REWARDING WHAT MATTERS: HUMAN-INSPIRED VISUAL SEARCH FOR FINE-GRAINED MLLM

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal large language models (MLLMs) have made substantial progress in visual understanding, yet they still struggle to resolve small but answer-critical details in high-resolution images. Exhaustively processing those images can preserve such details but incurs substantial visual-token cost, motivating selective visual inspection. In contrast, human visual search efficiently prioritizes information through complementary cognitive mechanisms: bottom-up processes that suppress irrelevant distractions and top-down cognitive control that directs attention toward task-relevant information. Inspired by these mechanisms, we introduce SAGE (Surprise-And-Goal-Guided Exploration), a training-free visual search method with a bottom-up surprise module and a top-down relevance module. The surprise module identifies more informative visual tokens relative to their surrounding context, while the relevance module selects regions that are most semantically relevant to the task query. Together, these modules enable efficient filtering of redundant visual content while preserving task-critical evidence. Experiments across four MLLM families demonstrate that our method consistently improves accuracy over ZoomEye, while requiring fewer visual tokens.

1 INTRODUCTION

Multimodal large language models (MLLMs) have made substantial progress in visual understanding (Liu et al., 2023; Dai et al., 2023), yet reliably resolving fine-grained details in high-resolution images remains challenging (Zhang et al., 2025a). One limitation arises from the spatial resolution at which images are encoded. For example, LLaVA-1.5 uses a CLIP vision encoder with 336×336 inputs (Liu et al., 2023); resizing a high-resolution image to this scale can obscure small, answer-critical details. More recent models mitigate this limitation through tiling-based high-resolution processing. LLaVA-NeXT employs AnyRes (Liu et al., 2024), while InternVL3 adopts a dynamic-resolution strategy using multiple image tiles (Zhu et al., 2025). These approaches combine a global image view with local tiles to preserve finer spatial information. However, such approaches increase computational complexity due to the expansion in visual tokens. However, processing additional tiles increases the visual token budget, while preserving more detail does not by itself determine which evidence requires closer inspection. Indeed, a model can attend to a relevant region without correctly resolving its details (Zhang et al., 2025a). These limitations motivate selective visual inspection: allocating additional computation to regions likely to contain answer-critical evidence.

The potential of selective inspection is illustrated by Zhang et al. (2025a), which show that cropping around ground-truth answer regions improves GPT-4o accuracy on small-target TextVQA examples. Guided visual search provides a framework for addressing this perceptual bottleneck. V* demonstrates the value of selective inspection for answering questions about fine details in high-resolution, cluttered images (Wu & Xie, 2024). Subsequent methods recover local evidence through recursive image decomposition (Wang et al., 2025b) or hierarchical crop exploration (Shen et al., 2025b). However, repeated regional processing can incur substantial computational cost. Other approaches prioritize regions using saliency-based cues (Ek et al., 2026; Gong et al., 2026) or internal model representations: ViCrop derives attention- and gradient-based importance maps (Zhang et al., 2025a), while FOCUS constructs object-relevance maps from cached visual and textual features (Zhong et al., 2025). Together, these approaches raise a central design question: how should a model prioritize a small set of regions for closer inspection?

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

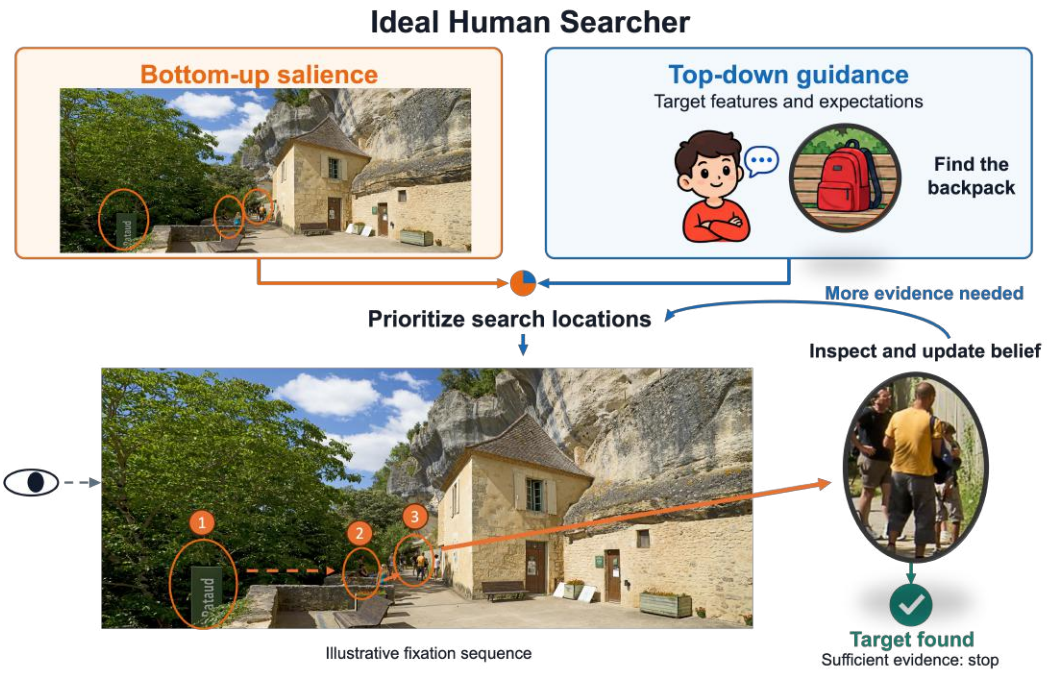


Figure 1: A conceptual account of ideal human visual search, in which bottom-up salience and top-down guidance direct inspection and subsequent belief updates (adapted from (Najemnik & Geisler, 2005)).

