SATLAERS: A RESOURCE-EFFICIENT REMOTE SENSING VLM FOR ONBOARD DISASTER UNDERSTANDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Efficient remote sensing vision-language models (RS-VLMs) can enable onboard first-look disaster understanding from the latest available high-resolution optical observation, providing multi-granularity information for rapid response. Recent RS-VLMs have made substantial progress in unified multi-task understanding and model efficiency, yet disaster-oriented models remain comparatively limited and are often specialized to individual tasks. We introduce SatLAERS, a granularity-aware adaptive RS-VLM for resource-efficient adaptation and onboard inference. Built on a frozen vision encoder and compact language backbone, SatLAERS allocates multimodal computation according to task granularity and sample complexity, combining lightweight multi-scale visual adaptation and query-adaptive visual routing with sparse low-rank experts. A progressive three-stage pipeline proceeds from efficient vision-language alignment and multi-granularity instruction tuning to budget-aware reinforcement learning, jointly balancing task performance and computational cost. We further introduce SatDISA, a single-observation disaster benchmark with 22,862 high-resolution images and 316,698 vision-language instructions across six disaster types and four semantic granularities, spanning image-level understanding, region-level interpretation, pixel-level grounding, and disaster-oriented situational understanding. SatLAERS completes domain adaptation on 8 NVIDIA RTX 3090 GPUs in 49.3 GPU-hours, outperforms the strongest comparable-scale adapted baseline by 1.55 points, and achieves 16.99 s end-to-end latency with 4.461 GB peak RAM on Jetson Orin NX Super 16 GB, supporting resource-constrained onboard deployment.

1 INTRODUCTION

High-resolution Earth observation is increasingly important for disaster monitoring and emergency response, enabling timely, wide-area situational awareness and supporting rapid assessment and decision-making Wang et al. (2025a). Advances in onboard computing are shifting satellite systems from the conventional acquisition–downlink–ground-processing paradigm toward onboard intelligence, allowing observations to be interpreted closer to the point of acquisition and reducing the latency and communication burden associated with raw-image transmission Giuffrida et al. (2022); Guerrisi et al. (2023); Victoria et al. (2026). During the early stages of a disaster, before complete multi-temporal observations become available, the latest high-resolution optical image often provides the most immediate source of first-look information. Onboard remote sensing vision-language models (RS-VLMs) can transform such observations into compact, semantically rich disaster information, enabling timely interpretation and response under constrained communication and computing resources.

RS-VLMs have rapidly expanded from image captioning and visual question answering to spatial grounding, pixel-level understanding, and broader geospatial reasoning, increasingly enabling unified remote sensing (RS) interpretation Kuckreja et al. (2023); Shabbir et al. (2025); Liu & Lian (2026). Disaster-oriented RS-VLM research, however, remains comparatively limited, with recent efforts largely focusing on disaster-specific assessment, benchmarks, and domain adaptation rather than unified resource-efficient onboard understanding Wang et al. (2025a; 2026b); Tehrani et al. (2026); Ayanzadeh et al. (2026). First-look disaster interpretation requires capabilities across multiple semantic granularities, spanning image-level understanding, region-level interpretation, pixel-level grounding, and disaster-oriented situational understanding. While recent general-purpose RS-

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.

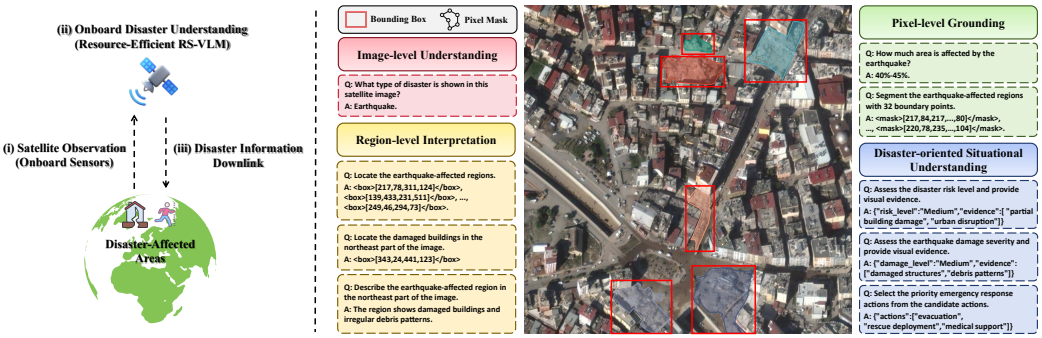

Figure 1: **Overview of SatLAERS for onboard disaster understanding.** SatLAERS supports unified image-, region-, pixel-, and situational-level understanding while adaptively allocating visual and language-side computation under constrained adaptation and inference budgets.

