IRSTD-AGENT: AGENTIC INFRARED SMALL TARGET DETECTION VIA ZOOM-GUIDED INTERACTION LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Infrared small-target detection plays an important role in maritime monitoring and aerial surveillance. Although multimodal large language models (MLLMs) offer promising capabilities for visual understanding, existing MLLM-based approaches struggle to precisely localize infrared small targets. In this paper, we propose **IRSTD-Agent**, an agentic framework for infrared small target detection through dynamic visual search. The framework enables an MLLM to adaptively determine where and at what scale to inspect an image and progressively gather fine-grained visual evidence for precise target localization. Five complementary visual tools (PROPOSAL, ZOOM, DETECT, DROP and REFINE) support object candidate discovery, adaptive observation, target localization, hypothesis rejection, and target extent refinement, together enabling a coordinated search process over original-resolution images. To teach the MLLMs to conduct this search, we introduce Zoom-guided Interaction Learning, which uses annotation-derived interaction trajectories to supervise tool selection and the corresponding arguments. Through extensive experiments on WideIRSTD-Full and IRSTD-1k datasets, we demonstrate that IRSTD-Agent outperforms the evaluated vision-language models and enhances the precise localization capabilities of MLLMs in IRSTD tasks.

1 INTRODUCTION

Infrared small target detection (IRSTD) aims to identify and accurately localize dim thermal targets in cluttered backgrounds, supporting applications such as maritime monitoring and aerial surveillance. Although specialized deep learning-based approaches have made substantial progress through feature enhancement, context aggregation, and background suppression (Li et al., 2023; Liu et al., 2024a; Yuan et al., 2026), they are typically developed as task-specific models, thereby limiting their generalization capabilities in visual understanding. With the development of multimodal large language models, it has become possible to possess capabilities similar to human visual intelligence and to handle complex visual understanding tasks (Li et al., 2025b; Yuan et al., 2025).

Existing efforts to improve the capability of MLLMs in IRSTD task can be broadly grouped into two categories: infrared-domain fine-tuning and cue-guided prompting. As shown in Figure 1(a), the first strategy fine-tunes MLLMs on infrared image–text pairs and task-specific instructions to improve their understanding of infrared imagery, such as Infrared-LLaVA and IRGPT (Jiang et al., 2024; Cao et al., 2025). However, they require substantial annotation effort and incur considerable computational overhead for model finetuning. More importantly, many state-of-the-art MLLMs (OpenAI et al., 2023; Team et al., 2023) are proprietary and cannot provide access to model parameters, making finetuning inapplicable to some of the strongest existing models and limiting its scalability to rapidly evolving MLLMs. Another strategy introduces auxiliary language cues to guide infrared small target localization, as shown in Figure 1(b). For example, MIRSAM (Zhang et al., 2025b) incorporates textual descriptions of target location and appearance through CLIP-based prompts to support infrared small target segmentation. Similarly, LGNet (Chib & Singh, 2026) leverages MLLMs to generate descriptions of infrared small target locations and incorporates these descriptions into a dedicated detection network to improve performance. **Although these strategies offer promising directions for adapting MLLMs to IRSTD without parameter updates, the effectiveness of precisely localizing infrared small targets remains to be established**.

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.

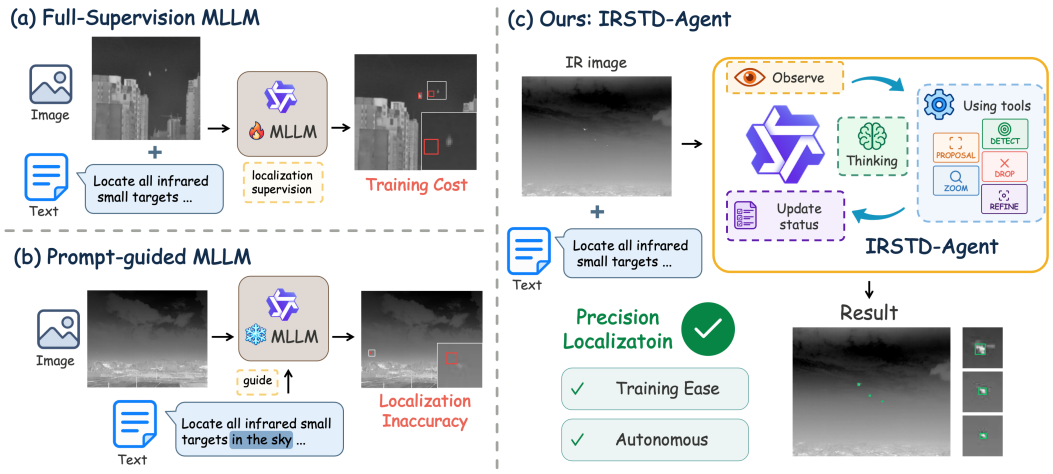

Figure 1: Motivation for IRSTD-Agent. (a) Full-supervision MLLMs adapt to infrared target localization through task-specific supervision, incurring annotation and training costs. (b) Prompt-guided MLLMs use auxiliary language cues without updating model parameters, but fail to precisely localize small infrared targets. (c) Our IRSTD-Agent employs dynamic visual search to acquire local evidence for target recognition and precise localization.

