GAZEVLM: ACTIVE VISION VIA INTERNAL ATTENTION CONTROL FOR MULTIMODAL REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Human visual reasoning is governed by active vision, a process where meta-cognitive control drives top-down goal-directed attention, dynamically routing foveal focus toward task-relevant details while maintaining peripheral awareness of the global scene. In contrast, modern Vision-Language Models (VLMs) process visual information passively, relying on the static accumulation of massive token contexts that dilute the visual evidence as the reasoning chain grows. Here we propose GazeVLM, a multimodal model that learns to exert this meta-cognitive oversight over its attention resources by generating gaze actions in its reasoning chain, in the form of <LOOK> tags specifying the coordinates of the region to inspect. During training, each <LOOK> block triggers a continuous suppression bias on the attention logits that dampens features outside the regions gazed at so far, and that lifts at the end of the block, restoring the global view. The model is trained with this suppression bias as a teaching signal, first by Supervised Fine-Tuning on curated gaze-reasoning traces, then through Group Relative Policy Optimization (GRPO) with rewards for correct answers and valid grounding. At deployment, the external suppression is removed and the model, running its unmodified forward pass, steers its own attention to the regions indicated by the <LOOK> blocks, retaining most of the effect of the bias and emulating top-down control of spatial attention. This model thus transitions between global spatial awareness and localized focal reasoning without relying on external agentic tools like cropping, and without adding visual tokens from re-encoded patches. Applied to 4B-parameter backbones, GazeVLM demonstrates strong high-resolution multimodal reasoning on HRBench-4k and HRBench-8k, surpassing its base models by about 4 points, and outperforming agentic multimodal pipelines built around thinking with images by 5 to 12 points.

1 INTRODUCTION

Despite the rapid evolution of Vision-Language Models (VLMs) and the successful adaptation of Chain-of-Thought (CoT) prompting to the multimodal domain (Zhang et al., 2023), multi-step visual reasoning remains bottlenecked by a fundamental imbalance between textual generation and visual perception. Modern VLMs process visual information passively; they rely on the static accumulation of massive token contexts. As VLMs engage in extended, autoregressive reasoning, this influx of textual data increasingly dominates the causal attention mechanisms, and the original visual features tend to be progressively underweighted. Consequently, the model often over-relies on linguistic priors and frequently hallucinates objects (Rohrbach et al., 2018; Biten et al., 2022; Li et al., 2023) or logical steps, drifting away from the spatial content of the image.

Current paradigms attempt to address this imbalance, but they largely inherit the same passive processing of visual information. One approach, grounded reasoning, forces models to output bounding box coordinates alongside textual claims (Peng et al., 2023; You et al., 2024; Man et al., 2025). While this enforces a superficial textual link to the image, it does not alter how the decoder attends to the pre-encoded visual features during generation. Another approach augments VLMs with external "zoom-in" tools (e.g., V* (Wu & Xie, 2024), Pixel Reasoner (Su et al., 2026), and Ground-R1 (Cao et al., 2025)). These frameworks isolate relevant details by explicitly cropping and re-encoding new image patches. While this improves local detail extraction, it incurs a substantial computational penalty: as the recent Adaptive-CoF analysis (Zhang et al., 2025) shows, these approaches can add

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.

nearly a thousand additional visual tokens to the context window during high-resolution reasoning. This sparse cropping acts as disjointed snapshots carrying no global position in the image, rather than a cohesive scene, which bloats the model's context with re-encoded pixels.

In human cognition, resolving complex visual queries is not a passive process of statically accumulating pixels or "cropping" the visual field. Instead, it relies on the top-down mechanism of active vision (Yarbus, 1967; Findlay & Gilchrist, 2003). This is fundamentally a meta-cognitive ability requiring observers to continuously maintain peripheral global awareness of the scene, monitor their understanding, and actively plan where to look and what to attend next (Rensink, 2000). They then dynamically route their focus of attention and foveal gaze to specific local regions, extracting granular details while tuning out background noise (Corbetta & Shulman, 2002). Humans intuitively distinguish between finding a target (global spatial awareness) and reasoning about its specific details (focal fixation). Our aim is to equip VLMs with this kind of meta-cognitive capability: an iterative, computationally efficient control loop that flexibly transitions between global and local views without reprocessing new image content.

In this work, we introduce GazeVLM, a visual reasoning framework that internalizes human-like active vision into the weights of a standard VLM. Rather than relying on external cropping tools, GazeVLM learns top-down control over how the decoder attends to the pre-encoded visual features. Using its global context, the model meta-cognitively assesses its need for information and generates special `<LOOK>` tokens paired with bounding box coordinates. During training, each `<LOOK>` token triggers a continuous, plateau-shaped suppression bias on the attention logits that dampens irrelevant visual areas, emulating covert spatial attention, while maintaining zero penalty for the targeted region. The closing `</LOOK>` token lifts the bias and restores access to the full visual context to dictate the next logical step. Crucially, this bias is a teaching signal rather than an inference-time mechanism: at deployment it is removed, and the model routes its own attention to the regions designated by its `<LOOK>` blocks while running the unmodified forward pass of its backbone.

To instill this top-down control, we adopt a two-stage training paradigm. First, we warm-start the model via supervised fine-tuning on a curated dataset demonstrating iterative human-like scanning. We then refine this policy via Group Relative Policy Optimization (GRPO) (Shao et al., 2024), using a reward system that incentivizes geometrically valid grounding while explicitly penalizing redundant or excessive gazing. In summary, our key contributions are:

- A curated gaze-reasoning dataset: we build an instructional dataset of verified gaze traces designed to train VLMs to transition from passive observation to active, multi-region visual navigation.
- A training signal for active vision: we introduce a gaze bias that modulates the VLM's attention logits over already-encoded visual features during training, teaching the model to autonomously shift focus from a global view to local details and back. It requires no change to the architecture and adds no new visual tokens.
- Internalization of the attention bias teaching signal: we show that at deployment time, the attention of the trained model running the unmodified forward pass (with the bias switched off) successfully focuses on the region in its `<LOOK>` block.
- Strong high-resolution reasoning without tool calling and re-encoded visual tokens: applied on 4B backbones, we show that GazeVLM lifts the performance of its base models by about 4 points on HRBench-4k and HRBench-8k, outperforming 7B zoom-in agents by 5 to 12 points, while generating about $5\times$ fewer output tokens.

2 Related Work

Vision-Language Model reasoning. Considerable research has focused on enhancing the reasoning capabilities of open-weight Vision-Language Models (VLMs) such as Qwen-VL (Bai et al., 2023; Wang et al., 2024; Bai et al., 2025b;a), LLaVA (An et al., 2025; Liu et al., 2024a), and InternVL (Zhu et al., 2025; Chen et al., 2024b). Following the success of Large Language Models (LLMs), a primary strategy has been the adaptation of Chain-of-Thought (CoT) prompting to the multimodal domain, encouraging models to articulate intermediate logical steps (Zhang et al., 2023).

To ensure these textual steps remain tethered to the image, a distinct line of work emphasizes grounded reasoning, requiring models to output bounding box coordinates (Peng et al., 2023; You et al., 2024; Man et al., 2025) or utilize superimposed visual markers (Yang et al., 2023). Another prominent direction involves augmenting VLMs with external tools, most notably "zoom-in" tools or visual search agents (e.g., V* (Wu & Xie, 2024), Pixel Reasoner (Su et al., 2026), Ground-R1 (Cao et al., 2025), DeepEyes (Zheng et al., 2026), Visual Perception Token (Yu et al., 2025)). While effective at capturing fine-grained details, these tool-use frameworks repeatedly crop and pass new image patches through the vision encoder, generating additional visual tokens at each reasoning step. Closer to our work, (Yu et al., 2025) add Visual Perception Tokens to the vocabulary, which the model generates to trigger its own perception. However, their region selection tokens, still crop and re-encode the selected region, and their vision re-Encoding tokens invoke an additional encoder (DINOv2 or SAM) through a learned projector; both appending new visual tokens.

