Starve to Perceive: Taming Lazy Perception in VLMs with Constrained Visual Bandwidth

Yuhuan Wu[1], Cong Wei[2], Fangzhen Lin[1], Wenhu Chen[2], and Haozhe Wang[1†]

[1] Hong Kong University of Science and Technology
[2] University of Waterloo

Abstract. Vision-Language Models (VLMs) deployed as situated agents in high-resolution visual environments require active perception — the ability to dynamically decide where to look through operations like zooming, cropping, and panning. However, current training paradigms produce models that mimic the surface form of such operations without functionally depending on their outputs, a phenomenon we term lazy perception. We trace this to a fundamental learning asymmetry: when coarse global views combined with language priors suffice for moderate accuracy, the model has no incentive to learn harder multi-step visual search. If a model can succeed without actively looking, it will never learn to look. This motivates Starve to Perceive, a training paradigm that constrains visual bandwidth — restricting each observation to a tight token budget so that no single view suffices for task completion, thereby strongly incentivizing strategic and efficient active perception. Despite requiring no auxiliary losses, reward shaping, or architectural changes — serving as a minimal, plug-in modification to standard post-training pipelines — models trained under perceptual starvation achieve substantial gains of 5% average relative improvement across diverse benchmarks. Our codes and data will be publicly available at https://github.com/WhuanY/Starve2Perceive.

1 Introduction

Vision-Language Models (VLMs) are evolving from passive encoders of static images into situated agents that must operate in complex, high-resolution visual environments. Applications such as autonomous navigation [3], GUI manipulation [26], and fine-grained document analysis [6,19] demand visual understanding at a level of detail that no single-pass encoding can efficiently provide. Just as biological vision relies on saccadic eye movements to sequentially sample high-acuity information from a vast visual field [8], effective visual agents must possess *active perception*: the capacity to dynamically determine *where* to look and *what* to examine through deliberate operations such as zooming, cropping, and panning. We argue that active perception is not an optional enhancement but a foundational capability for the next generation of vision-language systems.

Corresponding Author: jasper.whz@outlook.com

Despite growing recognition of this need, current training paradigms fail to produce models that genuinely rely on active visual operations. Supervised fine-tuning (SFT) on tool-use trajectories teaches models to mimic the surface form of perception actions — generating syntactically correct zoom or crop commands — without internalizing functional dependence on their outputs. Standard reinforcement learning (RL), while more flexible, does not resolve this deficiency. Prior works provide concrete diagnostic evidence of a pervasive lazy perception phenomenon: (i) models trained with existing methods maintain near-identical accuracy when intermediate visual observations are replaced with blank inputs, revealing that apparent multi-step visual reasoning is largely illusory [27]; and (ii) models seamlessly continue generating correct answers after tool calls are deliberately corrupted, indicating reliance on language priors rather than visual evidence [32].

The root cause is a fundamental asymmetry in learning difficulty. Neural networks follow the path of steepest descent toward loss reduction [21]. When a coarse global view combined with strong language priors suffices to achieve moderate accuracy, the model has no gradient-based reason to invest in the harder strategy of multi-step visual search. Active perception, under standard training, is a *dominated strategy*: it incurs additional computation, introduces the risk of selecting uninformative regions, and yields marginal reward improvements that are overwhelmed by the simpler passive baseline. The critical implication is that **if the model can succeed without actively looking, it will never learn to look**. Conversely, this asymmetry points directly to a remedy: the passive shortcut must be made structurally unavailable during training.

This analysis motivates our central thesis: *to learn active perception, a model must be trained under conditions where passive observation is insufficient.* We propose **Starve to Perceive**, a training paradigm grounded in a simple mechanism we term *constrained visual bandwidth*: at each reasoning step, the visual observation returned to the model is rescaled to fit within a tight token budget, ensuring that the information available from any single view is too impoverished for reliable task completion. Under this regime, the model faces a stark choice: either learn to strategically direct its limited visual budget toward task-relevant regions through deliberate zoom, crop, and pan operations, or fail. Active perception ceases to be one strategy among many and becomes the *only viable* strategy. The mechanism requires no auxiliary losses, reward shaping, or architectural modifications, and integrates directly into standard SFT-then-RL post-training pipelines as a reusable, plug-in component.

Crucially, we find that models trained under this perceptual starvation regime acquire active visual reasoning skills that robustly transfer to unconstrained test-time settings: under standard evaluation, our model surpasses the strongest tool-augmented baseline with 3% higher overall score than the strongest baseline at the same scale. Under test-time resource constraints, our model degrades only marginally due to robust active visual operations, whereas rival baselines suffer drops much larger. Visual bandwidth restriction yields a natural efficiency dividend: fewer visual tokens per step reduce the computational cost of both for-

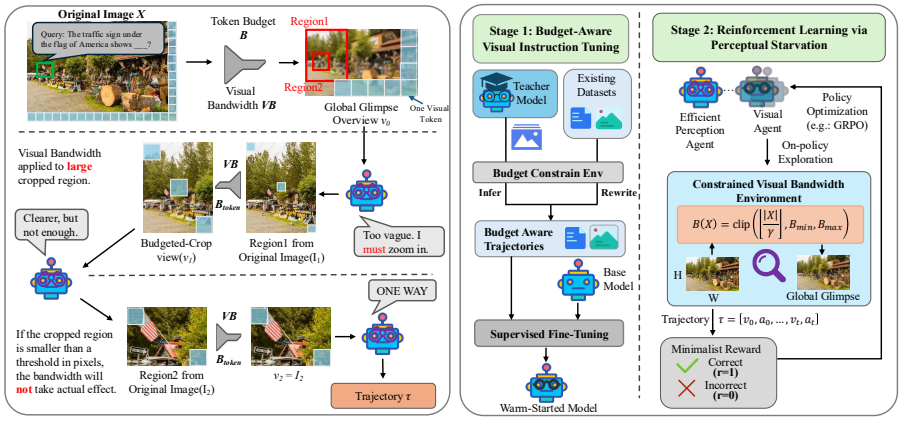


Fig. 1: Overview of **Starve to Perceive**. (a) A Visual Bandwidth (parametrized by B) limits the upper bound of both the global image and cropped regions (b) Two-stage training: Budget-Aware Visual Instruction Tuning initializes exploration under token constraints; Reinforcement Learning with Perceptual Starvation train the model via self-collected trajectories under visual constrain to learn active perception.

ward passes and gradient computation, cutting total training time by 50%, and opens up a path toward deploying active visual agents in compute-constrained settings, including on-device applications and real-time systems [37].