To address these challenges, we propose an **IRSTD-Agent**, a framework that formulates IRSTD as a dynamic visual search process, as illustrated in Figure 1(c). Instead of relying on a single fixed observation, our approach enables an MLLM to actively acquire local visual evidence and adapt its search through sequential tool interactions. Given an infrared image, the agent identifies candidate regions and adaptively determines where and at what scale to inspect, gathering evidence to localize targets and reject background clutter. To support this process, we introduce five complementary visual tools: **PROPOSAL** for candidate discovery, **ZOOM** for adaptive observation, **DETECT** for target localization, **DROP** for clutter rejection, and **REFINE** for target extent refinement. By coordinating these tools, our IRSTD-Agent systematically examines object candidate regions and extracts fine-grained visual evidence from original-resolution images, thereby achieving the recognition and precise localization of infrared small targets through an adaptive search process. To teach the MLLM to guide this search, we introduce **Z**oom-guided **I**nteraction **L**earning (**ZIL**), which uses annotation-derived interaction trajectories to supervise tool selection and the corresponding arguments. Through ZIL, the MLLM learns to select informative views, localize targets, and reject distractors based on the acquired visual evidence. Our contributions are summarized as follows:

- We introduce IRSTD-Agent, the first agentic framework for infrared small target detection, enabling the model to actively acquire local observations and use visual tools for adaptive visual search and precise target localization in cluttered infrared scenes.
- We design five active visual tools (PROPOSAL, ZOOM, DETECT, DROP and REFINE) for IRSTD-Agent to perform dynamic visual search and introduce zoom-guided interaction learning to teach MLLMs to select appropriate search tools and their parameters based on visual evidence, using annotation-derived interaction trajectories as supervision.
- Extensive experiments conducted on the WideIRSTD-Full and IRSTD-1K datasets demonstrate that IRSTD-Agent consistently outperforms existing MLLM-based methods and accurately localizes infrared small targets across various infrared image scenarios.

2 Related Work

Infrared Small-Target Detection. Modern IRSTD methods develop specialized architectures and supervision for weak-target representation and localization. DNANet (Li et al., 2023) strengthens hierarchical feature propagation, MSHNet (Liu et al., 2024a) introduces scale- and location-sensitive supervision, and NS-FPN (Yuan et al., 2026) improves multi-scale feature interaction

through frequency-guided feature fusion. Other studies explore unified encoder–decoder architectures (Wu et al., 2023), target-adapted convolutions (Yang et al., 2025), and learning from sparse point annotations (Yu et al., 2025). AuxDet (Shi et al., 2026) further studies omni-domain IRSTD by incorporating auxiliary acquisition metadata.Recent work has also introduced vision-language priors into IRSTD. SAIST (Zhang et al., 2025a) integrates textual guidance with SAM for infrared small-target segmentation. Overall, these approaches form specialized solutions tailored to infrared small-target detection and are typically deployed through predefined detection pipelines.

Vision-Language Models for Visual Localization. Recent VLMs have extended multimodal understanding to spatial localization and detection. General-purpose MLLMs such as Qwen3.5/3.8 (Qwen Team, 2026a;b) and InternVL3.5 (Wang et al., 2025) can follow visual localization instructions and produce coordinate-based predictions. Earlier grounding models explicitly align natural-language expressions with image regions. MDETR (Kamath et al., 2021) performs end-to-end text-conditioned object detection through joint vision-language reasoning, while GLIP (Li et al., 2022) unifies object detection and phrase grounding within a grounded language-image pre-training framework.Models such as LocateAnything (Wang et al., 2026), SenseNova-Vision (Han et al., 2026), Grounding DINO (Liu et al., 2024b), and Florence-2 (Xiao et al., 2024) further develop language-conditioned grounding and structured spatial prediction.

Visual Chain-of-Thought and Structured Perception Pipelines. Recent work incorporates visual evidence acquisition into multimodal reasoning along two main directions. One line extends chain-of-thought reasoning with intermediate visual interactions, allowing models to identify, revisit, or zoom into relevant regions during reasoning in a more deliberate and iterative manner. Representative approaches include V*, Visual CoT, DeepEyes, and Pixel Reasoner (Wu & Xie, 2024; Shao et al., 2024; Zheng et al., 2026; Su et al., 2025), which study such capabilities across grounding, fine-grained perception, and multimodal reasoning tasks. Another line organizes visual processing through structured workflows. ZoomEye performs hierarchical image exploration through tree search (Shen et al., 2025), ZoomEarth develops an adaptive cropping framework for high-resolution remote-sensing tasks (Liu et al., 2026), and AgentDet coordinates specialized roles through a shared-blackboard detection pipeline (Li et al., 2026).

3 IRSTD-AGENT

IRSTD-Agent formulates infrared small-target detection as an dynamic visual search process over the original-resolution image. Rather than predicting all targets from a single resized observation, the agent repeatedly selects visual tools to discover candidate regions, acquire higher-resolution evidence, localize visible targets, and reject unsupported hypotheses. Figure 2 summarizes this process: the current view, candidate list, and interaction history inform each tool call, whose execution updates the state for subsequent decisions. Target localizations are accumulated across search branches, while unsupported candidates are rejected. We first describe how the agent selects and executes these tools, and then detail each tool and the annotation-derived interaction learning.

3.1 AGENT FORMULATION

Figure 2 summarizes the framework, linking tool-based visual search and the evidence workspace with annotation-derived interaction supervision.