Attention steering and hallucination in VLMs. Improving the reasoning of VLMs is closely linked to the reduction of multimodal hallucinations. VLMs are notoriously prone to "object hallucination" (Rohrbach et al., 2018; Biten et al., 2022; Li et al., 2023), where there is an over-reliance on the language prior during generation with respect to the actual visual evidence. Techniques such as Visual Contrastive Decoding (VCD) (Leng et al., 2024) attempt to penalize language-driven hallucinations by contrasting output logits between original and explicitly distorted visual inputs. More recently, Attention-Steerable Contrastive Decoding (ASCD) (Wang et al., 2026) showed that similar hallucination reduction can be achieved more efficiently by directly manipulating the model's internal attention scores, dynamically amplifying logical text-heads while dampening spurious visual tokens. However, these methods are primarily designed to calibrate a single, direct answer and do not account for the dynamic, multi-step nature of complex reasoning.

In contrast to these test-time interventions, or to the injection of zoom-in visual tokens, our work introduces a training strategy that adds a bias to the decoder's attention logits over already-encoded visual features. The bias is applied only during training: at deployment the model runs the unmodified forward pass of its backbone. Inspired by human gaze scanpaths (Yarbus, 1967; Findlay & Gilchrist, 2003), our method enables the VLM to autonomously steer its own visual focus, seamlessly transitioning between global context and local details throughout its reasoning trace. This allows the model to fixate precisely on task-relevant regions without the overhead of physically manipulating or re-encoding the image.

3 Gaze Reasoning

3.1 Gaze actions

We introduce GazeVLM, a visual reasoning method that enables a VLM to autonomously focus on specific regions of an image, extract relevant information and return to a global view, all through attention-steering of the pre-encoded visual tokens.

The architecture of GazeVLM keeps the standard VLM design (e.g. Qwen-VL) unchanged, and comprises a vision encoder, a projector and an LLM. GazeVLM reasons by interleaving text with *gaze actions* as shown in Figure 1, and is trained with an attention bias that ties each action to the model's attention. Crucially, the bias is a training signal only and is removed at deployment. Gaze actions are `<LOOK>` tokens that the model can emit and are annotated with a bounding box indicating the region to focus on. Concretely, the model emits: `<LOOK at="[target]" bbox=[x1, y1, x2, y2]> [observation] </LOOK>`.

As it reasons, we allow the model to also generate chains-of-thought with `<THINK> [thought] </THINK>` at the beginning of the reasoning and between different `<LOOK>`s. The model then provides the final answer in `<ANSWER>` *final answer* `</ANSWER>` tags.

3.2 Gaze bias as a training signal

Training and deployment policies. Given a vision-language query (I, x), where I is the input image and x is the question, at deployment GazeVLM samples the policy $y_t \sim \pi_\theta(\cdot|I, x, h_t)$

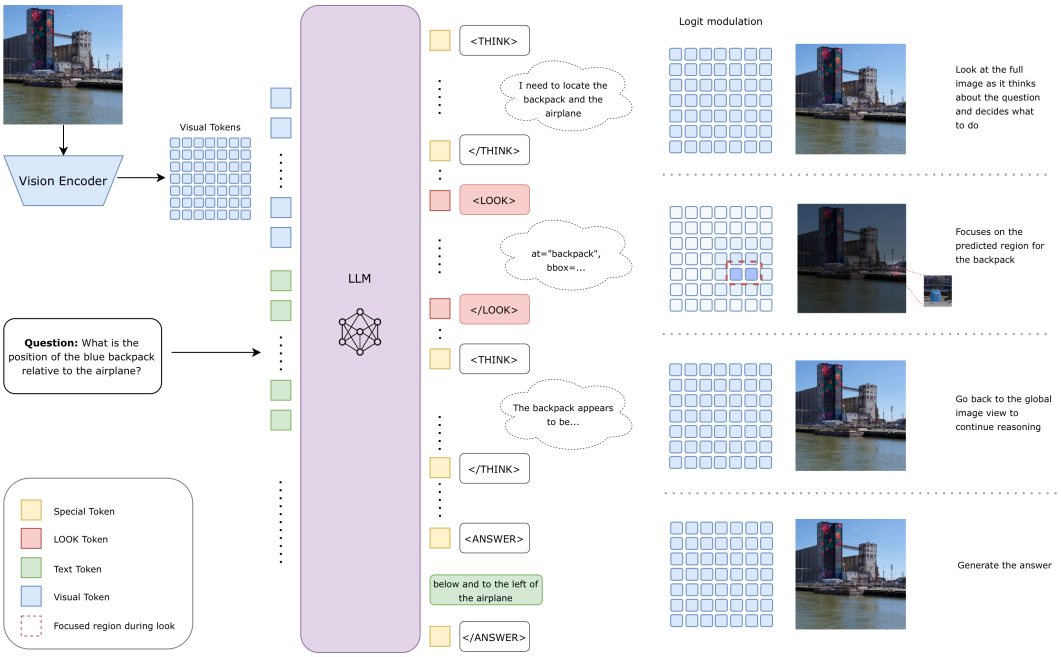


Figure 1: Architectural overview of the GazeVLM reasoning pipeline. Upon receiving a visual query, the model iteratively interleaves logical reasoning <THINK> with active visual navigation. At training, generating a <LOOK> token triggers top-down logit modulation, applying a continuous suppression bias to task-irrelevant visual tokens while preserving the focal region. Once local extraction is complete, the bias is lifted, restoring the global visual context to inform the next reasoning step until reaching the final <ANSWER> generation. The bias is not triggered at deployment.

parametrized by model weights θ with the unmodified forward pass of its base VLM, and where h_t is the generated history of previous steps.

During training, the forward pass is modified by a bias $\beta(B_t)$ (described below) operating on the attention logits, where B_t denotes the bounding boxes of regions of interest according to the trace h_t. Accordingly, the token at the t-th generation step is sampled as $y_t \sim \pi_\theta^{\beta(B_t)}(\cdot|I, x, h_t)$ from the biased policy $\pi_\theta^{\beta(B_t)}$. As a result, the intuition is that during the reasoning process, the policy can decide to focus on key specific areas on the image while generating text grounded on these areas. When no focus is active, the model has full access to the global visual context. The iterative reasoning process terminates when a designated end token is generated.

Attention suppression bias. The vision encoder maps input image I to visual tokens V. Each visual token $v_i \in V$ is assigned a normalized 2D coordinate $p_i = (c_i, r_i) \in [0, 1]^2$ of the image patch it encodes.

To enforce focus on a bounding box $b = [x_1, y_1, x_2, y_2]$, instead of re-encoding a cropped image of the box content as other works do, we define $D(p_i, b)$ to be the Euclidean distance from p_i to the box:

$$D(p_i, b) = \sqrt{\max(x_1 - c_i, c_i - x_2, 0)^2 + \max(y_1 - r_i, r_i - y_2, 0)^2}. \quad (1)$$

The suppression bias on token v_i is then

$$\beta_i(b) = -\alpha_s \frac{D^2(p_i, b)}{2\,\sigma^2} \quad (2)$$

where σ sets the spatial falloff and α_s scales the suppression strength. This bias is added to the attention logits (Equation (3)), leaving tokens in the box untouched, while attention to tokens outside is dampened by a Gaussian factor which falls exponentially to zero moving away from the bounding box. Figure 2 illustrates the effect of the bias on a single high-resolution example. Given oracle boxes on the two objects in the question, the base Qwen3-VL-4B, with no gaze training, never answers correctly with unmodified attention ($\alpha_s = 0$); as α_s grows its accuracy rises and levels off at about 55% from $\alpha_s = 4$, the value we use throughout.

