EFFICIENTVIS: ADAPTIVE PARALLEL VISUAL SEARCH FOR HIGH-RESOLUTION REMOTE SENSING VISUAL GROUNDING

Anonymous authors
Paper under double-blind review

ABSTRACT

High-resolution remote sensing visual grounding (RSVG) requires both global spatial context and fine-grained evidence for small targets. Full-scene inference preserves context but loses local details after resizing, whereas iterative visual search recovers such details through sequential crop-and-ground calls at substantial computational cost. We propose EfficientViS, an adaptive parallel framework for efficient high-resolution RSVG. EfficientViS consists of two components: 1) One-shot Complementary Region Planning (OCRP), which predicts diverse candidate regions from a single global image–text representation before local grounding; and 2) Adaptive Candidate Parallel Inference (ACPI), which selects a query-adaptive observation budget and packs selected crops into a single specialist forward pass for efficient grounding. We further introduce a two-stage training strategy: Stage I trains a crop-robust specialist grounder to improve local grounding, while Stage II applies supervised-to-PPO training to optimize the accuracy–cost trade-off. Extensive experiments on three benchmarks (RSVG-HR, AVVG, and XLRS-Bench) demonstrate that EfficientViS delivers 4.3–4.8× higher throughput than the corresponding full-scene specialist baselines and more than 14× higher throughput than sequential search.

1 INTRODUCTION

High-resolution remote-sensing visual grounding (RSVG) aims to localize a region described by a free-form referring expression in large aerial or satellite imagery (Sun et al., 2022; Zhan et al., 2023). The task is challenging because the referred target may occupy only a small fraction of the scene, and the expression may depend on scene-level context and long-range spatial relations. Specialist grounders typically resize the full image to a fixed input size, preserving the global layout but suppressing the fine-grained evidence needed for small or visually similar targets. Increasing the input resolution, constructing multi-scale representations, or retaining more visual tokens can alleviate this detail loss (Tong et al., 2024; Luo et al., 2025; Wang et al., 2025a), but at the cost of processing substantial query-irrelevant content.

Visual search provides a more selective solution by using a global view to identify promising regions and then acquiring high-resolution local observations (Mnih et al., 2014; Wu & Xie, 2024; Sun et al., 2025; Li et al., 2025; Su et al., 2025; Zheng et al., 2026). Recent remote-sensing methods follow this paradigm through adaptive crop-and-zoom perception, hierarchical search, progressive refinement, or active perception (Liu et al., 2026; Zhang et al., 2026; Xu et al., 2026; Wang et al., 2026; Zhu et al., 2026). However, these methods typically interleave region selection with local grounding, so each new observation depends on earlier local results. This sequential dependency makes inference latency grow

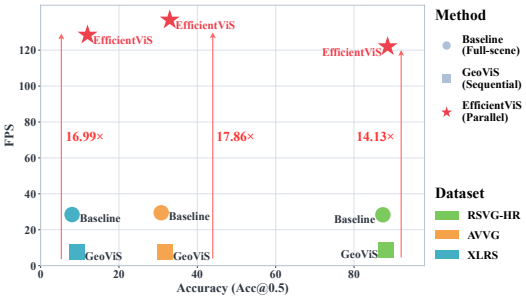

Figure 1: Accuracy–throughput comparison on RSVG-HR, AVVG, and XLRS-Bench.

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.

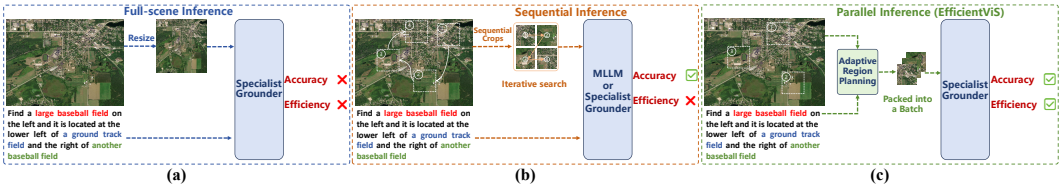

Figure 2: Three paradigms for large-scene visual grounding and their performance comparison. (a) Ordinary resizing preserves the global scene but may discard critical details of tiny targets, resulting in grounding failures. (b) Sequential visual search recovers fine-grained evidence through iterative crop-and-ground steps, at the cost of multiple dependent model searches. (c) EfficientViS replaces sequential search with one-shot complementary region planning and adaptive budget allocation, packs the selected regions into a batch, and grounds them in parallel with a specialist model.

with the number of observations, prevents parallel local grounding, and may introduce redundant searches over overlapping regions.

We observe that such local feedback is not always necessary for region planning. For many referring expressions, the global image–text representation can already provide sufficient semantic and spatial cues to propose multiple plausible target regions, while high-resolution crops verify local detail. This motivates a simple decomposition: determine where to look from the global representation, and perform local verification only after the complete observation plan is formed. We further formalize in Section 3.2 when the additional planning value of local feedback is limited.

Based on this insight, we propose **EfficientViS**, an adaptive parallel region-inference framework for high-resolution RSVG. EfficientViS contains two components. **One-shot Complementary Region Planning (OCRP)** predicts an ordered set of diverse candidate regions from a single global image–text representation, without iterative local feedback. **Adaptive Candidate Parallel Inference (ACPI)** estimates a query-adaptive observation budget, retains the required candidates, and packs their crops into a single specialist forward pass. We train the framework in two stages: crop-robust specialist adaptation first improves local grounding, and supervised-to-PPO planner optimization then learns region selection and budget allocation under the accuracy–cost trade-off. Figures 1 and 2 show that EfficientViS replaces detail-losing global resizing and latency-intensive sequential search with one-shot adaptive region planning and parallel verification, improving the accuracy–throughput trade-off across all three benchmarks.

Experiments on RSVG-HR, AVVG, and XLRS-Bench demonstrate that EfficientViS consistently improves grounding accuracy while delivering more than $4\times$ the throughput of full-scene specialist baselines and more than $14\times$ the throughput of sequential search. Our main contributions are summarized as follows:

- We recast high-resolution grounding as a feedback-free observation allocation problem and derive an information-theoretic bound characterizing when local feedback provides limited additional planning value, which motivates separating region coverage from local verification.
- We introduce **EfficientViS**, combining OCRP for one-shot complementary region planning with ACPI for query-adaptive budgeting and parallel local grounding in a single packed specialist pass.
- We develop a two-stage training strategy combining crop-robust grounding and supervised-to-PPO optimization, and validate EfficientViS on three high-resolution RSVG benchmarks with two specialist grounders, achieving consistent accuracy gains and substantially higher throughput.

2 Related Work

2.1 Remote-Sensing Visual Grounding

Remote-sensing visual grounding has evolved from two-stream vision–language architectures with box regression toward stronger cross-modal interaction, relation modeling, multiscale alignment, language-conditioned refinement, and query-based decoding (Sun et al., 2022; Zhan et al., 2023; Wang et al., 2024; Choudhury et al., 2024; Hang et al., 2024; Liu et al., 2025; Zhao et al., 2024; Li et al., 2024; Lan et al., 2024; Ding et al., 2025; Zhao et al., 2025). GeoGround further broadens the

grounding formulation to horizontal boxes, oriented boxes, and segmentation masks, and introduces AVVG for fine-grained aerial-vehicle grounding (Zhou et al., 2025). Remote-sensing MLLMs instead extend general multimodal backbones to support a broader range of geospatial perception and reasoning tasks (Kuckreja et al., 2024; Zhang et al., 2024; Luo et al., 2024; Pang et al., 2025; Soni et al., 2025; Ou et al., 2025). In contrast to broad multi-task MLLMs, EfficientViS targets efficient high-resolution localization with specialist grounders.

2.2 High-Resolution Perception and Visual Search

High-resolution vision–language systems mitigate detail loss through larger visual contexts, dynamic pyramids, patch selection, or token pruning (Luo et al., 2025; Wang et al., 2025a). Although these approaches preserve more fine-grained information, they may still spend substantial computation on query-irrelevant regions. Visual search offers a more selective alternative by acquiring local evidence through crop, zoom, or region-selection operations (Mnih et al., 2014; Wu & Xie, 2024; Sun et al., 2025; Li et al., 2025; Su et al., 2025; Zheng et al., 2026). In remote sensing, ZoomEarth adopts adaptive crop-and-zoom perception, GeoViS performs reward-guided hierarchical search, and VLAS and GeoSearcher progressively refine candidate regions, while GeoVista first constructs a global multi-region plan and then performs branch-wise verification and evidence tracking (Liu et al., 2026; Zhang et al., 2026; Xu et al., 2026; Wang et al., 2026; Zhu et al., 2026). Prior visual-search methods acquire local evidence selectively, but typically retain iterative or branch-wise execution. EfficientViS instead uses OCRP to generate complementary candidates from global image–text features and ACPI to select an adaptive budget subset, enabling all selected crops to be grounded jointly in one packed specialist pass.

