe (Lai et al., 2025) for visual search in high resolution images.

**Fixation alignment.** We measure spatial correspondence between model grounding points and human fixations. We form a Gaussian kernel density estimate over model coordinates with standard deviation $\sigma$ and uniform background weight $\epsilon$, then compute the mean log-likelihood of human fixations relative to that of uniform spatial baseline (following Kümmerer et al. (2022)). For each condition, we fit $(\sigma_c, \epsilon_c)$ via grid search on 336 calibration examples. We evaluate on a separate 362 trials. The metric does not capture fixation order, duration, or the model's internal attention. See Appendix A.6 for additional details.

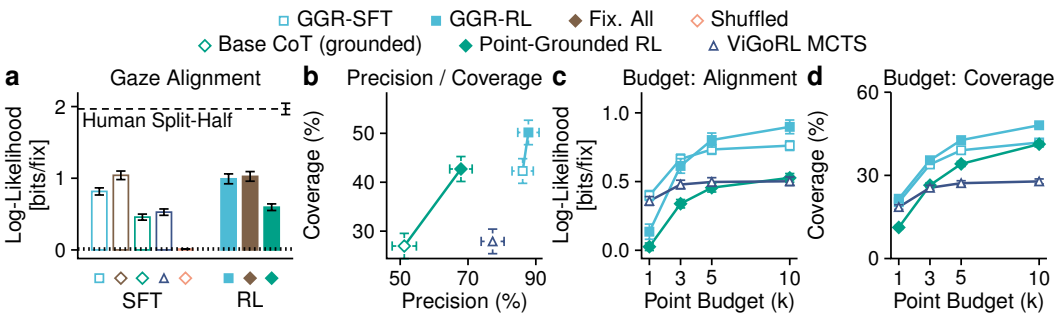


Figure 3: **Gaze alignment and coverage.** (a) Condition-fitted alignment after SFT (open bars) and after RL (filled bars). (b) Coverage versus precision at a radius of 100 normalized coordinate units. (c,d) Alignment and coverage at point budgets $k \in \{1, 3, 5, 10\}$. Reported is mean $\pm$ SEM.

### 5.2 Selective Human-Gaze Supervision Improves Reasoning

Table 1 reports results after supervised finetuning and subsequent reinforcement learning. We draw the following conclusions:

**Selective fixation supervision leads to strong reasoning performance and greater gaze alignment.** After SFT, GGR-SFT achieves the highest score among the SFT conditions in Table 1 (62.8), exceeding Qwen3-VL-8B-Instruct by 7.3 points and the ViGoRL-MCTS score by 4.4 points, with higher gaze alignment than both (0.816 versus 0.458 and 0.527 bits/fix, respectively). GGR-RL achieves the highest task score among the conditions in Table 1, reaching 67.3 overall, compared with 66.7 for Point-Grounded RL and 65.7 for Ungrounded RL. Along with performance, its gaze alignment increases to 0.992 bits/fix, substantially exceeding Point-Grounded RL (0.596).

**Selecting a subset of human fixations is more effective than supervising every fixation.** After SFT, GGR-SFT outperforms Fix. All on all six benchmarks (62.8 versus 55.9 overall). This advantage persists for GGR-RL on five of six benchmarks, although the overall gap narrows from 6.9 to 2.0 points (67.3 versus 65.3). The performance advantage holds despite lower gaze alignment than Fix. All at both stages, supporting selective fixation guidance without requiring trace imitation.

**Better reasoning and greater human-gaze alignment can coexist, but stronger alignment alone does not imply better performance.** Fix. All and Fix. + Think-Aloud are more gaze-aligned than GGR at both stages despite lower task scores. After RL, their alignment reaches 1.026 and 1.024 bits/fix, respectively, versus 0.992 for GGR-RL, while their overall scores remain lower (65.3 and 63.3 versus 67.3). Thus, spatial similarity to human gaze alone is insufficient to explain the performance benefits of gaze-guided supervision.

**Task-specific fixation guidance matters.** Shuffling fixations across examples reduces the score from 62.8 to 55.8 and alignment from 0.816 to 0.012 bits/fix, supporting task-specific gaze guidance.

### 5.3 Gaze-Guided Supervision Improves Alignment Coverage and Precision

In this section, we analyze how gaze-guided supervision affects human fixation alignment compared to baselines which do not use gaze supervision. We extract the coordinates generated by each model and examine fixation alignment. We find:

**Gaze-guided supervision increases gaze alignment.** Gaze-guided supervised finetuning yields higher gaze alignment of grounded reasoning: GGR-SFT and Fix. All reach 0.816 and 1.041 bits/fix, respectively, compared with 0.012 for Shuffled, 0.527 for ViGoRL-MCTS, and 0.596 for Point-Grounded RL (Figure 3a; Table 1). Alignment remains high after RL, with GGR-RL and Fix. All reaching 0.992 and 1.026 bits/fix, respectively.