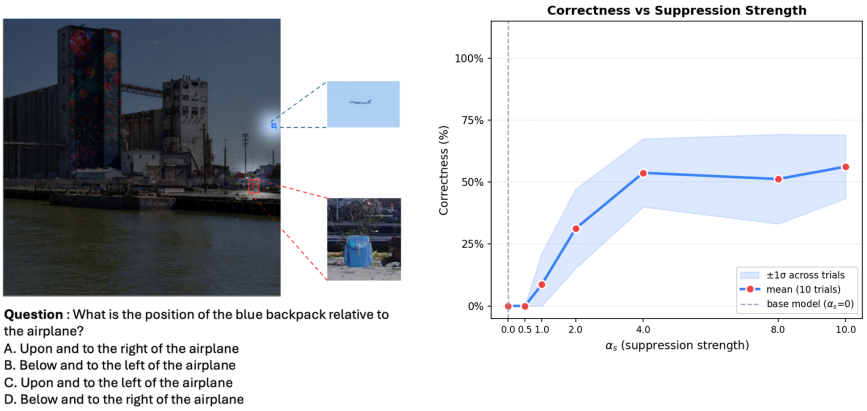


Figure 2: Effect of the suppression bias on one HRBench-4k question. The base Qwen3-VL-4B, without gaze training, is given oracle boxes on the airplane and the backpack; accuracy over 10 trials of 8 rollouts each, as a function of the suppression strength α_s (Equation (2)).

Gaze accumulation and release. Questions that compare or relate multiple objects require evidence from multiple regions. Previous works using zoom-in as a tool retain that evidence by appending to the context the tokens that encode each relevant cropped region. GazeVLM on the other hand accumulates evidence across multiple gaze actions by aggregating their corresponding bounding boxes. In particular, after the model has generated a sequence of K target bounding boxes $B = \{b_1, b_2, \ldots, b_K\}$ within its reasoning trace, we apply a bias that is computed as the element-wise maximum across all accumulated gaze actions: $\beta_i(B) := \max_{b \in B} \beta_i(b)$. This ensures that all visual tokens within any previously inspected area remain fully activated. As soon as a closing token `</LOOK>` is generated, the bias is released, so that the model attends to the full image until the next `<LOOK>` token.

Gaze bias through attention logit modulation. The gaze bias is added to the attention logits of all layers and heads of the language decoder. Given a query q_t and a visual key k_i, the pre-softmax attention score $s(q_t, k_i)$ is modulated additively:

$$\tilde{s}(q_t, k_i) = s(q_t, k_i) + \mathbf{1}_{\text{text}}(q_t) \cdot \beta_i(B_t), \tag{3}$$

where $s(q_t, k_i) = \frac{q_t \cdot k_i}{\sqrt{d}}$ with d the head dimension, and $\mathbf{1}_{\text{text}}(q_t)$ is 1 for queries from text tokens and 0 for queries from visual tokens. This design ensures that the gaze bias only affects how the language model attends to the image. By isolating the bias to text-to-vision queries, we direct the LLM's focus towards the region inside B_t, while preserving the internal representations of the global scene.

Importantly, at deployment we set $\beta \equiv 0$, meaning that Equation (3) reduces to the standard attention of the base VLM.

3.3 Training

We employ a two-stage training pipeline that includes a warm-start SFT training followed by reinforcement learning of the SFT model using GRPO (Shao et al., 2024).

Warm-start Instruction Tuning. We want to integrate into existing VLMs a new reasoning paradigm where the model can reason, focus on an area through our gazing mechanism, and then zoom-out into the global image to continue to reason until it reaches the final answer. To equip an existing Qwen model such as Qwen3-VL-Instruct (Bai et al., 2025a) to follow this pattern, which it learns during RL, we first need to help the model understand how it should do that through SFT (Supervised Fine-Tuning).

Dataset generation. We design our data generation pipeline so that our data (reasoning traces) display the gazing mechanism that we want to teach the model to follow. Our specialized dataset is sourced from: GQA (Hudson & Manning, 2019), ChartQA (Masry et al., 2022), PlotQA (Methani et al., 2020), and InfoVQA (Mathew et al., 2022). We leverage Qwen3-VL-235B-A22B as our expert

model to synthesize trajectories that explicitly utilize our gazing mechanism. Comprehensive details on dataset generation and heuristic filtering are provided in Section A.1.

We train the base model on these synthesized traces using standard cross-entropy loss, activating the continuous suppression bias during the forward pass so the model learns the visual consequences of its <LOOK> coordinates.

Learning gaze actions with Reinforcement Learning. Starting from the SFT model, we optimize the policy with GRPO (Shao et al., 2024). Inspired by other works in RL (Su et al., 2026; Pathak et al., 2017), we craft a gaze-driven reward to encourage the model to *look* at relevant areas of the image during its reasoning. More precisely, for a given sample consisting of an image I, a question x, and a ground truth answer y^*, we sample a group of G distinct completions $\{c_1, c_2, \ldots, c_G\}$ under the policy π_θ^β with the bias active. Each completion c_i is then scored by a composite reward function $R(c_i)$ designed to enforce answer correctness, visual navigation, and grounding. For a completion c, we specify the total reward as

$$R(c) = R_{\text{correct}}(c) + R_{\text{format}}(c) + R_{\text{bbox}}(c) + R_{\text{overlap}}(c) + R_{\text{excess}}(c) + R_{\text{len}}(c). \tag{4}$$

A comprehensive breakdown of each specific reward component, the penalty thresholds, and the formal GRPO objective are detailed in Section A.2.

4 EXPERIMENTS

Experimental Setup and Baselines. We build GazeVLM upon the Qwen3-VL-4B-Instruct architecture. The model is trained on our highly curated, gaze reasoning traces and optimized via GRPO. We evaluate GazeVLM against multiple VQA benchmarks spanning complex chart and logical reasoning (MathVista (Lu et al., 2024), ChartQA (Masry et al., 2022)), general reasoning (MMBench (Liu et al., 2024b), MMStar (Chen et al., 2024a), CV-Bench (Tong et al., 2024)), and high-resolution benchmarks (HRBench-4k, HRBench-8k (Wang et al., 2025)). We compare our results against a wide range of models, including closed-source models, vanilla open-source VLMs, SFT-trained reasoning models (Vision-R1-LlamaV-CI (Huang et al., 2026), SPARC (Avogaro et al., 2026)), and state-of-the-art RL-trained agents (Vision-R1 (Huang et al., 2026), Pixel-Reasoner (Su et al., 2026), DeepEyes (Zheng et al., 2026), and Ground-R1 (Cao et al., 2025)). Full details regarding dataset sizes, training settings, and baseline evaluation protocols are provided in Section A.1 and Section A.3.

Main results. Figure 3 provides a qualitative illustration of how GazeVLM training reshapes visual attention, on a representative sample from HRBench-4k. We visualize the decoder attention of the Qwen3-VL-4B-Instruct base model (left) and of its GazeVLM-finetuned counterpart (right), averaged over the tokens generated in response to the query *"Where is the water bottle placed relative to the person in the image?"*. The base model's attention is diffuse, and it answers incorrectly that *"The bottle is in front of the person"*. GazeVLM, by contrast, focuses tightly on the water bottle, within the bounding box that it emitted in its own <LOOK> trace, and answers correctly that *"The bottle is placed to the right of the person"*.