Contributions. We identify visual bandwidth restriction as a principled and effective mechanism for learning functional reliance in VLMs, and instantiate this insight as *Starve to Perceive* — a mechanism that integrates into standard SFT-then-RL pipelines with minimal modifications. We provide the first evidence that active perception skills learned under constrained training transfer to unconstrained evaluation, achieving state-of-the-art accuracy while remaining robust under strict token budgets — a property that promotes functionally grounded visual reasoning and mitigates superficial tool-use mimicry.

2 Method

Vision-Language Models typically process visual information in a single, unconstrained pass. When the entirety of a high-resolution image is available, the path of steepest descent during optimization naturally favors exploiting coarse global context and strong language priors. To overcome this "lazy perception" phenomenon, we propose a framework that fundamentally alters the model's observation space. By strictly limiting the visual bandwidth available at any single interaction step, we shift active perception from an optional enhancement to an absolute prerequisite for task success. The overall framework can be seen at Fig 1.

2.1 The Principle of Perceptual Starvation

Current models are trained by drinking from a "firehose" of high-resolution image data. We propose forcing the model to perceive the environment through a "straw" – a sequence of strictly low-resolution, low-bandwidth glimpses.

This design provides a profound optimization pressure that mimics the Information Bottleneck principle [25]. In standard settings, we want the model's final reasoning state Z_T to maximize mutual information with the correct answer Y, denoted as $I(Z_T; Y)$. However, by restricting the maximum token count per glimpse, we introduce a strict upper bound on the channel capacity between the original high-resolution image X and the model's internal state.

Because the "pipe" is intentionally narrowed, the current glimpse v_t can only carry a tiny fraction of the total spatial information. If the policy fills this limited capacity with irrelevant background "noise," the predictive power $I(Z_T; Y)$ mathematically drops to zero. Consequently, the only viable mathematical solution to maximize the objective is to learn a policy that actively filters out noise before it enters the context window. By artificially inducing perceptual starvation, active attention transforms from a "nice-to-have" feature into a survival mechanism necessary to solve the task.

2.2 Training with Constrained Visual Bandwidth

In this section, we formalize interaction under constrained visual bandwidth and describe how the model answers image queries by progressively acquiring task-relevant visual evidence.

Notations. Let an input be (X, Q), where X is an image and Q is a text query. We denote $\mathcal{D}(\cdot, B)$ as a deterministic downsampling operator that maps any view to a representation whose visual-token count does not exceed a preset budget B. Let v_t be the returned visual observation at step t, a_t the spatial action (e.g., a bounding box relevant to the image), Z_t the model's internal state, and $H_t = \{Q, v_0, a_1, v_1, \ldots, a_{t-1}, v_{t-1}\}$ the interaction history.

Multi-Turn Interaction. Inference starts from a global, budgeted overview: $v_0 = \mathcal{D}(X, B)$. At each step t, the policy predicts a action conditioned on the current state and query: $a_t = \pi_\theta(Z_{t-1}, Q)$. When Z_{t-1} contains sufficient information for the policy to infer answer, a_t may be just the Y. Otherwise, it represents a crop request.

Crucially, to avoid compounding information loss across turns, cropping is always executed on the original image X in native pixel space, and the budget is enforced only after cropping:

$$v_t = \mathcal{D}(\mathrm{Crop}(X, a_t), B). \tag{1}$$

This mechanism guarantees that while the model perceives a much smaller spatial area, it does so at a significantly higher effective resolution, seamlessly trading spatial coverage for fine-grained evidence. This constrained observation sequence constitutes the training environment that both SFT (§2.3) and RL (§2.4) operate within.

2.3 Budget-Aware Visual Instruction Tuning

Base VLMs lack the inherent ability to emit syntactically correct spatial actions or reason over degraded, bandwidth-constrained images. Because standard supervised fine-tuning (SFT) utilizes full-resolution images, it introduces a severe distribution shift. We bootstrap the policy using a Budget-Aware visual instruction tuning stage, sourcing trajectories for fine-tuning via two complementary methods:

- **Teacher Model Distillation:** We deploy a strong teacher model within our constrained environment to generate multi-turn trajectories. We apply rigorous rejection sampling to retain only trajectories that produce correct final answers and exhibit strict token-efficiency.
- **Trajectory Rewriting:** We convert existing high-quality tool-use datasets into budget-aware formats by applying "Visual Reconstruction" (downsampling all intermediate views to the budget B) and "Textual Grounding Realignment" (remapping spatial coordinates in the reasoning text to match the constrained image space). This can be achieved by a rewriting model(i.e., GPT-4o) via explicit prompting and post-hoc rule-based filtering. We will disclose the rewriting details in the supplementary material.

Given a collected trajectory $\tau = (v_0, a_1, v_1, \ldots, a_T, Y)$ generated under the budget B, we optimize the model parameters θ by minimizing the negative log-likelihood of the actions and the final answer:

$$\mathcal{L}_{SFT}(\theta) = -\mathbb{E}_{\tau}\left[\sum_{t=1}^{T} \log \pi_\theta(a_t|H_t) + \log \pi_\theta(Y|H_{T+1})\right] \tag{2}$$

2.4 Reinforcement Learning via Perceptual Starvation

While SFT provides a necessary behavioral prior, it is fundamentally limited by the perception strategies of the teacher. To develop genuine functional dependence on its own visual observations, the model must autonomously explore the constrained environment through reinforcement learning.