VLMs increasingly unify subsets of these capabilities, they are not explicitly designed for early-stage disaster understanding from the latest single observation under onboard resource constraints. This motivates a compact RS-VLM that jointly supports multi-task, multi-granularity disaster understanding and adaptive resource allocation for efficient adaptation and onboard inference.

To address this challenge, we introduce SatLAERS, a granularity-aware adaptive RS-VLM for resource-efficient adaptation and onboard disaster understanding. SatLAERS combines a frozen SigLIP2-L/16 Tschannen et al. (2025) vision encoder with a compact Qwen3-1.7B Yang et al. (2025) language backbone, preserving pretrained visual and linguistic knowledge while concentrating adaptation on lightweight trainable modules. A multi-scale visual adapter constructs multi-scale visual representations, while query-adaptive visual routing dynamically selects visual representations according to the semantic granularity and information requirements of each query. On the language side, sparse low-rank experts augment the shared backbone with task-adaptive capacity while activating only a small subset of additional parameters for each token. SatLAERS is trained progressively through efficient vision–language alignment, unified multi-granularity instruction tuning, and budget-aware reinforcement learning. The final stage explicitly encourages the model to balance task utility against computational cost, enabling adaptive allocation of visual- and language-side computation across heterogeneous disaster tasks.

To support this study, we introduce SatDISA, a large-scale benchmark for single-observation, multi-granularity disaster understanding, comprising 22,862 high-resolution optical images and 316,698 vision-language instructions across six representative disaster types: wildfire, flood, earthquake, landslide, tsunami, and volcanic eruption. SatDISA spans four semantic granularities: image-level understanding, region-level interpretation, pixel-level grounding, and disaster-oriented situational understanding, with visual-necessity filtering, hard-sample augmentation, and cross-granularity consistency validation to improve instruction quality. We benchmark SatLAERS against representative general-purpose, RS, and reasoning VLMs in both task performance and computational efficiency. SatLAERS achieves a favorable overall performance-efficiency trade-off, with an overall score of 57.31, a 1.55-point improvement over the strongest comparable-scale adapted baseline, and domain adaptation completed in 49.3 GPU-hours on 8 NVIDIA RTX 3090 GPUs. On Jetson Orin NX Super 16 GB, SatLAERS achieves 16.99 s end-to-end latency with 4.461 GB peak RAM, further demonstrating its suitability for resource-constrained onboard inference.

Our main contributions are summarized as follows:

- We propose SatLAERS, a granularity-aware adaptive RS-VLM that jointly enables resource-efficient adaptation and onboard inference through task- and sample-dependent multimodal computation.
- We introduce SatDISA, a single-observation disaster benchmark comprising 22,862 high-resolution optical images and 316,698 vision-language instructions across six disaster types and four semantic granularities, enabling unified evaluation of first-look disaster understanding.

- Extensive experiments demonstrate that SatLAERS achieves a favorable performance–efficiency trade-off, outperforming comparable-scale compact VLMs while completing domain adaptation on only 8 NVIDIA RTX 3090 GPUs and maintaining efficient inference on Jetson Orin NX Super 16 GB.

2 Related Work

2.1 Onboard Intelligence for Earth Observation

Onboard intelligence is increasingly shifting Earth observation from the conventional acquisition–downlink–ground-processing paradigm toward direct in-orbit interpretation, reducing communication overhead and observation-to-information latency for time-sensitive applications. Early systems primarily employed lightweight task-specific models for cloud filtering Giuffrida et al. (2020), image compression, and disaster mapping Wang et al. (2025a), as demonstrated by missions such as Φ-Sat-1 Giuffrida et al. (2022) and Φ-Sat-2 Guerrisi et al. (2023). More recent research has broadened satellite semantic interpretation beyond task-specific models, with geospatial foundation models such as Prithvi Jakubik et al. (2023) and compact RS-VLMs such as TinyRS-R1 Köksal & Alatan (2025) and NAVI-Orbital Victoria et al. (2026) providing more general semantic capabilities and motivating onboard-oriented adaptation. Such capabilities are particularly valuable for rapid disaster response, where heterogeneous information must be extracted from the latest available imagery under strict memory, computation, energy, and communication constraints. These developments motivate resource-efficient RS-VLMs that can provide flexible disaster understanding within practical onboard budgets.