Beyond this qualitative illustration, Table 1 reports our main results aggregated across a suite of VQA benchmarks. To rigorously isolate the contribution of our gaze mechanism from that of our data curation, we highlight two reference baselines. The first is the off-the-shelf base model Qwen3-VL-4B. The second is a computationally matched passive grounding baseline, denoted GazeVLM (w/o gaze bias): trained on the same curated dataset and the same GRPO pipeline, but operating entirely passively, emitting text and bounding box coordinates without modulating the underlying attention logits, in the spirit of grounded reasoning works such as Argus (Man et al., 2025).

This passive grounding baseline is already competitive, outperforming RL-trained agentic VLMs such as DeepEyes (Zheng et al., 2026) on HRBench-8k (74.4 vs. 72.6) and validating the quality of our curated traces. Activating our core architectural contribution, the dynamic top-down attention suppression, then delivers a further performance leap on top of this strong baseline, with consistent gains across chart and math reasoning, general reasoning, and high-resolution benchmarks. On complex logical reasoning, GazeVLM achieves state-of-the-art results within its parameter class,



(a) Vanilla Qwen3-VL-4B-Instruct (b) GazeVLM

Figure 3: **Internalized visual focus of GazeVLM.** We compare *(a)* vanilla Qwen3-VL-4B-Instruct and *(b)* GazeVLM on sample 32 of HRBench-4k, on the question "*Where is the water bottle placed relative to the person in the image?*". Both models run the same unmodified forward pass, as GazeVLM at *(b)* reflects the attention pattern internalized through training. The green box is the region GazeVLM designated in its own <LOOK> block. Heatmaps show the attention from each generated trace token to the visual tokens, averaged over tokens, decoder layers, and heads. Training shifts attention from a diffuse, content-driven distribution to one tightly focused on the region the model identifies as task-relevant.

Table 1: Comprehensive evaluation of GazeVLM against state-of-the-art vision-language models. We group benchmarks by their primary cognitive demands.

	Chart & Math Reasoning		*General Reasoning & Vision critical*			*High-Resolution*	
Method	**ChartQA**	**MathVista**	**MMBench**	**MMStar**	**CV-Bench**	**HRBench-4k**	**HRBench-8k**
Closed-source models							
OpenAI-o1	-	73.9	-	-	-	-	-
GPT-4o	-	-	-	-	66.0	59.0	55.5
Gemini2.5-Flash-Lite	-	70.3	-	-	64.9	67.2	-
Vanilla Open-source VLMs							
Qwen2.5-VL-7B	83.3	67.9	83.8	57.9	73.7	68.5	59.5
Qwen3-VL-4B	83.2	71.7	87.7	67.3	83.4	79.5	73.6
Qwen3.5-4B-Inst	78.8	55.7	83.8	58.3	81.1	78.6	74.0
Qwen3.5-4B-Think	83.1	68.8	88.0	65.4	84.5	83.0	78.4
Reasoning VLMs (SFT-based)							
Vision-R1-LlamaV-CI-11B	83.9	62.7	-	61.4	-	-	-
SPARC-4B (Avogaro et al., 2026)	-	-	-	-	-	70.5	68.4
RL-trained VLMs							
Vision-R1-7B (Huang et al., 2026)	-	73.5	-	-	-	-	-
Pixel-Reasoner-7B (Su et al., 2026)	-	-	-	-	-	74.0	66.9
DeepEyes (Zheng et al., 2026)	-	70.1	-	-	-	75.1	72.6
Ground-R1-7B (Cao et al., 2025)	-	-	-	-	-	75.0	71.1
GazeVLM (w/o gaze bias)	85.5	68.1	88.7	67.1	84.5	79.0	74.4
GazeVLM (Ours)	**85.8**	**73.3**	**89.2**	**69.0**	**85.4**	**83.4**	**78.0**
Δ (vs Qwen3-VL-4B)	+2.6	+1.6	+1.5	+1.7	+2.0	+3.9	+4.4
GazeVLM-Qwen3.5 (Ours)	**85.8**	**77.4**	**90.0**	**71.2**	**86.2**	**86.9**	**82.1**
Δ (vs Qwen3.5-4B-Think)	+2.7	+8.6	+2.0	+5.8	+1.7	+3.9	+3.7

pushing MathVista to 73.3 (+5.2% over the passive baseline) and MMBench to 89.2. The gains are most pronounced on the high-resolution HRBench datasets, where GazeVLM scores 83.4 on HRBench-4k (+3.9 over the base model) and 78.0 on HRBench-8k (+4.4).

5 Mechanism Analysis

This section tests whether the gains come from the gaze bias, and whether the trained model's attention follows its own gaze actions once the bias is removed.

Gaze bias at training. The passive grounding baseline, GazeVLM (w/o gaze bias), shares every ingredient of GazeVLM except the bias: the Qwen3-VL-235B-A22B teacher traces, the prompts, the box supervision, the SFT and GRPO recipe and the reward. Like standard grounding models

Table 2: Effect of training with gaze bias after SFT and GRPO. Models are evaluated with bias off.

		Benchmarks				
Training	**Train w/ Gaze**	**MathVista**	**MMStar**	**ChartQA**	**HRBench-4k**	**HRBench-8k**
Qwen3-VL-4B	✗	71.7	67.3	83.2	79.5	73.6
SFT	✗	68.7	65.9	83.7	79.8	76.0
	✓	68.4	65.7	83.8	80.0	73.5
GRPO	✗	68.1	67.1	85.5	79.0	74.4
	✓	**73.3**	**69.0**	**85.8**	**83.4**	**78.0**

Table 3: Attention–region agreement on 200 HRBench samples. IoU between the top-k attended visual tokens and the k tokens inside the emitted <LOOK> box (mean $\pm$ standard error). The base model is evaluated on GazeVLM traces. The condition "GazeVLM + bias" artificially applies the bias to check its limited additional contribution.

Model	**IoU**
Qwen3-VL-4B (base)	0.130 ± 0.009
GazeVLM	0.360 ± 0.018
GazeVLM **+ bias**	0.440 ± 0.018

(Peng et al., 2023; You et al., 2024; Man et al., 2025), it therefore emits boxes that do not affect its attention. After GRPO it trails GazeVLM on every benchmark in Table 2, by 5.2 points on MathVista and by 4.4 and 3.6 on HRBench-4k and HRBench-8k, where it stays within one point of the base model. GazeVLM and the baseline start GRPO from their own SFT checkpoints, trained with and without the bias respectively, but the gap is not inherited from SFT: after SFT alone, training with the bias brings no gain (68.4 vs. 68.7 on MathVista, 80.0 vs. 79.8 on HRBench-4k, 73.5 vs. 76.0 on HRBench-8k), and the gap opens only during GRPO.

Gaze bias at inference. Applying the bias at inference as well brings no gain: with it, GazeVLM scores 69.4 on MathVista, 82.5 on HRBench-4k and 76.5 on HRBench-8k, against 73.3, 83.4 and 78.0 without it, which is expected if the trained model already attends where the bias would steer it. All GazeVLM results are reported with the bias off, i.e. with unmodified forward pass of the VLM.

Is the teaching signal internalized? Here we show that the teaching signal is *internalized*, i.e. after training and with the bias off, the model's own gaze actions shift its attention to the named region as the bias would, and the model relies on them to answer successfully. We test this in two steps: by showing that attention agrees with the emitted box, and verifying that accuracy degrades close to base performance when the gaze actions are removed.