Resolution-Conditioned Token Budget: To ensure the bandwidth constraint acts as a consistent pressure mechanism – neither trivially easy for small images nor completely destructive for massive ones – we implement a resolution-conditioned token budget. Rather than fixing a static scalar, the budget B scales dynamically relative to the native token count of the source image $|X|$:

$$B(X) = \text{clip}\left(\left\lfloor \frac{|X|}{\gamma} \right\rfloor, B_{\min}, B_{\max}\right) \tag{3}$$

Here, the compression rate $\gamma \gg 1$ ensures a uniform degree of perceptual starvation across diverse inputs, while $B_{\min}$ and $B_{\max}$ bound the operational extremes.

The crucial advantage of our perceptual starvation formulation is the radical simplification of the reward signal. In unconstrained settings, models require dense, hand-crafted auxiliary losses to penalize excessive API calls or reward specific focus behaviors. In our framework, the constrained environment acts as a strict physical regularizer. As shown in Listing 1, integration into existing multi-turn VLM RL pipelines is minimal: we only replace the visual input with its budget-constrained counterpart (highlighted in green in the pseudo-code) in the rollout function. No architectural changes or specialized losses are required, making the method plug-and-play and easy to reuse across different training setups.

Listing 1: Comparison between a typical multi-turn VLM rollout (red) and our implementation (green). `B` is the token budget constant, which defines the maximum visual token count per image. `D` is the deterministic operator that apply image transformation. `pi` refers to the policy model.

```
- def vanilla_rollout(X, Q, T, pi):
+ def budgeted_rollout(X, Q, B, T, pi, D):
-    v0 = X
+    v0 = D(X, B)
     # first image uses a global glimpse under budget B
     Z = initialize_state(v0, Q)
     for t in range(1, T + 1):
         a_t = pi(Z, Q)
         if has_answer(a_t):
             break
         X_crop = crop(X, a_t)
-        v_t = X_crop
+        v_t = D(X_crop, B)
         # intermediate image is also constrained by budget B
         Z = update_state(Z, v_t)
     Y_hat = pi(Z, Q) # Final answer
     return Y_hat
```

Since the constrained environment already acts as a regularizer, we employ a minimalist, sparse reward function $R(\tau)$ for any trajectory τ generated during rollout:

$$R(\tau) = \begin{cases} 1, & \text{if } Y \text{ exactly matches } Y^* \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

Because the visual budget strictly forbids resolving fine details from the global glimpse v_0 alone, a $+1$ reward can realistically only be achieved if the intermediate actions a_t successfully isolate task-critical regions.

We define our training target as to maximize the expected return

$$\max_{\theta} J(\theta) = \mathbb{E}_{Q\sim D, \tau\sim\pi_\theta}[R(\tau)] \tag{5}$$

Under this objective, active multi-step visual reasoning ceases to be an optional strategy; it becomes the singular pathway to maximizing the reward. In practice,

Table 1: Main results on multimodal benchmarks regarding visual search and perception-intensive reasoning. MRWL and RWQA denote MME-RealWorld-Lite and RealWorldQA, respectively. Best and second best results in each group are highlighted in **bold** and underlined, respectively. * indicates models capable of calling tools. † denotes results reported from the other publications; all other entries are reproduced under our unified evaluation protocol.

Model	HR-Bench		V* Bench			$\text{VisualProbe}_{\text{test}}$			Perception-Intensive			Overall
	4K	8K	Attr	Pos	All	Easy	Med	Hard	MRWL	RWQA	TreeBench	
No Constrain Settings												
GPT-4o†	68.0	63.9	-	-	62.8	47.5	15.4	11.2	46.4	67.2	**46.9**	-
Gemini-2.5-Pro†	-	-	-	-	79.2	-	-	-	-	-	-	-
LLaVA-OneVision†	63.0	59.8	75.7	75.0	75.4	-	-	-	43.7	66.3	-	-
PixelReasoner*	71.13	69.25	85.22	80.26	83.25	58.4	29.6	28.8	47.21	**69.02**	39.0	60.10
DeepEyes*	71.50	67.88	82.61	85.53	83.77	63.83	35.07	32.07	50.81	68.89	37.5	61.77
ChainOfFocus*	71.63	67.25	86.09	**86.84**	**86.39**	56.74	32.46	38.67	48.98	64.45	39.75	61.75
Mini-o3*‡	57.75	47.75	72.17	80.26	75.39	19.86	25.37	19.81	32.15	52.55	30.37	46.68
Qwen2.5-VL†	65.25	63.0	73.9	67.1	71.2	39.0	26.0	23.9	43.0	68.5	37.0	52.53
+ BA-SFT*	72.50	65.88	85.22	73.68	80.62	49.64	32.46	33.97	46.95	67.32	40.24	58.95
+ BA-SFT+VanillaRL*	75.25	67.75	**87.83**	82.89	85.86	**69.50**	**36.57**	42.45	53.78	67.19	**40.25**	64.48
Ours*	**76.63**	**70.00**	86.96	84.21	85.86	68.09	34.70	**43.40**	**55.13**	67.71	37.78	**64.59**
Constrain Setting ($B = 256$)												
PixelReasoner*	49.38	43.13	46.96	65.79	54.45	-	-	-	29.60	55.03	-	-
DeepEyes*	57.13	51.25	60.87	64.47	62.30	37.58	13.43	15.09	40.85	63.01	35.30	45.57
ChainOfFocus*	59.63	53.75	55.70	68.40	60.70	38.30	14.93	14.15	37.83	60.00	34.32	45.25
Mini-o3*‡	51.63	40.88	53.04	64.47	57.59	17.02	17.54	0.90	27.36	48.63	25.19	36.75
Qwen2.5-VL	38.00	26.50	19.13	44.74	29.31	-	-	-	27.41	56.34	-	
+ BA-SFT*	59.25	56.13	55.65	64.47	59.16	47.51	19.78	14.15	39.55	60.65	36.29	46.60
+ BA-SFT+VanillaRL*	**66.25**	60.00	**66.09**	69.74	**67.54**	**58.16**	18.28	15.09	46.69	63.27	35.80	51.54
Ours*	66.13	**60.38**	62.61	**72.36**	66.49	**58.16**	**20.52**	**19.81**	**47.37**	**64.97**	**37.04**	**52.35**