Given an infrared image $I \in \mathbb{R}^{H\times W}$, IRSTD-Agent performs detection through successive visual-tool interactions. The agent is equipped with five tools,

$$\mathcal{U} = \{\texttt{PROPOSAL}, \texttt{ZOOM}, \texttt{DETECT}, \texttt{DROP}, \texttt{REFINE}\}. \tag{1}$$

These tools serve different stages of visual search: PROPOSAL discovers candidate locations, ZOOM acquires a new local observation, DETECT records visible targets, DROP rejects unsupported hypotheses, and REFINE estimates the spatial extent of an existing localization.

At interaction step t, the agent observes the current view $I[C_t]$, where $C_t = (X_t, Y_t, W_t)$ specifies the view center (X_t, Y_t) and window size W_t in the original image. The interaction state S_t maintains the information required to continue the search, including the current view, candidate

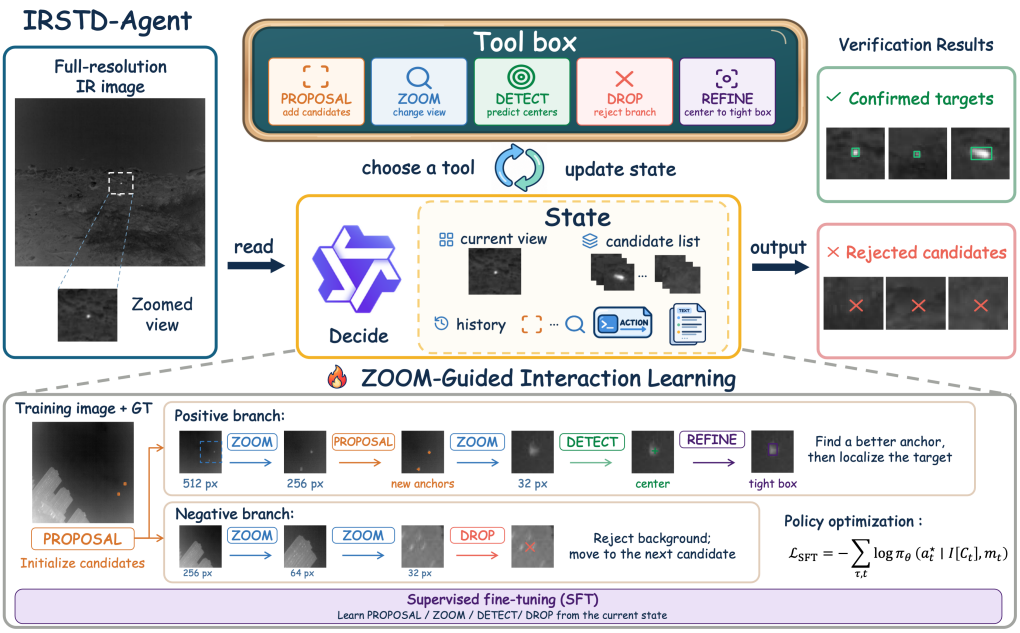

Figure 2: Overview of IRSTD-Agent. The policy interacts with an infrared image through visual tools, acquiring local observations and accumulating target localizations in an evidence workspace. Annotation-derived interaction trajectories supervise tool selection and arguments. Candidate discovery and appearance-guided refinement support the search and localization process.

hypotheses, interaction history, accumulated detections, and remaining budget. The model-visible context m_t is obtained by serializing the information in S_t that is exposed to the policy.

Conditioned on the current view and context, the search policy predicts a structured tool call

$$a_t \sim \pi_\theta(\cdot \mid I[C_t], m_t), \tag{2}$$

where a_t specifies the selected search tool together with its required arguments. The search policy selects from $\mathcal{U}_{\text{search}} = \{\texttt{PROPOSAL}, \texttt{ZOOM}, \texttt{DETECT}, \texttt{DROP}\}$. REFINE operates on an existing detection and is handled separately.

Once a tool is selected, it is executed on the original image and updates the interaction state:

$$S_{t+1} = \mathcal{E}(I, S_t, a_t), \tag{3}$$

where $\mathcal{E}$ denotes the tool executor. The updated state is then serialized into the context m_{t+1}, together with the latest interaction information, for the next agent decision.

3.2 Active Visual Tools

Each visual tool operates directly on native-resolution pixels from the original image I. Given the current view C_t, the tools either update the view, introduce candidate hypotheses, produce target localizations, or refine existing detections, as detailed below. Figure 3 illustrates how successive tool calls update the observations and accumulated evidence to produce target localizations and refined extents.

PROPOSAL: candidate discovery. PROPOSAL discovers potential target locations within the current search region during the active visual search process. Given the native-resolution crop $I[C_t]$, it returns a ranked set of candidate hypotheses:

$$\mathcal{Q}_t = \texttt{PROPOSAL}(I[C_t]) = \big((q_{tk}, s_{tk})\big)_{k=1}^{K_t}, \qquad s_{t1} \geq \cdots \geq s_{tK_t}. \tag{4}$$

Here q_{tk} denotes the k-th candidate location mapped to original-image coordinates, and s_{tk} is its proposal score. The returned candidates are search hypotheses rather than final detections. Internally, the tool combines complementary spatial and frequency responses. The spatial branch