For each <LOOK> block with box b, let $\mathcal{R}$ be the k visual tokens in b, and $\mathcal{H}$ the k visual tokens receiving the most attention from the text tokens of the block, averaged over layers and heads. We report $\mathrm{IoU}(\mathcal{H}, \mathcal{R}) = |\mathcal{H} \cap \mathcal{R}|/|\mathcal{H} \cup \mathcal{R}|$, which is 1 when the most-attended tokens are exactly those in the box. For the base model, which emits no gaze actions, we feed it GazeVLM's traces and measure against the same boxes. On 200 HRBench samples (100 each from 4k and 8k), Table 3 shows that the base model reaches an IoU of 0.130, while GazeVLM reaches a much higher IoU of 0.360, indicating that the model's attention concentrates on the named relevant region. Artificially applying the bias on top (the GazeVLM + bias condition in Table 3) raises the IoU only to 0.440, meaning that GazeVLM has almost completely internalized the effect of the external bias.

Since the model produces both the box and the attention, we investigate whether the gaze action drives the model's behavior by intervening on them. Specifically, we run GazeVLM with <LOOK> blocks disallowed at inference, so that it can generate <THINK> and <ANSWER> tags to reason, but is unable to produce an explicit observation step. Without <LOOK>, the accuracy drops from 82.0 to 76.0 on HRBench-4k and from 76.0 to 71.0 on HRBench-8k (Table 4). A careful inspection of the additional mistakes confirms that they mostly involve a failure in extracting fine details.

Table 4: Effect of removing <LOOK> blocks at inference. Accuracy on HRBench (100 samples each for 4k and 8k) with greedy decoding is reported, showing a drop when <LOOK> blocks are removed.

Model	HRBench-4k	HRBench-8k
GazeVLM	**82.0**	**76.0**
GazeVLM (w/o <LOOK>)	76.0	71.0

Table 5: Comparison of computational efficiency and visual routing behavior. Token generation and active vision interventions of GazeVLM (internal foveation) are compared against DeepEyes (external cropping) on high-resolution benchmarks. We also provide token counts for the base model. The *efficiency gain* refers to the comparison between GazeVLM and DeepEyes.

	HRBench-4k			*HRBench-8k*		
Method	**Tokens / Trace ↓**	**Calls / Trace ↓**	**Accuracy ↑**	**Tokens / Trace ↓**	**Calls / Trace ↓**	**Accuracy ↑**
Qwen3-VL-4B-Instruct	113.5	0.0	79.5	125.2	0.0	73.6
DeepEyes (ZOOM)	1325	1.81	75.1	1268	1.71	72.6
GazeVLM (<LOOK>)	**265.9**	**1.19**	**83.4**	**292.9**	**1.19**	**78.0**
Efficiency Gain	*4.98×*	*1.52×*	*+8.3%*	*4.33×*	*1.43×*	*+5.4%*

Computational efficiency. Table 5 compares GazeVLM with DeepEyes (Zheng et al., 2026), an RL-trained agent that zooms by cropping and re-encoding image regions, on HRBench-4k and HRBench-8k. Each DeepEyes zoom appends newly encoded visual tokens to the context, and its traces average 1,325 output tokens on HRBench-4k; GazeVLM encodes the image once and produces 266 output tokens per trace, about 5× fewer. GazeVLM also uses fewer visual actions (1.19 <LOOK>s vs. 1.81 zooms per trace) while scoring 8.3 points higher on HRBench-4k and 5.4 on HRBench-8k.

6 Discussion and Conclusion

GazeVLM departs from prevailing visual reasoning paradigms by learning top-down control over its own attention logits rather than invoking external cropping tools, implementing active vision *inside* the model as an operation on already-encoded visual features. A single attention-modulation primitive, gated by <LOOK>/</LOOK> tokens the policy emits itself, drives global-to-local-to-global scanning with no tool calls or extra visual tokens, and is trained end-to-end with GRPO under a geometrically grounded reward. Notably, once trained the model performs best at inference with the explicit bias turned off, consistent with the attention prior having been internalized into the weights and acted as a training-time regularizer. Empirically, GazeVLM matches or beats other RL-trained agentic VLMs across general, chart, math, and high-resolution benchmarks, with the largest gains where foveal control matters most (HRBench-4k +3.9, HRBench-8k +4.4), while emitting about 5× fewer output tokens per trace than zoom-in-based agents. More broadly, GazeVLM shows that a model can be taught self-directed control over its own visual processing, echoing the top-down routing of human gaze to task-relevant regions, and is a step toward broader meta-cognitive control of a model's own computations.

Limitations and broader impact. Because GazeVLM modulates attention over pre-encoded features and never re-encodes patches, when targets occupy only a handful of pixels at native resolution (e.g., V* (Wu & Xie, 2024)) the signal is lost during tokenization and no reweighting recovers it; a hybrid policy that falls back to re-encoding when foveation is insufficient is a natural extension. Our curated data also retains only successful <LOOK> traces, biasing the policy toward gazing even when a global view suffices; mixing in non-gazing traces and rewarding global answers when adequate should improve robustness. The explicit gaze trace offers built-in interpretability by exposing where the model attended at each step, but sharper localization is dual-use (document understanding and accessibility on one hand, surveillance and biometric tracking on the other), and <LOOK> blocks should be treated as an auditable reasoning aid rather than a certified attestation in high-stakes settings.

REFERENCES

Xiang An, Yin Xie, Kaicheng Yang, Wenkang Zhang, Xiuwei Zhao, Zheng Cheng, Yirui Wang, Songcen Xu, Changrui Chen, Didi Zhu, et al. LLaVA-OneVision-1.5: Fully open framework for democratized multimodal training. *arXiv preprint arXiv:2509.23661*, 2025.

Niccolò Avogaro, Nayanika Debnath, Li Mi, Thomas Frick, Junling Wang, Zexue He, Hang Hua, Konrad Schindler, and Mattia Rigotti. SPARC: Separating perception and reasoning circuits for test-time scaling of VLMs. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=c0lJ3rOBPu.

Sanghwan Bae, Jiwoo Hong, Min Young Lee, Hanbyul Kim, JeongYeon Nam, and Donghyun Kwak. Online difficulty filtering for reasoning oriented reinforcement learning. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 700–719, 2026.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-VL: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023.

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, et al. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025b.

Ali Furkan Biten, Lluís Gómez, and Dimosthenis Karatzas. Let there be a clock on the beach: Reducing object hallucination in image captioning. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 2473–2482. IEEE, 2022.

Meng Cao, Haoze Zhao, Can Zhang, Xiaojun Chang, Ian Reid, and Xiaodan Liang. Ground-r1: Incentivizing grounded visual reasoning via reinforcement learning. *arXiv preprint arXiv:2505.20272*, 2025. doi: 10.48550/arxiv.2505.20272.

Chongyan Chen, Yu-Yun Tseng, Zhuoheng Li, Anush Venkatesh, and Danna Gurari. Acknowledging focus ambiguity in visual questions. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1228–1238. IEEE, 2025.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *Advances in Neural Information Processing Systems*, 37:27056–27087, 2024a.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to GPT-4V? closing the gap to commercial multimodal models with open-source suites. *Science China Information Sciences*, 67 (12):220101, 2024b.

Maurizio Corbetta and Gordon L Shulman. Control of goal-directed and stimulus-driven attention in the brain. *Nature reviews neuroscience*, 3(3):201–215, 2002.