2.2 RS-VLMs for Disaster Understanding

RS-VLMs have rapidly evolved toward increasingly fine-grained and unified geospatial understanding. GeoChat Kuckreja et al. (2023) extends RS dialogue to region-level reasoning and visual grounding, while GeoPixel Shabbir et al. (2025) further enables pixel-level grounding in high-resolution imagery. More recently, RSUniVLM Liu & Lian (2026) unifies image-, region-, and pixel-level understanding together with multi-image analysis in a compact framework, reflecting a broader shift toward multi-granularity RS-VLMs. In contrast, disaster-oriented RS-VLM research remains comparatively limited. DisasterM3 Wang et al. (2025a) establishes a large-scale bi-temporal benchmark covering multi-hazard perception, damage assessment, relational reasoning, and disaster reporting; BDAChat Wang et al. (2026b) integrates temporal vision-language reasoning with object-level structural damage assessment from satellite imagery; DisasterInsight Tehrani et al. (2026) extends evaluation to building-function recognition, damage and disaster classification, counting, and structured reporting; and WildfireVLM Ayanzadeh et al. (2026) combines satellite-based wildfire detection with language-driven risk assessment. Despite these advances, few studies have jointly addressed compact, multi-granularity first-look disaster understanding from single observations under limited adaptation budgets and onboard resource constraints.

2.3 Efficient Vision-Language Modeling

Earlier efficient vision-language models, such as ViLT Kim et al. (2021), relied on jointly trained multimodal architectures, whereas modern VLMs increasingly reuse pretrained vision encoders and language models connected through lightweight interfaces. BLIP-2 Li et al. (2023) established a representative frozen-backbone paradigm, showing that strong multimodal capability can be obtained without fully updating large pretrained components. Building on this paradigm, subsequent work has improved efficiency from complementary perspectives. TokenPacker Li et al. (2024) reduces the visual-token overhead by compressing high-resolution multi-level features before LLM decoding, while AdaptVision Lin et al. (2025) further learns sample-adaptive visual acquisition through reinforcement learning. On the language side, Omni-SMoLA Wu et al. (2023) introduces low-rank experts to expand multimodal capacity with limited additional computation. More recently, SmartVL Wang et al. (2026a) begins to jointly coordinate visual-token and language-model computation under explicit resource budgets, reflecting a broader shift from static compression toward adaptive computation. SatLAERS extends this direction to RS by coupling granularity-aware visual adaptation and query-adaptive visual routing, sparse low-rank experts, and budget-aware reinforce-

ment learning, targeting efficient adaptation and onboard inference across heterogeneous disaster tasks.

3 SATDISA DATASET

To support resource-efficient RS-VLMs for onboard first-look disaster understanding, we introduce SatDISA, a single-observation benchmark integrating high-resolution optical satellite imagery with multi-granularity vision-language tasks, semantic labels, spatial annotations, and reasoning-oriented responses. SatDISA enables systematic evaluation from perception and spatial grounding to disaster-oriented reasoning.

Dataset Construction. SatDISA is constructed by integrating multiple publicly available disaster RS datasets, including xBD Gupta et al. (2019) for multi-hazard disaster scenes, CAS Xu et al. (2024), GSR2SD Zhang et al. (2023), and GVLM-CD for landslide observations, the Canada Wildfire Prediction Dataset for wildfire scenarios, Maxar Open Data for additional satellite observations, and Sen2MHD Ao et al. (2026) for complementary flood, wildfire, and volcanic eruption scenes. These sources provide diverse disaster scenarios, geographic regions, and imaging conditions under a unified single-observation high-resolution optical satellite setting. SatDISA contains 22,862 high-resolution optical satellite images and 316,698 vision-language instruction-response pairs covering six representative disaster types: wildfire (7,000), flood (4,500), landslide (5,500), earthquake (3,000), volcanic eruption (1,500), and tsunami (1,362). This composition provides a diverse benchmark environment for multimodal disaster understanding.

