Foveated Reasoning: Stateful, Action-based Visual Focusing for Vision-Language Models

Juhong Min Lazar Valkov Vitali Petsiuk
Hossein Souri Deen Dayal Mohan

AI Center – Mountain View, Samsung Electronics
https://juhongm999.github.io/FoveateR/

Abstract. Vision–language models benefit from high-resolution images, but the increase in visual-token count incurs high compute overhead. Humans resolve this tension via foveation: a coarse view guides "where to look", while selectively acquired high-acuity evidence refines "what to think". We introduce Foveated Reasoner, an autoregressive vision-language framework that unifies foveation and reasoning within a single decoding trajectory. Starting from a low-resolution view, the model triggers foveation only when needed, retrieves high-resolution evidence from selected regions, and injects it back into the same decoding trajectory. We train the method with a two-stage pipeline: coldstart supervision to bootstrap foveation behavior, followed by reinforcement learning to jointly improve evidence acquisition and task accuracy while discouraging trivial "see-everything" solutions. Experiments show that the method learns effective foveation policies and achieves stronger accuracy under tight visual-token budgets across multiple vision–language benchmarks.

Keywords: Foveation · Multimodal Reasoning · Vision-Language Model

1 Introduction

Large language models (LLMs) have driven substantial progress in natural language processing, and this momentum has naturally extended to vision–language models (VLMs). Despite good empirical performance, VLMs still face a fundamental bottleneck at the visual front-end: accuracy typically improves with high-resolution (high-res) input images, but the compute and memory costs grow rapidly as self-attention scales quadratically with the number of visual tokens. To stay within reasonable resource budgets, recent approaches [1, 20, 25, 26] operate on low-resolution (low-res) inputs, sacrificing fine-grained visual details as well as downstream accuracy. This uncomfortable resolution–efficiency trade-off is particularly problematic in practical deployments where budgets are tight, yet fine-grained visual understanding remains crucial.

A growing line of work attempts to ease this trade-off through *visual focusing*, a strategy that starts with a global low-res view and selectively "zooms in" to acquire local high-res evidence from task-relevant regions. As shown in Fig. 1, existing visual focusing methods largely fall into two streams: (a) multi-pass methods [4, 37, 43, 51, 64] first run a VLM on a low-res image to decide where to zoom in, then run it again on the high-res crop to answer, often involving multiple zoom-in passes. (b) text-grounded methods [6, 9, 42, 48, 62, 65] represents

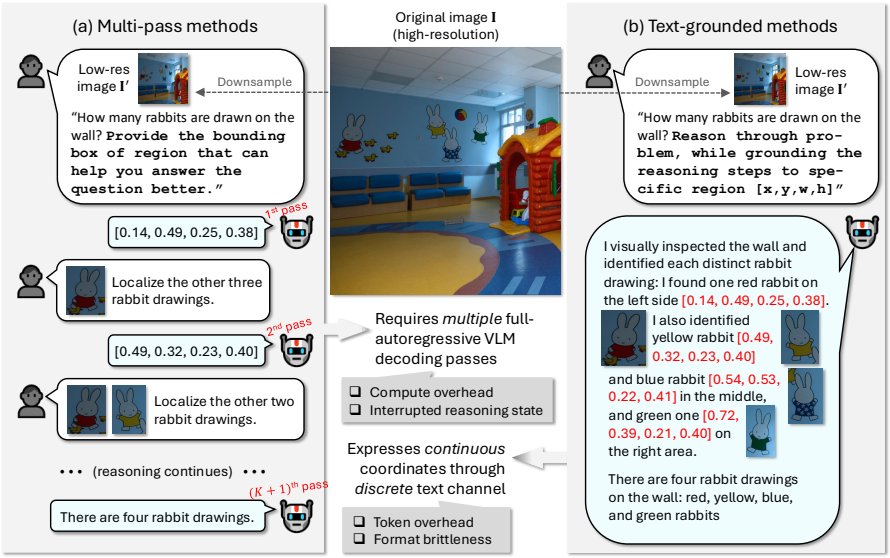

Fig. 1: Overview of prior visual focusing methods: (a) multi-pass (left) and (b) text-grounded (right). Given original high-res image (middle), both methods take downsampled image $\mathbf{I}'$ and progressively acquire task-relevant visual evidence from $\mathbf{I}$.

visual focus signals through free-form texts such as coordinate strings, tool calls, or specialized tokens [59], interleaving reasoning with the visual focus in their response. Although empirically effective, both streams pose practical limitations. In multi-pass methods, each zoom-in step triggers a new autoregressive run, incurring (i) *compute overhead* from multiple decoding passes, and often causing (ii) an *interrupted reasoning state* [43, 64] (*i.e.*, re-running the VLM on a new crop breaks hidden-state continuity across passes). In text-grounded methods, expressing where-to-look coordinates through discrete text channel introduces (iii) *token overhead* from repeated coordinate strings and (iv) *format brittleness* (*i.e.*, minor formatting deviations can make the output unparsable). These limitations raise a natural question: What do we miss in current VLM designs to overcome these shortcomings altogether?

One promising answer comes from the human visual system—*foveation* [56], in which low-acuity peripheral vision provides low-cost global context to guide where to look, while successive where-to-look decisions (*i.e.*, foveation steps) selectively acquire high-acuity visual evidence. Importantly, this evidence acquisition process is ***stateful*** and ***action-based***: as foveation unfolds, human vision maintains a persistent working memory of previously acquired evidence, thereby preserving an *uninterrupted reasoning state* rather than resetting it at each foveation step (as in multi-pass methods). Moreover, foveation is executed via *non-linguistic actions*—eye movements—rather than through explicitly verbalized coordinates (as in text-grounded methods). This contrast reveals a key gap: while prior methods emulate the *outcome* of foveation, they do not instantiate its *process*—a stateful, action-based evidence acquisition mechanism.

To bridge this gap, we introduce Foveated Reasoner (FoveateR), an autoregressive VLM pipeline that integrates *non-linguistic, action-based* foveation and textual reasoning in a *single, stateful* decoding trajectory. Starting from a coarse view, FoveateR acquires high-res evidence based on the model's evolving hidden state and injects the retrieved evidence into the ongoing generation process. By preserving an uninterrupted reasoning state and representing foveation as a non-linguistic action, FoveateR avoids (i) compute-heavy multiple passes and (ii) hidden state re-initializations, while eliminating both (iii) coordinate-text overhead and (iv) format brittleness, thereby more faithfully capturing key properties of human foveation in VLMs. Our contributions are threefold:

1. We introduce FoveateR, an autoregressive VLM that integrates action-based foveation with textual reasoning within a single, stateful decoding pass.
2. We show that our method learns effective foveation policies *without human-annotated foveation* trajectories via a two-stage training framework.
3. Experiments show FoveateR outperforms recent visual focusing methods on multiple benchmarks, verifying the efficacy of the proposed approach.

2 Preliminaries and Motivations

In this section, we begin by formalizing standard autoregressive VLMs and their fundamental limitations. We then discuss two dominant streams of prior visual focusing methods—multi-pass and text-grounded—and their key limitations.

2.1 Preliminaries: Autoregressive Vision-Language Modeling

We consider a VLM that generates a text response conditioned on a multimodal prompt consisting of an image and an instruction. Formally, the prompt is denoted by $p := (\mathbf{I}, \mathbf{x})$, where $\mathbf{I} \in \mathbb{R}^{H \times W \times 3}$ is an RGB image and $\mathbf{x} = (x_t)_{t=1}^{|\mathbf{x}|}$ is the input token sequence with $x_t \in \mathcal{V}$ for a finite vocabulary $\mathcal{V}$. Given p, a VLM parameterized by θ defines an autoregressive conditional distribution over output token sequences. Specifically, $\pi_\theta(\mathbf{y} \mid p)$ denotes the joint probability of generating the response token sequence $\mathbf{y} = (y_t)_{t=1}^{|\mathbf{y}|}$ conditioned on p:

$$\pi_\theta(\mathbf{y} \mid p) = \prod_{t=1}^{|\mathbf{y}|} \pi_\theta(y_t \mid p, y_{<t}), \tag{1}$$

where $y_{<t} = (y_1, \ldots, y_{t-1})$ denotes the tokens generated before step t.

Resolution-efficiency trade-off. Typical transformer-based VLMs [1, 20, 26] encode the image $\mathbf{I}$ into a sequence of N visual tokens. Empirically, increasing the input resolution improves model accuracy [4,13,37,43,45,51,64] but it comes at a steep efficiency cost. With ViT-style patchification, the token count scales as $N = \Theta(HW)$. Since self-attention forms an $N \times N$ matrix, the time/memory cost per layer scales quadratically in N: Time/Memory $= \mathcal{O}(N^2) = \Theta((HW)^2)$. This resolution-efficiency trade-off remains a fundamental limitation of standard VLMs, thus motivating *visual focusing* methods to start with low-res inputs and selectively acquire high-res evidence [4,6,9,19,37,42,43,45,48,51,54,59,62,64,65].

2.2 Prior visual focusing methods and their limitations

Prior visual focusing methods can be broadly categorized into two streams: (i) multi-pass [4, 37, 43, 45, 51, 64], and (ii) text-grounded [6, 9, 42, 48, 54, 59, 62, 65].