Human visual search offers a principle for this prioritization: combining bottom-up saliency with top-down guidance (Wolfe, 2010; 2020). Bottom-up signals highlight distinctive visual structure, while top-down guidance favors features associated with the sought target. These signals provide complementary information. A visually distinctive region may be irrelevant to the question, whereas semantically related context may appear relevant without containing the target itself. Their complementary roles motivate using both scene structure and task goals to guide region selection, as illustrated conceptually in Figure 1. Yet strong MLLMs remain weakly aligned with human search priorities (Lin et al., 2026). Specifically, the visual search paradigms overly rely on top-down guidance (e.g., Shen et al. (2025b); Zhang et al. (2025a); Zhong et al. (2025)). Some methods utilize bottom-up rely on bottom-up signals, but only for the token-reduction task (e.g., Ek et al. (2026); Gong et al. (2026)). We therefore ask: *Can bottom-up saliency and top-down guidance jointly guide effective visual inspection?*

To investigate this question, we introduce a training-free visual search framework that combines these two guidance signals. We jointly encode image tiles through a native-video representation and compute a scene-relative surprise signal from the departures of tile representations from their image-wide mean. This signal addresses *what is distinctive within the scene*. We complement it with a query-conditioned relevance signal derived from a target representation, addressing *what is relevant to the target*. Our key contribution is a training-free region-selection mechanism that couples bottom-up, scene-relative surprise computed from jointly encoded image tiles with top-down, query-conditioned relevance to generate a set of proposals for MLLM visual search.

Our contributions are threefold:

- We introduce a scene-relative measure of visual surprise for static images, computed across tiles jointly encoded through a video representation, and we develop a training-free search method that combines this bottom-up surprise signal with top-down relevance.
- We demonstrate consistent accuracy improvements over the base model and other training-free methods across four MLLM families while using fewer visual tokens.

- We constructed a new multiview benchmark and demonstrated that our method can be extended to multiview images.

2 Related Work

2.1 Fine-grained VQA Tasks

Fine-grained visual question answering (VQA) requires resolving local visual details that may be difficult to discern from a global image representation. Target objects often occupy only a small fraction of the image, making accurate localization and detailed inspection critical. For example, in V*Bench, approximately 69% of target objects occupy less than 0.1% of the image area Zhong et al. (2025). In addition to V*Bench, which contains images at approximately 2K resolution, we evaluate on HR-Bench-4K and HR-Bench-8K (Wang et al., 2025b) to assess performance at higher image resolutions.

Visual cropping is a common strategy for these tasks, enabling models to inspect selected regions at greater spatial detail. We organize cropping-based approaches according to their primary guidance mechanism: bottom-up methods prioritize regions based on image-driven signals, whereas top-down methods use question- or target-conditioned information to guide region selection.

2.2 Bottom-up visual guidance

Bottom-up visual guidance prioritizes regions or tokens using image-derived signals that are independent of the current query. PS3 learns saliency-based selection of high-resolution patches and also supports a separate mode conditioned on textual guidance (Shi et al., 2025). At the token level, LLaVA-PruMerge uses attention between the visual encoder's class token and spatial tokens to identify informative tokens, then aggregates information through token merging (Shang et al., 2025). These methods demonstrate the use of image-driven importance signals to allocate visual computation.

In explicit visual search, DeepScan adopts a bottom-up grounding procedure that recovers image-level evidence from question-conditioned local cues through hierarchical scanning and point-prompted segmentation (Li et al., 2026). Here, bottom-up refers to local-to-global evidence construction, rather than query-independent saliency. In video understanding, SURGE measures temporal surprise through token prediction errors and prioritizes unpredictable content for token retention; its selection can be further refined using query relevance (Ek et al., 2026).

2.3 Top-down visual search

Top-down visual search uses the question or target specification to guide the acquisition of visual evidence. Existing approaches differ in how this guidance is implemented: training-free methods construct search procedures from pretrained model outputs, whereas post-training methods learn additional localization capabilities or visual interaction policies.

Training-free methods. Training-free approaches recover question-relevant evidence without updating model parameters. One strategy organizes regional observations into a searchable representation. DC^2 recursively decomposes images, generates regional descriptions, and constructs a visual memory from which question-relevant patches are retrieved (Wang et al., 2025b). ZoomEye instead explores a hierarchy of image regions using target-conditioned confidence to prioritize inspection and answering confidence to determine when to stop (Shen et al., 2025b). Both approaches acquire detailed evidence through repeated regional processing, with computational cost depending on the extent of exploration. A second strategy derives crop proposals directly from internal model representations. ViCrop converts question-conditioned attention or gradient signals into spatial importance maps, exploiting the observation that models can localize relevant concepts internally despite answering incorrectly (Zhang et al., 2025a). FOCUS identifies target objects from the question and constructs object-relevance maps from cached visual and textual representations, which guide proposal generation and ranking (Zhong et al., 2025). These methods provide direct mechanisms for translating semantic relevance into regional inspection.

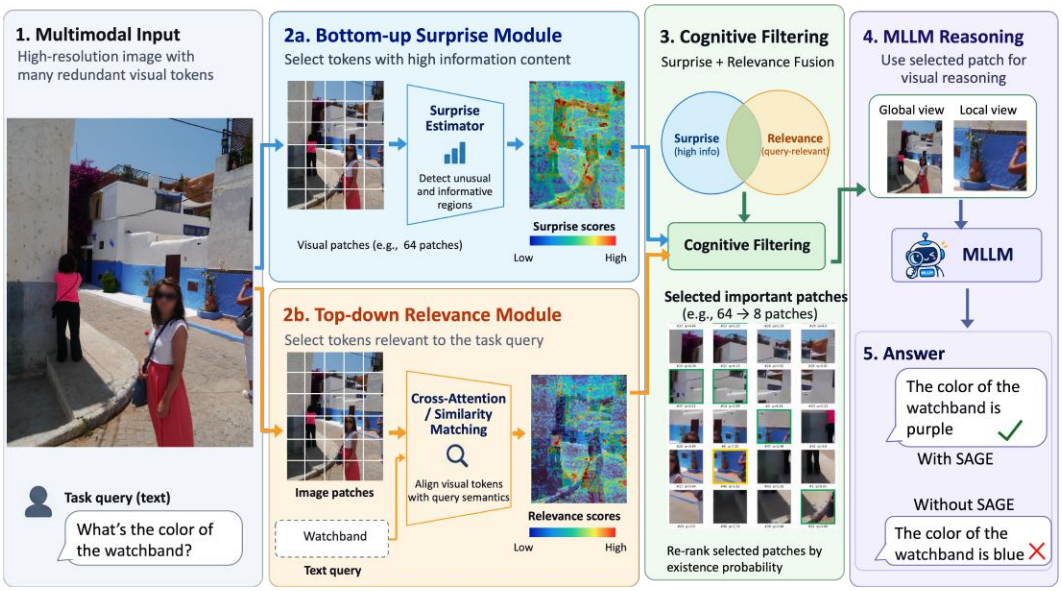


Figure 2: The executed propose–then–verify pipeline. A single native-video forward yields scene-relative surprise and target relevance over all tiles. Their product forms the routing heatmap; greedy peaks yield candidate windows. Existence probability alone selects the local crop supplied with the global image to the MLLM.