3 Method

3.1 Overview

EfficientViS decouples global region allocation from local verification. Operationally, OCRP constructs a set of complementary spatial hypotheses from a single global image–text representation, while ACPI selects query-adaptive regions of this plan and verifies the retained crops in one packed specialist pass. Thus, the complete observation plan and budget are fixed before any local grounding. As shown in Fig. 3, given an image $I \in \mathbb{R}^{H\times W\times 3}$ and a referring expression q, EfficientViS first plans where to observe and how many regions to retain, then grounds the selected crops jointly and maps the predictions back to the global scene.

We first apply image preprocessing to obtain the input

$$\bar{I} = \mathcal{P}_s(I), \tag{1}$$

where $\mathcal{P}_s$ applies letterbox padding and pixel normalization without globally resizing or downsampling the original image. Predictions in this preprocessed coordinate system are mapped back to the original image through the inverse geometric transformation for evaluation.

For global planning, EfficientViS reuses the specialist's visual encoder, text encoder, and cross-modal feature extractor. Specifically, we feed a downsampled global view and the referring expression through the specialist feature extractor,

$$z = F_\theta\big(\mathrm{Down}(\bar{I}), q\big), \qquad z \in \mathbb{R}^{H_a\times W_a\times d}, \tag{2}$$

where z denotes the fused global image–text feature map. EfficientViS then applies lightweight planning heads to z to predict the spatial score map and continuous location offsets used for complementary region generation. The same global representation is subsequently used by the utility head to determine the retained budget K^*. Once the region plan and K^* are fixed, the corresponding crops are extracted from $\bar{I}$, packed along the batch dimension, and processed jointly by G_ω. In this architecture, OCRP covers candidate-region generation, whereas ACPI covers adaptive budget selection and packed local grounding. This design performs global planning once and reduces fine-grained local grounding to a single packed pass.

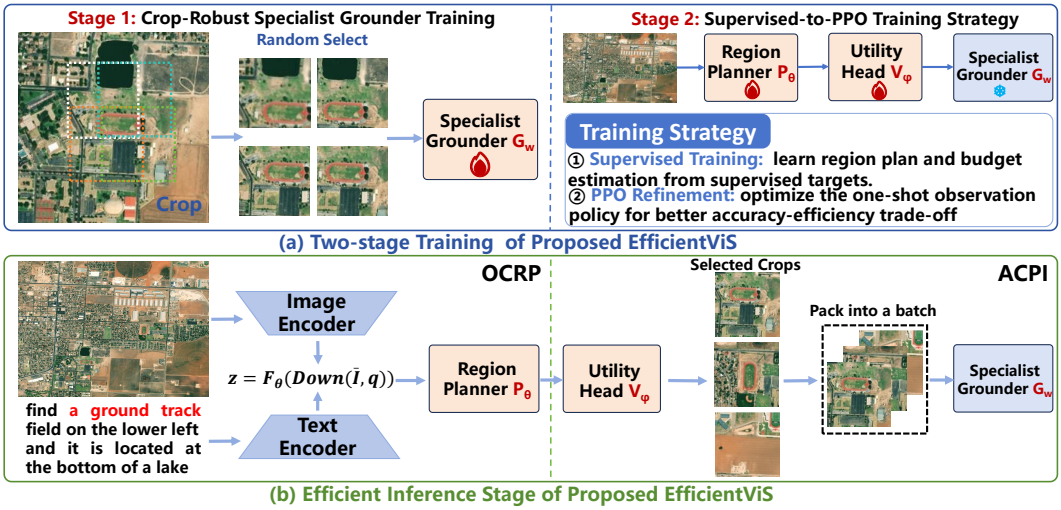

Figure 3: Overview of EfficientViS. During training, Stage I trains a crop-robust specialist grounder, while Stage II first trains the planner and utility head with supervised objectives and then refines them with PPO. During inference, OCRP generates complementary candidate regions from the global image–text representation, and ACPI selects an adaptive budget and grounds the selected crops jointly in one packed specialist pass.

3.2 Motivation and Theoretical Analysis

Sequential visual search lets the j-th spatial decision depend on local evidence $O_{<j}$ from earlier crop-and-ground steps. EfficientViS instead resolves the complete region plan from the global image–text state Z before observing any local crop. This does not assume that local evidence is unimportant for final discrimination; EfficientViS asks whether that evidence must recursively steer later region decisions.

Let $R_j^\dagger$ denote a latent target-relevant region whose observation would be useful at step j. The residual spatial information supplied by previous local observations is

$$\mathcal{I}_j = I(R_j^\dagger; O_{<j} \mid Z). \tag{3}$$

Using the conditional mutual-information identity, Pinsker's inequality, and Jensen's inequality yields

$$\mathbb{E}\Big[D_{\mathrm{TV}}\Big(p(R_j^\dagger \mid Z, O_{<j}), p(R_j^\dagger \mid Z)\Big)\Big] \le \sqrt{\frac{\mathcal{I}_j}{2}}, \tag{4}$$

and, for any planning utility normalized to $[0, 1]$,

$$0 \le V_{\mathrm{fb},j}^* - V_{\mathrm{ff},j}^* \le \sqrt{2\mathcal{I}_j}. \tag{5}$$

Thus, when $\mathcal{I}_j$ is small, even an ideal feedback-conditioned planner has only bounded additional decision value on average. The complete derivation, the Bayes-value bound, and its scope for learned sequential policies are given in Appendix A.3.

This perspective also separates candidate-region coverage from local grounding. Let C_K be the event that the retained K crops contain sufficient target evidence and S the event of successful grounding. Then

$$P(S) = P(C_K)P(S \mid C_K) + P(C_K^c)P(S \mid C_K^c). \tag{6}$$

For crop-based grounding, $P(S \mid C_K^c)$ is typically small. Complementary planning and adaptive budgeting therefore target coverage efficiency $P(C_K)$ under a small observation budget, while crop-robust specialist adaptation improves $P(S \mid C_K)$ once the target is covered.

Finally, a sequential search with K observations has local-inference dependency depth K, whereas EfficientViS determines K^* before local grounding and evaluates all retained crops together:

$$D_{\mathrm{local}}^{\mathrm{seq}} = K, \qquad D_{\mathrm{local}}^{\mathrm{EfficientViS}} = 1. \tag{7}$$

3.3 ONE-SHOT COMPLEMENTARY REGION PLANNING (OCRP)

A large remote-sensing scene may contain several visually or spatially plausible candidates for the same referring expression. Rather than committing to a single location, the planner constructs a small set of complementary spatial hypotheses from the global representation z. This process depends only on global image–text evidence and does not require feedback from the local grounder.

Specifically, the planner projects the fused feature map into a spatial score map $\ell \in \mathbb{R}^{H_a \times W_a}$ and four offset maps (u^x, v^x, u^y, v^y). These outputs are used to generate an ordered region plan $P = (p_1, \ldots, p_M)$ through region selection and continuous location refinement.

Region selection. At step j, the policy is conditioned on the global representation and the previously selected regions. Let $m_{<j}$ denote the suppression mask induced by $P_{<j} = (p_1, \ldots, p_{j-1})$. The region-selection policy is

$$\pi_j(c) = \mathrm{Cat}(\ell + m_{<j}), \qquad c_j = (y_j, x_j) \sim \pi_j. \tag{8}$$

By reducing the scores of previously covered neighborhoods, the suppression mechanism discourages repeated observations of the same area and encourages different regions to represent complementary spatial hypotheses. At inference, the highest-scoring cell is selected instead of sampling.

Location refinement. Each selected grid cell is refined by a continuous within-cell offset. During training the offset is sampled from a learned Beta distribution to support policy exploration; at inference its mean is used for deterministic region centers.

Region plan. At inference, let $c_j = \arg\max_c[\ell_c(z) + m(c; c_{<j})]$. Since every region is generated from the same global representation without local-grounder feedback, the complete plan can be determined before local inference begins. The region plan provides multiple candidate regions, but evaluating all of them is unnecessary for every query. We therefore estimate how much of the plan should be retained.

3.4 ADAPTIVE CANDIDATE PARALLEL INFERENCE (ACPI)

The amount of local evidence required for reliable grounding varies across queries. A visually distinctive target may be resolved from a single region, whereas a small target surrounded by repeated structures may benefit from several complementary observations. EfficientViS therefore selects the region budget adaptively rather than using a fixed number of crops.

For each planned center p_j, let h_j denote the corresponding feature sampled from the fused global feature map, let $\tilde{p}_j = [p_j^x, p_j^y]$ denote its normalized geometry, and a_j denote the corresponding planner confidence. Because the crop extent is fixed under a given experimental setting, the center coordinates are sufficient to describe the planned region. We combine h_j, $\tilde{p}_j$, and a_j using a lightweight region encoder f_r, implemented as a linear layer followed by GELU and LayerNorm:

$$e_j = f_r([h_j; \tilde{p}_j; a_j]), \qquad \bar{e}_k = \frac{1}{k}\sum_{j=1}^{k} e_j. \tag{9}$$

Because the region plan is ordered, each $k \in \{1, \ldots, M\}$ naturally defines a candidate region set $P_{1:k}$. The aggregated descriptor $\bar{e}_k$ summarizes the first k planned regions.

A utility head V_φ then evaluates each candidate budget jointly with the global image–text state:

$$v_k = V_\varphi(z, \bar{e}_k), \qquad K^* = \arg\max_{1 \le k \le M} v_k. \tag{10}$$

V_φ is implemented as a LayerNorm MLP followed by a linear output layer. The selected K^* determines how many planned regions are passed to the local grounder.