‡ Mini-o3 exhibits notably lower performance as the model has a strong bias toward producing many reasoning turns before reaching an answer.

we optimize the return with GRPO [22] algorithm with a clipped surrogate objective.

3 Experiments

In this section, we first outline the training and evaluation settings. We then examine the effectiveness of our model trained under the constrained visual bandwidth. Finally, we share the key insights for incentivizing precise visual perception for Multi-modal Large Language Models.

3.1 Setup

Implementation Details. We chose Qwen2.5-VL-7B [5] as our base policy model. For training data, we reuse image-query sources from prior related works [10, 28, 32, 48] and filter them by two criteria: (i) image in query must have sufficient high-resolution, because large images widen the gap between the

constrained view and the original view, placing greater demand on precise localization. (ii) queries should be of moderate difficulty, so that the reinforcement-learning signal remains informative rather than saturated. For Instruction Tuning phase, our data synthesis pipeline yield 7k budget-aware trajectories. For reinforcement learning, 19k training queries are drawn from the same curated sources after applying the resolution and difficulty filters above. The default hyperparameters we choose for RL training is $\gamma = 6.25, B_{min} = 169, B_{max} = 1337$. We implement a partial version of DAPO [40]. We present more details for dataset filtering criteria, training data composition and hyperparameter choice in our supplementary materials.

Benchmarks. We classified the selected benchmarks into three families. The first is *ultra-high-resolution understanding*: HR-Bench [36] supplies natural images at 4K and 8K resolution that stress a model's capacity to resolve fine-grained details. $\text{VisualProbe}_{\text{test}}$ [10] stratifies questions into Easy, Medium, and Hard splits, enabling us to evaluate model performance at different difficulty levels. The second family is V* Bench [38], a well-established visual search benchmark with two subtasks: attributes recognition and spatial relation reasoning. We report the subtask accuracies and overall accuracy. The third family evaluates *perception-intensive reasoning*: MME-RealWorld-Lite [43], RealWorldQA [39], and TreeBench [28] mainly composed of real-world images, charts and table images and at non-trivial resolutions with a wide range of question types, posing not only perceptual but also general reasoning challenges to models.

Baselines. We primary compare with model families that utilize zoom-in tool-calls: DeepEyes [48] is trained via an end-to-end manner. PixelReasoner [32] and Chain-of-Focus [42] first finetune on synthesized data and adopt specialized reward schema to invoke active perception. Mini-o3 [10] goes further from these works by meticulously designing an over-turn masking strategy to foster extra-long turn interactions. We also include Qwen2.5-VL [5], LLaVA-OneVision [11] and GPT-4o as baselines to evaluate the performance of the model without any tool call.

Evaluation Protocol. We evaluated our model and other baseline models under two environments: an *Unconstrained* setting, where the full-resolution image is available, and a *Constrained* setting, where every observation is subject to a small token budget $B = 256$, and the inference pipeline is the same as 2.2. In both cases, we use greedy decoding, limit the interaction to at most 3 turns, and set the maximum context length (including visual tokens) at 30k tokens. These restrictions ensure a fair comparison between methods and focus the evaluation not only on accuracy but also on the efficacy of perception rather than exhaustive exploration.

3.2 Main Results

Results on Unconstrained Settings. As shown in Table 1, our best model trained under constrain consistently outperforms all 7B-scale competitive base-

Table 2: Performance Gain and Loss comparing to the baseline performance of Qwen2.5-VL-7B.

Model	Overall Score		Δ Gain	Δ Loss
	No-Constrain	Constrain	(No Constrain)	(Constrain)
Qwen2.5-VL-7B	52.53	–	–	–
DeepEyes	61.77	45.57	+9.24	-6.96
Chain-of-Focus	61.75	45.25	+9.22	-7.29
Ours	64.59	52.35	+12.06	-0.18

lines across many benchmark categories in the unconstrained setting by a clear margin. On high resolution visual search, our model leads HR-Bench at both resolutions. It achieves the highest VisualProbe Hard score (43.4 vs. 38.7 for Chain-of-Focus) and a competitive VisualProbe Easy score (68.1 vs. 63.8 for DeepEyes), indicating that visual constrain training substantially sharpens the model's ability to localize and reason under varying levels of perceptual difficulty. On perception-intensive reasoning, our model leads MME-RealWorld-Lite (55.1 vs. 50.8 for DeepEyes), further validating broad gains in scene understanding.

A key finding emerges from the train–test discrepancy in evaluation conditions: *perceptual capabilities acquired under the visual bandwidth* ***generalize*** *beyond the constrained regime.* Our best model is trained exclusively under the limited visual bandwidth, yet at inference time all the intermediate images are provided with no such restriction. Despite this mismatch, the model achieves the highest unconstrained average score(64.59) among all 7B models trained with all unconstrained pipeline(other highest: 61.77, DeepEyes). This result indicates that training under perceptual pressure does not confine the model to low-resolution reasoning; rather, the precision demanded by the bandwidth cultivates fine-grained localization skills that transfer effectively to standard, unconstrained evaluation.