**RL on the gaze subset model increases coverage while maintaining precision.** GGR-RL's gaze alignment increases by 0.168 bits/fix from its RL initializer (0.824 $\rightarrow$ 0.992), while Point-Grounded RL gains 0.138 bits/fix (0.458 $\rightarrow$ 0.596; Figure 3a). In Figure 3b, we analyze coverage and pre-

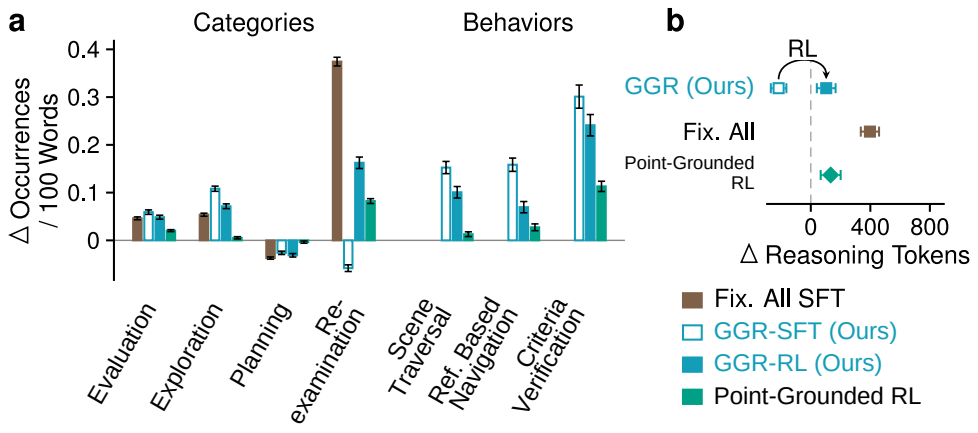


Figure 4: **The selective fixation model amplifies exploration and evaluation behaviors.** (a) Changes relative to the base model for four categories (evaluation, exploration, planning, re-examination) and three behaviors (scene traversal, reference-based navigation, criteria verification), in occurrences per 100 reasoning words (mean ± SEM). Blue outlined and filled bars show GGR after SFT and RL; green bars show Point-Grounded RL. Brown bars show Fix. All after SFT for the four categories. (b) Reasoning-length changes relative to the base model in tokens (mean ± SEM).

cision of fixations at radius 100 in normalized image coordinates. Point-Grounded RL increases coverage from 27.0% to 42.7% (+15.7 percentage points) and precision from 51.2% to 67.9% (+16.7 percentage points). For GGR-RL, the change from SFT initialization to RL step 400 is predominantly in coverage (42.3% → 50.2%, +7.9 percentage points), while precision remains high (86.1% → 87.8%, +1.7 percentage points).

**GGR retains alignment advantages when controlling for point budgets.** Figure 3c–d reports log-likelihood and coverage using up to the first $k \in \{1, 3, 5, 10\}$ unique points output by the model. GGR after both SFT and RL has higher coverage than Point-Grounded RL and ViGoRL-MCTS at all four budgets, and higher log-likelihood at $k \in \{3, 5, 10\}$. At $k = 10$, GGR-SFT reaches 0.761 bits/fix and GGR-RL reaches 0.898, compared with 0.527 for Point-Grounded RL and 0.502 for ViGoRL-MCTS; the corresponding coverage values are 41.9%, 48.1%, 41.3%, and 27.8%, respectively. At $k = 1$, GGR-SFT also leads in log-likelihood (0.398 versus 0.026 for Point-Grounded RL and 0.359 for ViGoRL-MCTS), but GGR-RL falls below ViGoRL-MCTS (0.135 versus 0.359 bits/fix), despite higher coverage (21.6% versus 18.6%).

### 5.4 Gaze-Guided Supervision Amplifies Exploration and Evaluation Behavior

To examine which behavioral changes arise during SFT and persist through RL, we analyze exploration, evaluation, and planning in the chains of thought of GGR and Fix. All before and after RL, with the base model (Qwen3-VL-8B-Instruct) and Point-Grounded RL as controls. Following Gandhi et al. (2025), for each behavior, we use a strong LLM (Qwen3.5-122B-A10B) to count occurrences in each chain of thought (held-out set; $N = 698$). Rates are expressed per 100 reasoning words, and category scores average the rates of their constituent behaviors. Figure 4 highlights GGR and Point-Grounded RL. An example behavioral contrast is provided in Figure 7. We find:

**GGR increases exploration and evaluation while reducing planning.** After SFT, GGR-SFT increases exploration by 0.108 and evaluation by 0.059 occurrences per 100 reasoning words relative to the base model, while planning decreases by 0.026 (Figure 4, left, outlined bars). At the individual-behavior level, GGR-SFT increases scene traversal (+0.152), reference-based navigation (+0.158), and criteria verification (+0.301) relative to the base model. Selective fixation supervision therefore strengthens these behaviors during SFT.

**RL preserves the behavioral gains from SFT.** Although the exploration and evaluation gains become smaller after RL, both remain above the base model and the RL baselines (Figure 4). Relative to the base model, exploration changes by +0.071 for GGR-RL, compared with +0.005 for Point-