After the region plan and budget have been determined, EfficientViS uses the selected regions to recover fine-grained visual evidence. Given K^*, the first K^* regions are cropped from the preprocessed image:

$$X_j = \mathrm{Crop}(\bar{I}, p_j), \qquad 1 \le j \le K^*. \tag{11}$$

Here, $\mathrm{Crop}(\bar{I}, p_j)$ extracts a fixed-size local window centered at p_j from the preprocessed image; all resulting crops use the same local input resolution. Rather than processing these observations

sequentially, we pack them along the batch dimension and invoke the specialist grounder once:

$$\mathcal{C}_{1:K^*} = G_\omega(\mathrm{Pack}(X_{1:K^*}), q). \tag{12}$$

In our experiments, G_ω is instantiated with LPVA (Li et al., 2024) or LQVG (Lan et al., 2024). Each crop returns local boxes and confidence scores, which are mapped back to global coordinates and fused with the corresponding planner confidence. The highest fused score defines the final prediction.

3.5 Two-Stage Training

Training proceeds in two stages while preserving the separation between global decision making and local verification. **Stage I** adapts the specialist to target-centered crops with bounded center perturbations, reducing the distribution shift between resized full scenes and planner-generated observations; the adapted specialist is then frozen. **Stage II** trains OCRP for complementary region planning and ACPI for query-adaptive budgeting, using supervised bootstrap for initialization followed by PPO refinement of the one-shot observation policy under the final accuracy–cost objective. The terminal reward combines final IoU, target coverage, and local-grounding quality with penalties for overlap, cropped area, and latency. Throughout Stage II, specialist predictions serve only as training-time reward signals and never feed back into OCRP region selection within the same rollout. Thus, PPO improves the one-shot observation-allocation policy without introducing iterative or closed-loop inference. Appendix A.1 and A.2 give the exact reward, score fusion, hyperparameters, and inference algorithm.

4 Experiments

4.1 Experimental Setup

Datasets and metrics. We evaluate EfficientViS on three remote-sensing visual-grounding benchmarks spanning substantially different image scales. RSVG-HR (Lan et al., 2024) contains 2,650 image–text pairs across seven object categories, with all images at a resolution of 1024×1024. AVVG (Zhou et al., 2025) consists of 26,465 referring instances collected from UAV imagery with an image width of approximately 4,000 pixels and a ground sampling distance of 0.007–0.04 m. XLRS-Bench (Wang et al., 2025b) further scales visual grounding to ultra-high-resolution remote-sensing imagery, comprising 1,400 source images with an average resolution of approximately 8500×8500, including 840 images at 10000×10000, and provides 12,619 manually annotated visual-grounding instances. Together, these benchmarks span 1K, 4K, and 8K imagery and stress relational reasoning, fine-grained target discrimination, and robustness to increasing scene extent and target sparsity. We report Acc@0.5 as the primary grounding metric.

Baselines. We compare general MLLMs (Qwen2.5-VL-3B (Bai et al., 2025), LLaVA-1.5-7B (Liu et al., 2024), GPT-4o (OpenAI, 2024)), remote-sensing MLLMs (GeoChat (Kuckreja et al., 2024), GeoViS (Zhang et al., 2026), GeoGround (Zhou et al., 2025)), specialist grounders (TransVG (Deng et al., 2021), VLTVG (Yang et al., 2022), LPVA (Li et al., 2024), LQVG (Lan et al., 2024)), and the same specialists equipped with GeoViS-Search or EfficientViS. GeoViS-Search preserves GeoViS's sequential search while sharing the specialist and other comparable modules with EfficientViS.

Implementation details. All EfficientViS variants use a 32×32 planning grid, $M = 4$ candidate regions, a 256×256 global planning view, and 256×256 local crops. Stage I runs for 50 epochs and Stage II for 20 epochs. Training uses batch size 8 and BF16 precision. All FPS values are measured on the same NVIDIA RTX 5090 with a common timing boundary that includes global planning, crop construction, packed local grounding, and score fusion but excludes data loading.

4.2 Quantitative Results

As shown in Table 1, EfficientViS consistently improves grounding performance across all three benchmarks. With LPVA, EfficientViS improves Acc@0.5 by 1.20, 2.23, and 2.97 on RSVG-HR, AVVG, and XLRS-Bench, respectively. Similar improvements are observed with LQVG, where EfficientViS achieves 88.58, 32.96, and 11.97, outperforming the original specialist by 1.21, 2.18, and 3.94, respectively.

Table 1: Acc@0.5 on three benchmarks. Overall is the mean over the three benchmarks and is reported only when all three results are available. Multimodal results are taken from prior benchmark reports and specialist/framework results are measured by us. Blank entries denote unreported metrics and green subscripts show absolute gains over the aligned specialist baseline. Best values are bold.

Method	Params	Mode	RSVG-HR	AVVG	XLRS	Overall
General multimodal models						
Qwen2.5-VL-3B	3B	Full-scene	29.40	–	–	–
LLaVA-1.5-7B	7B	Full-scene	–	15.05	0.09	–
GPT-4o	–	Full-scene	–	–	0.46	–
Remote-sensing multimodal models						
GeoChat	7B	Full-scene	–	11.52	0.14	–
GeoGround	7B	Full-scene	–	21.58	–	–
GeoViS	3B	Sequential	51.50	–	–	–
Specialist models						
TransVG	149.70M	Full-scene	46.26	21.89	3.97	24.04
VLTVG	152.20M	Full-scene	48.76	23.45	4.34	25.52
LPVA	163.80M	Full-scene	85.57	28.83	6.31	40.24
LQVG	166.38M	Full-scene	87.37	30.78	8.03	42.06
Specialist models + framework						
LPVA + GeoViS-Search	164.29M	Sequential	86.37	30.84	7.26	41.49
LPVA + EfficientViS	164.88M	**Parallel**	86.77$_{\uparrow 1.20}$	31.06$_{\uparrow 2.23}$	9.28$_{\uparrow 2.97}$	42.37$_{\uparrow 2.13}$
LQVG + GeoViS-Search	166.87M	Sequential	88.18	31.83	9.35	43.12
LQVG + EfficientViS	167.46M	**Parallel**	**88.58**$_{\uparrow 1.21}$	**32.96**$_{\uparrow 2.18}$	**11.97**$_{\uparrow 3.94}$	**44.50**$_{\uparrow 2.44}$

The gains with both LPVA and LQVG show that EfficientViS is not tied to a particular specialist. The largest absolute gains occur on XLRS-Bench: EfficientViS raises LPVA from 6.31 to 9.28 and LQVG from 8.03 to 11.97. This pattern is consistent with expectation: when full-scene resizing removes more local detail, selectively recovering high-resolution evidence becomes more valuable, without requiring full-resolution processing or multiple serial local passes.

Compared with GeoViS-Search under the same specialist backbone, EfficientViS also achieves consistently better performance. On XLRS-Bench, EfficientViS improves over GeoViS-Search by 2.02 with LPVA and 2.62 with LQVG, while maintaining gains on RSVG-HR and AVVG. Overall, LQVG+EfficientViS achieves the best results among the compared methods on all three benchmarks, supporting the effectiveness of adaptive complementary region planning for large-scene remote-sensing grounding.

4.3 Efficiency Analysis

Table 2 compares inference efficiency under the same hardware and timing protocol. GeoViS-Search performs MCTS-based exploration and adds 0.49M trainable parameters, but it repeatedly evaluates candidate regions as the search tree expands, resulting in multiple sequential perception steps and substantial inference latency. In contrast, EfficientViS determines the candidate regions and observation budget before local grounding and evaluates all retained crops in a single packed batch. EfficientViS introduces only 1.08M framework parameters while retaining the original specialist backbone, and its local inference is substantially cheaper: the specialist processes smaller crops, with fewer regions selected on average.

Consequently, EfficientViS consistently improves throughput across both specialist backbones. With LPVA, it achieves 140.76, 155.47, and 147.42 FPS on RSVG-HR, AVVG, and XLRS-Bench, corresponding to 4.46×, 4.75×, and 4.63× speedups over the original specialist. With LQVG, EfficientViS reaches 122.01, 136.84, and 128.46 FPS, yielding 4.31×, 4.64×, and 4.52× speedups, respectively. These results indicate that efficient high-resolution grounding depends not only on parameter count, but also on avoiding repeated local inference. By replacing sequential search with global planning followed by parallel local grounding, EfficientViS provides a consistently better accuracy–throughput trade-off across different backbones and image scales.

Table 2: Efficiency comparison on three benchmarks. Speedup is relative to the specialist baseline.

Method	Model Params	Framework Params	RSVG-HR		AVVG		XLRS	
			FPS ↑	Speedup ↑	FPS ↑	Speedup ↑	FPS ↑	Speedup ↑
LPVA	163.80M	–	31.56	1.00×	32.73	1.00×	31.84	1.00×
LPVA + GeoViS-Search	163.80M	0.49M	8.97	0.28×	8.14	0.25×	7.52	0.24×
LPVA + EfficientViS	163.80M	1.08M	**140.76**	**4.46**×	**155.47**	**4.75**×	**147.42**	**4.63**×
LQVG	166.38M	–	28.33	1.00×	29.47	1.00×	28.45	1.00×
LQVG + GeoViS-Search	166.38M	0.49M	8.63	0.30×	7.66	0.26×	7.56	0.27×
LQVG + EfficientViS	166.38M	1.08M	**122.01**	**4.31**×	**136.84**	**4.64**×	**128.46**	**4.52**×

Table 3: Component ablation (Acc@0.5).