Task Diversity. SatDISA organizes disaster understanding into four semantic granularities: image-level understanding (120,000 instructions), region-level interpretation (120,000 instructions), pixel-level grounding (55,000 instructions), and disaster-oriented situational understanding (21,698 instructions), covering progressively increasing spatial and reasoning requirements. Image-level tasks focus on global disaster recognition from scene context. Region-level tasks require localized evidence through object detection, grounding, and region description, supported by 45,000 unique bounding-box annotations. Pixel-level tasks further demand fine-grained spatial information for affected-area estimation and semantic segmentation, supported by 41,000 unique disaster-region masks. Situational-level tasks integrate visual observations, spatial evidence, and disaster knowledge for risk assessment, damage severity estimation, and emergency-response planning. This diversity in semantic objectives, spatial precision, and reasoning complexity enables comprehensive evaluation of RS-VLMs across perception, grounding, and disaster-oriented reasoning.

Quality Control and Consistency Validation. SatDISA applies expert-based annotation verification and cross-granularity consistency checking to maintain annotation quality across heterogeneous disaster understanding tasks. A team of eight RS and disaster-analysis experts reviewed spatial annotations, instruction-response pairs, and situational reasoning targets. Spatial annotation quality was quantified using intersection-over-union (IoU) for bounding boxes and Dice coefficient for disaster-region masks, achieving mean agreement scores of 0.82 and 0.88, respectively. Categorical situational annotations were evaluated using Cohen's kappa, reaching an agreement score of 0.79. Instruction-response pairs were further filtered according to semantic consistency, output-format validity, and alignment between visual evidence and target responses. These procedures provide reliable multi-granularity supervision for RS-VLM evaluation.

4 SATLAERS FRAMEWORK

4.1 FRAMEWORK OVERVIEW

Given an RS image I and instruction q, SatLAERS builds a resource-efficient RS-VLM by combining multi-scale visual adaptation, query-adaptive token routing, and sparse low-rank language adaptation. The framework dynamically allocates visual evidence under computation constraints and optimizes adaptive reasoning through progressive vision-language alignment, instruction tuning, and budget-aware reinforcement learning (Sec. 4.2, Sec. 4.3). SatLAERS supports image-, region-, pixel-, and situational-level disaster understanding within a unified autoregressive framework.

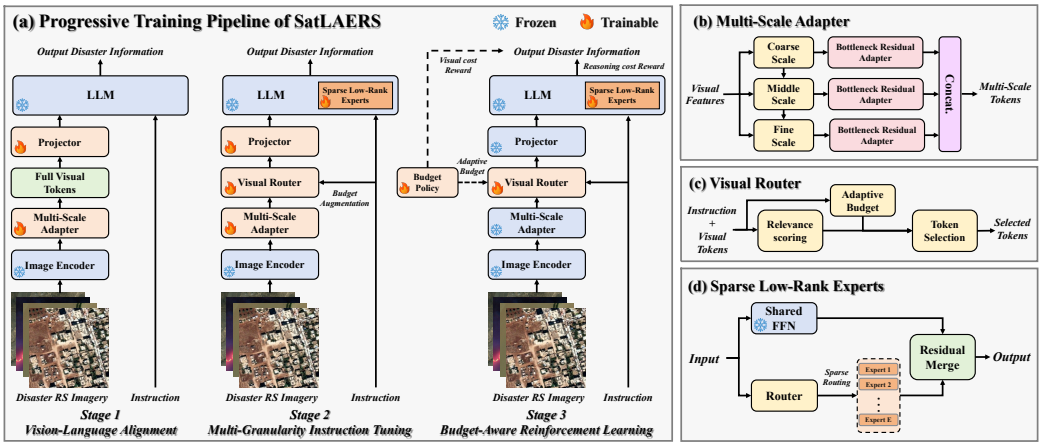

Figure 2: **Architecture and progressive training of SatLAERS.** (a) Progressive training from alignment and instruction tuning to budget-aware reinforcement learning. (b) Multi-scale visual adaptation with hierarchical feature fusion. (c) Query-adaptive routing for visual evidence allocation. (d) Sparse low-rank experts on the frozen language backbone.