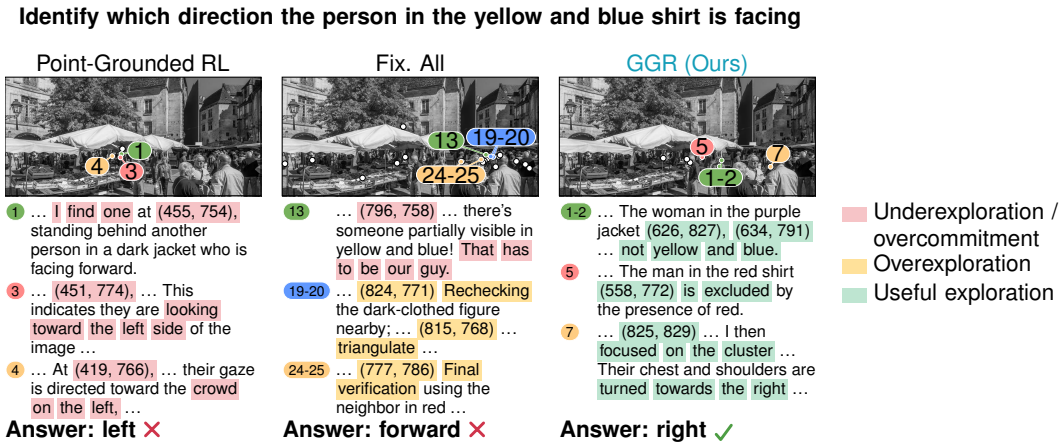


Figure 5: **Example of GGR eliciting useful grounded exploration compared to baselines.** Point-Grounded RL grounds few locations and overcommits to a location, Fix. All re-evaluates many locations leading to confusion, and GGR elicits useful exporation before finding the correct answer.

Grounded RL. Evaluation changes by $+0.048$ and $+0.020$, respectively. Planning remains below the base model ($-0.031$). The individual behaviors strengthened during SFT also remain elevated. GGR-RL exceeds Point-Grounded RL in scene traversal ($+0.088$ occurrences per 100 reasoning words), reference-based navigation ($+0.042$), and criteria verification ($+0.128$). These changes accompany comparable reasoning relative to Point-Grounded RL (Figure 4, right). Thus, RL retains the increased behaviors established by SFT at a comparable length to Point-Grounded RL.

**Annotating every fixation with reasoning leads to excessive examination.** Fix. All SFT led to higher re-examination of image regions: $+0.374$ occurrences per 100 reasoning words relative to the base model, compared with $-0.058$ for GGR-SFT, $+0.162$ for GGR-RL, and $+0.082$ for Point-Grounded RL (Figure 4a). This excessive examination may lead to confusion as shown in Figure 5.

## 6 Discussion - Conclusion

In this work, we investigated how human gaze grounding can be transferred to modern vision-language models. Across chart understanding, counting, and visual search, we find that gaze is most useful when used as selective supervision for grounded reasoning rather than a direct signal to imitate. Our findings motivate GGR, which selects task-relevant locations and uses a strong model to annotate reasoning at the selected locations. GGR obtains strong task performance on new examples while staying more human-gaze-aligned compared to grounded reasoning baselines even after subsequent RL without gaze supervision.

Our results support the idea that human gaze transfers to modern VLMs best when its computational role is adapated to the model. Human fixations acquire high-resolutions crops of the visual scene through foveation and reduce working memory demands (Hayhoe & Ballard, 2005). In modern VLMs, generating a coordinate does not on its own acquire a new view, so verbalizing fixations may impose unnatural perceptual steps in the model. This could help explain the greater re-examiniation and lower accuracy of the Fix. All condition. Future work should isolate if such mechanism is responsible for the advantage of GGR over the Fix. All condition.

The persistence of gaze-guided SFT behavior through RL suggests that gaze-guided initialization can influence how models optimize task performance. Although our RL does not reward alignment with human gaze, selectively initialized models retain greater gaze alignment and different exploration and verification behaviors at competitive task accuracy. This is consistent with prior work showing that initial reasoning behaviors can shape subsequent learning (Gandhi et al., 2025). Our results extend this question to spatial evidence selection. Whether these differences actively support performance or primarily reflect retained training patterns remains future work.

### AI USE STATEMENT

We used generative AI to construct synthetic reasoning-trace training data and to support behavioral coding of model outputs. Specifically, Qwen3.5 models annotated image regions at human fixation locations, generated grounded reasoning traces for our method and comparison conditions, and assessed candidate traces for answer correctness and reasoning quality (Section 4.1 and Appendix A.2). For behavioral analysis, Qwen3.5-122B-A10B counted explicitly expressed reasoning behaviors in model-generated traces (Section 5.4).

Additionally, we used Whisper to transcribe participants' think-aloud recordings. We used AI tools to edit software code and to in some cases, perform sentence-level editing of the manuscript for readability. The authors checked the AI-assisted code and its outputs. We take responsibility for the final content of this work, including its text, code, analyses, claims, and AI-assisted artifacts.

### ETHICS STATEMENT

This work involves research with human participants. The eye-tracking study was conducted under a protocol approved by an institutional review board (IRB). All participants provided informed consent before participation and were informed that participation was voluntary and that they could withdraw at any time. We recorded gaze and spoken responses while participants completed visual reasoning tasks, as described in Section 4.3 and the human data collection details in the appendix.

### REPRODUCIBILITY STATEMENT

We will release all models, code, and data upon publication to support reproduction. Section 4 describes the reasoning-trace construction and training pipeline, and Section 4.3 describes human gaze collection. Appendix A.2 provides annotation prompts, filtering procedures, and baseline construction details. Section 5.1 describes the evaluation benchmarks and metrics, while Appendix A.4 provides training hyperparameters, evaluation settings, and the behavioral coding procedure.

## REFERENCES

Dana H. Ballard, Mary M. Hayhoe, and Jeff B. Pelz. Memory representations in natural tasks. *Journal of Cognitive Neuroscience*, 7(1):66–80, 1995. doi: 10.1162/jocn.1995.7.1.66.

Dana H. Ballard, Mary M. Hayhoe, Polly K. Pook, et al. Deictic codes for the embodiment of cognition. *Behavioral and Brain Sciences*, 20(4):723–742, 1997. doi: 10.1017/s0140525x97001611.