Post-training methods. Post-training approaches develop visual search capabilities through additional supervision or reinforcement learning. V* and its SEAL architecture combine a trained localization model with an instruction-tuned VQA model, using target-specific and contextual cues to guide search and a visual working memory to retain retrieved evidence (Wu & Xie, 2024). More recent methods learn to interleave visual actions with reasoning. DeepEyes uses reinforcement learning without a cold-start supervised fine-tuning stage to learn coordinate-based cropping and zooming, incorporating the resulting views into subsequent reasoning (Zheng et al., 2026). Thyme-VL combines supervised fine-tuning with reinforcement learning to generate executable code for image manipulation and computation, enabling question-directed visual inspection within a broader tool-use framework (Zhang et al., 2025b).

3 Method

3.1 Problem Formulation

Let I be a high-resolution image, q a question, and c a target concept needed to answer q. Given a frozen MLLM f_θ, the objective is to select a local crop B_c without bounding-box supervision and answer using the full image together with $I[B_c]$. Figure 2 summarizes the executed pipeline: jointly encode all overlapping tiles, form one target-aware routing heatmap, extract heatmap peaks, verify target existence in a fixed number of windows, and answer from global and local evidence.

3.2 Image as Video Sequence

To expose fine-grained visual details throughout the image, we cover I with overlapping square tiles $\{X_t\}_{t=1}^{F}$. Each tile is extracted from a source region b_t with side length $s = \frac{\max(H,W)}{n}$, where n is set to 8, and the overlapping ratio is set to 25%. Processing these local views makes small structures more prominent at the model's input resolution, while overlap reduces the risk of truncating a target at tile boundaries.

We arrange the tiles in a fixed spatial order and treat each tile as a frame in the MLLM's native video interface, $\mathcal{V}(I) = \big(\mathcal{P}(X_1), \ldots, \mathcal{P}(X_F)\big)$, where $\mathcal{P}$ denotes the backbone's frame preprocessing.

Here, the frame index enumerates spatial views of a single image. We retain the source region b_t of each tile to map the resulting spatial scores back to the original image. The video interface packs these views into a single input sequence and, for backbones with video-specific spatial pooling, reduces their visual token budget. For example, LLaVA-OneVision represents each video frame with 196 visual tokens, compared with 729 tokens per image in its multi-image configuration (Li et al., 2024). This provides a more compact input to the language model for the same collection of tiles.

3.3 Joint internal representations

For target c, we process the tile sequence with the prompt "Locate the object in the image: c". Let $\boldsymbol{h}_{t,j}^{(\ell)}$ be the hidden state of visual token j in tile t at layer ℓ. We average layers $\mathcal{L} = \{1, \ldots, n\}, \boldsymbol{h}_{t,j} = \frac{1}{|\mathcal{L}|} \sum_{\ell \in \mathcal{L}} \boldsymbol{h}_{t,j}^{(\ell)}$, where $j \in \{1, \ldots, T\}$ indexes the per-tile visual-token grid. Hidden states at the target phrase are averaged over the same layers to yield a contextual target feature $\boldsymbol{e}_c$. Thus, all F tile grids and the target feature are obtained from one video forward pass.

3.4 Bottom-up cross-tile surprise

Instead of SURGE, which measures temporal surprise in video frames (Ek et al., 2026), we formulate a spatial surprise signal for region search within a single image. We therefore replace prediction from temporal history with comparison against an image-specific reference constructed across tiles. This formulation measures how strongly a local representation differs from the visual statistics of the current scene.

Let $\boldsymbol{h}_{t,j} \in \mathbb{R}^d$ denote the target-independent visual feature at position j in tile t, with F tiles and T tokens per tile. For each within-tile position, we construct a reference by averaging the corresponding features across all tiles:

$$\boldsymbol{\mu}_j = \frac{1}{F} \sum_{t=1}^{F} \boldsymbol{h}_{t,j}. \tag{1}$$

We define spatial surprise as the squared deviation from this reference, normalized by the mean residual energy across the image:

$$S_{t,j} = \frac{\|\boldsymbol{h}_{t,j} - \boldsymbol{\mu}_j\|_2^2}{\frac{1}{FT} \sum_{t'=1}^{F} \sum_{j'=1}^{T} \|\boldsymbol{h}_{t',j'} - \boldsymbol{\mu}_{j'}\|_2^2 + \epsilon}, \tag{2}$$

where $\epsilon > 0$ ensures numerical stability.

The reference $\boldsymbol{\mu}_j$ is estimated from the current image in feature space. Here, j identifies a relative position within the tile token grid; it does not imply correspondence between objects or absolute image coordinates. Both the reference and normalization aggregate the supplied tile features symmetrically. The normalization controls the overall score magnitude while preserving the ordering of token-level deviations. This scene-relative signal provides bottom-up guidance for allocating the region-proposal budget. A high score indicates visual distinctiveness which requires more attention.

3.5 Top-down target relevance

Internal visual–text similarity provides an established localization signal, as demonstrated by FOCUS (Zhong et al., 2025). Unlike FOCUS's direct visual–text similarity, we define target relevance as the deviation from the mean similarity at the same within-tile position across the image's tiles, emphasizing target affinity relative to the current scene. Specifically, we measure how much a token's target similarity exceeds the average similarity at the same within-tile position across all tiles. This introduces an image-specific reference into target-guided localization.