John M Findlay and Iain D Gilchrist. *Active vision: The psychology of looking and seeing*. Number 37. Oxford University Press, 2003.

Wenxuan Huang, Bohan Jia, Shaosheng Cao, Zheyu Ye, Zhe Xu, Yao Hu, Shaohui Lin, et al. Vision-r1: Incentivizing reasoning capability in multimodal large language models. In *International Conference on Learning Representations*, volume 2026, pp. 63794–63812, 2026.

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.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13872–13882. IEEE, 2024.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 292–305, 2023.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26286–26296. IEEE, 2024a.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. MMbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pp. 216–233. Springer, 2024b.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *International Conference on Learning Representations*, volume 2024, pp. 23439–23554, 2024.

Yunze Man, De-An Huang, Guilin Liu, Shiwei Sheng, Shilong Liu, Liang-Yan Gui, Jan Kautz, Yu-Xiong Wang, and Zhiding Yu. Argus: Vision-centric reasoning with grounded chain-of-thought. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14268–14280. IEEE, 2025.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2263–2279, 2022.

Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. InfographicVQA. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 2582–2591. IEEE, 2022.

Nitesh Methani, Pritha Ganguly, Mitesh M Khapra, and Pratyush Kumar. Plotqa: Reasoning over scientific plots. In *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1516–1525. IEEE, 2020.

Gyu-Min Park and Seong-Bae Park. GQA-Q2Q: A large-scale dataset for resolving entity ambiguity in visual question-answering via clarifying subquestion. Submitted to ICLR 2026. `https://openreview.net/forum?id=309pxoSvWr`, 2026.

Deepak Pathak, Pulkit Agrawal, Alexei A Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. In *International conference on machine learning*, pp. 2778–2787. PMLR, 2017.

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.

Ronald A Rensink. The dynamic representation of scenes. *Visual cognition*, 7(1-3):17–42, 2000.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 4035–4045, 2018.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y.K. Li, Yang Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. URL `https://arxiv.org/abs/2402.03300`.

Alex Su, Haozhe Wang, Weiming Ren, Fangzhen Lin, and Wenhu Chen. Pixel reasoner: Incentivizing pixel-space reasoning with curiosity-driven reinforcement learning. *Advances in Neural Information Processing Systems*, 38:8222–8251, 2026.

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal LLMs. In *Advances in Neural Information Processing Systems*, 2024.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

Wenbin Wang, Liang Ding, Minyan Zeng, Xiabin Zhou, Li Shen, Yong Luo, Wei Yu, and Dacheng Tao. Divide, conquer and combine: A training-free framework for high-resolution image perception in multimodal large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 7907–7915, 2025.

Yujun Wang, Jinhe Bi, Soren Pirk, Yunpu Ma, et al. Ascd: Attention-steerable contrastive decoding for reducing hallucination in mllm. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 10306–10314, 2026.

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

Jianwei Yang, Hao Zhang, Feng Li, Xueyan Zou, Chunyuan Li, and Jianfeng Gao. Set-of-mark prompting unleashes extraordinary visual grounding in GPT-4V. *arXiv preprint arXiv:2310.11441*, 2023.

Alfred L Yarbus. *Eye movements and vision*. Plenum Press, 1967.

Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. Ferret: Refer and ground anything anywhere at any granularity. In *International Conference on Learning Representations (ICLR)*, volume 2024, pp. 57153–57180, 2024.

Runpeng Yu, Xinyin Ma, and Xinchao Wang. Introducing visual perception token into multimodal large language model. *arXiv preprint arXiv:2502.17425*, 2025.

Xintong Zhang, Zhi Gao, Bofei Zhang, Pengxiang Li, Xiaowen Zhang, Yang Liu, Tao Yuan, Yuwei Wu, Yunde Jia, Song-Chun Zhu, et al. Adaptive chain-of-focus reasoning via dynamic visual search and zooming for efficient vlms. *arXiv preprint arXiv:2505.15436*, 2025. URL https://arxiv.org/abs/2505.15436.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. *Transactions on Machine Learning Research*, 2023. URL https://openreview.net/forum?id=y1pPWFVfvR.

Ziwei Zheng, Minghao Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, and Chao Shen. Deepeyes: Incentivizing "thinking with images" via reinforcement learning. In *International Conference on Learning Representations*, volume 2026, pp. 126775–126798, 2026.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

A APPENDIX

A.1 DETAILS ON OUR CURATED DATASET FOR SFT AND GRPO

In this section, we provide comprehensive details regarding the data generation and filtering pipeline used to train GazeVLM. We use four seed datasets: GQA (Hudson & Manning, 2019), ChartQA (Masry et al., 2022), PlotQA (Methani et al., 2020), and InfoVQA (Mathew et al., 2022).

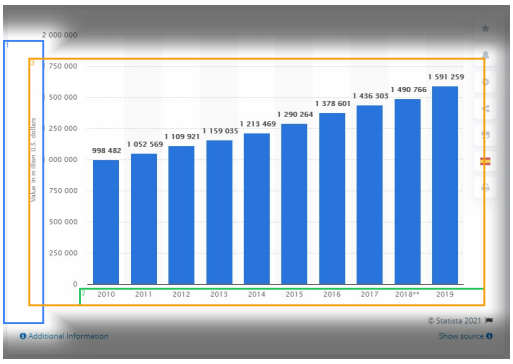
To ensure that a sample requires multi-region reasoning, we leverage Qwen3-VL-235B-A22B to filter out questions solvable via one-shot perception. Furthermore, we explicitly filter the GQA dataset to remove subjective, ambiguous, or multi-answer questions (e.g., generic "who" or "where" queries) (Chen et al., 2025; Park & Park, 2026).

For the remaining rigorous queries, we synthesize 10 candidate reasoning trajectories per sample using Qwen3-VL-235B-A22B as our expert model. To prevent erroneous trajectories from degrading our policy, we apply two strict filtering heuristics:

- **Structural validity**: Firstly, we verify that the traces strictly respect the format that we provided. We filter out the rollouts which do not satisfy this criterion and also the ones which have the incorrect final answer. We follow up to attribute a score related to the amount of <LOOK> operations which are performed in the trajectory. Indeed, we encourage traces where the model looks two or three times in different areas, but we start penalizing traces where there are more than five looks as typically it becomes more verbose unnecessarily. We also encourage traces where the model thinks, meaning where the <THINK> tokens are used. Overall, a score is given to each candidate trace and the traces are sorted with respect to that for the following step.
- **Spatial grounding check**: We proceed with verifying the grounding claims of the best scored candidate traces. Since our mechanism is not naturally present in the model, we need to make sure that the model does not hallucinate as it generates <LOOK> tokens with their observations. When the model says it focuses on an area in its trace, it generates bounding box coordinates and what it wants to look at. We ensure that the claims given within <LOOK> are correct, using the predicted bounding box and super-imposing it on the image. We then prompt Qwen3-VL-235B-A22B to verify that the description is contained within the highlighted region. If a single <LOOK> operation contains incorrect claims, we prompt the model to refine the coordinates until satisfaction; otherwise, the candidate is discarded. The first candidate (by score) to pass the grounding check is accepted and the others are discarded.

Following this pipeline, we extracted a highly curated subset of 11,080 structurally perfect trajectories for our initial Supervised Fine-Tuning (SFT) phase. Examples of generated traces are provided in Figure 4.

GRPO Dataset Preparation. For our RL phase, we require a balanced distribution of difficulty. Using the SFT-trained model, we generate 8 distinct rollouts for the training set and evaluate the empirical success rate. Following Bae et al. (2026), we filter out excessively easy (near 100% success) and impossible (0% success) samples. This results in a highly calibrated subset of 4,453 samples utilized strictly for GRPO optimization.