Marcel Binz, Elif Akata, Matthias Bethge, et al. A foundation model to predict and capture human cognition. *Nature*, 644(8078):1002–1009, August 2025. ISSN 1476-4687. doi: 10.1038/s41586-025-09215-4.

Nicholas Budny, Kia Ghods, Declan Campbell, et al. Visual serial processing deficits explain divergences in human and vlm reasoning. arXiv preprint arXiv:2509.25142, 2026.

Declan Campbell, Sunayana Rane, Tyler Giallanza, et al. Understanding the limits of vision language models through the lens of the binding problem. *Advances in Neural Information Processing Systems*, 37:113436–113460, 2024.

Meng Cao, Haoze Zhao, Can Zhang, et al. Ground-r1: Incentivizing grounded visual reasoning via reinforcement learning. *arXiv preprint arXiv:2505.20272*, 2025.

Patricia A. Carpenter and Priti Shah. A model of the perceptual and conceptual processes in graph comprehension. *Journal of Experimental Psychology: Applied*, 4(2):75–100, 1998. doi: 10.1037/1076-898X.4.2.75.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, et al. Training verifiers to solve math word problems. arXiv preprint arXiv:2110.14168, 2021.

Matt Deitke, Christopher Clark, Sangho Lee, et al. Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 91–104. IEEE, 2025.

Thomas Fel, Ivan F. Rodriguez Rodriguez, Drew Linsley, et al. Harmonizing the object recognition strategies of deep neural networks with humans. In *Advances in Neural Information Processing Systems*, volume 35, pp. 9432–9446. Curran Associates, Inc., 2022.

Kanishk Gandhi, Ayush Chakravarthy, Anikait Singh, et al. Cognitive behaviors that enable self-improving reasoners, or, four habits of highly effective stars. arXiv preprint arXiv:2503.01307, 2025.

Robert Geirhos, Kristof Meding, and Felix A. Wichmann. Beyond accuracy: quantifying trial-by-trial behaviour of cnns and humans by measuring error consistency. In *Advances in Neural Information Processing Systems*, volume 33, pp. 13890–13902, 2020.

Elizabeth R Grant and Michael J Spivey. Eye movements and problem solving: Guiding attention guides thought. *Psychological science*, 14(5):462–466, 2003.

Udith Haputhanthri, Declan Campbell, Rim Assouel, et al. Binding visual features point by point. arXiv preprint arXiv:2605.25427, 2026.

Mary Hayhoe and Dana Ballard. Eye movements in natural behavior. *Trends in Cognitive Sciences*, 9(4):188–194, 2005. doi: 10.1016/j.tics.2005.02.009.

Marcel Adam Just and Patricia A. Carpenter. Eye fixations and cognitive processes. *Cognitive Psychology*, 8(4):441–480, 1976. doi: 10.1016/0010-0285(76)90015-3.

Mohamed Amine Kerkouri, Marouane Tliba, Aladine Chetouani, et al. Matched outcomes, divergent gaze: How foveated mllms search compared to humans. arXiv preprint arXiv:2608.16514, 2026.

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.

Xin Lai, Junyi Li, Wei Li, et al. Mini-o3: Scaling up reasoning patterns and interaction turns for visual search. arXiv preprint arXiv:2509.07969, 2025.

Michael F. Land and Mary Hayhoe. In what ways do eye movements contribute to everyday activities? *Vision Research*, 41(25–26):3559–3565, 2001. doi: 10.1016/S0042-6989(01)00102-X.

Kinhei Lee, Peiyuan Jing, Zhenxuan Zhang, et al. Seeing through experts' eyes: a foundational vision-language model trained on radiologists' gaze and reasoning. *npj Artificial Intelligence*, 2026.

Yiwei Li, Zihao Wu, Yanjun Lv, et al. Thinking with gaze: Sequential eye-tracking as visual reasoning supervision for medical vlms. *arXiv preprint arXiv:2603.06697*, 2026.

Hunter Lightman, Vineet Kosaraju, Yura Burda, et al. Let's verify step by step. *arXiv preprint arXiv:2305.20050*, 2023.

Drew Linsley, Dan Shiebler, Sven Eberhardt, et al. Learning what and where to attend. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019.

Ji Ma, Wei Suo, Peng Wang, et al. Understanding and mitigating hallucinations in multimodal chain-of-thought models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 40224–40234, June 2026.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, et al. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. arXiv preprint arXiv:2203.10244, 2022.

Ahmed Masry, Mohammed Saidul Islam, Mahir Ahmed, et al. Chartqapro: A more diverse and challenging benchmark for chart question answering. arXiv preprint arXiv:2504.05506, 2025.

Lukas Muttenthaler, Klaus Greff, Frieda Born, et al. Aligning machine and human visual representations across abstraction levels. *Nature*, 647(8089):349–355, 2025.

Roni Paiss, Ariel Ephrat, Omer Tov, et al. Teaching clip to count to ten. arXiv preprint arXiv:2302.12066, 2023.

Tingting Qiao, Jianfeng Dong, and Duanqing Xu. Exploring human-like attention supervision in visual question answering. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Alec Radford, Jong Wook Kim, Tao Xu, et al. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 28492–28518. PMLR, 2023.

Ali Salamatian, Amirhossein Abaskohi, Wan-Cyuan Fan, et al. Chartgaze: Enhancing chart understanding in lvlms with eye-tracking guided attention refinement. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 12104–12124, 2025.