Let $\boldsymbol{e}_c$ denote the representation of target concept c. We first compute visual–target cosine similarity: $C_{t,j} = \frac{\boldsymbol{h}_{t,j}^\top \boldsymbol{e}_c}{\|\boldsymbol{h}_{t,j}\|_2 \|\boldsymbol{e}_c\|_2}$. We then center the scores across tiles at each token position and retain positive

residuals:

$$R_{t,j} = \left[C_{t,j} - \frac{1}{F} \sum_{u=1}^{F} C_{u,j} \right]_+ . \tag{3}$$

This operation subtracts the similarity component shared across tiles and emphasizes locations with above-reference target affinity.

The construction has a direct geometric interpretation. Writing $\tilde{\boldsymbol{h}}_{t,j}$ and $\tilde{\boldsymbol{e}}_c$ for the unit-normalized visual and target features, respectively, Eq. 3 is equivalent to

$$R_{t,j} = \left[\left(\tilde{\boldsymbol{h}}_{t,j} - \frac{1}{F} \sum_{u=1}^{F} \tilde{\boldsymbol{h}}_{u,j} \right)^{\top} \tilde{\boldsymbol{e}}_c \right]_+ . \tag{4}$$

Thus, relevance measures the positive projection of a cross-tile feature deviation onto the target direction.

This formulation connects top-down relevance with bottom-up surprise through a common principle of scene-relative comparison. Surprise measures the magnitude of deviations in visual feature space, while relevance measures target-aligned deviations after feature normalization. Both signals are constructed from the same tile population and jointly guide the region-proposal budget. Their fusion allows proposal generation to incorporate both visual distinctiveness and target-specific evidence.

3.6 Heatmap construction and region proposals

For each target concept, we reshape the fused token scores of each tile into its spatial token grid and bilinearly interpolate the grid to the tile's size in the original image. We place the resulting score map at the tile's source coordinates and average the contributions at pixels covered by multiple tiles. Candidate windows have the same side length as the initial image tiles, and are shifted when necessary to remain within the image boundaries. We iteratively select the highest remaining heatmap peak and construct a window centered on it. A candidate is retained only if its intersection-over-union with every previously retained window is below $\tau_{\mathrm{IoU}} = 0.35$. After each attempt, we suppress the heatmap scores within the candidate window, regardless of whether it is retained, to avoid repeatedly selecting the same area. We retain up to 8 candidate crops for existence-based reranking. The frozen MLLM evaluates each candidate crop $I[w]$ using the prompt "Is there a c in this image? Answer yes or no."Let ℓ_w^{yes} and ℓ_w^{no} denote the next-token logits aggregated by log-sum-exp over the accepted token variants of "yes" and "no", respectively. We compute

$$p_w^{\mathrm{exist}} = \frac{\exp(\ell_w^{\mathrm{yes}})}{\exp(\ell_w^{\mathrm{yes}}) + \exp(\ell_w^{\mathrm{no}})}, \qquad w_c^{\star} = \arg\max_{w \in \mathcal{W}_c} p_w^{\mathrm{exist}}, \tag{5}$$

where $\mathcal{W}_c$ is the candidate set for target c. The selected region is converted into the final crop B_c. For questions involving multiple concepts, we apply the localization procedure separately to each target. The selected regions are marked on the global image, and a single zoomed view is constructed to cover them. We denote this pair of visual inputs by $\mathcal{U} = \mathrm{ComposeViews}(I, \{B_c\}_{c \in \mathcal{C}})$.

We score each candidate answer using its mean token-level negative log-likelihood. For a scored answer continuation $y = (y_1, \ldots, y_{L_y})$, the prediction is

$$\hat{y} = \arg\min_{y \in \mathcal{Y}} \left[-\frac{1}{L_y} \sum_{i=1}^{L_y} \log p_\theta \left(y_i \mid y_{<i}, q, \mathcal{U} \right) \right]. \tag{6}$$

4 Experiments

4.1 Main Benchmark

We evaluate SAGE on V*Bench (Wu & Xie, 2024) and the 4K and 8K variants of HR-Bench (Wang et al., 2025b). V*Bench targets fine-grained attribute recognition and spatial reasoning in visually crowded images at approximately 2K resolution, whereas HR-Bench evaluates fine-grained single-instance perception (FSP) and cross-instance perception (FCP) using 4K crops and full 8K images. Table 1 reports multiple-choice accuracy across four open-source backbones.

Table 1: Main results on fine-grained high-resolution benchmarks. Attribute/Spatial denote the V* splits; FSP and FCP denote fine-grained single- and cross-instance perception. All entries are accuracy (%). The best result within each open-source backbone block is bold.

Method	V* Bench			HR-Bench 4K			HR-Bench 8K		
	Attribute	Spatial	Overall	FSP	FCP	Overall	FSP	FCP	Overall
Closed-source MLLMs									
GPT-4o (Achiam et al., 2023)	72.2	60.5	67.5	66.8	63.3	65.0	60.8	58.5	59.6
GPT-5.6-Luna	81.7	81.6	81.7	82.5	62.3	72.4	68.0	53.8	60.9
Post-training methods									
DeepEye-7B (Zheng et al., 2026)	91.3	88.2	90.1	91.3	59.0	75.1	86.8	58.5	72.6
Thyme-7B (Zhang et al., 2025b)	83.5	80.3	82.2	91.0	63.0	75.1	86.5	57.5	72.0
LLaVA-OV-7B variants									
LLaVA-OV-7B (Li et al., 2024)	75.7	75.0	75.4	72.0	54.0	63.0	67.3	52.3	59.8
ZoomEye (Shen et al., 2025b)	93.9	84.2	90.1	84.5	54.3	69.4	**86.8**	51.3	69.0
FOCUS (Zhong et al., 2025)	**93.9**	89.5	92.2	85.8	55.5	70.6	82.5	55.3	68.9
VisLens He et al. (2026)	87.0	73.7	81.7	82.0	56.8	69.4	77.8	**59.5**	68.6
SAGE (Ours)	**93.9**	**96.1**	**94.8**	**87.8**	**59.0**	**73.4**	85.5	55.0	**70.3**
Qwen2.5-VL-3B variants									
Qwen2.5-VL-3B (Bai et al., 2025)	80.9	71.1	77.0	82.8	49.0	65.9	80.5	45.3	62.9
ZoomEye (Shen et al., 2025b)	**88.7**	89.5	89.0	86.8	53.5	70.1	84.8	**52.0**	68.4
SAGE (Ours)	87.8	**92.1**	**89.5**	**87.5**	**54.8**	**71.1**	**86.8**	51.3	**69.0**
Qwen3-VL-8B variants									
Qwen3-VL-8B (Qwen Team, 2025)	89.6	86.8	88.5	91.3	66.0	78.6	84.5	64.0	74.3
ZoomEye (Shen et al., 2025b)	91.3	86.8	89.5	**92.5**	65.3	78.9	**90.8**	60.5	75.6
SAGE (Ours)	**93.9**	**88.2**	**91.6**	89.5	**71.0**	**80.3**	88.8	**66.0**	**77.4**
InternVL3.5-2B variants									
InternVL3.5-2B (Wang et al., 2025a)	67.0	63.2	65.5	76.5	54.3	65.4	55.0	**52.0**	53.5
ZoomEye (Shen et al., 2025b)	**80.9**	61.8	73.3	85.0	53.5	69.3	81.5	51.0	66.3
SAGE (Ours)	80.0	**84.2**	**81.7**	**85.8**	**55.3**	**70.5**	**84.0**	50.5	**67.3**