(i) Multi-pass visual focusing performs multiple autoregressive (AR) decoding passes to selectively acquire high-res evidence. Given a downsampled image $\mathbf{I}' = \text{Downsample}(\mathbf{I}) \in \mathbb{R}^{H' \times W' \times 3}$ and a pre-localization prompt $\mathbf{x}^{\text{loc}}$ (*e.g.*, "`<Question> Provide the bbox of the region helpful for answering the Question.`"), they first run a localization VLM π_{loc} to obtain a bounding box b:

$$b = [x, y, w, h] \sim \pi_{\text{loc}}(\cdot \mid p^{\text{loc}}), \qquad p^{\text{loc}} := (\mathbf{I}', \mathbf{x}^{\text{loc}}). \tag{2}$$

They then tokenize the corresponding high-res crop and run an answering VLM π_{ans} conditioned on the high-res evidence $\mathbf{V}$ and the original question $\mathbf{x}$:

$$\mathbf{y} \sim \pi_{\text{ans}}(\cdot \mid p^{\text{ans}}), \qquad p^{\text{ans}} := (\mathbf{V}, \mathbf{x}), \qquad \mathbf{V} = \text{Tokenize}(\text{Crop}(\mathbf{I}, b)), \tag{3}$$

Although empirically effective, there remain several limitations. (1) Most methods [4, 37, 43, 45, 51, 64] perform $K \geq 1$ localization passes which results in $K{+}1$ *full autoregressive runs* (including the final answering pass) as formulated below:

$$\underbrace{b_1 \sim \pi_{\text{loc}}(\cdot \mid p_1^{\text{loc}})}_{\text{pass 1}} \rightarrow \cdots \rightarrow \underbrace{b_K \sim \pi_{\text{loc}}(\cdot \mid p_K^{\text{loc}})}_{\text{pass } K} \rightarrow \underbrace{\mathbf{y} \sim \pi_{\text{ans}}(\cdot \mid p^{\text{ans}}(b_{1:K}))}_{\text{final } K{+}1 \text{ pass}}, \tag{4}$$

thus making multi-step focusing expensive. (2) As each pass starts a new decoding trajectory, which resets the VLM hidden states, it breaks the continuity of the model's internal reasoning state. (3) Several systems separate localization from answering ($\pi_{\text{loc}} \neq \pi_{\text{ans}}$) [4, 51], improving specialization but increasing total parameters and system complexity. While reusing the same model for both ($\pi_{\text{loc}} = \pi_{\text{ans}}$) [37, 42, 43, 45, 64] is more parameter-efficient, it couples two competing skills (localization *vs.* answering) into a single capacity budget, which can hurt either localization or answer quality unless model size is increased. (4) Some methods [4, 43, 64] use a fixed number of zoom-in steps K, rather than dynamically adapting it conditioned on task/question difficulty.

(ii) Text-grounded visual focusing methods [6, 9, 42, 59, 62] train VLMs to emit spatial references through natural language channel (*e.g.*, coordinate strings such as $[x, y, w, h]$, tool-call text, or augmented localization tokens[1]). Importantly, they interleave reasoning with the spatial references (*e.g.*, "the man is at (x_1, y_1); behind him at (x_2, y_2) there is a ball"). While conceptually appealing, casting visual focus as free-form text introduces some practical drawbacks: (1) The approach relies on a *text-to-geometry contract*: coordinates must follow strict conventions (*e.g.*, normalization, coordinate frame, ordering, delimiters, rounding), and deviations can yield non-executable outputs or incorrect region selection. (2) Learning continuous spatial values through a discrete language channel (*i.e.*, vocabulary set $\mathcal{V}$) quantizes the action space, which can potentially impair fine-grained control over where to zoom in. (3) Emitting spatial references as output tokens introduces *token overhead*, since coordinates/tool-call arguments consume sequence budget in addition to reasoning tokens.

[1] [59] augments the vocabulary with ∼2K localization tokens. While this reduces format brittleness, it still discretizes inherently continuous foveation actions, introducing quantization constraints and extra token-engineering overhead.

Motivation for the proposed approach. Existing visual focusing methods improve the accuracy–efficiency trade-off, but remain limited by either multi-pass design or text-grounded foveation. Human foveation is effective not only because it is coarse-to-fine, but also because it performs *stateful* evidence acquisition through *non-linguistic actions* within a single ongoing process. Prior methods, however, emulate the selective zoom-in without instantiating this core process, leading to repeated decoding, broken reasoning state, or token overhead. This motivates a *stateful, single-pass* and *non-linguistic, action-based* visual focusing mechanism, which we present in the next section.

3 Proposed Method

We present FoveateR, a vision-language generation framework that acquires a dynamic amount of high-res visual evidence within a single autoregressive decoding pass. As shown in Fig. 2, we interpret the decoder as an *agent* that interacts with a partially observable *environment* (the underlying high-res image). Based on an internal *state* of the agent, it chooses an *action* at each time step: either emitting the next text token or requesting additional high-res evidence from a foveated region. We start by formalizing this as a partially observable Markov Decision Process (POMDP).

3.1 FoveateR as a Partially Observable MDP

We interpret FoveateR decoding as a sequential decision process under partial observability: the agent must answer a question given only a low-res global view, and may acquire (partially observable) local high-res evidence from the original image.

Environment (latent state). Each episode starts by providing the agent with $\mathbf{I}'$ and an instruction/question token sequence $\mathbf{x}$, which requests a structured response (*e.g.*, reasoning in `<think>` and the final answer in `<answer>`). The environment also contains the corresponding underlying high-res image $\mathbf{I}$, which remains fixed throughout the episode. The instruction $\mathbf{x}$ is fully observed, whereas the fine-grained content of $\mathbf{I}$ is only partially observable. An episode terminates when decoding finishes, *i.e.*, when the model outputs `<answer>` $\cdots$ `</answer>` and then `<EOS>`.

Observations. At each decoding step t, the environment returns an observation o_t. The initial observation is the low-res view and the instruction: $o_1 = (\mathbf{I}', \mathbf{x})$. For $t \geq 2$, o_t is either (i) the previously emitted text token y_{t-1} or (ii) a newly revealed high-res evidence. In case (i), the environment simply echoes the token back to the agent.[2]

Agent memory. The agent maintains a memory buffer M_t that stores the entire history of interactions with the environment (*e.g.*, the initial observation $(\mathbf{I}', \mathbf{x})$, all previously generated text tokens $y_{<t}$, and any high-res evidence revealed from the environment). In practice, M_t is implemented by the Transformer's KV-cache, which stores the key/value of all previously processed tokens. The memory is initialized as an empty sequence $M_0 = \emptyset$ and updated by concatenating the latest observation:

$$M_t = M_{t-1} \oplus o_t. \tag{5}$$

As a result, M_t is exactly the input history used to compute the next decoder state.

[2] While the "echoing" is a non-standard convention as it adds no new external information to the agent, we use it for readability; it makes the autoregressive feedback loop explicit by indicating that each action is conditioned on the previous observations.

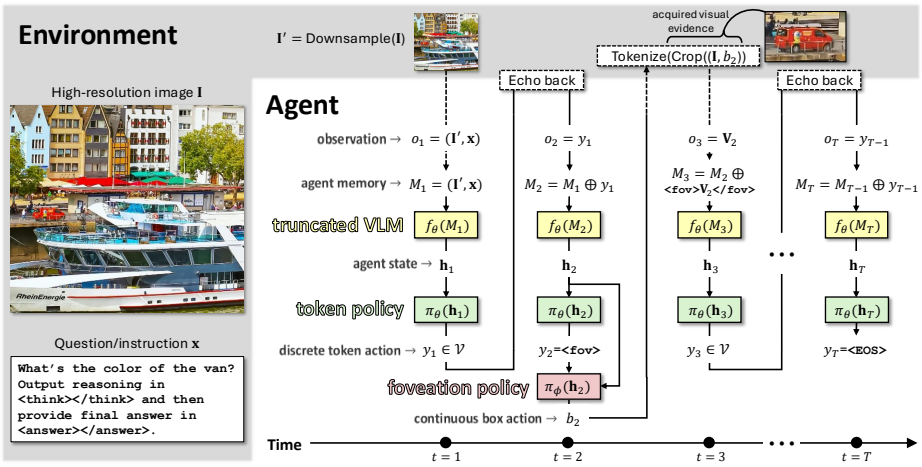

Fig. 2: The autoregressive pipeline of the proposed approach where $T = |\mathbf{y}|$.

Agent state. Since the agent never directly observes the full high-res image $\mathbf{I}$, it must integrate information over time. We summarize the current memory M_t using the decoder's hidden state $\mathbf{h}_t = f_\theta(M_t) \in \mathbb{R}^d$ where f_θ is the VLM truncated up to layer ℓ. Intuitively, $\mathbf{h}_t$ is the agent's "belief summary": a compact summary of what it believes about the unobserved details of $\mathbf{I}$ for the subsequent next action decision.

Actions. At each step t, the agent selects one of two action modes: emitting the next text token or acquiring additional high-res evidence. We parameterize these modes with (i) a token policy π_θ for discrete token generation and (ii) a foveation policy π_ϕ for continuous region selection, both conditioned on the current agent state $\mathbf{h}_t$. Concretely, the token policy samples either normal text in $\mathcal{V}$ or special trigger token `<fov>`:

$$y_t \sim \pi_\theta(\cdot \mid \mathbf{h}_t), \qquad y_t \in \mathcal{V} \cup \{\texttt{<fov>}\}. \tag{6}$$

If $y_t \in \mathcal{V}$ (the emitted token is normal text) under our observation convention, the environment echoes it back as $o_{t+1} = y_t$. If $y_t =$ `<fov>`, the agent instead requests new high-res evidence by sampling a continuous box using the foveation policy,

$$b_t \sim \pi_\phi(\cdot \mid \mathbf{h}_t), \qquad b_t = [x_t, y_t, w_t, h_t]. \tag{7}$$

Given b_t, the environment deterministically returns the corresponding evidence tokens $\mathbf{V}_t = \text{Tokenize}(\text{Crop}(\mathbf{I}, b_t)) \in \mathbb{R}^{m_t \times d}$ (where m_t depends on the crop size), which is injected back into the same autoregressive stream as a delimited block `<fov>` $\mathbf{V}_t$ `</fov>` (equivalently, $o_{t+1} = \mathbf{V}_t$), enabling subsequent decoding steps to condition on the newly acquired details without restarting the model.

This design reflects the role separation in human visual cognition: textual reasoning and spatial evidence acquisition are tightly coupled through a shared internal state, yet they operate in different output spaces (discrete linguistic actions *vs.* continuous geometric actions). Accordingly, FoveateR uses a shared agent state $\mathbf{h}_t$ for communication between reasoning and focusing, while delegating continuous box prediction to a dedicated policy net π_ϕ. Importantly, this head is *neither* a separate autoregressive localization model π_{loc} as in multi-pass pipelines, *nor* does it express localization through text tokens as in text-grounded methods; instead, it is a state-conditioned continuous action predictor executed within the same decoding trajectory.