Akanksha Saran, Ruohan Zhang, Elaine Schaertl Short, et al. Efficiently guiding imitation learning agents with human gaze. *arXiv preprint arXiv:2002.12500*, 2020.

Gabriel Sarch, Linrong Cai, Qunzhong Wang, et al. Vero: An open rl recipe for general visual reasoning. *arXiv preprint arXiv:2604.04917*, 2026a.

Gabriel Sarch, Snigdha Saha, Naitik Khandelwal, et al. Grounded reinforcement learning for visual reasoning. arXiv preprint arXiv:2505.23678, 2026b.

Ekta Sood, Fabian Kögel, Florian Strohm, et al. 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. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.conll-1.3.

Ilia Sucholutsky, Lukas Muttenthaler, Adrian Weller, et al. Getting aligned on representational alignment. *Transactions on Machine Learning Research*, 2025.

Jayant Sravan Tamarapalli, Rynaa Grover, Nilay Pande, et al. Countqa: How well do mllms count in the wild? arXiv preprint arXiv:2508.06585, 2025.

Hugo Touvron, Louis Martin, Kevin Stone, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Haozhe Wang, Alex Su, Weiming Ren, et al. Pixel reasoner: Incentivizing pixel-space reasoning with curiosity-driven reinforcement learning. arXiv preprint arXiv:2505.15966, 2025.

Niklas Wilming, Torsten Betz, Tim C. Kietzmann, et al. Measures and limits of models of fixation selection. *PLoS ONE*, 6(9):e24038, 2011. doi: 10.1371/journal.pone.0024038.

Zhongxing Xu, Chengzhi Liu, Qingyue Wei, et al. More thinking, less seeing? assessing amplified hallucination in multimodal reasoning models. *Advances in Neural Information Processing Systems*, 38:82878–82905, 2026.

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

Eric Zelikman, Yuhuai Wu, Jesse Mu, et al. Star: Bootstrapping reasoning with reasoning. *Advances in Neural Information Processing Systems*, 35:15476–15488, 2022.

Kaichen Zhang, Bo Li, Peiyuan Zhang, et al. Lmms-eval: Reality check on the evaluation of large multimodal models. In *Findings of NAACL*, 2025a.

Yi-Fan Zhang, Huanyu Zhang, Haochen Tian, et al. Mme-realworld: Could your multimodal llm challenge high-resolution real-world scenarios that are difficult for humans? arXiv preprint arXiv:2408.13257, 2025b.

Chujie Zheng, Shixuan Liu, Mingze Li, et al. Group sequence policy optimization. arXiv preprint arXiv:2507.18071, 2025.

Chunzheng Zhu, Yangfang Lin, Shen Chen, et al. Medeyes: Learning dynamic visual focus for medical progressive diagnosis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 13916–13924, 2026.

# A APPENDIX

A.1 Additional Human Data Collection Details . . . . . 13
A.2 Model Annotation Details . . . . . 13
A.3 Baseline Details . . . . . 14
A.4 Training and Evaluation Details . . . . . 15
A.5 Evaluation Protocol. . . . . 16
A.6 Fixation Alignment Metric. . . . . 16
A.7 Additional Experiments . . . . . 16
A.8 Additional Trial Examples . . . . . 18

## A.1 ADDITIONAL HUMAN DATA COLLECTION DETAILS

English-speaking participants with normal or corrected-to-normal vision were recruited under an institutional review board-approved protocol. All participants provided informed consent and were informed that participation was voluntary and that they could withdraw at any time. Stimuli were presented using PsychoPy on an Asus VG27AQ LCD monitor operating at $2560 \times 1440$ pixels and 120 Hz. Participants viewed the display from approximately 89.5 cm with their head stabilized using a chin rest. Right-eye gaze was recorded using an SR Research EyeLink 1000 Plus desktop system at 1,000 Hz. Calibration and validation were performed before each session using the standard nine-point EyeLink array, with drift-correction trials interleaved every 10–20 task trials.

Participants were instructed to verbalize their reasoning and final answer before proceeding to the next trial. Trial audio was recorded and transcribed using Whisper large-v3 (Radford et al., 2023), with word-level timestamps. These verbalizations were not used to generate training traces because participants varied substantially in how extensively they explained their reasoning.

Eye-tracking recordings were processed at the fixation-event level. Finalized EDF files were converted to an ASCII event representation using EyeLinkIO (version 0.3.0) and the EyeLink EDF API, retaining the fixation onset times, durations, and mean screen-space gaze positions recorded by EyeLink. No additional offline fixation detection, duration filtering, or fixation merging was applied. Recorded synchronization markers aligned the PsychoPy and tracker clocks, and fixation times were expressed relative to trial onset. Fixations ending before trial onset were discarded, while those spanning trial onset were truncated to their within-trial duration.

To place gaze from different viewing states in a common reference frame, screen-space fixation positions were mapped to original-image coordinates by inverting the stimulus display geometry and applying the logged zoom level and viewport center active at fixation onset. Fixations mapped outside the original image were discarded, and retained coordinates were rounded to image pixels. No additional interpolation of missing gaze samples or explicit blink-window filtering was performed. Trials explicitly flagged as bad were excluded; participant-level exclusions were made only when manual inspection indicated particularly poor eye-tracking data quality.