Variant	RSVG-HR	AVVG	XLRS
Baseline	87.37	30.78	8.03
+ planning	87.78	31.14	8.62
+ complementarity	87.98	31.47	9.84
+ adaptive budget	88.18	31.76	10.12
+ PPO	**88.58**	**32.96**	**11.97**

Table 4: Budget ablation on RSVG-HR.

Budget	Avg. K	Acc@0.5	FPS ↑
Fixed $K = 1$	1.00	87.78	**167.24**
Fixed $K = 2$	2.00	88.18	118.53
Fixed $K = 3$	3.00	**88.58**	92.71
Fixed $K = 4$	4.00	**88.58**	77.36
Adaptive	**1.63**	**88.58**	122.01

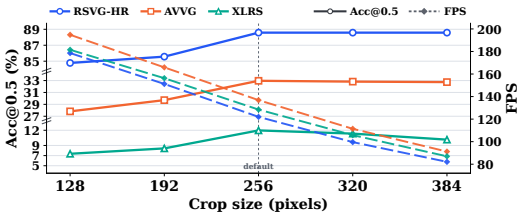

Figure 4: Crop-size accuracy–FPS ablation.

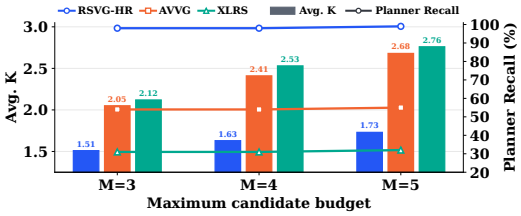

Figure 5: Candidate-capacity ablation.

4.4 Ablation and Diagnostic Analysis

The main results establish EfficientViS's accuracy–throughput advantage. We now isolate the mechanisms behind it: component-wise planning gains, fixed versus adaptive observation budgets, sensitivity to crop scale and candidate capacity, and the PPO terms that shape the final accuracy–cost trade-off. All ablations use LQVG with the remaining training and inference settings fixed.

Table 3 shows a consistent component-wise progression across all three benchmarks. Learned planning improves over the full-scene baseline; complementary suppression further improves the result by discouraging redundant hypotheses; adaptive budgeting adds another gain; and PPO reaches 88.58/32.96/11.97 Acc@0.5 on RSVG-HR/AVVG/XLRS-Bench. Each successive component improves Acc@0.5 in this ablation sequence.

Table 4 then isolates how many local observations are needed. On RSVG-HR, fixed $K = 3$ and $K = 4$ both reach 88.58 Acc@0.5, but throughput drops from 92.71 to 77.36 FPS as the fourth crop is added. The adaptive policy matches the same peak accuracy with only 1.63 crops on average and reaches 122.01 FPS. EfficientViS therefore preserves a sufficiently large candidate capacity for hard queries without paying that cost uniformly.

Figures 4 and 5 examine the two capacity hyperparameters. In Fig. 4, increasing the crop side from 128 to 256 pixels raises Acc@0.5 on three benchmarks; larger crops add little accuracy while FPS continues to fall. We therefore use 256×256 crops. In Fig. 5, increasing the maximum candidate cap from $M = 3$ to $M = 5$ raises the average retained budget on all three benchmarks but changes planner recall only slightly, indicating diminishing returns beyond $M = 4$. At the default $M = 4$, the adaptive policy retains 1.63/2.41/2.53 crops on average on RSVG-HR/AVVG/XLRS-Bench. Thus, ACPI uses fewer observations on RSVG-HR and more on the higher-resolution benchmarks while remaining below the maximum budget on average. Based on these ablations, we use a crop size of 256 and a maximum candidate budget of 4.

As shown in Table 5, using only the IoU reward yields a relatively large region budget and lower Acc@0.5. Adding coverage and local-quality rewards progressively improves localization, showing that these terms provide useful training signals for learning the one-shot region policy. The cost terms

Table 5: PPO reward ablation on RSVG-HR. R_I, R_C, and R_L denote IoU, coverage, and local-quality rewards; C_O, C_R, and C_T denote overlap, region, and latency costs.

Setting	R_I	R_C	R_L	C_O	C_R	C_T	Avg. $K \downarrow$	FPS ↑	Acc@0.5
IoU only	✓	✗	✗	✗	✗	✗	2.86	79.42	88.18
+ Coverage	✓	✓	✗	✗	✗	✗	2.52	88.17	88.38
+ Local quality	✓	✓	✓	✗	✗	✗	2.47	97.15	**88.58**
+ Overlap cost	✓	✓	✓	✓	✗	✗	2.08	108.42	**88.58**
+ Region cost	✓	✓	✓	✓	✓	✗	1.78	116.74	**88.58**
+ Latency cost	✓	✓	✓	✓	✓	✓	**1.63**	**122.01**	**88.58**

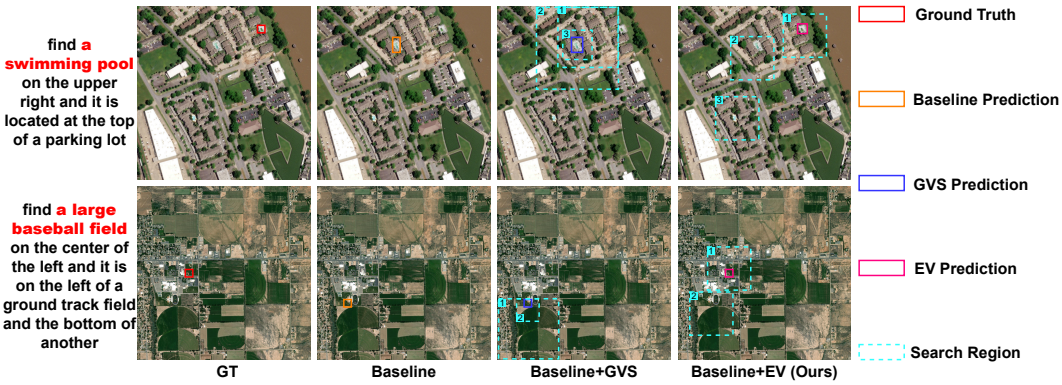

Figure 6: Qualitative comparisons on RSVG-HR among the baseline (LQVG), GeoViS-Search, and EfficientViS. EV: EfficientViS, GVS: GeoViS-Search.

mainly regulate efficiency: overlap and region costs progressively reduce redundant observations while preserving the best accuracy, and the latency cost further encourages efficient execution. With the full objective, EfficientViS uses only 1.63 regions on average and reaches 122.01 FPS at 88.58 Acc@0.5. This suggests that the reward and cost terms play important roles in balancing localization quality and computational efficiency.

Figure 6 shows how these design choices appear in practice. Full-scene LQVG can lose discriminative detail after global resizing, whereas sequential GeoViS-Search may propagate an early mistake into later search decisions. EfficientViS instead commits to a complementary set of global hypotheses and verifies the retained crops jointly afterward. These examples illustrate the distinct roles of planning and local grounding: planning governs coverage, diversity, and budget allocation, while the local specialist provides discrimination without feeding back into subsequent observation decisions.

5 Conclusion

We presented **EfficientViS**, an efficient parallel framework for high-resolution remote-sensing visual grounding. EfficientViS factorizes high-resolution grounding into two complementary components: OCRP constructs diverse spatial hypotheses from global image–text evidence, while ACPI selects a query-adaptive budget and verifies the retained high-resolution crops in a single packed specialist pass. By fixing both the region plan and observation budget before local grounding, this design decouples observation allocation from local discrimination and changes local-inference dependency to a single parallel stage. The feedback-free analysis provides a conditional bound on the value of local feedback: if previous local observations contain little residual information about the next useful spatial hypothesis beyond the global state, even an ideal feedback-conditioned planner has only bounded additional decision value. Removing this dependency also eliminates a pathway through which intermediate local errors can affect later region selections. Across two specialist grounders and three scene scales, EfficientViS consistently improves localization accuracy while maintaining substantially higher throughput than sequential search. Overall, EfficientViS establishes one-shot global planning followed by adaptive parallel grounding as an effective alternative to sequential visual search, recovering fine-grained local evidence without sacrificing the efficiency of specialist models.

AI USE STATEMENT

Large Language Models (LLMs) were used in this work exclusively for language polishing and improving the clarity of writing. No LLM was used for generating scientific ideas, experimental design, or data analysis. All technical content, results, and interpretations presented in this paper are solely the work of the authors. The authors take responsibility for the content of this paper, including the text refined with the assistance of LLMs.

REPRODUCIBILITY STATEMENT

We provide the model formulation, training and inference procedures, and evaluation protocol. Detailed implementation and optimization settings are provided in Appendix A. Upon acceptance, we will release the code, configurations, and model checkpoints to facilitate reproducibility.