With LLaVA-OV-7B, SAGE achieves overall accuracies of 94.8%, 73.4%, and 70.3% on V*Bench, HR-Bench 4K, and HR-Bench 8K, respectively. These results improve the unmodified backbone by 19.4, 10.4, and 10.5 percentage points (pp), and outperform the strongest comparison method in the corresponding block by 2.6, 2.8, and 1.3 pp. The improvements transfer consistently across the tested backbones. SAGE improves over the corresponding unmodified backbone in all 12 backbone–benchmark combinations and obtains the best overall accuracy in every open-source backbone block. Our method can also adapt to different image resolutions from 2K to 8K.

For completeness, we also report DeepEyes-7B (Zheng et al., 2026) and Thyme-7B (Zhang et al., 2025b), which employ additional model post-training. DeepEyes is optimized end-to-end with reinforcement learning, while Thyme uses supervised fine-tuning followed by reinforcement learning; their results are therefore not directly comparable to SAGE's training-free setting.

4.2 Qualitative Results

Figure 3 presents three qualitative examples that illustrate SAGE's crop selection and its internal representations. In each example, the target object occupies only a small fraction of a cluttered full image, a setting in which fine-grained recognition can be difficult when the full image is passed directly to the MLLM. The figure compares the crops selected by ZoomEye and SAGE for a backpack, car, and scarf. We further visualize the combined surprise-and-relevance heatmap for LLaVA-OneVision-7B and derive the corresponding target object, delineated by a bounding box. The bounding box encloses the connected region around the heatmap global maxima that exceeds an activation threshold of 0.2. Across all three examples, the heatmap concentrates on target-relevant regions, yielding crops that more effectively isolate the target object from surrounding visual clutter.

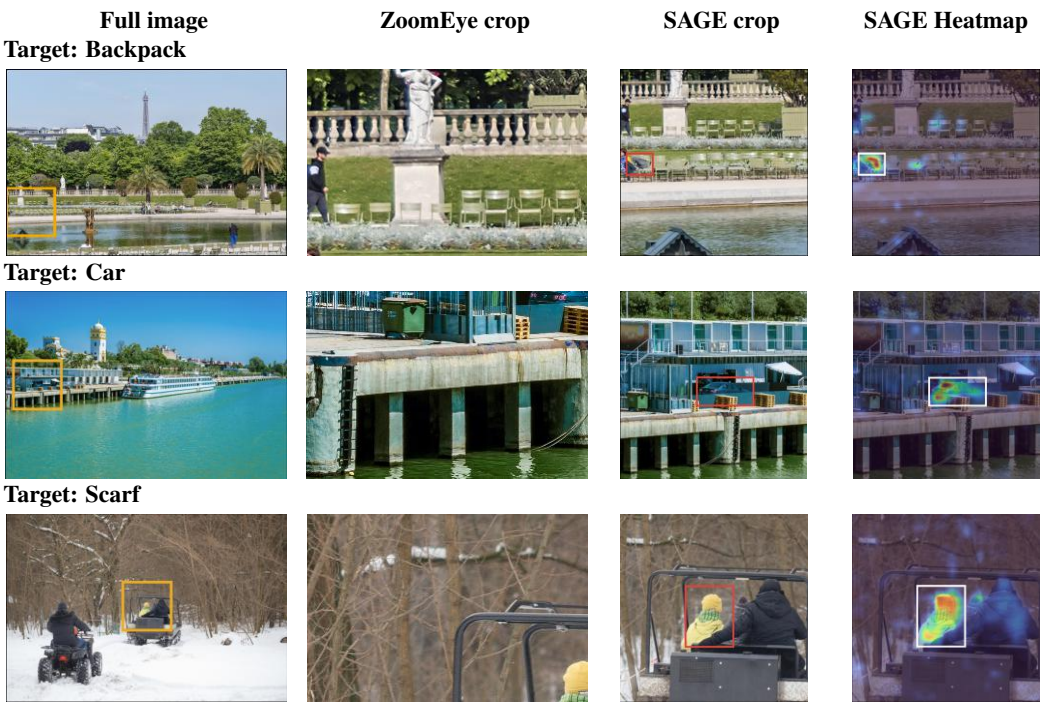


Figure 3: Qualitative examples of the proposal stage for three small target concepts. Each row shows the target name above four views: the full image; note that the orange box shows the selected search region by SAGE and is not part of the dataset; the ZoomEye crop, the SAGE crop, and the SAGE heatmap with the heatmap-selected region (white). The examples illustrate localization under changes in target size, scene clutter, and image composition.

Table 2: Efficiency on the V*Bench with LLaVA-OV-7B. Runtime is mean end-to-end latency per question; visual tokens are accumulated across sequential localization forwards.