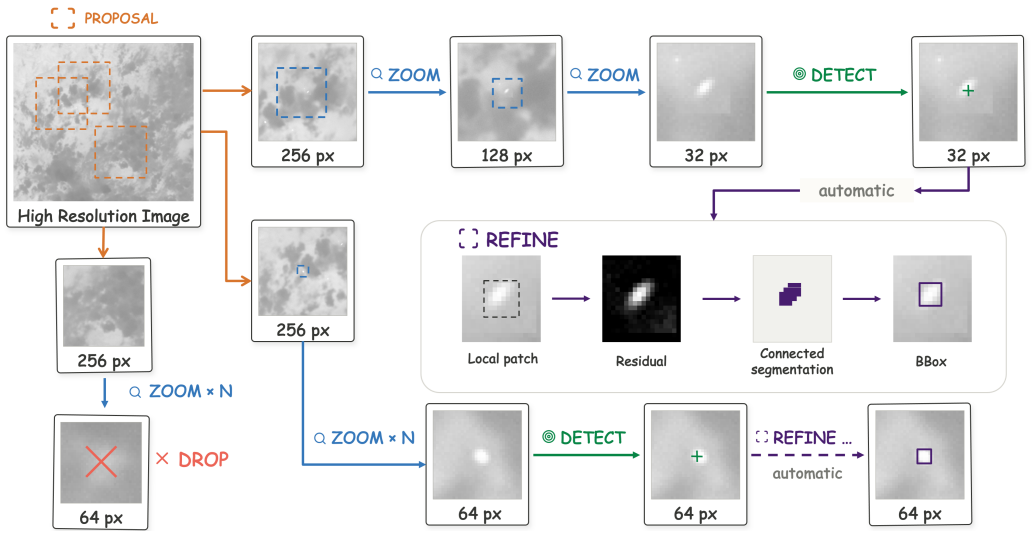

Figure 3: Visual-tool interaction in IRSTD-Agent. The agent iteratively selects tools based on the current view and accumulated interaction context, while tool execution updates the search process and acquired evidence. Detection and refinement produce the final target localizations and extents.

aggregates multi-scale morphological top-hat filtering, difference-of-Gaussians, local standardized contrast, and patch contrast, while the frequency branch uses FFT-based spectral-residual saliency. Their normalized responses are fused as

$$R_t = \lambda R_{\mathrm{sp},t} + (1-\lambda) R_{\mathrm{fft},t}. \tag{5}$$

Candidate peaks are then ranked and spatially suppressed to form $\mathcal{Q}_t$.

ZOOM: evidence acquisition. ZOOM predicts a new crop-relative center (x_t, y_t) and window size w_t for the next observation during the active visual search process. These parameters are applied to the current view C_t to update the search region:

$$C_{t+1} = \texttt{Zoom}(C_t;\, x_t, y_t, w_t)\,. \tag{6}$$

Here (x_t, y_t) specifies where to move within the current view, while w_t controls the scale of the next view. The updated region $I[C_{t+1}]$ is then used in the next agent interaction.

DETECT: target localization. DETECT enables IRSTD-Agent itself to localize the targets visible in the current view. Given the current image region $I[C_t]$, the agent predicts a structured set of normalized target centers:

$$\mathcal{P}_t = \texttt{DETECT}(I[C_t]) = \{p_{tj}\}_{j=1}^{n_t}, \tag{7}$$

where p_{tj} denotes the normalized center of the j-th target predicted by IRSTD-Agent within the current view. The predicted centers are converted to the original-image coordinate system and recorded as image-level detections. The current search branch is then terminated.

DROP: hypothesis rejection. DROP is an argument-free action selected when the current visual evidence does not support the presence of a target:

$$u_t = \texttt{DROP} \quad \Longrightarrow \quad \text{terminate the current search branch without producing a detection.} \tag{8}$$

The rejection is recorded in the interaction history, while no image-level detection is added.

REFINE: appearance-based extent refinement. Given a detected center p_i, REFINE uses its location to recover the target extent from the original image. It first initializes a local box around the center, suppresses the local background, and applies adaptive thresholding to isolate target-supporting responses. The connected region associated with the detected center is then extracted, and its spatial extent is used to obtain the refined bounding box. We denote this image-space transformation as

$$\tilde{b}_i = \texttt{Refine}(I, p_i), \tag{9}$$

where $\tilde{b}_i$ denotes the refined bounding box. The foreground pixels of the extracted connected region define a binary mask M_i, which is mapped back to the original-image coordinates. Taking the union of these masks produces the image-level segmentation map used for visualization and the evaluation in Appendix B.2. Thus, the mask follows the image-derived target region rather than a filled detection box. If no valid region is obtained, the initialized local box is preserved and no mask is added to the segmentation map.

3.3 ZOOM-GUIDED INTERACTION LEARNING

We construct interaction supervision from standard target annotations to teach IRSTD-Agent how to select visual tools and their arguments during search in a supervised manner. Each training trajectory consists of a sequence of agent interactions,

$$\tau = \left\{ \left(I[C_t], m_t, a_t^\star \right) \right\}_{t=0}^{T}, \tag{10}$$

where $I[C_t]$ is the current view, m_t is the accumulated interaction context, and $a_t^\star$ is the supervised tool call at step t. Successive actions therefore connect what the agent currently observes with how it should continue the search.

Annotation-derived interaction trajectories. The lower panel of Figure 2 illustrates supervision through positive and negative search branches: positive trajectories acquire evidence for target localization, whereas negative trajectories terminate with DROP after inspecting background candidates.