4.2 Granularity-Aware Adaptive Computation

Multi-scale visual adaptation. The frozen SigLIP2-L/16 Tschannen et al. (2025) encoder produces dense patch-level features, from which adaptive pooling constructs coarse, middle, and fine branches at spatial resolutions of 8×8, 16×16, and 32×32, respectively. These branches provide complementary global, regional, and local spatial context. Let F_0 denote the shared encoder feature map and $s \in \{c, m, f\}$ denote the three scales. Multi-scale construction, coarse-to-fine fusion, and bottleneck refinement are jointly expressed as

$$\begin{aligned} F_s &= \mathrm{Pool}_s(F_0), \\ \bar{F}_c &= F_c, \qquad \bar{F}_s = F_s + \alpha_s \,\mathrm{Up}(\bar{F}_{s^-}), \quad s \in \{m, f\}, \\ \tilde{F}_s &= \bar{F}_s + \gamma_s W_2^s \sigma\big(W_1^s \,\mathrm{LN}(\bar{F}_s)\big)\,, \\ V &= \mathrm{Concat}_s\left(\tilde{F}_s + e_s\right), \end{aligned} \tag{1}$$

where s^- denotes the preceding coarser scale, $\mathrm{Pool}_s(\cdot)$ and $\mathrm{Up}(\cdot)$ denote scale-specific adaptive pooling and spatial upsampling, α_s and γ_s are learnable residual scaling factors, σ is GELU Hendrycks & Gimpel (2016), and e_s is the scale embedding. The resulting sequence V preserves complementary spatial abstractions for subsequent adaptive routing.

Query-adaptive visual routing. Let $V = \{v_i\}_{i=1}^N$ denote the multi-scale visual tokens and $Q = \{q_j\}_{j=1}^L$ denote instruction embeddings from the language embedding layer. Token relevance and the visual-budget distribution are computed as

$$\begin{aligned} a_i &= \mathrm{LSE}_j\left(\frac{(W_q q_j)^\top (W_v v_i)}{\sqrt{d_r}}\right), \\ \pi_b(K|I,q) &= \mathrm{Softmax}\left(f_\theta(z_v, z_q)\right), \qquad K \in \{K_{\min}, \ldots, N\}, \end{aligned} \tag{2}$$

where d_r is the routing dimension and z_v and z_q are pooled visual and textual representations. The relevance score a_i determines which tokens are informative, while π_b determines how many tokens are retained. During instruction tuning, K is sampled from granularity-dependent priors; during reinforcement learning, the budget policy is optimized for instance-specific allocation.

Given K, the token-selection policy samples a subset $\mathcal{S}$ of size K from the relevance scores using a Plackett–Luce formulation Plackett (1975); Luce (1979) during training. At inference time,

deterministic Top-K selection is applied under the predicted budget. The router therefore jointly determines how much visual evidence is required and which evidence is preserved.

Sparse low-rank language adaptation. The dense Qwen3-1.7B Yang et al. (2025) backbone remains frozen, while sparse low-rank residual experts are inserted into the `gate_proj`, `up_proj`, and `down_proj` transformations of its feed-forward blocks. For a frozen linear transformation Wx, the adapted output is

$$y = Wx + g_{e^\star}(x)\frac{\alpha}{r}B_{e^\star}A_{e^\star}x, \qquad e^\star = \arg\max_e g_e(x), \tag{3}$$

where A_e and B_e parameterize a rank-r residual expert and $g_e(x)$ is the token-level routing probability. The frozen dense pathway is always preserved, while only the routed low-rank expert contributes additional token-dependent adaptation. An auxiliary load-balancing objective aligns expert routing probabilities with realized token assignments to maintain stable expert utilization. We use the standard routing-balance objective Fedus et al. (2021) $\mathcal{L}_{\text{bal}} = E\sum_{e=1}^{E}\bar{p}_e f_e$, where $\bar{p}_e$ is the mean routing probability and f_e is the fraction of tokens assigned to expert e.