REFERENCES

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025. doi: 10.48550/arXiv.2502.13923.

Shabnam Choudhury, Pratham Kurkure, Priyanka Talwar, and Biplab Banerjee. CrossVG: Visual grounding in remote sensing with modality-guided interactions. In *IEEE International Geoscience and Remote Sensing Symposium*, pp. 2858–2862, 2024. doi: 10.1109/IGARSS53475.2024.10642183.

Jiajun Deng, Zhengyuan Yang, Tianlang Chen, Wengang Zhou, and Houqiang Li. Transvg: End-to-end visual grounding with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1769–1779, 2021.

Yueli Ding, Di Wang, Ke Li, Xiaohong Zhao, and Yifeng Wang. Visual grounding of remote sensing images with multi-dimensional semantic-guidance. *Pattern Recognition Letters*, 189:85–91, 2025. doi: 10.1016/j.patrec.2025.01.013.

Renlong Hang, Siqi Xu, and Qingshan Liu. A regionally indicated visual grounding network for remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–11, 2024. doi: 10.1109/TGRS.2024.3490847.

Kartik Kuckreja, Muhammad Sohail Danish, Muzammal Naseer, Abhijit Das, Salman Khan, and Fahad Shahbaz Khan. GeoChat: Grounded large vision-language model for remote sensing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27831–27840, June 2024.

Meng Lan, Fu Rong, Hongzan Jiao, Zhi Gao, and Lefei Zhang. Language query-based transformer with multiscale cross-modal alignment for visual grounding on remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–13, 2024. doi: 10.1109/TGRS.2024.3407598.

Geng Li, Jinglin Xu, Yunzhen Zhao, and Yuxin Peng. Dyfo: A training-free dynamic focus visual search for enhancing lmms in fine-grained visual understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9098–9108, 2025.

Ke Li, Di Wang, Haojie Xu, Haodi Zhong, and Cong Wang. Language-guided progressive attention for visual grounding in remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–13, 2024. doi: 10.1109/TGRS.2024.3423663.

Biao Liu, Xu Liu, Lingling Li, Licheng Jiao, Fang Liu, Xinyu Sun, and Youlin Huang. Language-guided hybrid representation learning for visual grounding on remote sensing images. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pp. 1557–1566, 2025. doi: 10.24963/ijcai.2025/174.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26296–26306, June 2024.

Ruixun Liu, Bowen Fu, Jiayi Song, Kaiyu Li, Wanchen Li, Lanxuan Xue, Hui Qiao, Weizhan Zhang, Deyu Meng, and Xiangyong Cao. ZoomEarth: Active perception for ultra-high-resolution geospatial vision-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 34877–34888, June 2026.

Junwei Luo, Zhen Pang, Yongjun Zhang, Tingzhu Wang, Linlin Wang, Bo Dang, Jiangwei Lao, Jian Wang, Jingdong Chen, Yihua Tan, and Yansheng Li. Skysensegpt: A fine-grained instruction tuning dataset and model for remote sensing vision-language understanding. *arXiv preprint arXiv:2406.10100*, 2024. doi: 10.48550/arXiv.2406.10100.

Junwei Luo, Yingying Zhang, Xue Yang, Kang Wu, Qi Zhu, Lei Liang, Jingdong Chen, and Yansheng Li. When large vision-language model meets large remote sensing imagery: Coarse-to-fine text-guided token pruning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9206–9217, October 2025.

Volodymyr Mnih, Nicolas Heess, Alex Graves, and Koray Kavukcuoglu. Recurrent models of visual attention. In *Advances in Neural Information Processing Systems*, volume 27, pp. 2204–2212, 2014.

OpenAI. Hello GPT-4o. OpenAI, May 2024. Published May 13, 2024.

Ruizhe Ou, Yuan Hu, Fan Zhang, Jiaxin Chen, and Yu Liu. GeoPix: A multimodal large language model for pixel-level image understanding in remote sensing. *IEEE Geoscience and Remote Sensing Magazine*, 13(3): 324–337, 2025. doi: 10.1109/MGRS.2025.3560293.

Chao Pang, Xingxing Weng, Jiang Wu, Jiayu Li, Yi Liu, Jiaxing Sun, Weijia Li, Shuai Wang, Litong Feng, Gui-Song Xia, and Conghui He. VHM: Versatile and honest vision language model for remote sensing image analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 6381–6388, 2025. doi: 10.1609/aaai.v39i6.32683.

Sagar Soni, Akshay Dudhane, Hiyam Debary, Mustansar Fiaz, Muhammad Akhtar Munir, Muhammad Sohail Danish, Paolo Fraccaro, Campbell D Watson, Levente J Klein, Fahad Shahbaz Khan, and Salman Khan. EarthDial: Turning multi-sensory earth observations to interactive dialogues. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14303–14313, 2025.

Alex Su, Haozhe Wang, Weiming Ren, Fangzhen Lin, and Wenhu Chen. Pixel reasoner: Incentivizing pixel space reasoning via curiosity-driven reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 38, pp. 9130–9159, 2025. doi: 10.52202/085713-0280.

Guangyan Sun, Mingyu Jin, Zhenting Wang, Cheng-Long Wang, Siqi Ma, Qifan Wang, Tong Geng, Ying Nian Wu, Yongfeng Zhang, and Dongfang Liu. Visual agents as fast and slow thinkers. In *International Conference on Learning Representations*, 2025.

Yuxi Sun, Shanshan Feng, Xutao Li, Yunming Ye, Jian Kang, and Xu Huang. Visual grounding in remote sensing images. In *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 404–412, 2022. doi: 10.1145/3503161.3548316.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9568–9578, 2024.

Dianyu Wang, Yidan Zhang, Peirong Zhang, Xuyang Li, Xiaoxuan Liu, and Lei Wang. GeoSearcher: Anchor-guided progressive reasoning for remote sensing visual grounding with process supervision. *arXiv preprint arXiv:2607.01050*, 2026. doi: 10.48550/arXiv.2607.01050.

Fengxiang Wang, Mingshuo Chen, Yueying Li, Di Wang, Haotian Wang, Zonghao Guo, Zefan Wang, Shan Boqi, Long Lan, Yulin Wang, Hongzhen Wang, Wenjing Yang, Bo Du, and Jing Zhang. Geollava-8k: Scaling remote-sensing multimodal large language models to 8k resolution. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 159185–159218. Curran Associates, Inc., 2025a. doi: 10.52202/085713-5317.

Fengxiang Wang, Hongzhen Wang, Zonghao Guo, Di Wang, Yulin Wang, Mingshuo Chen, Qiang Ma, Long Lan, Wenjing Yang, Jing Zhang, Zhiyuan Liu, and Maosong Sun. XLRS-Bench: Could your multimodal LLMs understand extremely large ultra-high-resolution remote sensing imagery? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14325–14336, June 2025b.

Fuyan Wang, Chunlei Wu, Jie Wu, Leiquan Wang, and Canwei Li. Multistage synergistic aggregation network for remote sensing visual grounding. *IEEE Geoscience and Remote Sensing Letters*, 21:1–5, 2024. doi: 10.1109/LGRS.2024.3360473.

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*, pp. 13084–13094, 2024.

Nuo Xu, Kelu Yao, Rong Yang, and Chao Li. Visual-language active search for wide-area remote sensing imagery. *Pattern Recognition*, 175:113106, 2026. doi: 10.1016/j.patcog.2026.113106.

Li Yang, Yan Xu, Chunfeng Yuan, Wei Liu, Bing Li, and Weiming Hu. Improving visual grounding with visual-linguistic verification and iterative reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9499–9508, 2022.

Yang Zhan, Zhitong Xiong, and Yuan Yuan. RSVG: Exploring data and models for visual grounding on remote sensing data. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–13, 2023. doi: 10.1109/TGRS.2023.3250471.

Peirong Zhang, Yidan Zhang, Luxiao Xu, Jinliang Lin, Zonghao Guo, Fengxiang Wang, Xue Yang, Kaiwen Wei, and Lei Wang. GeoViS: Geospatially rewarded visual search for remote sensing visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14335–14345, June 2026.

Wei Zhang, Miaoxin Cai, Tong Zhang, Yin Zhuang, and Xuerui Mao. EarthGPT: A universal multimodal large language model for multisensor image comprehension in remote sensing domain. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–20, 2024. doi: 10.1109/TGRS.2024.3409624.

Enyuan Zhao, Ziyi Wan, Ze Zhang, Jie Nie, Xinyue Liang, and Lei Huang. A spatial-frequency fusion strategy based on linguistic query refinement for RSVG. *IEEE Transactions on Geoscience and Remote Sensing*, 62: 1–13, 2024. doi: 10.1109/TGRS.2024.3471082.

Yichen Zhao, Yaxiong Chen, Ruilin Yao, Shengwu Xiong, and Xiaoqiang Lu. Context-driven and sparse decoding for remote sensing visual grounding. *Information Fusion*, 123:103296, 2025. doi: 10.1016/j.inffus.2025.103296.