For positive target hypotheses, we construct target-preserving ZOOM actions that progressively expose the annotated target at finer spatial scales, followed by DETECT supervision once sufficient target evidence is visible. Proposal-offset examples place the target away from the initial candidate location, preventing the agent from simply copying the proposal position and encouraging localization from the visual evidence itself.

Negative hypotheses are supervised with DROP when the inspected region contains background rather than an annotated target. For views containing multiple visible targets, DETECT supervision includes all valid targets in current view, allowing a single branch to return multiple detections. PROPOSAL supervision is included in the same interaction to teach candidate discovery during search.

The resulting interactions are organized into complete trajectories $\tau = \{(I[C_t], m_t, a_t^\star)\}_{t=0}^{T}$. After trajectory construction and sampling, the retained trajectories form the supervised fine-tuning set

$$\mathcal{D}_{\mathrm{SFT}} = \{\tau_i\}_{i=1}^{N}. \tag{11}$$

Together, these trajectories teach the agent when to propose, where and how far to zoom, when to detect, and when to reject a hypothesis.

Policy optimization. The search policy is optimized over $\mathcal{D}_{\mathrm{SFT}}$ by

$$\mathcal{L}_{\mathrm{SFT}} = -\sum_{\tau \in \mathcal{D}_{\mathrm{SFT}}} \sum_{t} \log \pi_\theta \left(a_t^\star \mid I[C_t], m_t \right). \tag{12}$$

This objective encourages the policy to assign higher probability to the correct tool call $a_t^\star$ given the current view and interaction context, thereby learning both tool selection and its corresponding arguments. Each supervised action is serialized as a structured tool call, while prompt tokens are excluded from the loss during supervised fine-tuning of the policy.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Datasets. We study two complementary settings. WideIRSTD-Full (Li et al., 2025a; Shi et al., 2026) evaluates wide-area target search on its complete 2,000-image test set, preserving each image's native resolution and allowing multiple targets or no targets per image. Its images span a wide range of resolutions and acquisition domains, introducing substantial variation in both scale and scene and making wide-area target search particularly challenging. IRSTD-1k (Zhang et al., 2022) is evaluated using its official split of 800 training images and 201 test images at 512×512 resolution. Because of overlapping in two datasets, the settings are evaluated independently.

Table 1: Visual tool ablation on IRSTD-1k datasets. Checkmarks indicate enabled tools. P, R, and F1-scores are reported in percentages. The best results are highlighted in **bold**.

PROPOSAL	ZOOM	DETECT	DROP	REFINE	P	R	F1
✓	✓	✓			3.61	77.44	6.90
✓		✓	✓	✓	23.89	86.87	37.47
✓	✓	✓	✓		63.78	69.36	66.45
✓	✓	✓	✓	✓	**81.09**	**75.08**	**77.97**

Implementation Details. IRSTD-Agent is built on Qwen3.5-27B, with separate checkpoints independently fine-tuned on WideIRSTD-Full and IRSTD-1k using the zoom-guided SFT. IRSTD-1k uses its independently trained checkpoint, with no weights transferred from WideIRSTD-Full. We apply LoRA with rank 16, $\alpha = 32$, and dropout 0.05, while freezing the vision encoder and multimodal projector. Training is performed for one epoch with a learning rate of 1×10^{-5}, cosine decay, 5% warmup, BF16 computation, and gradient checkpointing on NVIDIA A100 80GB GPUs.

Evaluation Metrics. MLLM base models are evaluated using their native boxes without additional refinement or NMS. Grounding methods, Visual Reasoning and Active Search methods, and IRSTD-Agent use center-based refinement and box NMS. We report precision (P), recall (R), and F1-scores for target detection. Following prior IRSTD evaluations (Li et al., 2023; Chen et al., 2022), a prediction is matched to an annotation if their box centers are within 3 pixels in the original image. Matches are established one-to-one using the same criterion for all methods. We additionally report the mean box IoU over matched targets to evaluate localization extent.indicating that adaptive evidence acquisition is important for reliable target verification. Finally, REFINE further improves the precision–recall balance, raising F1-scores from 66.45% to 77.97%.

4.2 Ablation Studies

All ablations below cover the official IRSTD-1k datasets. Tool and budget experiments use the main-table Qwen3.5-27B checkpoint.

Visual tools. Table 1 shows that the tools play complementary roles in the search process. Without DROP, the agent retains many candidate hypotheses, resulting in high recall (77.44%) but very low precision (3.61%).Introducing DROP substantially suppresses false detections, increasing precision to 63.78% and F1-scores to 66.45%. Replacing adaptive ZOOM with a fixed-scale observation strategy yields higher recall (86.87%) but much lower precision (23.89%) and F1-scores (37.47%),

Zoom-guided Interaction learning. Table 2 compares supervision settings and reports the full system as a reference. Under the same native tool-call execution protocol, the model without SFT (w/o SFT) attains 5.80 precision, 78.11 recall, and 10.80 F1-scores; SFT without the negative branch (w/o negative branch) yields 6.72 precision, 78.79 recall, and 12.38 F1-scores. Positive-only supervision thus provides limited improvement in this setting. The full zoom-guided SFT achieves 81.09 precision and 77.97 F1-scores, delivering the best performance.

Table 2: Ablation of zoom-guided interaction learning on IRSTD-1k datasets, P/R/F1-scores are percentages.