4.3 Progressive Budget-Aware Training

SatLAERS progressively introduces adaptive computation through vision-language alignment, multi-granularity instruction tuning, and budget-aware reinforcement learning.

Stage I: Vision-language alignment. The vision encoder and dense language backbone remain frozen, while the visual router, budget policy, and sparse experts are inactive. Only the multi-scale adapter and visual projector are optimized, and the complete multi-scale visual sequence is retained.

$$\mathcal{L}_{\text{align}} = -\frac{1}{T}\sum_{t=1}^{T}\log p(y_t|I, q, y_{<t}), \tag{4}$$

where $y = \{y_t\}_{t=1}^{T}$ is the target response. This stage establishes the visual-language interface before introducing adaptive routing.

Stage II: Multi-granularity instruction tuning. The vision encoder and dense Qwen3 backbone remain frozen, while the multi-scale adapter, projector, relevance scorer, sparse low-rank experts, and expert routers are optimized. The learned budget policy is bypassed in Stage II; instead, K is sampled from granularity-dependent priors, exposing coarse-grained tasks to compact visual budgets and spatially fine-grained tasks to richer visual evidence.

For a sampled budget K, multiple visual subsets are drawn from the relevance-based selection policy. Their teacher-forced response likelihoods provide relative supervision for token selection. The objective is

$$\mathcal{L}_{\text{SFT}} = \mathcal{L}_{\text{AR}} - \lambda_s A_{\text{SFT}}\log\pi_{\text{sel}}(\mathcal{S}|I, q, K) + \lambda_b\mathcal{L}_{\text{bal}}, \tag{5}$$

where $\mathcal{L}_{\text{AR}}$ is the autoregressive instruction-following loss and π_{sel} is the token-selection policy. For multiple subsets sampled for the same image-instruction pair, A_{SFT} is computed by contrasting the teacher-forced response log-likelihood of each subset against the mean likelihood of the remaining subsets. $\mathcal{L}_{\text{bal}}$ denotes the expert load-balancing objective. This stage jointly learns instruction following, visual evidence selection, and sparse expert specialization under varying visual budgets.

Stage III: Budget-aware reinforcement learning. The vision encoder, multi-scale adapter, projector, and dense Qwen3 backbone are frozen, while the relevance scorer, budget policy, sparse low-rank experts, and expert routers remain trainable. For each image-instruction pair, SatLAERS samples a visual budget K, a visual subset $\mathcal{S}$, and an autoregressive response. The sequence-level reward is

$$R = U - \lambda_v \frac{K}{N} - \lambda_r \frac{N_r}{N_{\max}}, \tag{6}$$

where $U \in [0, 1]$ is the normalized task-specific utility computed from the same scoring rule used for the corresponding SatDISA evaluation task, K/N measures visual-token consumption, and $N_r/N_{\max}$ measures reasoning cost. Reasoning length is not controlled by a separate predictor but emerges from the autoregressive policy under this utility-cost objective.

For each input, GRPO Shao et al. (2024) samples G trajectories and optimizes

$$\begin{aligned} A_i &= \frac{R_i - \bar{R}}{\sigma_R + \epsilon}, \qquad \rho_i(\theta) = \frac{\pi_\theta(\tau_i)}{\pi_{\theta_{\text{old}}}(\tau_i)}, \\ \mathcal{L}_{\text{GRPO}} &= -\frac{1}{G} \sum_{i=1}^{G} \min[\rho_i A_i,\ \text{clip}(\rho_i, 1-\epsilon_c, 1+\epsilon_c) A_i] + \lambda_b \mathcal{L}_{\text{bal}}. \end{aligned} \tag{7}$$

Here $\bar{R}$ and σ_R are group reward statistics, ϵ_c is the clipping coefficient, and $\pi_\theta(\tau_i)$ jointly models the sampled budget, visual subset, and generated response.