Results on Constrained Settings. Table 1 further reveals a critical asymmetry in how models capable of zooming-in respond when visual information is limited. We contextualize key data points in Table 2. Under unconstrained evaluation, all zooming-in methods yield substantial gains over the non-tool baseline Qwen2.5-VL-7B, suggesting that active visual exploration is broadly beneficial. Yet under constrained evaluation, a stark divergence emerges. DeepEyes and Chain-of-Focus, both trained **without a visual bandwidth constraint** and performs relatively well under no constrain, suffer gains-to-losses reversals of -6.96 and -7.29 points relative to their untrained base model. In contrast, our model, trained under visual bandwidth constraints, limits this loss to just -0.18 points, preserving nearly all of the baseline's predictive power.

Another key finding emerges from the huge contrast: *optimizing visual perception performance in an unconstrained setting* ***generalizes poorly*** *to a constrained setting.* Models trained with abundant visual tokens learn to use zooming-

Table 3: Standard SFT vs. Budget-Aware SFT. Both variants are trained on content-equivalent trajectories; the sole difference is whether the visual-budget constraint is present.

	HRBench		V*			MRWL	RWQA	TreeBench
	4K	8K	dir_attr	rela_pos	Overall			
Standard SFT	68.75	59.88	77.39	73.68	75.92	41.06	62.09	38.52
BA-SFT	**72.50**	**65.88**	**85.22**	73.68	**80.62**	**46.95**	**67.32**	**40.24**
	+3.75	+6.00	+7.83	0.00	+4.70	+5.89	+5.23	+1.72

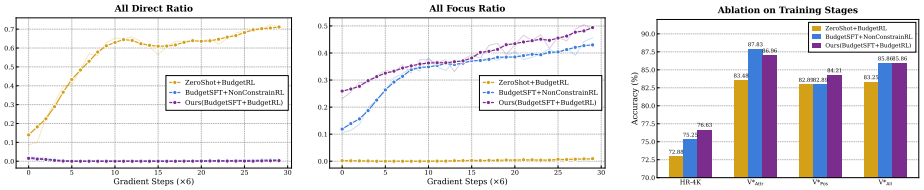
in as a *supplement* to already-rich perception—a superficial enhancement that collapses once the underlying richness is removed. Our model, by necessity, learns to use zooming-in as a *survival* mechanism: the sole means of recovering fine-grained detail that the budget otherwise forbids. The results suggest that effective perception strategies must be forged under the very constraints they are meant to overcome.

3.3 Further Analysis

Having established that training under visual constraints improves both unconstrained accuracy and constrained robustness, we conduct a staged ablation to isolate the role of each phase: Budget-Aware Visual Instruction Tuning and Reinforcement Learning with Visual Bandwidth. The four findings below form one mechanism chain: budget-aware SFT creates an active-perception prior, this prior is required for RL to avoid collapse, constrained RL then strengthens that prior, and the strengthened behavior finally translates into benchmark gains.

Budget-Aware Tuning vs. Standard Tuning. To determine whether the gains from Budget-Aware Visual Instruction Tuning stem from the specific reasoning trajectories or the *budget-constrained* framing itself, we perform a controlled trajectory ablation. We synthesize a "reversed" version of our budget-aware data: for each trajectory, we restore the full-resolution image and map all cropping coordinates back to the original pixel space (verified via rigorous rule-based checks). This ensures both datasets contain identical visual reasoning content, differing only in the presence of bandwidth constraints.

As shown in Table 3, when evaluated in a **non-constrained** setting, the model fine-tuned on budget-constrained data consistently outperforms the one trained on standard (reversed) trajectories across all benchmarks. This suggests that the budget constraint functions as a vital active perception catalyst, forcing the model to internalize the functional dependency between local visual cues and task success. Consequently, we employ the budget-aware checkpoint as the initialization for the RL phase.



(a) All Direct ratio dynamics. **(b)** All Focus ratio dynamics. **(c)** Final Performance of Budget Ablation Across Training Stages.

Fig. 2: Training Dynamics and Final Performance of Budget Ablation Across Training Stages. "All Direct Ratio" measures the proportion of queries where the model consistently bypasses visual grounding and directly answers across all sampled rollouts at a given policy checkpoint, while "All Focus Ratio" measures the proportion of queries for which the model consistently chooses to select regions across all sampled rollouts.

RL Under Constraints Necessitates Visual SFT. We investigate whether a visually constrained environment and accuracy-based rewards are sufficient to elicit active perception from scratch. We compare RL training starting from the base model (*ZeroShot + BudgetRL*) against our SFT-warmed-start model. As illustrated by the yellow lines in Fig. 2a and Fig. 2b, the base model's strategy rapidly collapses: it learns to bypass the costly active perception mechanism, relying instead on language priors to answer questions blindly. By the end of training, active perception is rarely invoked. In contrast, SFT-initialized models (blue and purple lines) maintain and progressively refine their reliance on active perception. This divergence highlights that budget-aware SFT is essential to provide the necessary behavioral prior for successful RL under visual bandwidth constraints.

Visual Constraints as an Incentive for Active Perception. To validate whether the visual bandwidth bottleneck effectively incentivizes active perception, we compare RL dynamics in two environments: one without visual constraints (*NonConstrainRL*) and one with them (*BudgetRL*), both starting from the same warm-start checkpoint.

As shown in Fig. 2b, while both models exhibit an upward trend in the *All Focus Ratio*, the model trained under visual constraints (purple, "Ours") consistently maintains a significantly higher ratio throughout the training process. By the end of RL, our model achieves an *All Focus Ratio* of approximately 0.50, compared to 0.40 for the unconstrained variant. This indicates that while SFT provides an initial inclination toward active perception, the explicit visual constraint during RL imposes a stronger optimization pressure. This pressure prevents the model from defaulting to "lazy" low-resolution heuristics, instead compelling it to actively ground its reasoning in high-resolution local evidence.