Reward. The learning signal combines reward terms for (i) task success from the final answer and (ii) structural correctness of well-formed `<think>`/`<answer>` outputs, following [8]. During reinforcement learning, these rewards jointly optimize the token policy π_θ and the foveation policy π_ϕ. We refer to Sec. 3.3 for details.

3.2 Interleaved Token-and-Foveation Policy

Under the partially observable MDP formulation, FoveateR follows a joint policy over two types of actions: it primarily takes *discrete token* actions or *continuous foveation* actions (only when needed). Specifically, the backbone token policy π_θ generates the next token y_t. Whenever $y_t =$ `<fov>`, the foveation policy samples a box $b_t \sim \pi_\phi(\cdot \mid \mathbf{h}_t)$ and retrieves evidence $\mathbf{V}_t$ that is injected into the same context for future decoding steps. We define the set of foveation time steps in the generated sequence as $\mathcal{T} := \{t \mid y_t =$ `<fov>`$\}$. FoveateR then defines the following joint distribution over the discrete token action sequence $\mathbf{y}$ and the continuous foveation actions $\{b_t\}_{t\in\mathcal{T}}$:

$$\pi_{\theta,\phi}(\mathbf{y}, \{b_t\}_{t\in\mathcal{T}} \mid p) = \prod_{t=1}^{|\mathbf{y}|} \Big[\pi_\theta(y_t \mid p, y_{<t}, \{\mathbf{V}_\tau\}_{\tau<t}) \cdot \pi_\phi(b_t \mid \mathbf{h}_t)^{\mathbb{I}[y_t=\texttt{<fov>}]}\Big], \tag{8}$$

where $p := (\mathbf{I}', \mathbf{x})$, $\mathbb{I}[\cdot]$ is indicator function which equals 1 if its argument is true and 0 otherwise, and the injected evidence satisfies $\mathbf{V}_t = \text{Tokenize}(\text{Crop}(\mathbf{I}, b_t))$ for $t \in \mathcal{T}$. Eq. (8) highlights that FoveateR acquires high-res evidence only when triggered within a single autoregressive trajectory, with an adaptive number of foveation steps $|\mathcal{T}|$.

We now relate FoveateR to a standard autoregressive VLM, whose conditional generation distribution is defined in Eq. (1). Intuitively, a standard VLM takes only discrete token actions conditioned on a *static visual context*, whereas FoveateR additionally permits "event-triggered" continuous foveation actions conditioned on an *evolving visual context*. This relationship can be formalized as Proposition 1, which shows that standard AR VLMs are a special case of FoveateR: it preserves conventional next-token modeling while enabling *optional*, state-dependent evidence acquisition. This coupling leads to a mixed discrete-continuous learning problem, as both the (discrete) token policy π_θ and the (continuous) foveation policy π_ϕ must be jointly optimized. To this end, we employ a two-stage training pipeline that first bootstraps `<fov>` usage and then uses reinforcement learning to improve task performance as well as foveation behavior.

Proposition 1: Standard AR VLMs as a Special Case of FoveateR

Consider the joint policy of $\pi_{\theta,\phi}$ in Eq. (8). If the foveation trigger `<fov>` is never produced, i.e.,

$$\pi_\theta(y_t = \texttt{<fov>} \mid p, y_{<t}, \{\mathbf{V}_\tau\}_{\tau<t}) = 0 \quad \text{at every decoding step } t, \tag{9}$$

then FoveateR reduces exactly to the standard autoregressive VLM,

$$\pi_{\theta,\phi}(\mathbf{y}, \{b_t\}_{t\in\mathcal{T}} \mid p) = \pi_\theta(\mathbf{y} \mid p) = \prod_{t=1}^{|\mathbf{y}|} \pi_\theta(y_t \mid p, y_{<t}). \tag{10}$$

Proof. Under Eq. (9), $\mathcal{T} = \emptyset$ for any generated trajectory, so $\mathbb{I}[y_t =$ `<fov>`$] = 0$ for all t. Hence the foveation factor in Eq. (8) evaluates to 1 at every step, and the joint factorization collapses to the standard autoregressive product in Eq. (1). □

3.3 Training Foveated Reasoner

We initialize FoveateR from a pretrained instruction-following VLM (*e.g.*, Qwen2.5-VL [2]), which provides the truncated VLM f_θ as well as the token policy π_θ. We augment this backbone with a lightweight foveation policy π_ϕ, implemented as an MLP with nonlinearities which maps the hidden state to continuous box actions. Since the pretrained backbone is not trained to use `<fov>` tags and the foveation policy is newly initialized, we train FoveateR in two stages: (i) coldstart supervised finetuning (SFT) to introduce `<fov>` usage and box prediction, and (ii) reinforcement learning (RL) to further improve task performance and learn task-effective foveation trajectories.

(1) Coldstart SFT. The goal of this stage is to teach a pretrained VLM how to use the foveation mechanism in its decoding stream. However, standard visual reasoning supervision does not specify when `<fov>` should be triggered, which region should be queried, or how to integrate retrieved evidence. Coldstart dataset construction is therefore required to convert such samples into explicit teacher-forcing targets.

Dataset construction. Starting from a standard visual reasoning dataset [43], where each sample is a triplet $(\mathbf{I}, \mathbf{x}, \mathbf{a})$ of an image, a question/instruction, and a target answer, we construct a teacher-forcing target sequence $\tilde{\mathbf{y}} = (\tilde{y}_t)_{t=1}^{|\tilde{\mathbf{y}}|}$ that explicitly interleaves reasoning text with foveation segments. Given $(\mathbf{I}, \mathbf{x})$, a pretrained reasoning VLM [2] first generates an intermediate reasoning text (*i.e.*, a step-by-step rationale before the final answer), which is then split into S reasoning sentences $\{l_{(s)}\}_{s=1}^{S}$. For each sentence $l_{(s)}$, we obtain a bounding box $b_{(s)} = \text{Localize}(\mathbf{I}, l_{(s)})$, which localizes the image region described by that sentence [2]. From each box, we extract its corresponding high-res tokens $\mathbf{V}^\star_{(s)} := \text{Tokenize}(\text{Crop}(\mathbf{I}, b_{(s)}))$. We then build the following interleaved teacher-forcing target, where foveation and reasoning are interleaved within `<think>` tags:

$$\begin{aligned}\tilde{\mathbf{y}} := \mathbf{x}\ \texttt{<think>}\ \texttt{<fov>}\ \mathbf{V}^\star_{(1)}\ \texttt{</fov>}\ l_{(1)}\ \cdots\ \texttt{<fov>}\ \mathbf{V}^\star_{(S)}\ \texttt{</fov>}\ l_{(S)} \\ \texttt{</think>}\ \texttt{<answer>}\ \mathbf{a}\ \texttt{</answer>}.\end{aligned} \tag{11}$$

By placing foveation $\mathbf{V}^\star_{(s)}$ before each reasoning sentence $l_{(s)}$, the model is trained to first acquire task-relevant visual evidence and then produce reasoning based on that evidence. This design is consistent with a broader pattern in human cognition, where newly acquired evidence should be incorporated before subsequent reasoning updates.

Training objectives. Coldstart SFT jointly supervises (i) next-token generation by π_θ and (ii) box prediction by π_ϕ. For next-token prediction, we apply cross-entropy loss only to non-visual tokens in $\tilde{\mathbf{y}}$ (*i.e.*, textual tokens and special tag `<fov>`), excluding injected visual tokens $\mathbf{V}^\star_{(*)}$ because these are not predicted by the model (but deterministically retrieved from the environment given a `<fov>` trigger and the corresponding box). For box prediction, we define a set of foveation trigger positions as $\mathcal{T}_{\text{fov}}(\tilde{\mathbf{y}}) := \{t \mid \tilde{y}_t = \texttt{<fov>}\}$. If the `<fov>` token at position $t \in \mathcal{T}_{\text{fov}}(\tilde{\mathbf{y}})$ corresponds to sentence $l_{(s)}$, we assign the box supervision target $b^\star_t := b_{(s)}$. The resulting cold-start objective combines token prediction and box regression:

$$\mathcal{L}_{\text{cold}} = \underbrace{-\sum_{t \in \mathcal{M}(\tilde{\mathbf{y}})} \log \pi_\theta(\tilde{y}_t \mid p, \tilde{y}_{<t}, \{\mathbf{V}^\star_\tau\}_{\tau<t})}_{\mathcal{L}_{\text{LM}}} + \underbrace{\lambda_{\text{box}} \sum_{t \in \mathcal{T}_{\text{fov}}(\tilde{\mathbf{y}})} \|g_\phi(\mathbf{h}_t) - b^\star_t\|_1}_{\mathcal{L}_{\text{box}}}, \tag{12}$$

where $\mathcal{M}(\tilde{\mathbf{y}})$ denotes the set of non-visual-token positions, and $g_\phi(\mathbf{h}_t)$ is box predicted from the same foveation policy as the conditional box policy $\pi_\phi(\cdot \mid \mathbf{h}_t)$ (*e.g.*, by taking its mean $g_\phi(\mathbf{h}_t) := \mathbb{E}_{b \sim \pi_\phi(\cdot|\mathbf{h}_t)}[b]$). In short, this objective bootstraps both `<fov>` usage and box prediction by foveation policy, serving as initialization prior to RL.

(2) RL finetuning with GRPO. The coldstart stage provides useful initialization for foveation behavior, but its supervision is not necessarily optimal for end-task performance. Both the reasoning $l_{(*)}$ and the foveation labels $b^\star_*$ are "pseudo" labels from a pretrained VLM, which can bias FoveateR toward a teacher-induced foveation/reasoning behavior. In practice, however, the ideal amount of reasoning and foveation is problem-dependent: some examples can be solved directly from the coarse view without any foveation (*e.g.*, identifying the color of a salient object), whereas others may demand multiple foveations with no intermediate reasoning (*e.g.*, zooming into a few small text regions to read the answer). To train this problem-adaptive behavior, we further optimize the model using RL.