## A.2 MODEL ANNOTATION DETAILS

**Three-Stage Reasoning Annotation.** We construct the annotation context through three stages. First, the annotation model assigns a visual description to each candidate fixation in $F$, identifying the object or region at that location. Second, conditioned on the question $q$, it annotates the task relevance of these regions, such as identifying a target, ruling out a distractor, or checking for an omitted object. Third, the model uses the image, question, and annotated regions to generate candidate reasoning traces. We exclude the participant's raw think-aloud transcript from this generation stage and withhold the reference answer $a^*$, requiring the model to derive its answer from the visual evidence.

**Rejection and Visual Verification.** We filter candidate traces using answer checks and image-grounded assessments of reasoning quality. The checks remove candidates with incorrect answers,

invalid point references, excessive repetition, or commentary about the annotation process. A subsequent visual assessment checks whether the trace identifies the correct objects and attributes, supports its claims with appropriate spatial references, and avoids counting errors, contradictions, or unsupported conclusions. Accepted traces are converted into training targets containing the grounded reasoning and final answer.

**Listing A.1: Visual Reasoning Trace Prompt**

Write a task-directed visual reasoning trace from the image, question, and annotated human fixation points.

---

**POINT SELECTION**

- Select useful supplied points. Omit others and use them in any semantic order.
- Cite at least one supplied point using its `<N>` placeholder.
- Cite individual points, not coordinate ranges or grouped IDs.
- Describe the scene directly; do not narrate the human scanpath or annotation process.

---

**REASONING**

- Ground observations in the image and keep them relevant.
- Verify object identity, attributes, counts, and relationships.
- A correction or one verification pass is acceptable. Avoid repeated recounting, irrelevant exploration, and loops.
- Never mention reference answers, judges, or previous attempts.

---

**OUTPUT**

Return only the reasoning trace followed by exactly one `\boxed{answer}`. Do not include `<think>` tags or text after the answer.

{Few-shot examples}

Image: {image}
Question: {question}
Annotated fixations:
<1> {visual label} — {task-related function}
<2> {visual label} — {task-related function}
...

### A.3 Baseline Details

**Fix. + Think-Aloud.** We process the raw fixation sequence and think aloud into a single reasoning chain. We create a flattened list of (x, y) coordinates and interleave the think aloud in between the coordinate list at locations that align based on the Whisper transcriptions. Each flattened list + think aloud + final answer is converted into a training target.

**ViGoRL-MCTS.** Following Sarch et al. (2026b), we generated spatially grounded reasoning trajectories with Qwen3.5-122B-A10B using Monte Carlo Tree Search (MCTS), with 16 simulations per example, four children per expansion, and four rollouts per child. The teacher received the image and question and generated reasoning steps grounded in image coordinates, while reference answers were used to score completed rollouts and guide search. We retained answer-correct trajectories and constructed correction examples by pairing unsuccessful and successful continuations from the same node. After deduplication, we selected up to 25 trajectories per example.

| Setting | Configuration |
|---|---|
| Learning rate | $5 \times 10^{-6}$ |
| Schedule | Cosine decay with 3% warmup |
| Weight decay | 0 |
| Gradient clipping | Max norm 1.0 |
| Training epochs | 5 |
| Batch size | 64 |
| Maximum sequence length | 32,768 tokens |
| Image pixel bounds | 921,600–8,388,608 |
| Precision / parallelism | BF16 / DeepSpeed ZeRO-3 across 8 GPUs |

Table 2: SFT training configuration.

| Setting | Configuration |
|---|---|
| Learning rate | 20-step warmup to $2 \times 10^{-5}$; linear decay to $7 \times 10^{-6}$ at step 400 |
| Prompt batch / optimization minibatch | 256 / 128 |
| Responses per prompt | 8 |
| GSPO clipping | $\epsilon_{\text{low}} = 3 \times 10^{-4}$, $\epsilon_{\text{high}} = 4 \times 10^{-4}$ |
| Rollout sampling | Temperature 1.0, top-$p$ 1.0 |
| Maximum prompt / response length | 16,384 / 14,336 tokens |
| Precision / parallelism | BF16 / FSDP2 across 8 GPUs |

Table 3: RL training configuration.

### A.4 Training and Evaluation Details

**SFT Training Protocol.** All baselines and our models use the system prompt in Listing A.2. We fine-tune using supervised next-token prediction on the reasoning trace and final answer. Training uses full-parameter fine-tuning of the language model with the vision encoder frozen.

**Listing A.2: SFT and RL System Prompt**

You are a helpful, conversational assistant tasked with answering a question about an image. Generate a reasoning trace with reasoning grounded in points in the image ((x, y) coordinates) to help you answer.

1. **Reasoning**: A detailed, free-flowing chain of thought enclosed in `<think>` and `</think>` tags, with reasoning grounded in points in the image ((x, y) coordinates).
   While reasoning, explicitly model these cognitive behaviors:
   - **Planning**: Outline possible approaches before committing.
   - **Exploration**: Consider multiple image regions or interpretations.
   - **Evaluation**: Compare alternatives and verify against visual evidence.
   - **Re-examination**: Revisit earlier ideas if they might still be viable.
   - **Error Correction**: Acknowledge mistakes if grounded reasoning reveal an incorrect path.
2. **Response**: After reasoning and reaching a conclusion, generate a clear, conversational response. After the response, indicate your final answer using `\boxed{}`.