Q: What was the value of property, plant and equipment in service between 2010 and 2019?

I need to find the value of property, plant and equipment in service for each year from 2010 to 2019 as shown in the bar chart. I will examine the chart's axes to understand the…

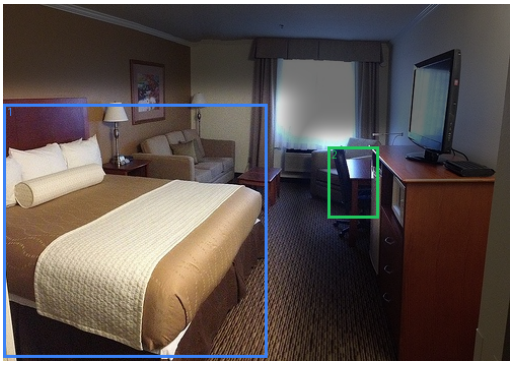
1. *y-axis label* [0.00, 0.10, 0.08, 0.90]
Value in million U.S. dollars

2. *x-axis labels* [0.15, 0.80, 0.95, 0.85]
2010, 2011, 2012, 2013, 2014, 2015, 2016, 2017, 2018**, 2019

3. *bar values 2010-2019* [0.05, 0.15, 0.95, 0.85]
2010: 998 482, 2011: 1 052 569, 2012: 1 109 921, 2013: 1 159 035, 2014: 1 213 469, 2015: 1 290 264, 2016: 1 378 601, 2017: 1 436 303, 2018**: 1 490 766, 2019: 1 591 259

The chart displays the value of property, plant and equipment in service annually from 2010 to 2019 in million U.S. dollars. I have extracted the specific values for each year. …

Answer: `1591259` ✓ GT: `1591259`



Q: Is the chair to the right or to the left of the bed?

I need to locate both the bed and the chair in the image to determine their relative positions. I will first identify the bed, then find the chair and compare their horizontal p…

1. *bed* [0.00, 0.28, 0.52, 0.99]
large bed with brown headboard, white pillows, beige quilted blanket, positioned on left side of room

2. *chair* [0.64, 0.40, 0.74, 0.60]
black office-style chair with wheels, positioned near desk and window

The bed is clearly on the left side of the image, occupying the left half of the room. The chair is located further to the right, near the desk and window. Therefore, the chair …

Answer: `right` ✓ GT: `right`

Figure 4: What was the value of property, plant and equipment in service between 2010 and 2019? *(top)*; Is the chair to the right or to the left of the bed? *(bottom)*.

Trace Generation and System Prompts. To synthesize the SFT reasoning traces, we prompted the expert Qwen3-VL-235B-A22B model with a strict set of guardrails. Crucially, we enforced a "No Oracle" rule, ensuring the expert model discovers the answer sequentially through `<LOOK>`s rather than hallucinating spatial coordinates from a global prior.

The exact instruction prompt provided to the expert model is as follows:

Expert Model Prompt (SFT Trace Generation)

```
You are a visual reasoning agent that examines specific image regions
step-by-step to answer a question.  You do NOT know the answer in
advance -- you must search the image to find it.
GUARDRAILS:
- NO ORACLE: In your first <THINK>, state what you need to find, NOT
where things are or what the answer is.  You discover locations by
looking.
- STRUCTURE FIRST: For charts/graphs, <LOOK> at the legend and axes
before data points.  For documents, read headers before details.  For
scenes, find the main subject before searching around it.
- TIGHT CROPS: Bounding boxes must tightly frame the target.  No large
boxes covering most of the image.
- FACTS IN <LOOK>, LOGIC IN <THINK>:  Inside <LOOK>, report only what
you see (objects, text, colors, numbers).  All reasoning, comparison,
and math goes in <THINK>.
```

During actual SFT and GRPO training (as well as test-time inference), the GazeVLM 4B model is initialized with the following lightweight system prompt to define the syntactic interaction loop:

GazeVLM System Prompt

```
You are a visual reasoning assistant that examines specific parts of
images to answer questions.
Always start by thinking about what you need to examine:  <THINK>
[plan] </THINK>
When you need to examine a specific region, use:  <LOOK at="[target]"
bbox=[x1, y1, x2, y2]> [Observation] </LOOK>
Use <THINK> tags for reasoning between observations, and give your
final answer in:  <ANSWER> [your answer] </ANSWER>
```

A.2 Mathematical Breakdown of the GRPO Reward Formulation

As introduced in Section 3.3, our GRPO reward function balances positive reinforcement for correctness and grounding against strict penalties for inefficiency. Here, we define the exact scalar values for each component of Equation 4.

Correctness and grounding. We want to prevent the model from guessing the right answer without navigating the image. In order to enforce that, we blend together the correctness reward with a visual grounding indicator, $\mathbf{1}_{\text{gaze}}(c)$. This indicator equals 1 if the completion c contains at least one <LOOK> block with a geometrically valid bounding box (area $\in [0.005, 0.95]$) of the image and a non-empty observation; otherwise, it is 0. Let $\hat{y}$ be the parsed answer obtained from c. We define the correctness reward as:

$$R_{\text{correct}}(c) = \begin{cases} +1.5 & \text{if } \hat{y} \approx y^* \text{ and } \mathbf{1}_{\text{gaze}}(c) = 1 \\ +0.7 & \text{if } \hat{y} \approx y^* \text{ and } \mathbf{1}_{\text{gaze}}(c) = 0 \\ -0.2 & \text{if } \hat{y} \not\approx y^* \text{ and } \mathbf{1}_{\text{gaze}}(c) = 1 \\ -0.5 & \text{if } \hat{y} \not\approx y^* \text{ and } \mathbf{1}_{\text{gaze}}(c) = 0 \\ -1.0 & \text{if no ANSWER} \end{cases}$$

Formatting and validity bonuses. We apply other positive rewards with regard to formatting where $R_{\text{format}}(c) = 0.15$ if the completion strictly contains all the different tokens. We also reward the model with regard to the validity of the predicted bounding boxes during the <LOOK> blocks. Let N_L be the total number of LOOK blocks and N_{valid} be the subset of geometrically valid bounding box coordinates. We give a proportional validity bonus $R_{\text{bbox}}(c) = 0.10 \times \frac{N_{\text{valid}}}{N_L}$ or $R_{\text{bbox}}(c) = 0$ if we do not have any <LOOK> block.

Efficiency penalties. A model can very easily start applying many <LOOK> blocks because of how much we reward it to do so. Nevertheless, this can become useless and redundant especially if the bounding boxes start consistently overlapping and are big regions of the image. To mitigate this issue, we penalize redundant gazing by applying an overlap penalty, depending on the mean pairwise IoU across all generated bounding boxes: $R_{\text{overlap}}(c) = -0.15 \times \overline{\text{IoU}}(c)$, with $\overline{\text{IoU}}(c) = \frac{1}{N_L(N_L-1)} \sum_{i \neq j} \text{IoU}(b_i, b_j)$. Finally, we also penalize excessive looping with $R_{\text{excess}}(c) = -0.15$ if $N_L \geq 11$, and rambling with a linearly scaling length penalty $R_{\text{len}}(c) = -0.05 \times \frac{\max(0, |c|_{\text{words}} - w_0)}{w_0}$, where $|c|_{\text{words}}$ denotes the word count of completion c and we set $w_0 = 500$ words.