RL for token policy. We optimize the token policy π_θ with a group-relative policy optimization (GRPO) objective, following the work of [8, 44]. Specifically, for each prompt p, we sample a group of G trajectories $\{\mathbf{y}^{(i)}\}_{i=1}^{G} \sim \pi_{\theta,\phi(\text{old})}(\cdot \mid p)$ and assign a reward $r^{(i)} = r^{(i)}_{\text{acc}} + r^{(i)}_{\text{fmt}}$, where $r^{(i)}_{\text{acc}} = 1$ if the extracted final answer is correct, and $r^{(i)}_{\text{fmt}} = 1$ if the output follows the required structural format (*e.g.*, well-formed `<think>` and `<answer>` tags); otherwise each term is 0. Instead of using an external value function, GRPO computes a relative advantage within each sampled group as follows: $A^{(i)} = r^{(i)} - 1/G \sum_{j=1}^{G} r^{(j)}$, centering the reward at the mean, so above-average trajectories obtain positive advantage, while below-average ones are penalized by negative advantage. We then optimize π_θ using the GRPO objective (omitting PPO-style clipping for brevity):

$$\mathcal{L}_{\text{GRPO-tok}} = -\mathbb{E}\left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|\mathcal{M}^{(i)}|}\sum_{t\in\mathcal{M}^{(i)}}\frac{\pi_{\theta,\phi}(y_t^{(i)} \mid p)}{\pi_{\theta,\phi(\text{old})}(y_t^{(i)} \mid p)}A^{(i)} - \beta\,\mathbb{D}_{\text{KL}}\right], \tag{13}$$

where $\mathcal{M}^{(i)} = \mathcal{M}(\mathbf{y}^{(i)})$ is the set of non-visual-token positions in trajectory i, $\mathbb{D}_{\text{KL}}$ is a regularization term [8] that prevents overly large updates of $\pi_{\theta,\phi}$ by keeping it close to the reference model (*e.g.*, coldstart model) and $\beta > 0$ controls the regularization.

RL for foveation policy. We optimize the foveation policy π_ϕ with a GRPO objective analogous to the token-policy update (Eq. (13)), but using only *accuracy advantage* $A^{(i)}_{\text{acc}} = r^{(i)}_{\text{acc}} - 1/G \sum_{j=1}^{G} r^{(j)}_{\text{acc}}$ because format reward r_{fmt} is irrelevant to the foveation prediction. At each foveation step $t \in \mathcal{T}(\mathbf{y}^{(i)})$, the foveation policy samples a box $b_t^{(i)} \sim \pi_\phi(\cdot \mid \mathbf{h}_t^{(i)})$, and we denote the corresponding action likelihood (probability density) by $\pi_\phi(b_t^{(i)} \mid \mathbf{h}_t^{(i)})$. We then optimize π_ϕ using a GRPO objective over the sampled foveation actions as follows:

$$\mathcal{L}_{\text{GRPO-fov}} = -\mathbb{E}\left[\frac{1}{G}\sum_{i=1}^{G}A^{(i)}_{\text{acc}}\sum_{t\in\mathcal{T}(\mathbf{y}^{(i)})}\frac{\pi_\phi(b_t^{(i)} \mid \mathbf{h}_t^{(i)})}{\pi_{\phi(\text{old})}(b_t^{(i)} \mid \mathbf{h}_t^{(i)})}\right]. \tag{14}$$

As a result, the foveation policy is trained to prefer box selections that improve end-task answer accuracy under the current group-relative comparison.

Foveated region regularization. Given the two GRPO objectives only (Eqs. (13) & (14)), the model may converge to an undesirable trivial solution: it can increase reward by simply enlarging the foveated box (*e.g.*, increasing its width and height), thereby acquiring larger high-res regions. In the extreme, it observes the entire high-res image $\mathbf{I}$, which undermines the intended resolution–efficiency trade-off, constituting a form of reward hacking. To discourage this behavior, we introduce an additional objective that penalizes large foveated regions only when the model prediction is already

correct ($r_{\text{acc}}^{(i)} = 1$); notably, our foveation actions are *differentiable, continuous box parameters* (rather than linguistic coordinates), which allows us to directly regularize the queried area as follows:

$$\mathcal{L}_{\text{reg}} = \mathbb{E}\left[\frac{1}{G}\sum_{i=1}^{G} r_{\text{acc}}^{(i)} \sum_{t \in \mathcal{T}(\mathbf{y}^{(i)})} w_t^{(i)} \cdot h_t^{(i)}\right], \tag{15}$$

where $b_t^{(i)} = [x_t^{(i)}, y_t^{(i)}, w_t^{(i)}, h_t^{(i)}]$ denotes the sampled foveation action at step t in trajectory i, and the product $w_t^{(i)} \cdot h_t^{(i)}$ measures the size of the queried region. When a prediction is correct ($r_{\text{acc}}^{(i)} = 1$), this term penalizes large foveated regions and encourages the model to solve the task with less high-res evidence. When incorrect ($r_{\text{acc}}^{(i)} = 0$), the regularizer is inactive, allowing the model to use larger regions to acquire additional evidence. Overall, this regularization encourages FoveateR to learn accuracy-preserving yet evidence-efficient foveation, rather than trivially enlarging crops to improve reward.

Final objective. The RL-stage objective jointly optimizes the token policy (Eq. (13)), the foveation policy (Eq. (14)), and the foveated-region regularization (Eq. (15)):

$$\mathcal{L}_{\text{RL}} = \mathcal{L}_{\text{GRPO-tok}} + \lambda_{\text{fov}}\,\mathcal{L}_{\text{GRPO-fov}} + \lambda_{\text{reg}}\,\mathcal{L}_{\text{reg}}, \tag{16}$$

where λ_{fov} and λ_{reg} balance the contributions of $\mathcal{L}_{\text{GRPO-fov}}$ and $\mathcal{L}_{\text{reg}}$, respectively. We provide additional details of the proposed method in the supplementary material.

4 Experiments

4.1 Experimental setup

Datasets. For training and coldstart dataset generation, we use the training splits of the Visual CoT benchmark [43] (438K), RefCOCO/+/g [15] (321K), and ScienceQA [28] (6K), following the recent baseline [43], which yields a total of 765K $(\mathbf{I}, \mathbf{x}, \mathbf{a})$ triplet samples. For evaluation, we primarily report results on the validation splits of Visual CoT benchmark, which comprises 12 datasets spanning document/text/chart understanding (DocVQA [31], TextCaps [46], TextVQA [47], DUDE [18], SROIE [11], InfographicsVQA [30]), general VQA (Flickr30k [36], Visual7W [67]), relational reasoning (GQA [12], OpenImages [17], VSR [24]), and fine-grained understanding (CUB [49]), where we evaluate DUDE, SROIE, and Visual7W in a zero-shot setting following prior baselines [43, 64]. For consistency with [43], we adopt the same GPT-based evaluation metric [34]. We additionally evaluate on V* Bench [51], a multiple-choice benchmark designed to probe fine-grained visual perception, including subsets such as direct attribute, GPT4V-hard, and OCR, using the same MCQ evaluation protocol in [64].

Implementation details. We build FoveateR on top of Qwen2.5-VL [2] (3B and 7B) as the backbone VLM, and compare against both recent visual focusing methods [43, 59, 60, 64] and standard VLMs [23, 26]. We train FoveateR in two stages: coldstart for 3 epochs and RL finetuning for 70k steps, using the same batch size of 16 with learning rates of $3e$-5 and $1e$-6, respectively. For RL, we use task-specific rewards: (i) LLM-based string matching for VQA tasks [43], which returns a score in $[0, 1]$ based on semantic agreement between the predicted and ground-truth (GT) answers; (ii) IoU-based accuracy for grounding tasks [15], assigning 1.0 if the IoU between the predicted and GT boxes is ≥ 0.5 and 0.0 otherwise; and (iii) multiple-choice accuracy for ScienceQA [28], assigning 1.0 if the predicted option matches the GT and 0.0 otherwise. Additional implementation details are provided in the supplementary.

Table 1: Comparison on the Visual CoT benchmark [43]. Method subscripts denote visual focusing type: multi-pass ($*_{\text{multi}}$), text-grounded ($*_{\text{text}}$). uses GT box indicates use of human-annotated boxes for "visual focusing supervision", and visual token cnt reports the number of visual tokens the model uses, directly impacting inference cost.

Method	uses GT box	visual token cnt(↓)	Doc/Text/Chart						General		Relation			Fine
			DocV	TxtC	TxtV	Dude*	Sroi*	Info	F30k	V7W*	GQA	OI	VSR	CUB
224^2 input image resolution:														
Sphinx^{13B} [23]	✗	1734	19.8	55.1	53.2	0.00	7.1	35.2	60.7	55.8	58.4	46.7	61.3	50.5
VPT$^{2B}_{\text{text}}$ [59]	✓	≥375	12.5	53.7	46.6	10.3	-	-	**68.0**	-	60.6	**84.2**	65.7	**89.2**
VisCoT$^{7B}_{\text{multi}}$ [43]	✓	1,152	35.5	61.0	71.9	27.9	34.1	35.6	67.1	58.0	61.6	83.3	68.2	55.6
FoveateR^{3B}	✗	$242.3_{\pm 168.7}$	61.1	70.0	76.8	48.9	65.7	42.6	57.4	52.9	56.7	77.7	69.4	81.7
FoveateR^{7B}	✗	$253.8_{\pm 162.7}$	**73.9**	**76.4**	**83.5**	**55.1**	**73.8**	**51.2**	59.1	**58.6**	**64.5**	83.7	**74.4**	87.4
336^2 input image resolution:														
Llava1.5^{7B} [26]	✗	576	24.4	59.7	58.8	29.0	13.6	40.0	58.1	57.5	53.4	41.2	57.2	53.0
Llava1.5^{13B} [26]	✗	576	26.8	61.5	61.7	28.7	16.4	42.6	62.0	58.0	57.1	41.3	59.0	57.3
UV-CoT$^{7B}_{\text{multi}}$ [64]	✗	1,152	28.3	-	71.1	25.3	22.7	19.8	64.9	45.5	56.8	-	55.3	-
VisCoT$^{7B}_{\text{multi}}$ [43]	✓	1,152	47.6	67.5	77.5	38.6	47.0	32.4	66.8	55.8	63.1	82.2	61.4	55.9
DocThink^{3B} [60]	✓	official code unavailable	46.0	66.3	74.6	21.3	48.6	35.5	66.4	57.2	48.6	48.5	62.5	-
DocThink^{7B} [60]	✓		57.9	68.2	80.2	40.8	49.5	34.7	**67.4**	58.0	54.6	54.2	65.6	-
FoveateR^{3B} (ours)	✗	$307.0_{\pm 156.0}$	74.0	74.0	83.0	58.2	75.4	49.0	60.1	56.1	60.5	77.6	68.1	82.7
FoveateR^{7B} (ours)	✗	$322.5_{\pm 162.7}$	**83.3**	**78.3**	**86.2**	**62.7**	**83.2**	**60.0**	60.8	**61.3**	**68.2**	**85.5**	**73.9**	**88.6**