**RL Training Protocol.** We use the system prompt in Listing A.2. We train with GSPO for 400 training steps. Each batch contains 256 prompts balanced across chart, counting, and search tasks. Advantages do not use standard-deviation normalization. We do not apply KL regularization nor an entropy bonus. We use full-finetuning of all parameters. RL hyperparameters are detailed in Table 3.

The reward combines task-specific answer correctness (weight $0.95$) and valid point-grounded reasoning format (weight $0.05$); invalid formatting also zeros the correctness component. For counting, the correctness reward is $2^{-|\hat{a}-a^*|}$ for absolute errors below four and zero otherwise.

| Setting | Configuration |
|---|---|
| Chart benchmarks | ChartQA, ChartQA Pro |
| Counting benchmarks | CountQA, CountBenchQA |
| Search benchmarks | MME-RealWorld-Lite (Vero subsample), Visual Probe |
| Decoding | Temperature 1.0, top-$p$ 0.95, top-$k$ 20, presence penalty 1.5 |
| Output budget | 16,384 tokens |
| Image preprocessing | Min 2M and max 10M |

Table 4: Evaluation configuration.

| Category | Coded behaviors |
|---|---|
| Planning | Scope definition, task decomposition, and viewpoint calibration. |
| Exploration | Systematic scene traversal, reference-based navigation, grouping, enumeration, and pattern-based scaling. |
| Evaluation | Criteria verification, reasoning about uncertain evidence, cross-instance verification, consensus-based resolution, fragment integration, and constraint relaxation. |
| Re-examination | Re-examination of earlier observations and verification of provisional or final results. |

Table 5: Behavioral coding categories. Counts describe behaviors expressed in generated reasoning, rather than independently verified visual actions.

### A.5 Evaluation Protocol.

We detail our evaluation protocol in Table 4. We use lmms-eval (Zhang et al., 2025a) for all evaluations.

**Behavioral Analysis.** We analyze reasoning traces from held-out gaze evaluations using Qwen3.5-122B-A10B as a text-only behavioral coder, with temperature zero and thinking disabled. Each rollout is coded separately for 16 behaviors organized into four categories (Table 5). The coder counts distinct, explicitly expressed occurrences. The behaviors are summarized in Table 5.

### A.6 Fixation Alignment Metric.

For example $i$, let $h_{ij}$ denote human fixation $j$, and let $p_{irkt}$ denote grounding point $k$ from model response $r$ at checkpoint $t$. Coordinates lie in the normalized image plane $[0, 1000]^2$. After excluding invalid and out-of-bounds coordinates, we construct

$$q_{cirt}(x) = (1 - \epsilon_c)\frac{1}{M_{irt}} \sum_{k=1}^{M_{irt}} \mathcal{N}(x; p_{irkt}, \sigma_c^2 I) + \epsilon_c u, \qquad u = 10^{-6},$$

where $c$ indexes training condition and $M_{irt}$ is the number of valid reported points.

We score each response by its mean log-likelihood relative to uniform:

$$A_{cirt} = \frac{1}{N_i} \sum_{j=1}^{N_i} \log_2 \frac{q_{cirt}(h_{ij})}{u}.$$

We jointly fit $(\sigma_c, \epsilon_c)$ separately for each condition using 336 calibration examples. The candidate grid is

$$\sigma \in \{25, 35, 50, 70, 100, 140, 200\}, \qquad \epsilon \in \{0.01, 0.03, 0.1, 0.2, 0.4, 0.7, 0.9\}.$$

Selection maximizes calibration-set mean alignment.

### A.7 Additional Experiments

Figure 6 provides additional plots of GGR and Point-Grouded RL across RL steps for unique points output, human fixation coverage, and human fixation precision on the held out set. Figure 7 provides additional behavior and category comparisons across SFT and RL.

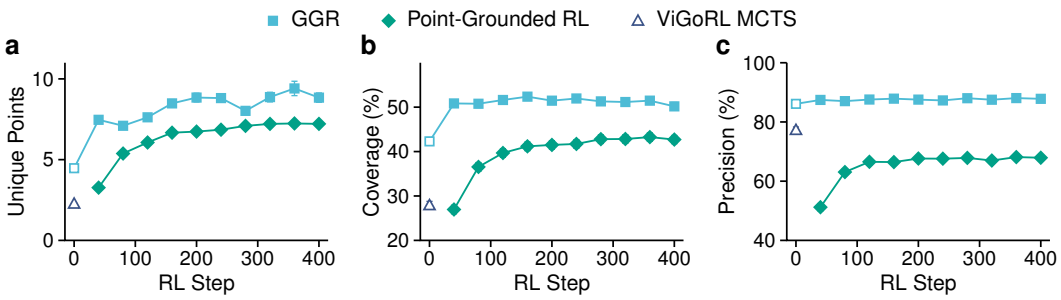


Figure 6: Spatial behavior across RL training. Reported is mean ± SEM across images. Open markers denote SFT or pre-RL initialization; filled markers denote RL.