Training	P	R	F1
w/o SFT	5.80	78.11	10.80
w/o negative branch	6.72	78.79	12.38
Ours	**81.09**	**75.08**	**77.97**

Action budget. Figure 4 varies both model-call and tool-call caps over B={8,16,32,64,96,128}, with the checkpoint and proposal/detection caps fixed. F1-scores increases from 69.28 to 79.45 and recall from 56.57 to 77.44. However, F1-scores rises by 0.93 points from 64 to 96 despite a substantial increase in inference time,and by 0.61 points from 96 to 128,indicating diminishing returns.These results indicate that $B = 64$ provides a favorable trade-off between detection performance and action budget.

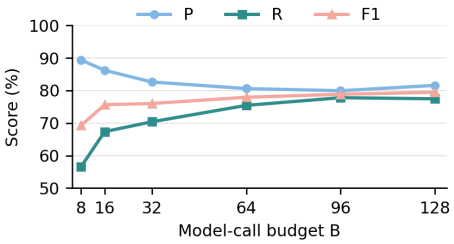

Figure 4: Inference-budget ablation on IRSTD-1k with $B = 8, 16, 32, 64, 96, 128$, varying the maximum tool-call budget.

Table 3: Detection results on WideIRSTD-Full and IRSTD-1k. The shared Method and Venue columns are followed by dataset-specific precision, recall, F1-scores, and IoU; values are percentages.

Method	Venue	WideIRSTD-Full				IRSTD-1k			
		P↑	R↑	F_1 ↑	IoU↑	P↑	R↑	F_1 ↑	IoU↑
Open-Source MLLMs									
InternVL3.5-38B	'25	10.92	4.04	5.89	34.32	6.49	4.04	4.98	41.84
Qwen3.5-9B	'26	7.14	5.16	5.99	36.07	23.85	20.88	22.26	40.84
Qwen3.5-27B	'26	13.77	6.85	9.15	38.68	41.04	34.68	37.59	44.54
Qwen3.8-27B	'26	30.03	15.02	20.02	43.14	63.79	49.83	55.95	48.54
LocateAnything-3B	'26	32.22	17.06	22.31	50.03	79.02	38.05	51.36	54.40
SenseNova-Vision-7B-MoT	'26	3.43	2.48	2.88	42.00	28.07	5.39	9.04	59.67
VLM-Based Grounding Models									
Florence-2-large-ft	CVPR '24	6.90	3.66	4.78	64.39	11.43	14.81	12.90	68.22
Grounding DINO-B	ECCV '24	9.33	5.50	6.92	66.85	10.93	11.45	11.18	69.85
YOLO-World	CVPR '24	13.48	6.57	8.83	67.99	6.82	75.08	12.51	69.65
Rex-Omni	CVPR '26	16.04	39.06	22.74	67.01	48.37	75.08	58.84	70.34
Visual Reasoning and Active Search Methods									
ZoomEye	EMNLP '25	19.01	34.74	24.58	65.56	48.40	**81.48**	60.73	69.07
RSCoVLM-det	RS '26	46.38	22.35	30.17	63.42	77.70	72.73	75.13	69.70
ZoomEarth	CVPR '26	15.97	4.37	6.87	64.72	44.62	29.29	35.37	69.79
DeepEyes	ICLR '26	4.34	4.30	4.32	62.28	26.90	28.62	27.73	69.01
VisionReasoner	ICLR '26	18.78	21.98	20.25	66.63	37.73	54.88	44.72	68.33
IRSTD-Agent	**Ours**	**71.63**	**34.83**	**46.87**	**67.07**	**81.09**	75.08	**77.97**	**71.56**

4.3 Comparison with SOTA Methods

Baseline Setup. *Open-Source MLLMs* are evaluated by directly prompting them to detect infrared targets from the full image and return bounding boxes. For models with their own coordinate conventions, the outputs are normalized to a common image-coordinate format before evaluation; for example, Qwen predictions follow its native 0–1000 coordinate protocol and are mapped back to the original image size. *VLM-Based Grounding Models* support multi-object localization or open-vocabulary detection, so they are directly queried with an infrared-target prompt and evaluated on all returned target boxes in the full image. *Visual Reasoning and Active Search Methods* are adapted to IRSTD through method-specific search, cropping, or reasoning interfaces. In particular, ZoomEye uses Qwen3.5-27B for hierarchical visual search before producing detections. Other methods receive infrared-target queries and structured detection outputs; Appendix A details their training, inference, and output-processing differences.

WideIRSTD-Full datasets. Direct full-image perception remains challenging on WideIRSTD-Full because tiny targets occupy only a small fraction of large infrared images and their weak responses can be lost during visual compression. As shown in Table 3, Qwen3.5-27B and Qwen3.8-27B achieve 9.15 and 20.02 F1-scores, while LocateAnything-3B reaches 22.31. Grounding models show a similar limitation; Rex-Omni attains 39.06 recall but only 22.74 F1-scores. In contrast, search-based methods benefit from acquiring local evidence. ZoomEye reaches 24.58 F1-scores, RSCoVLM-det reaches 30.17, and IRSTD-Agent achieves 46.87 F1-scores with 71.63 precision. Compared with direct Qwen3.5-27B inference using the same backbone, this is a 37.72-point F1-scores gain, supporting the importance of active local observation in wide-area IRSTD.