4.2 Experimental results and analyses

Comparison with previous approaches. Table 1 compares FoveateR against recent visual focusing methods [43, 59, 64] and standard VLM baselines [23, 26, 60] under two low-res settings, $H', W' \in \{224, 336\}$. Across both 3B and 7B backbones, FoveateR outperforms prior methods on most benchmarks while consuming fewer visual tokens. Specifically under the 336^2 setting, our FoveateR^{3B} (i) uses only ∼307 visual tokens on average (*vs.* 1,152 in [43]) (ii) *without* using GT-box supervision for foveation policy training, while outperforming GT-trained prior visual focusing methods [43,59,64] on every document understanding task. We attribute these gains to our RL objectives: $\mathcal{L}_{\text{GRPO-fov}}$ learns a *goal-driven* foveation policy that directly optimizes end-task accuracy—rather than mimicking GT-box targets—thereby selecting regions most useful for the final answer, while the foveated-region regularizer $\mathcal{L}_{\text{reg}}$ penalizes large foveations to reduce the number of acquired visual tokens.

Table 2: Zero-shot results on V* [51]. Results of [35, 43, 55, 64] are from [64].

Method (336^2 input)	Attr	GPT4	OCR
MiniCPM^{8B} [55]	32.2	11.8	10.0
OmniLMM^{12B} [35]	32.6	11.8	16.7
VisCoT$^{7B}_{\text{multi}}$ [43]	33.0	11.8	59.3
UV-CoT$^{7B}_{\text{multi}}$ [64]	35.2	17.6	67.7
FoveateR^{3B} (ours)	**56.5**	**76.5**	**80.0**

Since FoveateR adaptively varies its visual budget across samples, we report both the mean and standard deviation of the visual token count (*e.g.*, $307.0_{\pm 156.0}$ for 3B at 336^2). The *huge variance* is expected: FoveateR uses little or no extra evidence for easy cases, while allocating more foveations to visually demanding cases as shown in Fig. 4. This problem-dependent allocation is learned during the RL stage (Eqs. (13)-(15)): by rewarding end-task success, the policy is encouraged to request additional evidence only when it improves the final answer, and to stop foveating otherwise, achieving a favorable accuracy-efficiency balance in expectation.

From Tab. 2, we observe a similar trend on V* Bench [51], which features substantially higher-resolution images (average 2246×1582) than the Visual CoT benchmarks. Even in this visually demanding regime, our 3B model under the 336^2 setting outperforms prior 7-12B methods, highlighting that the proposed evidence acquisition effectively scales to higher-res inputs by foveating only task-relevant regions.

Table 3: Resolution-efficiency ablation. From a low-res baseline (b), we quantify how it recovers accuracy via selective foveation. Oracle is Qwen2.5-VL with original res.

Model	input img res.	λ_{reg}	DocV [31] (**1787×2101**)				TxtV [47] (**941×820**)				V7W [67] (**556×455**)			
			N_{fov}	ρ	T_{fov}	acc	N_{fov}	ρ	T_{fov}	acc	N_{fov}	ρ	T_{fov}	acc
(a) oracle^{7B} [2]	org	-	-	100%	-	90.4	-	100%	-	87.8	-	100%	-	48.2
(b) baseline^{7B} [2]	336^2	-	-	0%	-	55.1	-	0%	-	77.2	-	0%	-	46.5
(c) ours$^{7B}_{\text{cold}}$	336^2	-	35.9	3%	2.3	64.2	73.9	7%	1.4	82.2	32.2	8%	1.1	58.3
(d) ours$^{7B}_{\text{cold+RL}}$	336^2	1.0	62.5	5%	2.2	61.5	48.8	4%	1.5	81.2	44.3	4%	1.3	57.2
(e) ours$^{7B}_{\text{cold+RL}}$	336^2	0.5	269.3	22%	2.3	81.5	218.9	18%	1.7	85.4	61.8	15%	1.2	61.0
(f) ours$^{7B}_{\text{cold+RL}}$	336^2	0.2	354.8	25%	2.3	83.3	263.5	21%	1.7	86.2	81.0	17%	1.3	61.3
(g) ours$^{7B}_{\text{cold+RL}}$	336^2	0.0	377.8	28%	2.4	74.8	319.6	24%	2.0	85.0	92.0	19%	1.6	59.3

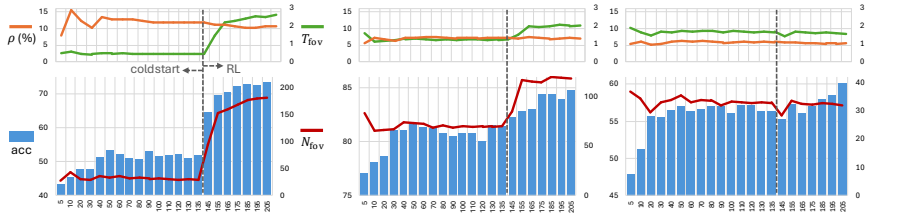

Fig. 3: Training dynamics of FoveateR^{3B}. The x-axis denotes training steps, and the gray vertical line indicates the stage transition from coldstart to RL.

Ablation study. Table 3 reports an ablation on our model and baseline [2], measuring accuracy with foveation usage statistics—N_{fov} (additional visual tokens acquired via foveation), ρ (fraction of the high-res image area revealed to the model), and T_{fov} (number of foveation steps)—on three benchmarks: DocV [31] (high-res documents), TxtV [47] (comparatively low-res text-centric images), and V7W [67] (general VQA). Fig. 3 further visualizes the dynamics of these metrics during training.

Comparing (a) with (b), we observe sharp accuracy drops, highlighting aggressive downsampling severely limits fine-grained perception, particularly for text-heavy tasks (*e.g.*, DocV and TxtV). Comparing the baseline (b) with our coldstart model (c), we observe large gains across all three datasets, even though the model only foveates a small portion of the image ($\rho = 3\%$–8%) with a limited number of extra visual tokens ($N_{\text{fov}} = 32$–74, $T_{\text{fov}} \approx 1$–2). This suggests that coldstart is sufficient to bootstrap *where-to-look* behavior: reading just a few task-relevant regions often recovers the accuracy or even outperforms the oracle (58.3 *vs.* 48.2 on V7W). The trend is consistent with Fig. 3, where accuracy improves rapidly in initial coldstart (from step 5 to 40) while the foveation usage (N_{fov}, ρ, T_{fov}) remains conservative.

Moreover, we study RL stage with different λ_{reg}, which penalizes excessive number of visual tokens acquired via foveation (cf. Sec. 3): With a strong regularization ($\lambda_{\text{reg}} = 1.0$, (d)), the model under-utilizes foveation (small ρ and N_{fov}) and fails to improve accuracy. Relaxing it ($\lambda_{\text{reg}} = 0.5$ and 0.2, (e)–(f)) allows more foveated evidence ($\rho \approx 15\%$–25%), leading to large accuracy gains on all the three benchmarks. This is also reflected in Fig. 3: after the coldstart→RL transition, DocV and TxtV show a sharp increase in N_{fov} and T_{fov}, with a noticeable jump in accuracy, spending more visual budget when fine-grained reading is necessary. Removing the regularization entirely ($\lambda_{\text{reg}} = 0.0$, (g)) further increases foveation usage but degrades accuracy, suggesting that *unconstrained evidence acquisition* can be inefficient, introducing distractions. In our experiments, we set $\lambda_{\text{reg}} = 0.2$ as a good compromise.

Table 4: Coldstart target ablation under 336^2 setup. We mark fov-only results with ⇑ (⇓) if they are $> 2\%p$ higher (lower) than the interleaved counterpart, $\sim$ otherwise.

Model	cold start trg	Doc/Text/Chart						General		Relation			Fine
		DocV	TxtC	TxtV	Dude	Sroi	Info	F30k	V7W	GQA	OI	VSR	CUB
(a) $\text{ours}^{\text{3B}}_{\text{cold}}$	$\tilde{\mathbf{y}}$	55.4	71.6	81.4	38.4	57.8	46.0	59.3	55.5	61.2	78.9	65.1	79.5
(b) $\text{ours}^{\text{3B}}_{\text{cold}}$	$\tilde{\mathbf{y}}_{\text{fov}}$	53.8 (∼)	72.8 (∼)	80.8 (∼)	40.8 (⇑)	56.7 (∼)	43.6 (⇓)	58.3 (∼)	57.2 (∼)	64.1 (⇑)	81.7 (⇑)	67.9 (⇑)	79.5 (∼)
(c) $\text{ours}^{\text{3B}}_{\text{cold+RL}}$	$\tilde{\mathbf{y}}$	74.0	74.0	83.0	58.2	75.4	49.0	60.1	56.1	60.5	77.6	68.1	82.7
(d) $\text{ours}^{\text{3B}}_{\text{cold+RL}}$	$\tilde{\mathbf{y}}_{\text{fov}}$	77.3 (⇑)	75.9 (∼)	85.6 (⇑)	58.5 (∼)	74.4 (∼)	47.9 (∼)	59.3 (∼)	56.9 (∼)	62.5 (∼)	71.4 (⇓)	71.7 (⇑)	84.6 (∼)

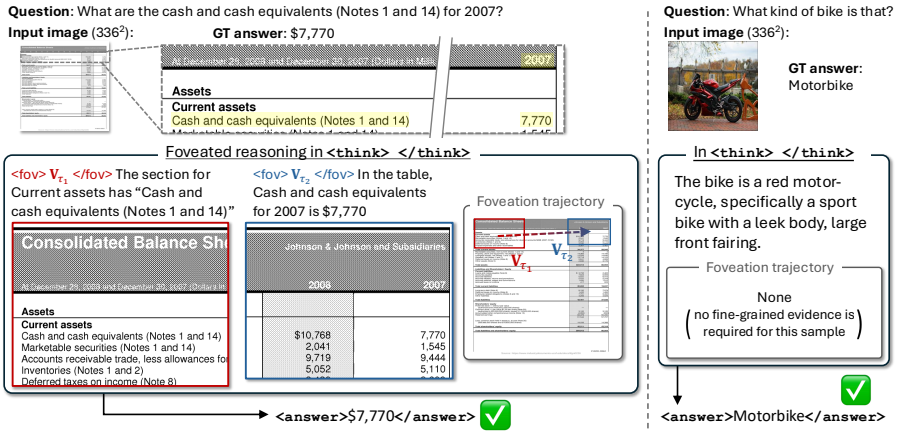

Fig. 4: Visually demanding cases (left; *e.g.*, documents; more foveations/N_{vis}) *vs.* less demanding cases (right; *e.g.*, centered dominant objects; fewer or no foveations/N_{vis}).