Training Dynamics Translate to Downstream Task Performance. The behavioral differences observed in Figures 2a and 2b directly manifest in final benchmark accuracy, as shown in Figure 2c. The model trained without

Table 4: E2E inference latency per sample under identical evaluation infrastructure. **Vanilla** denotes the base model inferring without token budget. **Rel. Perf.** denotes the relative performance of **ours** to that of **Vanilla**.

Benchmark	Ours ($B_{\text{token}}=256$)	Vanilla	Speedup	Rel. Perf. (Acc@1)
HRBench8K	2.21 s	11.26 s	5.10×	96.3% ↓
V*	1.85 s	5.08 s	2.75×	93.4% ↓
TreeBench	1.97 s	5.38 s	2.74×	100.1% ↑
MRWL	1.97 s	5.90 s	3.00×	110.2% ↑

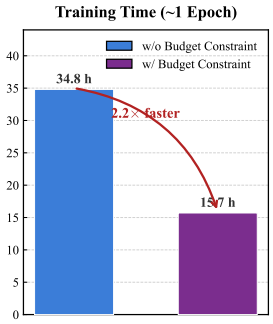


Fig. 3: RL training cost comparison.

Budget-Aware SFT (*ZeroShot + BudgetRL*), which collapses toward direct answering during RL, achieves the lowest performance across all high-resolution visual search benchmarks. The model trained without the visual constraint during RL (*BA-SFT + VanillaRL*), which exhibits weaker active-perception pressure despite a healthy SFT initialization, yields intermediate scores. Our full pipeline—Budget-Aware SFT followed by RL under the visual bandwidth—achieves the best or competitive accuracy across splits, confirming that the two training stages are complementary rather than redundant: instruction tuning establishes the behavioral foundation for active perception, and the visual constraint during RL amplifies it into measurable task-level gains.

3.4 Efficiency Analysis.

Beyond task performance, our budget-constrained paradigm yields substantial efficiency gains at both inference and training time.

Inference efficiency. Table 4 compares the end-to-end inference latency per sample between our budget-constrained agent ($B_{\text{token}} = 256$) and the vanilla unconstrained baseline, measured under identical evaluation infrastructure. By dynamically allocating visual tokens within a fixed budget, our method achieves **2.7–5.1× speedup** across all benchmarks, with the largest gain on HRBench8K where the high-resolution input incurs the heaviest visual token overhead for the unconstrained agent. Notably, this acceleration comes at *minimal* accuracy trade-off: on two of four benchmarks (TreeBench and MRWL), the constrained agent even *surpasses* the baseline, suggesting that the budget constraint acts as an implicit regularizer that discourages redundant visual exploration.

Training efficiency. The efficiency advantage also manifests during reinforcement learning. Fig. 3 plots the total training time cost one epoch of RL training. Compared to our budget-constrained RL, training data under no-constrain for

one epoch costs 2.2× more time with the same training configuration(19k training data of high resolution images, 6 A800 GPUs), primarily due to the reduced number of visual tokens processed. These empirical results highlight the potential of a budget-aware approach and may inspire both the research and industrial communities to consider adopting a new multimodal post-training paradigm.

4 Related Work

4.1 Vision-Language Models

Modern vision-language understanding builds on a representation foundation laid by contrastive learning at web scale. CLIP [20] and ALIGN [9] show that aligning visual and textual embeddings through large-scale image-text pairing produces highly generalizable features, motivating a subsequent line of work that couples frozen large language models (LLMs) with pre-trained vision encoders via lightweight connectors [2, 12, 18]. A pivotal catalyst in this progression is visual instruction tuning introduced by LLaVA [18], which reframes multimodal learning as an instruction-following problem and enables LLMs to process interleaved image-text inputs. Both open-source community [4, 49] and proprietary systems [7] scaled and refined this paradigm by incorporating extensive supervised fine-tuning with reinforcement learning from human or automated feedback. Despite the progress, modern VLMs are prone to hallucinations when they are asked for vague and small visual details [15,16], and tend to rely disproportionately on textual priors rather than genuinely grounding answers in image content [1, 17, 46]. This gap between strong language-side reasoning and brittle visual perception forms a central motivation for our work.

4.2 Tool Augmented Visual Reasoning

To address visual hallucination and over-reliance on textual priors, recent research has explored equipping VLMs with external tools for image manipulation, advancing visual reasoning from static image-text understanding to dynamic interactive interaction. Early works adopted training-free approaches: SEAL [38] uses guided visual search directed by LLMs to focus on task-relevant regions, while ZoomEye [23] employs a tree-search algorithm to mimic human-like zooming behavior. More recently, learning-based methods have emerged, training models to integrate tool manipulation into their perceptual processes. For instance, DeepEyes [48] employs end-to-end reinforcement learning to acquire tool-use behaviors, while other works [32, 42] adopt a commonly used SFT-RL two-stage training pipeline. Despite these advances, recent studies [27,32] reveal that tool-augmented models still suffer from *lazy perception* – relying on textual shortcuts rather than genuinely leveraging visual tools for reasoning. To this end, we propose a new training paradigm that incentivizes genuine active perception through a constrained visual environment.

5 Conclusion

We presented **Starve to Perceive**, a training paradigm that enforces constrained visual bandwidth throughout both supervised fine-tuning and reinforcement learning, transforming active perception from an optional strategy into a necessary survival mechanism. Our experiments demonstrate that perceptual skills forged under token starvation transfer robustly to unconstrained evaluation—achieving the highest overall score (64.59) among all 7B tool-augmented baselines—while degrading only 0.18 points under strict budgets, compared to near 7-point drops for baselines trained without visual constraints. The paradigm simultaneously yields over 50% reduction in RL training time and 2.7–5.1× inference speedup, requires no architectural changes or auxiliary losses, and integrates into standard post-training pipelines as a plug-and-play component. These findings suggest that superior visual reasoning emerges not from input density but from the optimization pressure to extract information efficiently, and we hope this perspective encourages the community to treat visual bandwidth as a deliberate design variable for building perceptually grounded vision-language agents.