Method	Acc. (%)	time(s/q)↓	throughput(q/h)↑	VRAM (GB)↓	Visual tokens↓	Recall(%)
ZoomEye	90.10	56.72	63.5	19.39	234.0k	75.65
SAGE (Ours)	**94.76**	**31.89**	**112.9**	**19.35**	**68.5k**	**95.92**
Δ vs. ZoomEye	+4.66 pp	−43.8%	1.78×	−0.2%	−70.7%	+20.27

4.3 Efficiency and localization accuracy

We evaluate both the computational efficiency and localization quality of our method on the complete V*Bench. Efficiency is measured using LLaVA-OneVision-7B on a single NVIDIA A40 GPU. As shown in Table 2, our surprise–attention routing reduces the average latency from 56.72 to 31.89 seconds per sample, corresponding to a 1.78× increase in throughput. It also reduces the mean localization cost from 234.0k to 68.5k visual tokens, a reduction of 70.7%, while requiring comparable peak GPU memory. This is because ZoomEye iteratively expands the search tree and re-encodes the image or candidate crops at multiple steps.

We measure localization using proposal recall based on ground-truth coverage. A ground-truth object is successfully localized if at least one proposed region covers at least 50% of its area. Our method attains 95.92% proposal recall over 245 annotated objects. In contrast, the recall of ZoomEye is only 75.65%, which means that despite the high accuracy, ZoomEye answers the question based on the wrong crop (e.g., the backpack example in Fig 3). ZoomEye's lower proposal recall despite relatively high VQA accuracy indicates that answer accuracy alone does not establish successful localization.

Table 3: Ablation of bottom-up surprise and top-down target relevance module using LLaVA-OV-7B.

Condition	V* Acc.	HR4K Acc.
Bottom-up only	85.86	66.38
Top-down only	92.67	70.75
SAGE (ours)	**94.76**	**73.38**
LLaVA-OV-7B vanilla	75.39	63.00
Random guess	35.99	25.00

Table 4: Qwen3-VL-8B results on the five-region multiview location benchmark. SAGE can localize small transient objects in multi-view images better.

	Multiview benchmark	
Method	**VQA**	**Loc.**
Vanilla	46.43	58.93
ZoomEye	48.21	46.43
SAGE(Ours)	**51.79**	**64.29**

4.4 Ablation Study

We conduct an ablation study to verify the contribution of the bottom-up surprise and top-down relevance modules in region proposal in Table 3. Both bottom-up and top-down modules can increase the fine-grained VQA accuracy, which establishes a strong foundation. Our proposed SAGE method can leverage the advantages of both approaches and contributes to the strongest performance.

4.5 Additional Multi-View Transient-Object Benchmark

Our main benchmarks study visual search within a single image. To test whether selective search extends to multi-image reasoning, we introduce a paired-view benchmark built from Lu et al. (2026). Each example contains two captures of the same scene and asks where a transient object appears in one of the views. A transient object is defined as an object that changes position, pose, state, or presence between captures. We annotated a 56-scene subset with bounding boxes of the transient object location. We evaluate vanilla Qwen3-VL-8B, ZoomEye, and our method with the same backbone. Our method achieves 51.79% VQA accuracy, compared with 48.21% for ZoomEye and 46.43% for vanilla. More details are shown in Section A.5.

5 Limitations

The current implementation assumes that a target phrase can be extracted from the benchmark annotation or question. Inaccurate target phrase extraction will affect the top-down relevance module. Moreover, as shared by all the training-free methods, the upper bound of our method depends on the base model. We noticed that although our method selects a correct crop, the final VQA is still wrong, as demonstrated in our failure analysis in Section A.1.

6 Conclusion

We presented SAGE, a training-free method for visual search in high-resolution images. SAGE combines scene-relative visual surprise with target-conditioned relevance to identify regions that are both distinctive within the image and relevant to the question. Across the evaluated backbones and high-resolution VQA benchmarks, SAGE improves overall accuracy over both the unmodified models and ZoomEye. Our efficiency evaluation also demonstrates lower inference latency. The additional grounding and multiview experiments suggest that the selected visual evidence can benefit tasks beyond single-image question answering. In additional, our results also highlight the importance of evaluating where a model looks alongside what it answers. Improving fine-grained visual understanding therefore requires progress in both evidence selection and reasoning over that evidence, and SAGE provides a practical way to improve the former without updating model weights.

Reproducibility Statement

We provide a detailed description of our method and implementation details. The supplementary material includes executable launch scripts, saved per-example predictions, evaluation outputs, and

visualization code. No training data are required. All code and datasets mentioned in this paper will be made public.

AI USE STATEMENT

Generative AI tools were used to assist with language editing, retrieval, and discovery (e.g., finding related work). All AI-assisted text and related works were reviewed by the authors. The authors take responsibility for the final content, claims, and artifacts.

REFERENCES

Josh Achiam et al. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Shuai Bai et al. Qwen2.5-VL technical report. *arXiv preprint*, 2025.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in neural information processing systems*, 36:49250–49267, 2023.

Sannara Ek, Chong Tang, Robert Mullins, Dirk Koch, Alex Weddell, and Jagmohan Chauhan. Surge: Surprise-guided token reduction for efficient video understanding with vlms. In *International Conference on Learning Representations*, 2026.

Shizhan Gong, Minda Hu, Qiyuan Zhang, Chen Ma, and Qi Dou. Saliency-r1: Enforcing interpretable and faithful vision-language reasoning via saliency-map alignment reward. *arXiv preprint arXiv:2604.04500*, 2026.

Jingyi He, Sanghwan Kim, and Zeynep Akata. Vislens: Single-pass interpretable visual search for multimodal llms. *arXiv preprint arXiv:2608.30705*, 2026.

Bo Li et al. LLaVA-OneVision: Easy visual task transfer. *Transactions on Machine Learning Research*, 2024.

Yangfu Li, Hongjian Zhan, Jiawei Chen, Yuning Gong, Qi Liu, and Yue Lu. Deepscan: A training-free framework for visually grounded reasoning in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19155–19164, 2026.

Jiaying Lin, Shuquan Ye, Dan Xu, Wanli Ouyang, and Rynson W. H. Lau. Do mllms exhibit human-like perceptual behaviors? hvsbench: A benchmark for mllm alignment with human perceptual behavior. In *CVPR Findings*, 2026.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024. URL https://llava-vl.github.io/blog/2024-01-30-llava-next/.