GRPO objective. During training, we update our policy π_θ by minimizing the following objective:

$$L(\theta) = -\frac{1}{G} \sum_{i=1}^{G} A_i \log \pi_\theta(c_i | I, x) + \beta \, \text{KL}(\pi_\theta \,\|\, \pi_{\text{ref}}), \quad \text{with } A_i = \frac{R(c_i) - \mu_G}{\sigma_G + \epsilon}, \tag{5}$$

where π_{ref} is the frozen reference policy initialized from the SFT model and β controls the KL regularization strength; A_i is the relative advantage, with μ_G and σ_G the mean and standard deviation of rewards $\{R_1, R_2, \ldots, R_G\}$ and $\epsilon = 10^{-8}$.

A.3 Experimental Details and Evaluation Settings

Training specifics and architecture. We train GazeVLM atop the Qwen3-VL-4B-Instruct architecture using Low-Rank Adaptation (LoRA) to maintain computational efficiency. We apply LoRA adapters ($r = 64, \alpha = 256$, dropout$= 0.05$) across all linear projection layers (`q_proj, k_proj, v_proj, o_proj, gate_proj, up_proj, down_proj`).

The initial SFT phase runs for 3 epochs with a learning rate of 1×10^{-5}. The subsequent GRPO phase is executed for 1 epoch with a learning rate of 5×10^{-6} and generating 8 rollouts per prompt. During GRPO, the KL divergence penalty (β) is dynamically scheduled, decaying linearly from 0.04 to 0.01 over the first 50% of training steps to allow the policy to safely drift from the SFT prior as the correctness signal matures.

For our `<LOOK>` operations, we lock the suppression strength at $\alpha_s = 4$ and the spatial plateau falloff at $\sigma = 0.25$. All our experiments are run on 8 $\times$ A100-80GB GPUs.

Evaluation settings. We evaluated our model and other baselines on multiple visual reasoning benchmarks; from general reasoning, to high-resolution reasoning, chart-image reasoning, hallucination, and others. More specifically, we report our main results on MathVista (Lu et al., 2024), MMBench (Liu et al., 2024b), MMStar (Chen et al., 2024a), CV-Bench (Tong et al., 2024), ChartQA (Masry et al., 2022), and HRBench (Wang et al., 2025). Success in these benchmarks demonstrates the robustness of a model on a span of multiple image reasoning problems. We use an external model `gpt-oss-120b` to judge the answers provided by our model and other baselines and report the accuracy on the different benchmarks.

We compare our results against a wide range of models. These include closed-source models such as GPT-4o, Gemini2.5-Flash-Lite; vanilla open-source VLMs: Qwen2.5-VL (Bai et al., 2025b), Qwen3-VL (Bai et al., 2025a), and Qwen3.5-4B. We also compare against reasoning VLMs which use SFT such as Vision-R1-LlamaV-CI (Huang et al., 2026) and SPARC (Avogaro et al., 2026). Finally, we include RL-trained VLMs work such as Vision-R1 (Huang et al., 2026), Pixel-Reasoner (Su et al., 2026), DeepEyes (Zheng et al., 2026), and Ground-R1 (Cao et al., 2025).

A.4 Additional Ablation Study

Impact of reinforcement learning (GRPO). Does the model require RL to internalize active vision, or is Supervised Fine-Tuning (SFT) sufficient? In Table 2, we compare the SFT-only variant of our model trained with the gaze bias against the fully optimized GRPO policy. The results indicate a substantial performance margin in favor of GRPO across all benchmarks (e.g., 4.9 points on MathVista). While SFT successfully imparts the syntactic formatting of the LOOK tokens via behavioral cloning, it is insufficient for complex reasoning. The policy optimization provided by GRPO is essential for teaching the model autonomous visual navigation. By utilizing our bespoke reward function in Equation (4) to penalize redundant gazing and reward geometrically valid grounding, GRPO forces the model to transition from merely mimicking instructional traces to executing highly efficient, goal-directed foveal fixations.

Judge validation. In our work, we score answers with `gpt-oss-120b` under greedy decoding rather than string matching, because model outputs often combine the selected option with its text (e.g. "B. Iceland" against the reference "B"), which exact matching after parsing marks as incorrect. To verify that the judge does not change our conclusions, we rescored all HRBench predictions of the base model and GazeVLM with string matching and compared per-sample decisions (Table 6). The two scorers agree on 99.8% of samples, and the gain of GazeVLM over the base model is almost the same under both (+4.12 with the judge, +4.25 with string matching, averaged over both splits). The few disagreements are formatting cases of the kind above.

A.5 Qualitative Results

To better understand the mechanistic effect of our learned gaze bias on the model's internal attention, we provide a controlled qualitative analysis in Figure 5 on sample 120 from HRBench-4k. Our goal is to isolate the direct effect of the suppression bias on decoder attention, independently of any downstream change in the generated trace or final answer.

Table 6: Judge (`gpt-oss-120b`) vs. string matching on all HRBench predictions.

Split	Method	n	Judge acc.	String acc.	Agreement (%)
HRBench-4k	Qwen3-VL-4B	800	79.50	79.38	99.88
	GazeVLM	800	83.38	83.25	99.88
	Δ		+3.88	+3.88	
HRBench-8k	Qwen3-VL-4B	800	73.62	73.38	99.75
	GazeVLM	800	78.00	78.00	99.75
	Δ		+4.38	+4.62	
Overall	Δ	1,600	+4.12	+4.25	99.8



(a) Input with emitted bounding box (b) Attention with bias OFF (c) Attention with bias ON

Figure 5: **Effect of the learned gaze bias on decoder attention.** We consider GazeVLM on sample 120 from HRBench-4k, with question "*Which direction is the river flowing relative to the clock tower?*". The model is decoded greedily once with the bias inactive (standard GazeVLM inference); the resulting trace is then re-fed through two analysis forward passes over the same prompt and trace, differing only in whether the gaze bias is activated at inference. *(a)* the input image with the bounding box the model emitted in its own <LOOK> block (in green). *(b)-(c)* decoder attention to visual tokens with the bias disabled (OFF) and applied to the emitted bounding box (ON). Heatmaps show the mean attention from every generated trace token to visual tokens, averaged across decoder layers and heads, and share a single $5/99$ percentile color scale. The generated trace and decoded answer are identical across *(b)* and *(c)*; only the analysis forward differs. This isolates the attention bias as the proximate cause of the model's focus, independent of any change in the output sequence.

To this end, we adopt a two-pass analysis protocol. First, we decode GazeVLM greedily allowing the model to autonomously emit its reasoning trace, including the <LOOK> block with its self-predicted bounding box (shown in green in Figure 5(a)). We then re-feed this exact prompt and trace through two additional forward passes that differ *only* in whether the gaze suppression bias is applied to the emitted bounding box at inference. Figure 5(b) shows the decoder attention with the bias OFF (standard GazeVLM), while Figure 5(c) shows the same attention with the bias ON at inference. Both heatmaps display the mean attention from every generated trace token to the visual tokens, averaged across decoder layers and heads.

Because the trace and decoded answer are identical across (b) and (c), the difference between the two heatmaps should originate from the bias itself. As shown in Figure 5(c), activating the bias visibly concentrates attention onto the targeted region, demonstrating that our suppression mechanism functions as intended at the attention level: it tightly steers the model's foveal focus toward the region the policy has independently identified as task-relevant, while smoothly dampening peripheral visual tokens. This confirms that the gaze bias acts as a clean intervention on attention.

A.6 AI USE STATEMENT

In this work, we used generative AI tools for helping with code editing, data generation, evaluation, research, and to improve grammar and readability of the manuscript. All AI-assisted outputs and edits were reviewed and tested by the authors.