References

1. Agrawal, A., KV, G., Aralikatti, R., Jagatap, G., Yuan, J., Kamarshi, V., Fanelli, A., Huang, F.: Towards mitigating hallucinations in large vision-language models by refining textual embeddings. arXiv preprint arXiv:2511.05017 (2025)
2. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al.: Flamingo: a visual language model for few-shot learning. Advances in neural information processing systems **35**, 23716–23736 (2022)
3. Anderson, P., Wu, Q., Teney, D., Bruce, J., Johnson, M., Sünderhauf, N., Reid, I., Gould, S., van den Hengel, A.: Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments (2018), https://arxiv.org/abs/1711.07280
4. Bai, J., Bai, S., Chen, K., Du, M., Fan, Y., Fan, Z., Ge, W., Liu, D., Men, R., Ren, X., et al.: Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. arXiv preprint arXiv:2308.12966 (2023)
5. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., Lin, J.: Qwen2.5-vl technical report (2025), https://arxiv.org/abs/2502.13923
6. Feng, H., Liu, Q., Liu, H., Tang, J., Zhou, W., Li, H., Huang, C.: Docpedia: Unleashing the power of large multimodal model in the frequency domain for versatile document understanding (2024), https://arxiv.org/abs/2311.11810
7. Gemini Team, Anil, R., Borgeaud, S., Alayrac, J.B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., Millican, K., Silver, D., et al.: Gemini: A family of highly capable multimodal models. arXiv preprint arXiv:2508.11630 (2025), https://arxiv.org/abs/2312.11805
8. Ibbotson, M., Krekelberg, B.: Visual perception and saccadic eye movements. Current opinion in neurobiology **21**(4), 553–558 (2011)

9. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: International conference on machine learning. pp. 4904–4916. PMLR (2021)
10. Lai, X., Li, J., Li, W., Liu, T., Li, T., Zhao, H.: Mini-o3: Scaling up reasoning patterns and interaction turns for visual search (2025), https://arxiv.org/abs/2509.07969
11. Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Zhang, P., Li, Y., Liu, Z., Li, C.: Llava-onevision: Easy visual task transfer (2024), https://arxiv.org/abs/2408.03326
12. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: International conference on machine learning. pp. 19730–19742. PMLR (2023)
13. Li, J., Xu, B., Chen, S., Li, J., Lei, J., Zhao, H., Zhang, D.: Iag: Input-aware backdoor attack on vlm-based visual grounding. arXiv preprint arXiv:2508.09456 (2025)
14. Li, J., Zhang, D., Wang, X., Hao, Z., Lei, J., Tan, Q., Zhou, C., Liu, W., Yang, Y., Xiong, X., et al.: Chemvlm: Exploring the power of multimodal large language models in chemistry area. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 415–423 (2025)
15. Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, W.X., Wen, J.R.: Evaluating object hallucination in large vision-language models. arXiv preprint arXiv:2305.10355 (2023)
16. Li, Y., Zhou, K., Zhao, W.X., Fang, L., Wen, J.R.: Analyzing and mitigating object hallucination: A training bias perspective. arXiv preprint arXiv:2508.04567 (2025)
17. Liu, C., Xu, Z., Wei, Q., Wu, J., Zou, J., Wang, X.E., Zhou, Y., Liu, S.: More thinking, less seeing? assessing amplified hallucination in multimodal reasoning models. arXiv preprint arXiv:2505.21523 (2025)
18. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36**, 34892–34916 (2023)
19. Masry, A., Long, D.X., Tan, J.Q., Joty, S., Hoque, E.: Chartqa: A benchmark for question answering about charts with visual and logical reasoning (2022), https://arxiv.org/abs/2203.10244
20. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
21. Rumelhart, D.E., Hinton, G.E., Williams, R.J.: Learning representations by back-propagating errors. nature **323**(6088), 533–536 (1986)
22. Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y.K., Wu, Y., Guo, D.: Deepseekmath: Pushing the limits of mathematical reasoning in open language models (2024), https://arxiv.org/abs/2402.03300
23. Shen, H., Zhao, K., Zhao, T., Xu, R., Zhang, Z., Zhu, M., Yin, J.: ZoomEye: Enhancing multimodal LLMs with human-like zooming capabilities through tree-based image exploration. In: Christodoulopoulos, C., Chakraborty, T., Rose, C., Peng, V. (eds.) Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. pp. 6602–6618. Association for Computational Linguistics, Suzhou, China (Nov 2025). https://doi.org/10.18653/v1/2025.emnlp-main.335, https://aclanthology.org/2025.emnlp-main.335/
24. Sheng, G., Zhang, C., Ye, Z., Wu, X., Zhang, W., Zhang, R., Peng, Y., Lin, H., Wu, C.: Hybridflow: A flexible and efficient rlhf framework. arXiv preprint arXiv: 2409.19256 (2024)