Cheng-You Lu, Yi-Shan Hung, Wei-Ling Chi, Hao-Ping Wang, Charlie Li-Ting Tsai, Yu-Cheng Chang, Yu-Lun Liu, Thomas Do, and Chin-Teng Lin. Df3dv-1k: A large-scale dataset and benchmark for distractor-free novel view synthesis. *arXiv preprint arXiv:2604.13416*, 2026.

Jiri Najemnik and Wilson S. Geisler. Optimal eye movement strategies in visual search. *Nature*, 434:387–391, 2005.

Qwen Team. Qwen3-VL technical report. *arXiv preprint*, 2025.

Yuzhang Shang, Mu Cai, Bingxin Xu, Yong Jae Lee, and Yan Yan. Llava-prumerge: Adaptive token reduction for efficient large multimodal models. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22857–22867. IEEE, 2025.

Haozhan Shen, Peng Liu, Jingcheng Li, Chunxin Fang, Yibo Ma, Jiajia Liao, Qiaoli Shen, Zilun Zhang, Kangjia Zhao, Qianqian Zhang, et al. Vlm-r1: A stable and generalizable r1-style large vision-language model. *arXiv preprint arXiv:2504.07615*, 2025a.

Haozhan Shen, Kangjia Zhao, Tiancheng Zhao, Ruochen Xu, Zilun Zhang, Mingwei Zhu, and Jianwei Yin. Zoomeye: Enhancing multimodal llms with human-like zooming capabilities through tree-based image exploration. In *Conference on Empirical Methods in Natural Language Processing*, 2025b.

Baifeng Shi, Boyi Li, Han Cai, Yao Lu, Sifei Liu, Marco Pavone, Jan Kautz, Song Han, Trevor Darrell, Pavlo Molchanov, et al. Scaling vision pre-training to 4k resolution. *arXiv preprint arXiv:2503.19903*, 2025.

Weiyun Wang et al. InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint*, 2025a.

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

Jeremy M Wolfe. Visual search. *Current biology*, 20(8):R346–R349, 2010.

Jeremy M. Wolfe. Visual search: How do we find what we are looking for? *Annual Review of Vision Science*, 6:539–562, 2020.

Penghao Wu and Saining Xie. V*: Guided visual search as a core mechanism in multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Jiarui Zhang, Mahyar Khayatkhoei, Prateek Chhikara, and Filip Ilievski. Mllms know where to look: Training-free perception of small visual details with multimodal llms. In *International Conference on Learning Representations*, 2025a.

Yi-Fan Zhang, Xingyu Lu, Shukang Yin, Chaoyou Fu, Wei Chen, Xiao Hu, Bin Wen, Kaiyu Jiang, Changyi Liu, Tianke Zhang, et al. Thyme: Think beyond images. *arXiv preprint arXiv:2508.11630*, 2025b.

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

Liangyu Zhong, Fabio Rosenthal, Joachim Sicking, Fabian Hüger, Thorsten Bagdonat, Hanno Gottschalk, and Leo Schwinn. Focus: Internal mllm representations for efficient fine-grained visual question answering. In *Advances in Neural Information Processing Systems*, 2025.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

A Appendix

A.1 Failure analysis

Our LLaVA-OneVision-7B configuration makes ten errors on V*Bench. We manually inspect all ten cases and assign each error to one of three mutually exclusive categories, summarized in Table 5. Only three errors (30%) are caused by an incorrectly localized crop. In six cases (60%), the selected crop contains the target, but the MLLM still predicts the wrong answer. The remaining case (10%) reflects an ambiguity in the benchmark example rather than a clear failure of either localization or answering.

Table 5: Manual categorization of the ten errors made by our main configuration on V*Bench.

Failure mode	Count	Share
Correct crop, wrong answer	6	60%
Wrong crop	3	30%
Dataset issue	1	10%
Total	10	100%

Figure 4 shows representative examples. In Figures 4(a)–(b), the target is preserved at high resolution, yet the model confuses a fine-grained color attribute and a human pose. These examples isolate the remaining bottleneck to visual recognition or answering after successful search. Figure 4(c) contains an ambiguous handbag attribute that is not reliably visible even in the local view, illustrating a limitation of the benchmark supervision. The true search failure in Figure 4(d) selects a region that does not contain the telephone. Overall, the error distribution indicates that crop localization accounts for a minority of the remaining mistakes; further gains will also require stronger fine-grained recognition and reasoning within a correctly localized crop.

A.2 Referring Expression Comprehension Task

Referring Expression Comprehension (REC) is a task that requires outputting the bounding box coordinates of the region given a sentence description. MLLMs with reinforcement learning fine-tuning (e.g., VLM-R1) have emerged and demonstrated promising performance on datasets such as RefCOCO Shen et al. (2025a). Despite the success in large object detection tasks, VLM-R1 still suffers from low mIOU when the object is small, such as in the V* benchmark, as shown in Table 6. With SAGE, a local crop containing the target object can be supplied to the VLM-R1, which can increase the mIOU from 0.50 to 0.68. This is further demonstrated by some qualitative examples in Fig. 5, where the VLM-R1 base model made mistakes in glove, helmet, and suitcase grounding tasks; while with SAGE, all those mistakes can be corrected.

Table 6: V*Bench accuracy and visual-grounding results for `Qwen2.5VL-3B-VLM-R1`. Higher is better for every metric. "ZoomEye + VLM-R1" uses LLaVA-OneVision-7B to search for a crop, then applies VLM-R1-3B for box prediction. Object mIoU is the mean IoU over 245 annotated objects. Box precision is the intersection divided by predicted-box area; target coverage is the intersection divided by ground-truth-box area.