Coldstart training with 'foveation-only' *vs.* 'interleaved foveation-reasoning'. Table 4 compares two coldstart target constructions: The first is an *interleaved* target $\tilde{\mathbf{y}}$ that alternates foveated evidences $\mathbf{V}^{\star}_{(*)}$ and rationales $l_{(*)}$ (Eq. (11)). The second is a *foveation-only* (fov-only) target $\tilde{\mathbf{y}}_{\text{fov}}$ that removes the intermediate rationales $l_{(*)}$ and supervises only the visual evidence $\mathbf{V}^{\star}_{(*)}$ acquisition behavior as follows:

$$\tilde{\mathbf{y}}_{\text{fov}} := \mathbf{x}\ \texttt{<think><fov>}\mathbf{V}^{\star}_{(1)}\texttt{</fov>}\cdots\texttt{<fov>}\mathbf{V}^{\star}_{(S)}\texttt{</fov></think>}\cdots . \tag{17}$$

Interestingly, the two coldstart variants, (a) and (b), achieve comparable performance across most benchmarks, suggesting that coldstart can learn an effective *where-to-look* policy even without intermediate narration $l_{(*)}$. Notably, the fov-only target (b) is even slightly stronger on relation datasets (GQA/OI/VSR). We attribute this to the nature of their questions (*e.g.*, "who is dressed in blue?" or "what is he wearing on his hand?"), which typically require only *one decisive region lookup*, making $\tilde{\mathbf{y}}_{\text{fov}}$ a more direct signal for evidence acquisition than relying on extraneous narration.

We further apply RL to both checkpoints, yielding (c) and (d). Even after RL, the fov-only checkpoint (d) performs comparably—and often better—than the interleaved checkpoint (c) on most datasets, implying accurate prediction does not necessarily require explicit rationales $l_{(*)}$. We hypothesize that, in such low-res input setup, what matters more is *acquiring the right visual evidence* $\mathbf{V}^{\star}_{(*)}$ through foveation: intermediate rationales can be redundant, often paraphrasing what is already visible without adding much additional structure, and may even inject noise by encouraging unnecessary narration. By contrast, learning an effective evidence-acquisition behavior appears more critical under low-res inputs, where success largely depends on capturing decisive high-res visual cues—*a picture is worth a thousand words.*

Table 5: Foveation localization quality on the VisCoT benchmark. N_{fov} and Acc. are averaged over all 12 VisualCoT datasets [43]; coverage and IoU@0.5 are shown for four representative ones: DocVQA, V7W, GQA, and CUB.

Model	GT box	T_{fov}	N_{fov}	reas.	GT-region coverage (%) ↑				IoU@0.5				Acc.
					DocV	V7W	GQA	CUB	DocV	V7W	GQA	CUB	
VisCoT$^{7B}_{\text{multi}}$	✓	fixed(1)	1152	✗	1.9	8.1	11.3	13.3	19.6	27.8	46.5	45.4	57.9
FoveateR$^{3B}_{\text{cold+RL}}$	✗	dynamic	258.1	✗	**11.0**	59.3	65.7	**60.0**	15.2	18.3	22.1	47.0	**68.8**
FoveateR$^{3B}_{\text{cold+RL}}$	✗	dynamic	242.3	✓	9.6	**63.3**	**68.5**	58.3	15.1	18.5	22.6	44.7	68.2

Table 6: End-to-end efficiency on V7W (single H100). Visual-token and latency figures are measured on V7W only; throughput is total processed tokens per second.

Model	Input res.	Fov.?	Reas.?	T_{fov}	# foveated vis. tokens	# text tokens	Throughput (tok/s)↑	VRAM (G)↓	Latency (s)↓
Qwen2.5-VL^{3B}	orig.	✗	✗	0	0	72.9	32.7	35	<u>2.23</u>
VisCoT^{7B}	336^2	✓	✗	1	576	0	58.9	<u>25</u>	9.78
FoveateR$^{3B}_{\text{cold+RL}}$	336^2	✓	✗	1.3	80.6	0	**76.0**	**24**	**1.06**
FoveateR$^{3B}_{\text{cold+RL}}$	336^2	✓	✓	1.2	76.6	121.9	<u>75.8</u>	<u>25</u>	2.62

4.3 Foveation Quality and End-to-End Efficiency

Foveation localization quality. A natural concern is whether the learned policy π_ϕ attends to task-relevant regions, or merely accumulates evidence until the answer is found. Since VisCoT predicts a single box whereas FoveateR may emit several, a per-box IoU@0.5 is not directly comparable; we therefore report *GT-region coverage*—the fraction of GT-region pixels covered by the union of predicted boxes—which measures how much of the GT evidence is collectively acquired and credits FoveateR's ability to recover from an initially imprecise box at later decoding steps. As shown in Tab. 5, despite a smaller backbone and *no* human-annotated box supervision, FoveateR covers far more GT evidence than VisCoT^{7B} (*e.g.*, $8.1 \rightarrow 63.3$ on V7W) with fewer foveated tokens, while improving average accuracy by over 10%p.

End-to-end efficiency. To confirm that the gains are not confined to visual-token count, we measure wall-clock latency (from tokenization to final prediction), peak VRAM, and throughput on a single H100 (Tab. 6). Relative to the multi-pass VisCoT^{7B}, FoveateR cuts visual tokens by ∼7× and latency by ∼9×, with lower peak memory and higher throughput at better accuracy—avoiding the repeated re-encoding that makes multi-pass pipelines costly. Moreover, comparing our two variants isolates the cost of emitting text: adding intermediate reasoning alone raises the text-token budget to 121.9 and roughly doubles latency ($1.06 \rightarrow 2.62$s). This indicates that text-grounded methods, which must additionally verbalize box coordinates at every foveation, would incur an even larger budget—a cost FoveateR avoids by acquiring evidence through non-linguistic, continuous actions. Together, these results show that FoveateR's efficiency stems from *acquiring the right evidence rather than more of it*, validating single-pass foveation as a practical path to high-resolution reasoning under tight budgets.

5 Related Work

Effective visual evidence acquisition has been studied for decades, from classic neural vision [33] to modern VLMs [43]. Below, we briefly review some recent relevant work.

Token reduction. A natural way to reduce visual cost is to start from a dense set of visual tokens and then *select*, *merge*, or *prune* tokens to form a compact representation [3, 7, 16, 21, 27, 39, 41, 53, 58]. One key drawback is that token reduction methods

mostly compress an "already-observed" representation: if fine-grained details are missing in the initial (often low-res) encoding, token selection alone cannot recover new information. Feeding a higher-res image can preserve such details, but it makes the initial encoding itself compute-heavy. Moreover, aggressive pruning may discard task-critical evidence, making performance sensitive to the selection policy/threshold.

Visual reasoning. A large class of VLMs perform visual reasoning, in which the model acquires task-relevant information by composing it through *textual reasoning about the visual content*. In practice, these models [1,5,10,20,25,31,38,50,52,61,63,66] typically operate in a *fixed-context* manner: the input image is encoded once into a static set of visual tokens, and all subsequent reasoning is conditioned on this unchanged visual context. As a result, performance is ultimately bounded by the level of detail preserved in the initial visual input; increasing its resolution may recover fine-grained details but it quickly brings back the resolution–efficiency trade-off.

Visual focusing. To overcome the fixed-context bottleneck, visual focusing methods enable the model to adaptively acquire additional evidence during inference, typically via coarse-to-fine procedures that revisit the high-res image. While empirically effective, there remain fundamental bottlenecks: (i) Multi-pass pipelines [4,13,19,37,43,45,51,54,64] can increase inference latency due to repeated re-encoding/re-prompting, often breaking hidden-state continuity across different passes. (ii) Text-grounded methods [6,9,42,48,54,59,62,65] often introduce token overhead and format-brittleness. In contrast to both streams, FoveateR differs along multiple axes. (1) It preserves an uninterrupted trajectory via a shared KV-cache instead of resetting hidden states across passes; (2) acquires all evidence in a *single* run, conditioning each foveation on prior evidence *and* queried positions; (3) emits continuous, non-linguistic box actions rather than brittle coordinate strings; and (4) jointly optimizes foveation and reasoning end-to-end without GT-box supervision. These instantiate the *process* of foveation—stateful, action-based evidence acquisition—rather than merely emulating its *outcome*.

Foveation in machine vision. Early work by Mnih *et al.* [33] explored a human-inspired foveation paradigm, where an agent sequentially selects glimpses and integrates them through an internal state. Although conceptually appealing, it was difficult to demonstrate as a practical solution at the time, due to limited model capacity, unstable optimization, and the lack of large-scale data [1, 26]. Subsequent machine vision research has adopted idea of foveation more broadly, *e.g.*, foveated/peripheral representations [14, 29, 40], eccentricity-aware pooling [22, 57], and peripheral-aware attention [32]. In this paper, we also draw inspiration from three key characteristics of foveation: (i) *coarse-to-fine* evidence acquisition mechanism, (ii) an *uninterrupted, stateful* inference process, and (iii) *continuous action-based* control. Based on these principles, we instantiate foveation as a practical vision–language system.