25. Tishby, N., Zaslavsky, N.: Deep learning and the information bottleneck principle (2015), https://arxiv.org/abs/1503.02406
26. Wang, B., Li, G., Zhou, X., Chen, Z., Grossman, T., Li, Y.: Screen2words: Automatic mobile ui summarization with multimodal learning (2021), https://arxiv.org/abs/2108.03353
27. Wang, C., Wang, H., Chen, X., Liu, J., Xue, T., Peng, C., Qi, D., Lin, F., Yan, Y.: From illusion to intention: Visual rationale learning for vision-language reasoning (2025), https://arxiv.org/abs/2511.23031
28. Wang, H., Li, X., Huang, Z., Wang, A., Wang, J., Zhang, T., Zheng, J., Bai, S., Kang, Z., Feng, J., Wang, Z., Zhang, Z.: Traceable evidence enhanced visual grounded reasoning: Evaluation and methodology (2025), https://arxiv.org/abs/2507.07999
29. Wang, H., Li, L., Qu, C., Xu, W., Zhu, F., Chu, W., Lin, F.: To code or not to code? adaptive tool integration for math language models via expectation-maximization. In: Findings of the Association for Computational Linguistics: ACL 2025. pp. 3060–3075 (2025)
30. Wang, H., Qu, C., Huang, Z., Chu, W., Lin, F., Chen, W.: Vl-rethinker: Incentivizing self-reflection of vision-language models with reinforcement learning. In: Advances in Neural Information Processing Systems (NeurIPS) (2025), spotlight
31. Wang, H., Que, H., Xu, Q., Liu, M., Zhou, W., Feng, J., Zhong, W., Ye, W., Yang, T., Huang, W., et al.: Reverse-engineered reasoning for open-ended generation. In: International Conference on Learning Representations (ICLR) (2026)
32. Wang, H., Su, A., Ren, W., Lin, F., Chen, W.: Pixel reasoner: Incentivizing pixel-space reasoning with curiosity-driven reinforcement learning (2025), https://arxiv.org/abs/2505.15966
33. Wang, H., Wei, C., Ren, W., Liu, J., Lin, F., Chen, W.: Rationalrewards: Reasoning rewards scale visual generation both training and test time. arXiv preprint arXiv:2604.11626 (2026)
34. Wang, H., Xu, Q., Liu, C., Wu, J., Lin, F., Chen, W.: Emergent hierarchical reasoning in llms through reinforcement learning. In: International Conference on Learning Representations (ICLR) (2026)
35. Wang, H., Xu, Q., Wang, C., Xue, T., Peng, C., Chen, W., Lin, F.: Bad seeing or bad thinking? rewarding perception for vision-language reasoning. In: International Conference on Machine Learning (ICML) (2026), spotlight
36. Wang, W., Ding, L., Zeng, M., Zhou, X., Shen, L., Luo, Y., Tao, D.: Divide, conquer and combine: A training-free framework for high-resolution image perception in multimodal large language models (2024), https://arxiv.org/abs/2408.15556
37. Wang, X., Huang, J., Abdalla, R., Zhang, C., Xian, R., Manocha, D.: Bi-vlm: Pushing ultra-low precision post-training quantization boundaries in vision-language models (2025), https://arxiv.org/abs/2509.18763
38. Wu, P., Xie, S.: V*: Guided visual search as a core mechanism in multimodal llms (2023), https://arxiv.org/abs/2312.14135
39. xAI: Grok-1.5 vision preview. https://x.ai/news/grok-1.5v (Apr 2024), accessed: 2024-08-27
40. Yu, Q., Zhang, Z., Zhu, R., Yuan, Y., Zuo, X., Yue, Y., Dai, W., Fan, T., Liu, G., Liu, L., Liu, X., Lin, H., Lin, Z., Ma, B., Sheng, G., Tong, Y., Zhang, C., Zhang, M., Zhang, W., Zhu, H., Zhu, J., Chen, J., Chen, J., Wang, C., Yu, H., Song, Y., Wei, X., Zhou, H., Liu, J., Ma, W.Y., Zhang, Y.Q., Yan, L., Qiao, M., Wu, Y., Wang, M.: Dapo: An open-source llm reinforcement learning system at scale (2025), https://arxiv.org/abs/2503.14476

41. Zhang, D., Lei, J., Li, J., Wang, X., Liu, Y., Yang, Z., Li, J., Wang, W., Yang, S., Wu, J., et al.: Critic-v: Vlm critics help catch vlm errors in multimodal reasoning. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 9050–9061 (2025)
42. Zhang, X., Gao, Z., Zhang, B., Li, P., Zhang, X., Liu, Y., Yuan, T., Wu, Y., Jia, Y., Zhu, S.C., Li, Q.: Adaptive chain-of-focus reasoning via dynamic visual search and zooming for efficient vlms (2025), https://arxiv.org/abs/2505.15436
43. Zhang, Y.F., Zhang, H., Tian, H., Fu, C., Zhang, S., Wu, J., Li, F., Wang, K., Wen, Q., Zhang, Z., Wang, L., Jin, R., Tan, T.: Mme-realworld: Could your multimodal llm challenge high-resolution real-world scenarios that are difficult for humans? (2025), https://arxiv.org/abs/2408.13257
44. Zhang, Y., Ma, J., Hou, Y., Bai, X., Chen, K., Xiang, Y., Yu, J., Zhang, M.: Evaluating and steering modality preferences in multimodal large language model, 2025a. URL https://arxiv. org/abs/2505.20977
45. Zhang, Y., Xu, M., Bai, X., Zhang, P., Xiang, Y., Zhang, M., et al.: Instruction anchors: Dissecting the causal dynamics of modality arbitration. arXiv preprint arXiv:2602.03677 (2026)
46. Zhang, Z., Wang, T., Gong, X., Shi, Y., Wang, H., Wang, D., Hu, L.: When modalities conflict: How unimodal reasoning uncertainty governs preference dynamics in mllms. arXiv preprint arXiv:2511.02243 (2025)
47. Zheng, Y., Zhang, R., Zhang, J., Ye, Y., Luo, Z., Feng, Z., Ma, Y.: Llamafactory: Unified efficient fine-tuning of 100+ language models. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations). Association for Computational Linguistics, Bangkok, Thailand (2024), http://arxiv.org/abs/2403.13372
48. Zheng, Z., Yang, M., Hong, J., Zhao, C., Xu, G., Yang, L., Shen, C., Yu, X.: Deepeyes: Incentivizing "thinking with images" via reinforcement learning (2025), https://arxiv.org/abs/2505.14362
49. Zhu, J., Wang, W., Chen, Z., Liu, Z., Ye, S., Gu, L., Tian, H., Duan, Y., Su, W., Shao, J., et al.: Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. arXiv preprint arXiv:2504.10479 (2025)