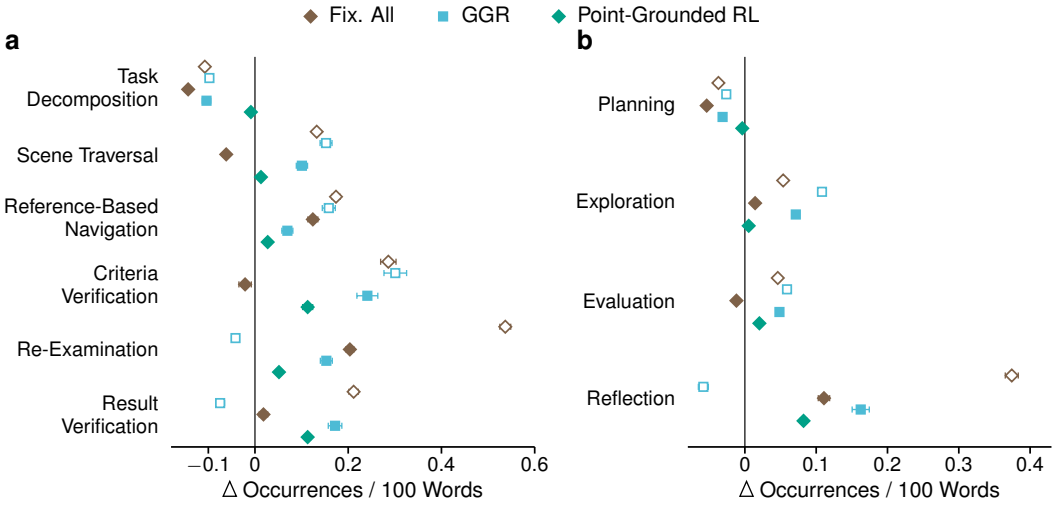


Figure 7: Full behavioral contrasts relative to base. Plotted is individual behaviors (left) and categories (right), in occurrences per 100 reasoning words. Reported is mean ± SEM across paired image differences. Open markers denote SFT and filled markers denote RL.

### A.8 Additional Trial Examples

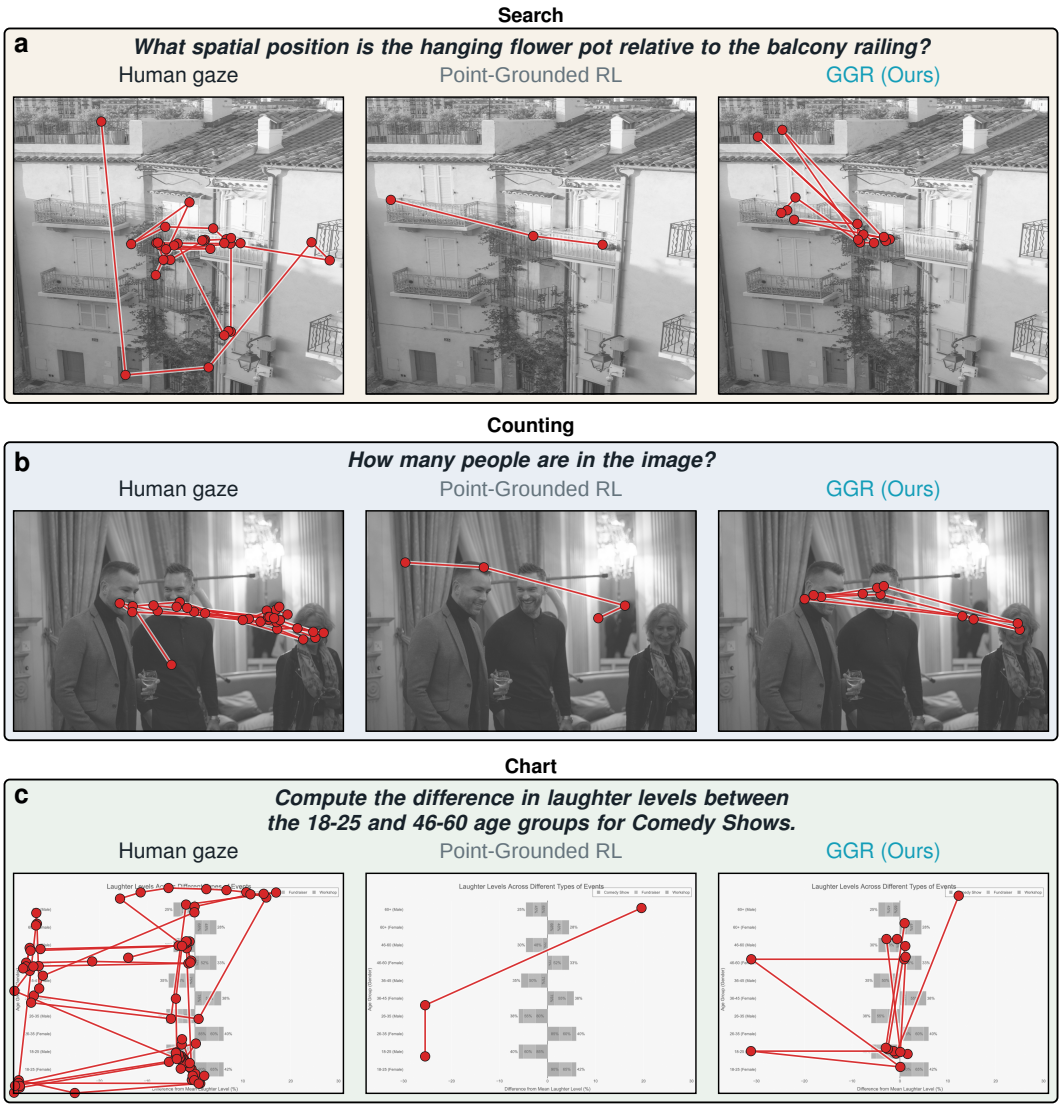


Figure 8: Human gaze and model grounding traces from Point-Grounded RL and GGR on additional held-out examples across our three datasets.