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

Yue Zhou, Mengcheng Lan, Xiang Li, Litong Feng, Yiping Ke, Xue Jiang, Qingyun Li, Xue Yang, and Wayne Zhang. Geoground: A unified large vision-language model for remote sensing visual grounding, 2025. URL https://arxiv.org/abs/2411.11904.

Jiashun Zhu, Ronghao Fu, Jiasen Hu, Jing Huang, Nachuan Xing, and Bo Yang. Geovista: Visually grounded active perception for vision-language understanding of ultra-high-resolution remote sensing images, 2026. URL https://arxiv.org/abs/2605.14475.

A APPENDIX

A.1 IMPLEMENTATION AND TRAINING DETAILS

This appendix specifies the implementation details used in the main experiments. Table 6 consolidates the common training, optimization, and inference settings.

Table 6: Training, optimization, and inference hyperparameters used in the main experiments.

Configuration	Setting	Configuration	Setting
Full-scene specialist input	1024^2	Planning grid	32^2
Global planning view	256^2	Local crop	256^2
Maximum candidates	4	Batch size	8
Precision	BF16	Fusion weight λ	0.25
Stage-I optimizer	AdamW	Stage-I epochs	50
Stage-I main / backbone LR	$1e{-4}$ / $5e{-5}$	Stage-I weight decay / clip	$5e{-4}$ / 0.1
Stage-II optimizer	AdamW	Stage-II bootstrap / PPO	5 / 15 epochs
Location-actor / reducer LR	$1e{-5}$	Budget actor–critic LR	$5e{-5}$
Stage-II weight decay / clip	$5e{-4}$ / 0.5	Stage-II LR milestones	10, 16
Location temperature	1.0	Suppression neighborhood	3×3
PPO clip ratio	0.20	PPO update steps	2
PPO critic coefficient	0.5	Initial entropy coefficient	$1e{-3}$
Bootstrap value / policy weights	1.0 / 0.5	Extra-crop cost	0.04
Direct cell / offset weights	1.0 / 0.5	Stage-I loss weights	2 / 20 / 2

A.1.1 STAGE-I CROP-ROBUST SPECIALIST ADAPTATION

The original specialist is trained on resized full-scene images but is later applied to planner-generated local crops. Stage I explicitly reduces this input-distribution mismatch. We construct target-centered crops and independently perturb the crop center along both spatial axes within a bounded range while keeping the referred target visible. The perturbation exposes the specialist to moderate proposal error, scale variation, and reduced context. After 50 epochs of adaptation, the specialist parameters ω are frozen and all subsequent optimization therefore uses a fixed local-evidence model.

Let $c^{\mathrm{gt}} = (c_x^{\mathrm{gt}}, c_y^{\mathrm{gt}})$ be the target center in the image coordinate system. During Stage I training, the crop center is sampled as

$$\tilde{c} = c^{\mathrm{gt}} + (\Delta_x, \Delta_y), \qquad \Delta_x, \Delta_y \overset{\text{i.i.d.}}{\sim} \mathcal{U}(-64, 64). \tag{13}$$

When the target extent along an axis does not exceed the 256-pixel crop size, the sampled center is clipped so that the complete target remains inside the crop; this constraint is relaxed only when the target itself is larger than the crop extent. Stage I uses a binary classification loss, an ℓ_1 box-regression loss, and a generalized-IoU loss with weights 2, 20, and 2, respectively. This stage exposes the specialist to controlled localization error while keeping the supervised local-grounding target valid.

A.1.2 PLANNER PARAMETERIZATION AND PREDICTION FUSION

For a selected planning cell $c_j = (y_j, x_j)$ and axis $d \in \{x, y\}$, the continuous offset head predicts

$$\alpha_j^d = \mathrm{softplus}(u_{c_j}^d) + 1, \qquad \beta_j^d = \mathrm{softplus}(v_{c_j}^d) + 1. \tag{14}$$

During policy training, $\delta_j^d \sim \mathrm{Beta}(\alpha_j^d, \beta_j^d)$ provides bounded within-cell exploration. At inference, we use the Beta mean, so region centers are deterministic while avoiding fixed-grid quantization.

After the single packed specialist pass, let s_{ij}^{local} denote the score of local candidate j from retained crop i, and let a_i be the corresponding planner confidence. The final score is

$$s_{ij} = s_{ij}^{\mathrm{local}} + \lambda\, \mathrm{sg}(a_i), \qquad \lambda = 0.25, \tag{15}$$

where $\mathrm{sg}(\cdot)$ denotes stop-gradient. Every local box is first mapped back to global image coordinates, after which the highest fused score is returned.

A.2 Planner Optimization and One-Shot Inference

This section details Stage II of the supervised-to-PPO training strategy. A supervised bootstrap first establishes stable spatial and budget priors, after which PPO refines the policy under the final accuracy–cost objective. Importantly, the frozen specialist evaluates sampled region plans to provide training signals, but its outputs are never incorporated into the planner state or used to condition subsequent region selections within the same rollout.

A.2.1 Supervised Planner Bootstrap

The ground-truth target center determines the planning-cell target and the within-cell offset target, optimized with cross-entropy and Huber losses, respectively. For candidate budget k, the utility-head target is

$$\hat{U}_k = \mathrm{IoU}_k + 0.10\,\mathrm{Cov}_k + 0.10\,Q_k^{\mathrm{local}} - 0.04(k-1), \tag{16}$$

where IoU_k is the final localization quality obtained from the first k planned regions, Cov_k measures target coverage, and Q_k^{local} summarizes the frozen specialist's local-grounding quality. The final term discourages retaining an additional crop unless it provides sufficient benefit.

The budget module uses separate policy and utility heads during training. For each $\mathcal{P}_{1:k}$, the budget-policy head produces a categorical logit used to sample the discrete budget action and to evaluate its PPO log-probability and entropy. The utility head predicts the end-to-end utility and serves as the value estimator. During supervised bootstrap, the utility head is optimized with Smooth-ℓ_1 regression to $\hat{U}_k$, while the policy head is trained by cross-entropy toward $k^\dagger = \arg\max_k \hat{U}_k$, with loss weights 1.0 and 0.5, respectively. Meanwhile, the spatial actor is initialized by cross-entropy on the ground-truth planning cell and Smooth-ℓ_1 regression on the mean of the within-cell Beta distribution, with weights 1.0 and 0.5. At inference, the training-only budget-policy head is not used for the decision; the deployed budget is selected deterministically from the utility predictions.

A.2.2 PPO Refinement and Rollout Semantics

Starting from the supervised initialization, PPO refines the region policy and utility head using the terminal objective

$$R = R^+ - R^-, \tag{17}$$

with

$$R^+ = R_{\mathrm{IoU}} + 0.10R_{\mathrm{cover}} + 0.10R_{\mathrm{local}}, \tag{18}$$

$$R^- = 0.10C_{\mathrm{overlap}} + 0.50C_{\mathrm{region}} + 0.25C_{\mathrm{latency}}. \tag{19}$$

The positive terms reward final localization, target-containing observations, and useful local evidence; the cost terms penalize overlap, unnecessary cropped area, and measured local-inference latency. The main-text ablation in Table 5 isolates how these terms move the policy from a high-cost IoU-only solution to the final accuracy–efficiency trade-off.

For each training example, the planner first determines a complete region plan from the global image–text state. Region selection depends only on the global state and no local specialist output is available during this process. After the plan is fixed and a candidate budget k is considered, the frozen specialist evaluates the corresponding $\mathcal{S}_k = \{p_1, \ldots, p_k\}$ to compute the terms in Eqs. (16)–(19). These local predictions are used only to evaluate the sampled plan and construct its training reward. PPO uses this reward only after the plan has been completed to update the probability of such plans in future rollouts, without introducing observation-conditioned visual search. At test time, sampling and policy updates are disabled: the global state first determines the ordered region plan and K^*, after which all retained crops are sent to the specialist for local grounding.

For a sampled plan, R_{IoU} is the IoU of the final candidate selected by the fused inference score, R_{cover} is the maximum fraction of the ground-truth box covered by an active crop, and R_{local} is the maximum candidate IoU available from the local observations. The overlap cost is computed from pairwise normalized crop-center distances as an exponentially decaying redundancy penalty, $\exp(-d/0.15)$, averaged over active non-diagonal pairs. The region cost is the summed active crop area normalized by the planning area. The latency term uses the measured local-execution latency, and clips the normalized cost to $[0, 2]$.

During PPO collection, spatial cells are sampled from the categorical location policy and each selected cell suppresses a 3×3 neighborhood before the next action. The horizontal and vertical within-cell offsets are sampled independently from the predicted Beta distributions. The budget-policy head samples $K \in \{1, \ldots, M\}$. For training efficiency, all M proposed crops are evaluated once in a packed batch; the resulting local predictions are reused to compute the return of every plan without additional specialist calls. The budget advantage is computed from the sampled value return and its predicted utility. Spatial actions receive marginal credit, $R_k - R_{k-1}$, so each additional observation is trained according to its incremental contribution. The utility critic is trained with Smooth-ℓ_1 loss on all value returns.