6 Conclusion

This paper argues that the resolution–efficiency trade-off requires not only "zooming in," but also stateful, action-driven evidence acquisition. We presented FoveateR, a stateful, single-pass autoregressive framework that triggers non-linguistic, continuous foveation actions on demand and integrates the retrieved high-resolution evidence into the on-going decoding stream, preserving hidden-state continuity. Trained with a two-stage pipeline, it learns task-adaptive foveation strategies and delivers strong performance with limited visual tokens on a broad set of vision–language benchmarks. We hope this work inspires future multimodal systems under strict compute budgets.

References

1. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., Ring, R., Rutherford, E., Cabi, S., Han, T., Gong, Z., Samangooei, S., Monteiro, M., Menick, J., Borgeaud, S., Brock, A., Nematzadeh, A., Sharifzadeh, S., Binkowski, M., Barreira, R., Vinyals, O., Zisserman, A., Simonyan, K.: Flamingo: a visual language model for few-shot learning. arXiv preprint arXiv:2204.14198 (2022) 1, 3, 15
2. 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. arXiv preprint arXiv:2502.13923 (2025) 8, 10, 12
3. Bolya, D., Fu, C.Y., Dai, X., Zhang, P., Feichtenhofer, C., Hoffman, J.: Token merging: Your vit but faster. In: International Conference on Learning Representations (ICLR) (2023) 14
4. Carvalho, M., Dias, H., Martins, B.: Cropvlm: Learning to zoom for fine-grained vision-language perception. arXiv preprint arXiv:2511.19820 (2025) 1, 3, 4, 15
5. Dai, W., Li, J., Li, D., Tiong, A.M.H., Zhao, J., Wang, W., Li, B., Fung, P., Hoi, S.: Instructblip: Towards general-purpose vision-language models with instruction tuning. arXiv preprint arXiv:2305.06500 (2023) 15
6. Fan, Y., He, X., Yang, D., Zheng, K., Kuo, C.C., Zheng, Y., Narayanaraju, S.J., Guan, X., Wang, X.E.: Grit: Teaching mllms to think with images. In: Advances in Neural Information Processing Systems (NeurIPS) (2025) 1, 3, 4, 15
7. Goyal, S., Choudhury, A.R., Raje, S.M., Chakaravarthy, V.T., Sabharwal, Y., Verma, A.: Power-bert: Accelerating bert inference via progressive word-vector elimination. In: Proc. International Conference on Machine Learning (ICML) (2020) 14
8. Guo, D., Yang, D., Zhang, H., Song, J., Wang, P., Zhu, Q., Xu, R., Zhang, R., Ma, S., Bi, X., Zhang, X., Yu, X., Wu, Y., Wu, Z.F., Gou, Z., Shao, Z., Li, Z., Gao, Z., Liu, A., Xue, B., Wang, B., Wu, B., Feng, B., Lu, C., Zhao, C., Deng, C., Ruan, C., Dai, D., Chen, D., Ji, D., Li, E., Lin, F., Dai, F., Luo, F., Hao, G., Chen, G., Li, G., Zhang, H., Xu, H., Ding, H., Gao, H., Qu, H., Li, H., Guo, J., Li, J., Chen, J., Yuan, J., Tu, J., Qiu, J., Li, J., Cai, J.L., Ni, J., Liang, J., Chen, J., Dong, K., Hu, K., You, K., Gao, K., Guan, K., Huang, K., Yu, K., Wang, L., Zhang, L., Zhao, L., Wang, L., Zhang, L., Xu, L., Xia, L., Zhang, M., Zhang, M., Tang, M., Zhou, M., Li, M., Wang, M., Li, M., Tian, N., Huang, P., Zhang, P., Wang, Q., Chen, Q., Du, Q., Ge, R., Zhang, R., Pan, R., Wang, R., Chen, R.J., Jin, R.L., Chen, R., Lu, S., Zhou, S., Chen, S., Ye, S., Wang, S., Yu, S., Zhou, S., Pan, S., Li, S.S., Zhou, S., Wu, S., Yun, T., Pei, T., Sun, T., Wang, T., Zeng, W., Liu, W., Liang, W., Gao, W., Yu, W., Zhang, W., Xiao, W.L., An, W., Liu, X., Wang, X., Chen, X., Nie, X., Cheng, X., Liu, X., Xie, X., Liu, X., Yang, X., Li, X., Su, X., Lin, X., Li, X.Q., Jin, X., Shen, X., Chen, X., Sun, X., Wang, X., Song, X., Zhou, X., Wang, X., Shan, X., Li, Y.K., Wang, Y.Q., Wei, Y.X., Zhang, Y., Xu, Y., Li, Y., Zhao, Y., Sun, Y., Wang, Y., Yu, Y., Zhang, Y., Shi, Y., Xiong, Y., He, Y., Piao, Y., Wang, Y., Tan, Y., Ma, Y., Liu, Y., Guo, Y., Ou, Y., Wang, Y., Gong, Y., Zou, Y., He, Y., Xiong, Y., Luo, Y., You, Y., Liu, Y., Zhou, Y., Zhu, Y.X., Huang, Y., Li, Y., Zheng, Y., Zhu, Y., Ma, Y., Tang, Y., Zha, Y., Yan, Y., Ren, Z.Z., Ren, Z., Sha, Z., Fu, Z., Xu, Z., Xie, Z., Zhang, Z., Hao, Z., Ma, Z., Yan, Z., Wu, Z., Gu, Z., Zhu, Z., Liu, Z., Li, Z., Xie, Z., Song, Z., Pan, Z., Huang, Z., Xu, Z., Zhang, Z., Zhang, Z.:

Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. Nature **645**(8081), 633–638 (Sep 2025). https://doi.org/10.1038/s41586-025-09422-z, http://dx.doi.org/10.1038/s41586-025-09422-z 7, 9
9. Huang, J., Tan, Z., Gong, S., Zeng, F., Zhou, J.T., Miao, C., Tan, H., Yao, W., Li, J.: Lav-cot: Language-aware visual cot with multi-aspect reward optimization for real-world multilingual vqa. arXiv preprint arXiv:2509.10026 (2025) 1, 3, 4, 15
10. Huang, S., Dong, L., Wang, W., Hao, Y., Singhal, S., Ma, S., Lv, T., Cui, L., Mohammed, O.K., Patra, B., Liu, Q., Aggarwal, K., Chi, Z., Bjorck, J., Chaudhary, V., Som, S., Song, X., Wei, F.: Language is not all you need: Aligning perception with language models. In: Advances in Neural Information Processing Systems (NeurIPS) (2023) 15
11. Huang, Z., Chen, K., He, J., Bai, X., Karatzas, D., Lu, S., Jawahar, C.V.: Icdar2019 competition on scanned receipt ocr and information extraction. In: 2019 International Conference on Document Analysis and Recognition (ICDAR). IEEE (2019) 10
12. Hudson, D.A., Manning, C.D.: Gqa: A new dataset for real-world visual reasoning and compositional question answering. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019) 10
13. Jiang, Y., Gu, J., Xue, T., Cheung, K.C., Molchanov, P., Yin, H., Liu, S.: Token-efficient vlm: High-resolution image understanding via dynamic region proposal. In: Proc. IEEE International Conference on Computer Vision (ICCV). pp. 24147–24158 (October 2025) 3, 15
14. Jonnalagadda, A., Wang, W.Y., Manjunath, B.S., Eckstein, M.P.: Foveater: Foveated transformer for image classification. arXiv preprint arXiv:2105.14173 (2022) 15
15. Kazemzadeh, S., Ordonez, V., Matten, M., Berg, T.L.: Referit game: Referring to objects in photographs of natural scenes. In: Conference on Empirical Methods in Natural Language Processing (EMNLP) (2014) 10
16. Kong, Z., Dong, P., Ma, X., Meng, X., Sun, M., Niu, W., Shen, X., Yuan, G., Ren, B., Qin, M., Tang, H., Wang, Y.: Spvit: Enabling faster vision transformers via soft token pruning. In: Proc. European Conference on Computer Vision (ECCV) (2022) 14
17. Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Malloci, M., Kolesnikov, A., Duerig, T., Ferrari, V.: The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. In: International Journal of Computer Vision (IJCV) (2020) 10
18. Landeghem, J.V., Tito, R., Łukasz Borchmann, Pietruszka, M., Józiak, P., Powalski, R., Jurkiewicz, D., Coustaty, M., Ackaert, B., Valveny, E., Blaschko, M., Moens, S., Stanisławek, T.: Document understanding dataset and evaluation (dude). In: Proc. IEEE International Conference on Computer Vision (ICCV) (2023) 10
19. Lee, J., Shin, W., Yang, S., Song, K.U., Lim, D., Kim, J., Kim, T.H., Kim, B.K.: Ergo: Efficient high-resolution visual understanding for vision-language models. In: International Conference on Learning Representations (ICLR) (2026) 3, 15
20. Li, J., Li, D., Savarese, S., Hoi, S.: BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In: Proc. International Conference on Machine Learning (ICML) (2023) 1, 3, 15
21. Liang, Y., Ge, C., Tong, Z., Song, Y., Wang, J., Xie, P.: Not all patches are what you need: Expediting vision transformers via token reorganizations. In: International Conference on Learning Representations (ICLR) (2022) 14