Metric	VLM-R1 basemodel	ZoomEye Crop + VLM-R1	SAGE Crop VLM-R1
MCQ accuracy (all 191 samples)	62.30%	90.58%	**94.8%**
Direct-attribute MCQ accuracy	69.57%	93.04%	**96.1%**
Relative-position MCQ accuracy	51.32%	86.84%	**93.9%**
Object mIoU	0.5010	0.3570	**0.6825**
Box precision	0.6046	0.4944	**0.8555**
Target coverage	0.6121	0.4261	**0.7371**
Grounding accuracy at IoU ≥ 0.50	56.54%	39.59%	**80.00%**

(a) Correct crop, wrong answer — Apple-logo color



(b) Correct crop, wrong answer — human pose



(c) Dataset issue — handbag color



(d) Wrong crop — telephone color



Figure 4: Representative failure cases from V*Bench. For each case, the left image shows the selected crop in red and the annotated target in green; the right image shows the selected local crop supplied to the MLLM. Cases (a) and (b) localize the target correctly but fail during downstream recognition or answering. Case (c) has ambiguous benchmark evidence. Case (d) is a genuine localization failure in which the selected crop misses the annotated telephone.

A.3 Multi-View Transient-Object Localization

Multiview benchmark. Our main benchmarks evaluate visual search within one image. We additionally test whether the method can use evidence across images by constructing a new multiview



Figure 5: Bounding box detection using the VLM-R1-3B model. The green box is the GT; the red box is the model prediction. With SAGE, a strong prior is set for the downstream visual grounding task, leading to a higher mIOU in small object detection.

benchmark from Lu et al. (2026). Unlike other fine-grained VQA benchmarks, this benchmark includes two images per question, and the model must reason across them to identify a transient object. A transient object is defined as an object that changes its physical position, pose, state, or presence between captures; differences caused solely by viewpoint, parallax, lighting, or framing are excluded. The two views are not assumed to be pixel-aligned. The ground truth bounding boxes of the transient objects are manually labelled, and 56 pairs are labelled in total. We map the center of its ground-truth box to one of five labels: the middle half of both image dimensions forms the central region, while locations outside it are assigned to one of four quadrants. All methods use Qwen3-VL-8B, the strongest backbone for cross-reasoning tasks in our main benchmark. Using the same five-region criterion, vanilla direct-box prediction, ZoomEye, and our method correctly localize 58.93%, 46.43%, and 64.29% of examples, respectively. Final question-answering accuracies are 46.43%, 48.21%, and 51.79%. Overall, our method has the strongest performance in this multiview reasoning task.



Figure 6: Representative images of our proposed multiview benchmark. Each set contains two images from two different views. The MLLM needs to reason between two images to identify a small transient object that is not present in the other view (e.g., a bird). The green box shows the GT annotation, and the red box shows the crop selected by our method.

Table 7: Statistics of the VQA datasets used in our evaluation.

Attribute	V*Bench	HRBench-4K	HRBench-8K	Multiview (ours)
Avg. width (px)	2,246	4,024	7,431	4,032
Avg. height (px)	1,582	3,503	5,358	3.024
Images (#)	191	200	200	112
QA pairs (#)	191	800	800	280

A.4 Qualitative Example of Relative Position

To answer relative position (spatial) questions, the model needs to identify the position of two objects from the same image. SAGE can identify small objects such as a soccer ball, which makes the model answer relative position questions of two objects easier. The heatmap visualization demonstrates that our method is interpretable and can focus on important objects.

A.5 Multi-View Transient-Object Localization



Figure 7: Representative examples of the heatmap of relative position. The green box is the ground truth bounding box and the red box is the selected crop from SAGE.

A.6 Implementation Details

Experiments ran on a server with two NVIDIA A40 GPUs (46,068 MiB each), two 16-core Intel Xeon Gold 6346 CPUs at 3.10 GHz, and 251 GiB system memory. Each efficiency run used one A40 without model or data parallelism. The environment used Linux 4.18, Python 3.10.20, PyTorch 2.6.0 with CUDA 12.4, cuDNN 9.1, Transformers 4.50.1, and NVIDIA driver 570.144. Runtime includes localization and final VQA; diagnostic image serialization was disabled. Peak memory is `torch.cuda.max_memory_allocated()` after model initialization.

We present the prompt templates used for target cue extraction, localization, and final question answering. Here, `<target>` denotes the target concept, `<question>` denotes the original question, and `<image>` denotes the model's visual input token. The cue extraction template is supplied with six question–response demonstrations, same as ZoomEye.

TARGET_CUE_PROMPT

Question: `<question>`
If you want to answer the question, which objects' information do you need?

TARGET_CUE_PROMPT

```
Question: <question>
If you want to answer the question, which objects' information do
you need?
```

The following six question–response pairs are used as in-context demonstrations. They are reproduced verbatim from the implementation.

Target-cue demonstrations

Demonstration 1
Question: What is the color of the boy's bag?
Response: To answer the question, I need know the location of the boy with a bag so that I can determine the color of the bag. So I need the information about the following objects: boy with a bag.

Demonstration 2
Question: Is the yellow car on the left or right side of the white car?
Response: To answer the question, I need know the location of the yellow car and the white car so that I can determine the positional relationship between the two of them. So I need the information about the following objects: white car and yellow car.

Demonstration 3
Question: Tell me the number on the black board.
Response: To answer the question, I need know the location of the black board so that I can determine the number on it. So I need the information about the following objects: black board

Demonstration 4
Question: Is the girl with pink hair on the left or right side of the man with backpack?
Response: To answer the question, I need know the location of the girl with pink hair and the man with backpack so that I can determine the positional relationship between the two of them. So I need the information about the following objects: girl with pink hair and man with backpack.

Demonstration 5
Question: What kind of animal is on the red sign?
Response: To answer the question, I need know the location of the red sign so that I can determine the kind of animal on it. So I need the information about the following objects: red sign.

Demonstration 6
Question: From the information on that advertising board, what is the type of this shop?
Response: To answer the question, I need know the location of the advertising board so that I can determine the type of the shop. So I need the information about the following objects: advertising board.

LOCALIZATION_PROMPT

```
<image>
Locate the object in the image: <target>
```

The localization prompt accompanies the spatial tile sequence. We extract the internal visual and target representations from this forward pass to construct the localization scores.

VQA_PROMPT

```
<image>
This is the main image, and the section enclosed by the red rectangle is the focus region.
<image>
This is the zoomed-in view of the focus region.
<question>
```