IRSTD-1k datasets. On IRSTD-1k, the fixed 512×512 resolution substantially reduces the scale mismatch between the original image and the model input. Table 3 shows that direct full-image models therefore become much stronger: Qwen3.8-27B reaches 55.95 F1-scores and Rex-Omni reaches

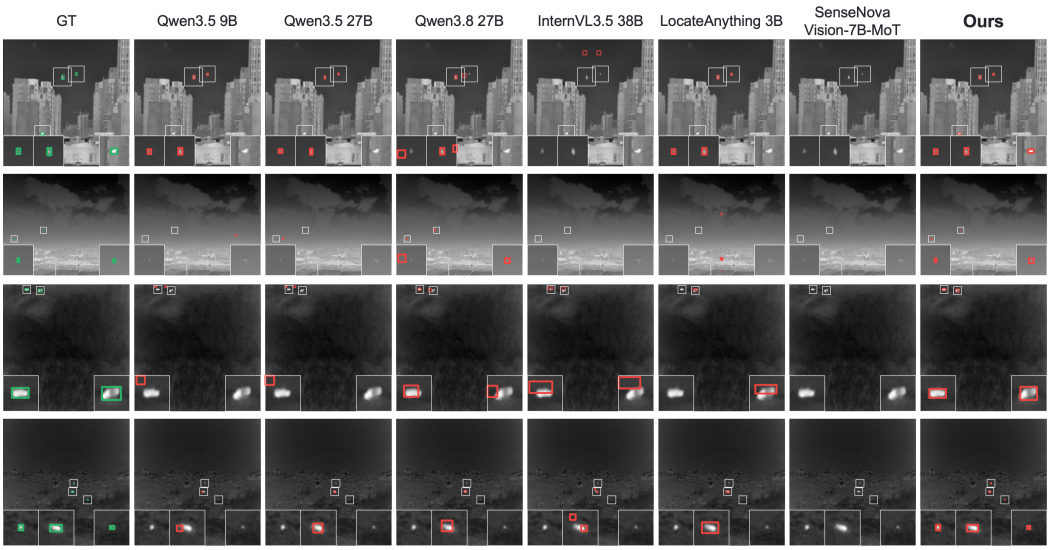

Figure 5: Selected qualitative multi-target detection examples from WideIRSTD-Full (top two rows) and IRSTD-1k (bottom two rows) datasets. Columns compare ground truth, six full-image VLM baselines, and our method. Green boxes indicate ground truth and red boxes indicate predictions. White boxes mark regions enlarged in the insets along the bottom of each image.

58.84. Active search remains beneficial, with ZoomEye reaching 60.73 F1-scores and RSCoVLM-det 75.13. IRSTD-Agent further achieves 81.09 precision, 75.08 recall, and 77.97 F1-scores. Compared with the WideIRSTD-Full results, this contrast suggests that active evidence acquisition is especially important when large image scale makes a single global observation insufficient.

4.4 Qualitative analysis.

Figure 5 presents representative multi-target detection results on WideIRSTD-Full (top two rows) and IRSTD-1k (bottom two rows). The cyan regions highlight target-containing areas and are enlarged in the bottom insets, where green boxes denote ground-truth targets and red boxes denote model predictions. Direct full-image MLLMs frequently miss faint targets or produce spatially offset detections. In contrast, IRSTD-Agent recovers more weak targets through local visual search and yields more accurate localization in these challenging regions. Nevertheless, the examples also show that precise localization remains difficult for extremely small and visually ambiguous targets.

5 Conclusion

In this work, we investigate the challenges of infrared small target detection with multimodal large language models and highlight the importance of actively acquiring fine-grained visual evidence for precise localization. We propose IRSTD-Agent, an agentic framework that formulates detection as a dynamic visual search process. Five complementary visual tools enable candidate discovery, adaptive observation, target localization, hypothesis rejection, and target extent refinement. We further introduce Zoom-guided Interaction Learning, which uses annotation-derived interaction trajectories to teach the model to select appropriate tools and their arguments. Extensive experiments on WideIRSTD-Full and IRSTD-1k demonstrate the effectiveness of our framework, demonstrating detection performance superior to other state-of-the-art MLLM-based methods.

Limitations and Future Work. Sequential tool interactions introduce additional inference overhead, and extremely small or visually ambiguous targets in IRSTD task remain challenging. Future work could investigate more efficient search policies and adaptive computation allocation to improve the balance between detection accuracy and computational cost.

REFERENCES

Zhe Cao, Jin Zhang, and Ruiheng Zhang. Irgpt: Understanding real-world infrared image with bi-cross-modal curriculum on large-scale benchmark. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 166–176. IEEE, 2025.

Gao Chen, Weihua Wang, and Sirui Tan. IRSTFormer: A hierarchical vision transformer for infrared small target detection. *Remote Sensing*, 14(14):3258, 2022. doi: 10.3390/rs14143258.

Pranav Singh Chib and Pravendra Singh. Leveraging language prior for infrared small target detection. *Neurocomputing*, pp. 133512, 2026.

Xiaoyang Han, Jianhua Li, Kewang Deng, Zukai Chen, Xuanke Shi, Sihan Wang, Boxuan Li, Linyan Wang, Siyi Xie, Xin You, Jinsheng Quan, Zhongang Cai, Haiwen Diao, Ziwei Liu, Lei Yang, Dahua Lin, and Quan Wang. Vision as unified multimodal generation. *arXiv preprint arXiv:2607.06560*, 2026.