We use the clipped PPO objective with ratio threshold 0.20 and critic coefficient 0.5. Each rollout is followed by two policy updates: the first uses the actual packed grounding rollout, whereas the second re-evaluates the sampled actions under the updated policy without rerunning the frozen specialist. The entropy coefficient starts at 10^{-3} and is linearly annealed to zero over the first eight PPO epochs. To stabilize the transition from supervised initialization, direct spatial supervision is retained with scale 0.20 at the beginning of PPO and linearly annealed to zero over the first five PPO epochs; direct budget supervision is retained for the first three PPO epochs. The remaining supervised term is combined with the PPO actor–critic objective using coefficient 0.25.

A.2.3 One-Shot Inference Algorithm

Algorithm 1 states the execution order explicitly.

Algorithm 1 EfficientViS inference: OCRP followed by ACPI

Require: image I, expression q, Encoder F_θ, utility head V_φ, specialist G_ω, maximum budget $M = 4$
Ensure: final global bounding box b^*

```
Preprocess I and form the fixed-resolution global planning view
Encode the global image–text state with F_θ
Initialize ordered plan P ← ∅ and suppression mask
for j = 1 to M do
    Select the highest-scoring unsuppressed planning cell
    Refine its center with the deterministic continuous offset
    Append the region and planner confidence to P
    Suppress the local neighborhood of the selected cell
for k = 1 to M do
    Form S_k = {p_1, ..., p_k} and evaluate its utility with V_φ
Select K* = arg max_k V_φ(S_k)
Crop and pack the first K* regions along the batch dimension
Run G_ω once on the packed crops and query q
Map all local boxes to global coordinates and fuse their scores using Eq. (15)
Return the highest-scoring global box b*
```

A.3 Full Analysis of Feedback-Free Planning

This section provides the complete derivation underlying Eqs. (4) and (5).

A.3.1 Setup and Residual Spatial Information

Let Z be the global image–text state, $O_{<j}$ the local evidence available to a sequential method before its j-th spatial decision, and $R_j^\dagger$ a latent target-relevant region whose observation would be useful at that decision. A feedback-conditioned planner may reason from $p(R_j^\dagger \mid Z, O_{<j})$, whereas a feedback-free planner has access only to $p(R_j^\dagger \mid Z)$. We define the residual information in previous local observations by

$$\mathcal{I}_j = I(R_j^\dagger; O_{<j} \mid Z). \tag{20}$$

We use natural logarithms in the mutual-information definition. If $\mathcal{I}_j$ is small, it means that, once the global state is known, previous local observations change the posterior over the next useful spatial region only weakly on average.

A.3.2 POSTERIOR-SHIFT BOUND

Proposition 1. For every decision step j,

$$\mathbb{E}_{Z,O_{<j}}\Big[D_{\mathrm{TV}}\Big(p(R_j^\dagger \mid Z, O_{<j}), p(R_j^\dagger \mid Z)\Big)\Big] \le \sqrt{\frac{\mathcal{I}_j}{2}}. \tag{21}$$

Proof. By the conditional-mutual-information identity,

$$\begin{aligned}\mathcal{I}_j &= I(R_j^\dagger; O_{<j} \mid Z)\\ &= \mathbb{E}_{Z,O_{<j}}\Big[D_{\mathrm{KL}}\Big(p(R_j^\dagger \mid Z, O_{<j}) \| p(R_j^\dagger \mid Z)\Big)\Big].\end{aligned} \tag{22}$$

For a fixed realization $(z, o_{<j})$, define

$$P_{z,o}(r) = p(R_j^\dagger = r \mid z, o_{<j}), \qquad Q_z(r) = p(R_j^\dagger = r \mid z). \tag{23}$$

Pinsker's inequality applies pointwise and gives

$$D_{\mathrm{TV}}(P_{z,o}, Q_z) \le \sqrt{\frac{1}{2} D_{\mathrm{KL}}(P_{z,o} \| Q_z)}. \tag{24}$$

Taking expectation over $(Z, O_{<j})$ yields

$$\mathbb{E}[D_{\mathrm{TV}}] \le \mathbb{E}\left[\sqrt{\tfrac{1}{2} D_{\mathrm{KL}}}\right]. \tag{25}$$

Because $x \mapsto \sqrt{x}$ is concave, Jensen's inequality gives

$$\mathbb{E}\left[\sqrt{\tfrac{1}{2} D_{\mathrm{KL}}}\right] \le \sqrt{\tfrac{1}{2}\,\mathbb{E}[D_{\mathrm{KL}}]}. \tag{26}$$

Substituting Eq. (22) into the right-hand side gives

$$\mathbb{E}[D_{\mathrm{TV}}] \le \sqrt{\frac{\mathcal{I}_j}{2}}, \tag{27}$$

which proves Eq. (21).

A.3.3 BAYES-VALUE BOUND

We next translate posterior shift into decision value. Let $u(a, r) \in [0, 1]$ be a normalized utility for selecting observation action a when the useful latent region is r. Define

$$\begin{aligned}V^*_{\mathrm{fb},j} &= \mathbb{E}_{Z,O_{<j}}\Big[\max_a \mathbb{E}[u(a, R_j^\dagger) \mid Z, O_{<j}]\Big],\\ V^*_{\mathrm{ff},j} &= \mathbb{E}_{Z}\Big[\max_a \mathbb{E}[u(a, R_j^\dagger) \mid Z]\Big].\end{aligned} \tag{28}$$

Here "fb" denotes feedback-conditioned planning and "ff" feedback-free planning.

Corollary 1. The Bayes-optimal expected gain from conditioning decision j on previous local evidence satisfies

$$0 \le V^*_{\mathrm{fb},j} - V^*_{\mathrm{ff},j} \le \sqrt{2\mathcal{I}_j}. \tag{29}$$

Proof. For the lower bound, let $a_Q(z)$ be an action attaining the maximum in the feedback-free value. A feedback-conditioned planner can always ignore $O_{<j}$ and choose $a_Q(z)$; therefore

$$V^*_{\mathrm{fb},j} \ge \mathbb{E}_{Z,O_{<j}}\Big[\mathbb{E}[u(a_Q(Z), R_j^\dagger) \mid Z, O_{<j}]\Big] \tag{30}$$

$$= \mathbb{E}_Z\Big[\mathbb{E}[u(a_Q(Z), R_j^\dagger) \mid Z]\Big] = V^*_{\mathrm{ff},j}, \tag{31}$$

where the equality uses the tower property.

For the upper bound, fix $(z, o_{<j})$ and write $P = P_{z,o}$, $Q = Q_z$, and $\delta = D_{\mathrm{TV}}(P, Q)$. Let a_P and a_Q maximize expected utility under P and Q, respectively. Since $0 \le u \le 1$, the variational characterization of total variation implies, for every action a,

$$|\mathbb{E}_P[u(a, R)] - \mathbb{E}_Q[u(a, R)]| \le \delta. \tag{32}$$

We then decompose the conditional advantage of a_P over a_Q under P:

$$\begin{aligned}
&\mathbb{E}_P[u(a_P, R)] - \mathbb{E}_P[u(a_Q, R)] && (33)\\
&= \underbrace{\mathbb{E}_P[u(a_P, R)] - \mathbb{E}_Q[u(a_P, R)]}_{\le \delta} + \underbrace{\mathbb{E}_Q[u(a_P, R)] - \mathbb{E}_Q[u(a_Q, R)]}_{\le 0} && (34)\\
&\quad + \underbrace{\mathbb{E}_Q[u(a_Q, R)] - \mathbb{E}_P[u(a_Q, R)]}_{\le \delta} \le 2\delta. && (35)
\end{aligned}$$

The middle term is non-positive because a_Q is optimal under Q. Averaging Eq. (35) over $(Z, O_{<j})$ gives exactly $V^*_{\mathrm{fb},j} - V^*_{\mathrm{ff},j}$ on the left: a_P realizes the feedback-conditioned optimum, while averaging the $a_Q(Z)$ term over $O_{<j}$ reduces to the feedback-free value by the tower property. Therefore,

$$V^*_{\mathrm{fb},j} - V^*_{\mathrm{ff},j} \le 2\,\mathbb{E}[D_{\mathrm{TV}}] \le 2\sqrt{\frac{\mathcal{I}_j}{2}} = \sqrt{2\mathcal{I}_j}, \tag{36}$$

where the second inequality is Proposition 1. This proves Eq. (29).

A.3.4 Scope of the Bound and Connection to EfficientViS

Corollary 1 is an expected, Bayes-optimal statement. However, even under Bayes-optimal decision rules, the realized gain

$$g_j = u(a_P, R_j^\dagger) - u(a_Q, R_j^\dagger) \tag{37}$$

may be negative. This is particularly relevant for learned sequential policies, where ambiguous crops, erroneous local predictions, or miscalibrated confidence can cause the policy to overreact to an early observation and propagate that error into subsequent region selections. In the qualitative comparison in Figure 6, an incorrect early local hypothesis under GeoViS-Search is followed by later regions that do not recover the referent, illustrating how errors can persist along a sequential search trajectory.

EfficientViS is designed around this separation between allocation and discrimination. OCRP handles allocation from global evidence, while ACPI performs adaptive local verification after the plan is decided. Local evidence remains fully available to the specialist for discrimination: once the candidate regions are fixed, high-resolution crops are used to identify the referred object. What EfficientViS removes is only the recursive dependence of future allocation decisions on intermediate local predictions. When $\mathcal{I}_j$ is small, Corollary 1 bounds the Bayes-optimal expected value lost by imposing this restriction. At the same time, eliminating the feedback loop prevents errors in intermediate local predictions from being recursively propagated into later region choices.