22. Lin, W., Feng, Y., Zhu, Y.: <scp>metasapiens:</scp> real-time neural rendering with efficiency-aware pruning and accelerated foveated rendering. In: Proceedings of the 30th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 1. p. 669–682. ASPLOS '25, ACM (Mar 2025). https://doi.org/10.1145/3669940.3707227, http://dx.doi.org/10.1145/3669940.3707227 15
23. Lin, Z., Liu, C., Zhang, R., Gao, P., Qiu, L., Xiao, H., Qiu, H., Lin, C., Shao, W., Chen, K., Han, J., Huang, S., Zhang, Y., He, X., Li, H., Qiao, Y.: Sphinx: The joint mixing of weights, tasks, and visual embeddings for multi-modal large language models. arXiv preprint arXiv:2311.07575 (2023) 10, 11
24. Liu, F., Emerson, G., Collier, N.: Visual spatial reasoning. In: Proc. Annual Meeting of the Association for Computational Linguistics (ACL) (2023) 10
25. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2024) 1, 15
26. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: Advances in Neural Information Processing Systems (NeurIPS) (2023) 1, 3, 10, 11, 15
27. Liu, Y., Gehrig, M., Messikommer, N., Cannici, M., Scaramuzza, D.: Revisiting token pruning for object detection and instance segmentation. In: Proc. Winter Conference on Applications of Computer Vision (WACV) (2024) 14
28. Lu, P., Mishra, S., Xia, T., Qiu, L., Chang, K.W., Zhu, S.C., Tafjord, O., Clark, P., Kalyan, A.: Learn to explain: Multimodal reasoning via thought chains for science question answering. In: Advances in Neural Information Processing Systems (NeurIPS) (2022) 10
29. Lukanov, H., König, P., Pipa, G.: Biologically inspired deep learning model for efficient foveal-peripheral vision. Frontiers in Computational Neuroscience **Volume 15 - 2021** (2021). https://doi.org/10.3389/fncom.2021.746204, https://www.frontiersin.org/journals/computational-neuroscience/articles/10.3389/fncom.2021.746204 15
30. Mathew, M., Bagal, V., Tito, R.P., Karatzas, D., Valveny, E., Jawahar, C.V.: Infographicvqa. In: Proc. Winter Conference on Applications of Computer Vision (WACV) (2022) 10
31. Mathew, M., Karatzas, D., Jawahar, C.V.: Docvqa: A dataset for vqa on document images. In: Proc. Winter Conference on Applications of Computer Vision (WACV) (2021) 10, 12, 15
32. Min, J., Zhao, Y., Luo, C., Cho, M.: Peripheral vision transformer. In: Advances in Neural Information Processing Systems (NeurIPS) (2022) 15
33. Mnih, V., Heess, N., Graves, A., Kavukcuoglu, K.: Recurrent models of visual attention. In: Advances in Neural Information Processing Systems (NeurIPS) (2014) 14, 15
34. OpenAI: Chatgpt (2025), accessed: 2025-04-05 10
35. OpenBMB: MiniCPM-o. https://github.com/OpenBMB/MiniCPM-o (2024), accessed: 2024-03-05 11
36. Plummer, B.A., Wang, L., Cervantes, C.M., Caicedo, J.C., Hockenmaier, J., Lazebnik, S.: Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In: Proc. IEEE International Conference on Computer Vision (ICCV) (2015) 10
37. Qi, J., Ding, M., Wang, W., Bai, Y., Lv, Q., Hong, W., Xu, B., Hou, L., Li, J., Dong, Y., Tang, J.: Cogcom: Train large vision-language models diving into details through chain of manipulations. arXiv preprint arXiv:2402.04236 (2024) 1, 3, 4, 15

38. Qin, Y., Wei, B., Ge, J., Kallidromitis, K., Fu, S., Darrell, T., Wang, X.: Chain-of-visual-thought: Teaching vlms to see and think better with continuous visual tokens. arXiv preprint arXiv:2511.19418 (2025) 15
39. Rao, Y., Zhao, W., Liu, B., Lu, J., Zhou, J., Hsieh, C.J.: Dynamicvit: Efficient vision transformers with dynamic token sparsification. In: Advances in Neural Information Processing Systems (NeurIPS) (2021) 14
40. Rosenholtz, R., Huang, J., Raj, A., Balas, B.J., Ilie, L.: A summary statistic representation in peripheral vision explains visual search. Journal of Vision **12**(4), 14–14 (04 2012). https://doi.org/10.1167/12.4.14, https://doi.org/10.1167/12.4.14 15
41. Ryoo, M.S., Piergiovanni, A., Arnab, A., Dehghani, M., Angelova, A.: Tokenlearner: What can 8 learned tokens do for images and videos? In: Advances in Neural Information Processing Systems (NeurIPS) (2021) 14
42. Sarch, G., Saha, S., Khandelwal, N., Jain, A., Tarr, M.J., Kumar, A., Fragkiadaki, K.: Grounded reinforcement learning for visual reasoning. In: Advances in Neural Information Processing Systems (NeurIPS) (2025) 1, 3, 4, 15
43. Shao, H., Qian, S., Xiao, H., Song, G., Zong, Z., Wang, L., Liu, Y., Li, H.: Visual cot: Unleashing chain-of-thought reasoning in multi-modal language models. In: Advances in Neural Information Processing Systems (NeurIPS) (2024) 1, 2, 3, 4, 8, 10, 11, 14, 15
44. 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. arXiv preprint arXiv:2402.03300 (2024) 9
45. 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: Conference on Empirical Methods in Natural Language Processing (EMNLP) (2025) 3, 4, 15
46. Sidorov, O., Hu, R., Rohrbach, M., Singh, A.: Textcaps: a dataset for image captioning with reading comprehension. In: Proc. European Conference on Computer Vision (ECCV) (2020) 10
47. Singh, A., Natarajan, V., Shah, M., Jiang, Y., Chen, X., Batra, D., Parikh, D., Rohrbach, M.: Towards vqa models that can read. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019) 10, 12
48. Su, A., Wang, H., Ren, W., Lin, F., Chen, W.: Pixel reasoner: Incentivizing pixel-space reasoning with curiosity-driven reinforcement learning. In: Advances in Neural Information Processing Systems (NeurIPS) (2025) 1, 3, 4, 15
49. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: Caltech-ucsd birds 200. Tech. Rep. CNS-TR-2011-001, California Institute of Technology (2011) 10
50. Wang, W., Lv, Q., Yu, W., Hong, W., Qi, J., Wang, Y., Ji, J., Yang, Z., Zhao, L., Song, X., Xu, J., Xu, B., Li, J., Dong, Y., Ding, M., Tang, J.: Cogvlm: Visual expert for pretrained language models. In: Advances in Neural Information Processing Systems (NeurIPS) (2024) 15
51. Wu, P., Xie, S.: V*: Guided visual search as a core mechanism in multimodal llms. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2024) 1, 3, 4, 10, 11, 15
52. Xu, G., Jin, P., Wu, Z., Li, H., Song, Y., Sun, L., Yuan, L.: Llava-cot: Let vision language models reason step-by-step. In: Proc. IEEE International Conference on Computer Vision (ICCV) (2025) 15
53. Xu, Y., Zhang, Z., Zhang, M., Sheng, K., Li, K., Dong, W., Zhang, L., Xu, C., Sun, X.: Evo-vit: Slow-fast token evolution for dynamic vision transformer. In: Proc. AAAI Conference on Artificial Intelligence (AAAI) (2022) 14

54. Yang, S., Li, J., Lai, X., Yu, B., Zhao, H., Jia, J.: Visionthink: Smart and efficient vision language model via reinforcement learning. In: Advances in Neural Information Processing Systems (NeurIPS) (2025) 3, 4, 15
55. Yao, Y., Yu, T., Zhang, A., Wang, C., Cui, J., Zhu, H., Cai, T., Li, H., Zhao, W., He, Z., Chen, Q., Zhou, H., Zou, Z., Zhang, H., Hu, S., Zheng, Z., Zhou, J., Cai, J., Han, X., Zeng, G., Li, D., Liu, Z., Sun, M.: Minicpm-v: A gpt-4v level mllm on your phone. arXiv preprint arXiv:2408.01800 (2024) 11
56. Yarbus, A.L.: Eye movements and vision. Springer (2013) 2
57. Ye, J., Meng, X., Guo, D., Shang, C., Mao, H., Yang, X.: Neural foveated super-resolution for real-time vr rendering. Computer Animation and Virtual Worlds **35**(4), e2287 (2024). https://doi.org/https://doi.org/10.1002/cav.2287, https://onlinelibrary.wiley.com/doi/abs/10.1002/cav.2287 15
58. Yin, H., Vahdat, A., Alvarez, J., Mallya, A., Kautz, J., Molchanov, P.: Adavit: Adaptive tokens for efficient vision transformer. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2022) 14
59. Yu, R., Ma, X., Wang, X.: Auto-controlled image perception in mllms via visual perception tokens. In: Proc. IEEE International Conference on Computer Vision (ICCV) (2025) 2, 3, 4, 10, 11, 15
60. Yu, W., Yang, Z., Liu, Y., Bai, X.: Docthinker: Explainable multimodal large language models with rule-based reinforcement learning for document understanding. In: Proc. IEEE International Conference on Computer Vision (ICCV) (2025) 10, 11
61. Zhang, R., Zhang, B., Li, Y., Zhang, H., Sun, Z., Gan, Z., Yang, Y., Pang, R., Yang, Y.: Improve vision language model chain-of-thought reasoning. In: Proc. Annual Meeting of the Association for Computational Linguistics (ACL) (2024) 15
62. 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. arXiv preprint arXiv:2505.15436 (2025) 1, 3, 4, 15
63. Zhang, Z., Zhang, A., Li, M., Zhao, H., Karypis, G., Smola, A.: Multimodal chain-of-thought reasoning in language models. arXiv preprint arXiv:2302.00923 (2024) 15
64. Zhao, K., Zhu, B., Sun, Q., Zhang, H.: Unsupervised visual chain-of-thought reasoning via preference optimization. In: Proc. IEEE International Conference on Computer Vision (ICCV) (2025) 1, 2, 3, 4, 10, 11, 15
65. Zheng, Z., Yang, M., Hong, J., Zhao, C., Xu, G., Yang, L., Shen, C., Yu, X.: Deepeyes: Incentivizing "thinking with images" via reinforcement learning. In: International Conference on Learning Representations (ICLR) (2025) 1, 3, 4, 15
66. Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: Minigpt-4: Enhancing vision-language understanding with advanced large language models. In: International Conference on Learning Representations (ICLR) (2024) 15
67. Zhu, Y., Groth, O., Bernstein, M., Fei-Fei, L.: Visual7w: Grounded question answering in images. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016) 10, 12