Shixin Jiang, Zerui Chen, Jiafeng Liang, Yanyan Zhao, Ming Liu, and Bing Qin. Infrared-llava: Enhancing understanding of infrared images in multi-modal large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 8573–8591, 2024.

Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr - modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1780–1790, October 2021.

Boyang Li, Chao Xiao, Longguang Wang, Yingqian Wang, Zaiping Lin, Miao Li, Wei An, and Yulan Guo. Dense nested attention network for infrared small target detection. *IEEE Transactions on Image Processing*, 32:1745–1758, 2023. doi: 10.1109/TIP.2022.3199107.

Boyang Li et al. Icpr 2024 competition on resource-limited infrared small target detection challenge: Methods and results. In *Pattern Recognition. Competitions*, volume 15334 of *Lecture Notes in Computer Science*, pp. 62–77. Springer, 2025a. doi: 10.1007/978-3-031-80139-6_5.

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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9098–9108. IEEE, 2025b.

Haolin Li, Yaohua Wang, Ze Yan, Lijie Wen, and Biqing Huang. Agentdet: A shared-blackboard multi-agent framework for zero-/few-shot object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 41626–41635, 2026.

Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10965–10975, June 2022.

Qiankun Liu, Rui Liu, Bolun Zheng, Hongkui Wang, and Ying Fu. Infrared small target detection with scale and location sensitivity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17490–17499, 2024a.

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*, pp. 34877–34888, 2026.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, pp. 38–55, 2024b. doi: 10.1007/978-3-031-72970-6_3.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Qwen Team. Qwen3.5-27B model card. Official model card, Hugging Face, 2026a. Accessed September 20, 2026.

Qwen Team. Qwen3.8-27B model card. Official model card, Hugging Face, 2026b. Accessed September 20, 2026.

Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 8612–8642. Curran Associates, Inc., 2024. doi: 10.52202/079017-0275.

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 *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 6602–6618. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.335.

Yangting Shi, Yinfei Zhu, Renjie He, Le Hui, Meng Cai, Ming-Ming Cheng, and Yimian Dai. Auxdet: Auxiliary metadata matters for omni-domain infrared small target detection. *IEEE Transactions on Geoscience and Remote Sensing*, 64:1–16, 2026. doi: 10.1109/TGRS.2026.3674761.

Alex Su, Haozhe Wang, Weiming Ren, Fangzhen Lin, and Wenhu Chen. Pixel reasoner: Incentivizing pixel space reasoning via curiosity-driven reinforcement learning. 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. 8222–8251. Curran Associates, Inc., 2025. doi: 10.52202/085713-0280.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Shihao Wang, Shilong Liu, Yuanguo Kuang, Xinyu Wei, Yangzhou Liu, Zhiqi Li, Yunze Man, Guo Chen, Andrew Tao, Guilin Liu, Jan Kautz, Lei Zhang, and Zhiding Yu. LocateAnything: Fast and high-quality vision-language grounding with parallel box decoding. *arXiv preprint arXiv:2605.27365*, 2026.

Weiyun Wang, Zhangwei Gao, Lixin Gu, et al. InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

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. doi: 10.1109/CVPR52733.2024.01243.

Xin Wu, Danfeng Hong, and Jocelyn Chanussot. Uiu-net: U-net in u-net for infrared small object detection. *IEEE Transactions on Image Processing*, 32:364–376, 2023. doi: 10.1109/TIP.2022.3228497.

Bin Xiao, Haiping Wu, Weijian Xu, Xiyang Dai, Houdong Hu, Yumao Lu, Michael Zeng, Ce Liu, and Lu Yuan. Florence-2: Advancing a unified representation for a variety of vision tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4818–4829, 2024.

Jiangnan Yang, Shuangli Liu, Jingjun Wu, Xinyu Su, Nan Hai, and Xueli Huang. Pinwheel-shaped convolution and scale-based dynamic loss for infrared small target detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 9202–9210, 2025. doi: 10.1609/aaai.v39i9.32996.

Chuang Yu, Jinmiao Zhao, Yunpeng Liu, Sicheng Zhao, Yimian Dai, and Xiangyu Yue. From easy to hard: Progressive active learning framework for infrared small target detection with single point supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2588–2598, 2025.

Maoxun Yuan, Bo Cui, Tianyi Zhao, Jiayi Wang, Shan Fu, Xue Yang, and Xingxing Wei. Unirgb-ir: A unified framework for visible-infrared semantic tasks via adapter tuning. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 2409–2418, 2025.

Maoxun Yuan, Duanni Meng, Ziteng Xi, Tianyi Zhao, Shiji Zhao, Yimian Dai, and Xingxing Wei. Seeing through the noise: Improving infrared small target detection and segmentation from noise suppression perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27783–27792, June 2026.

Mingjin Zhang, Rui Zhang, Yuxiang Yang, Haichen Bai, Jing Zhang, and Jie Guo. Isnet: Shape matters for infrared small target detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 877–886, 2022.

Mingjin Zhang, Xiaolong Li, Fei Gao, Jie Guo, Xinbo Gao, and Jing Zhang. Saist: Segment any infrared small target model guided by contrastive language-image pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9549–9558, June 2025a.

Mingjin Zhang, Qian Xu, Yuchun Wang, Xi Li, and Haojuan Yuan. Mirsam: Multimodal vision-language segment anything model for infrared small target detection. *Visual Intelligence*, 3:4, 2025b. doi: 10.1007/s44267-025-00075-0.

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 C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 126775–126798, 2026.