Finally, the coverage–discrimination identity in Eq. (6), which follows directly from the law of total probability, makes the division of labor between the two training stages explicit. Complementary planning and adaptive budgeting primarily increase the probability that the retained regions contain sufficient target evidence, whereas Stage-I crop adaptation improves conditional grounding accuracy given that such evidence has been retained.

A.4 Additional Ablation Analysis

The main text establishes the contribution of the principal planning components, compares fixed and adaptive budgets, analyzes crop size and maximum candidate capacity, and decomposes the PPO reward. The experiments below address two questions that require denser controls: whether the local specialist must be adapted to planner-generated crops, and whether the gains attributed to complementary planning remain after controlling explicitly for crop count.

A.4.1 Stage-I Adaptation and Crop Scale

Stage I addresses a deployment mismatch introduced by EfficientViS: the original specialist is trained on resized full-scene images but is applied to planner-generated local crops at inference. To isolate this effect from planner quality, we keep the trained planner, adaptive budget, crop coordinates, and final inference rule fixed and vary only the local-grounder training strategy.

Table 7: Stage-I crop adaptation with all planning components fixed.

Local-grounder training	RSVG-HR	AVVG	XLRS
No crop adaptation	87.58	31.66	10.03
Target-centered crops	88.18	32.34	11.18
+ bounded center perturbation	**88.58**	**32.96**	**11.97**

Table 7 reports the resulting Acc@0.5. Without crop adaptation, Acc@0.5 is 87.58/31.66/10.03 on RSVG-HR, AVVG, and XLRS-Bench. Target-centered adaptation raises these values to 88.18/32.34/11.18, corresponding to gains of 0.60/0.68/1.15 points. Adding bounded center perturbation provides a further 0.40/0.62/0.79 improvement and reaches the final 88.58/32.96/11.97. The consistent gains show that matching the local-crop distribution materially helps the specialist exploit planned observations, while perturbation further improves robustness to imperfect planner alignment. The larger improvement on XLRS-Bench is consistent with the stronger need to preserve discriminative local evidence in larger scenes.

A.4.2 Candidate Budget and Complementary Selection

To separate observation diversity from raw crop count, we compare vanilla Top-K, Top-K with complementary suppression, an adaptive policy before PPO refinement, and the full PPO-refined policy. Table 8 reports Acc@0.5 and FPS. Vanilla Top-K exhibits an accuracy–throughput trade-off: increasing the fixed budget from one to four crops raises Acc@0.5 by 1.40/0.97/0.98 on RSVG-HR, AVVG, and XLRS-Bench, but reduces throughput from 168.15/187.57/176.08 FPS to 79.36/86.75/81.45 FPS. Diversity matters even at the same nominal budget: at $K = 4$, complementary suppression improves accuracy over vanilla Top-K by 1.81/0.43/0.67 points. Adaptive budgeting then avoids applying the largest budget uniformly. PPO raises Acc@0.5 from 88.18/31.76/10.12 to 88.58/32.96/11.97 and increases FPS from 116.20/130.32/122.34 to 122.01/136.84/128.46. PPO refinement then delivers the final accuracy–efficiency balance. Overall, the results show that EfficientViS benefits from complementary evidence, query-adaptive budgets, and reward-driven policy refinement, not simply from observing more regions.

Table 8: Effect of candidate budget and complementary selection. Each benchmark reports Acc@0.5 and FPS; adaptive rows choose between one and four crops per query.

Selection	Budget	RSVG-HR		AVVG		XLRS	
		Acc@0.5	FPS	Acc@0.5	FPS	Acc@0.5	FPS
Top-K	Top-1	85.17	168.15	30.42	187.57	8.63	176.08
	Top-2	85.77	122.53	31.08	132.94	9.21	124.80
	Top-3	86.17	96.71	31.31	103.98	9.52	97.61
	Top-4	86.57	79.36	31.39	86.75	9.61	81.45
Top-K + suppression	Top-2	86.77	117.96	31.44	132.30	9.78	124.20
	Top-3	87.37	92.18	31.71	103.38	10.11	97.05
	Top-4	88.38	76.90	31.82	86.21	10.28	80.97
Adaptive, no PPO	1–4	88.18	116.20	31.76	130.32	10.12	122.34
EfficientViS (full)	**1–4**	**88.58**	**122.01**	**32.96**	**136.84**	**11.97**	**128.46**

A.5 Failure Modes and Future Work

OCRP can fail when its retained regions do not cover the ground-truth referent. As illustrated in Figure 7, such failures occur when the global planner assigns high scores to plausible but incorrect

regions, leaving the local specialist without sufficient visual evidence for accurate localization. These conditions are particularly challenging for extremely small targets, scenes with repeated distractors, and referring expressions whose discriminative spatial relations are difficult to resolve from the downsampled global view.

Complementary suppression and adaptive budgeting mitigate this failure mode by maintaining multiple spatial hypotheses and allocating additional observations when needed. However, neither mechanism can recover the target if it is absent from all retained regions. This reflects a trade-off between parallel inference and the corrective feedback available to sequential search: decoupling region planning from local grounding enables all selected crops to be processed in a single packed pass, but prevents local observations from revising the region plan. A possible extension is therefore to introduce a lightweight recovery stage, such as a confidence-triggered second search activated when the local predictions are uniformly weak or inconsistent. This would retain single-pass inference for most samples while allowing additional observations in cases where the initial region plan provides insufficient evidence.

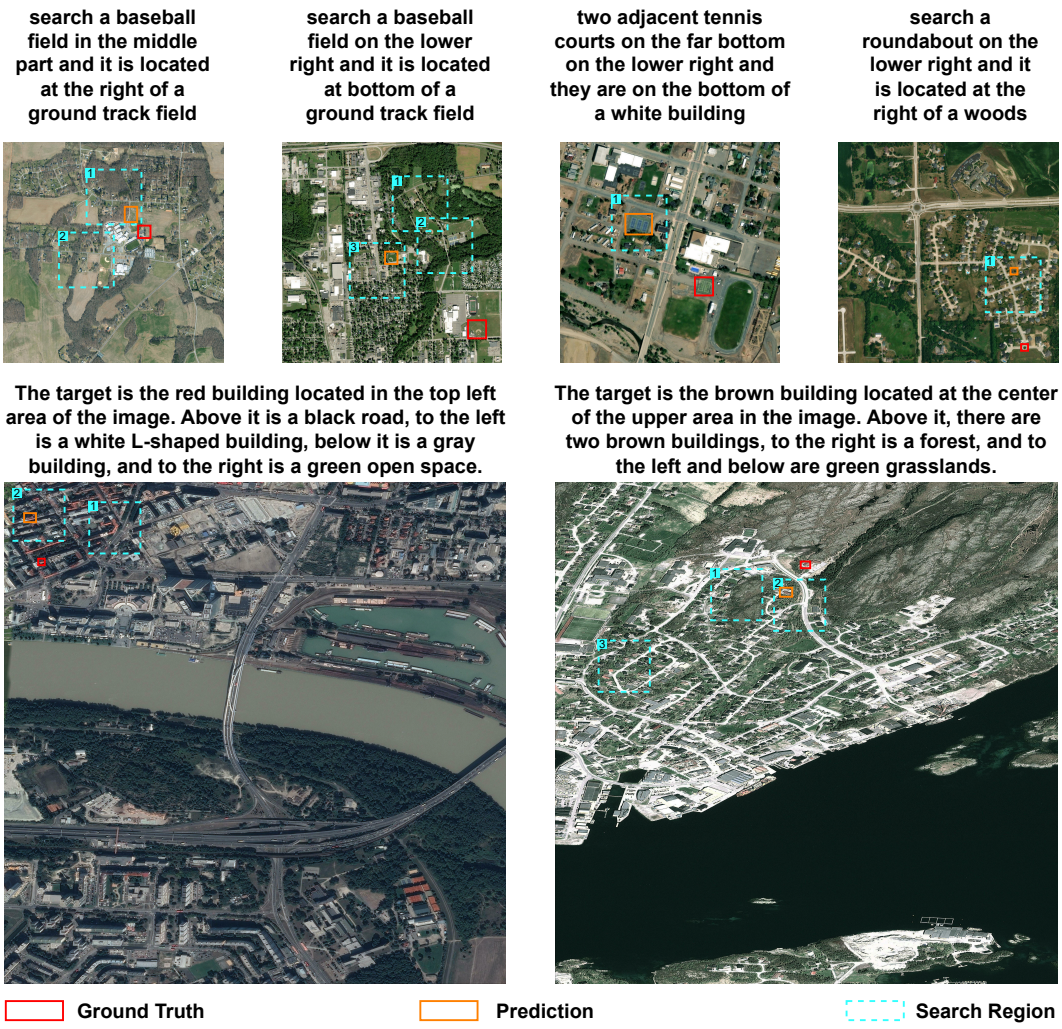

Figure 7: Representative EfficientViS failure cases. Errors arise when the retained global hypotheses miss or insufficiently cover the true referent.