5 Experiments

5.1 Experimental Setup

Baselines. We compare representative VLMs from four categories: general-purpose VLMs, RS-VLMs, reasoning-enhanced VLMs, and disaster-domain adapted VLMs. General-purpose baselines include SmolVLM2 Marafioti et al. (2025), Qwen3-VL Bai et al. (2025), and InternVL3.5 Wang et al. (2025b), while RS-VLM baselines include EarthDial Soni et al. (2024), GeoChat Kuckreja et al. (2023), and TEOChat Irvin et al. (2024). Reasoning-enhanced baselines include TinyRS-R1 Köksal & Alatan (2025) and RSThinker Liu et al. (2025). To evaluate domain adaptation, Qwen3-VL and TinyRS-R1 are further fine-tuned on SatDISA under the same instruction-tuning protocol. All models retain their original architectures, and spatial outputs are converted into unified coordinate and mask formats for evaluation.

Evaluation metrics. We adopt task-specific metrics according to the output format of each task. Image-level disaster classification is evaluated using Accuracy. Region-level detection and grounding are evaluated using bounding-box IoU@0.5, while region description is evaluated using BERTScore-F1 Zhang et al. (2019). Pixel-level affected-area estimation is evaluated using Accuracy, and semantic segmentation is evaluated using mIoU after rasterizing predicted polygons into masks. For situational-level understanding, risk assessment and damage severity estimation are evaluated by jointly considering the correctness of the structured decision and the semantic consistency of the generated evidence:

$$M_t = \alpha \text{Acc}_{\text{main}} + (1 - \alpha) \text{BERTScore}_{\text{evidence}},$$

where $\alpha = 0.7$. Emergency response planning is evaluated using multi-label F1 over the predicted action set. For overall comparison, task-specific scores are normalized to a 0–100 scale, averaged within each semantic granularity, and then averaged across the four granularity levels.

5.2 Main Results

Table 1 compares multi-granularity performance on SatDISA. SatLAERS achieves the best overall score of 57.31, outperforming the strongest off-the-shelf and SatDISA-adapted baselines by 9.07 and 1.55 points, respectively. Compared with the strongest adapted baseline, SatLAERS improves region-level and pixel-level averages by 2.20 and 2.14 points, while also achieving gains in image-level understanding and situational reasoning. These results demonstrate that granularity-aware adaptive computation improves both semantic understanding and spatially demanding disaster interpretation.

Table 1: Multi-granularity performance comparison on SatDISA. Scores are reported on a 0–100 scale. Bold and underlined values indicate the best and second-best results, respectively.

Model	Params	Image-level	Region-level				Pixel-level			Situational-level				Overall
		Cls.	Det.	Grd.	Desc.	Avg.	Area	Seg.	Avg.	Risk	Sev.	Plan	Avg.	
Off-the-shelf General-purpose VLMs														
SmolVLM2-2.2B	2.2B	51.83	0.09	0.05	82.46	27.53	7.86	0.04	3.95	40.57	43.16	44.28	42.67	31.50
Qwen3-VL-2B-Instruct	2.0B	64.17	6.31	6.64	88.93	33.96	5.22	0.23	2.73	50.68	52.14	52.49	51.77	38.16
InternVL3.5-2B	2.3B	65.06	0.56	0.34	89.27	30.06	4.47	0.15	2.31	51.36	53.28	52.74	52.46	37.47
Off-the-shelf RS-VLMs														
EarthDial-4B-RGB	4.0B	64.37	7.92	8.64	86.73	34.43	10.67	4.83	7.75	43.62	45.87	40.26	43.25	37.45
GeoChat	7.0B	67.86	2.05	7.52	91.06	33.54	8.92	1.26	5.09	53.86	54.17	52.58	53.54	40.01
TEOChat	7.0B	72.38	0.28	11.68	92.47	34.81	12.36	2.85	7.61	60.47	61.56	62.08	61.37	44.04
Reasoning-enhanced VLMs														
TinyRS-R1	2.0B	72.86	8.94	9.87	92.08	36.96	15.44	5.68	10.56	58.87	60.48	61.29	60.21	45.15
RSThinker	10.0B	75.46	10.72	11.95	93.18	38.62	18.63	7.91	13.27	64.08	65.67	67.08	65.61	48.24
Disaster-domain Adapted VLMs														
Qwen3-VL-2B (SatDISA-tuned)	2.0B	81.47	24.76	26.81	94.18	48.58	28.42	13.67	21.05	63.57	65.08	65.27	64.64	53.93
TinyRS-R1 (SatDISA-tuned)	2.0B	<u>82.06</u>	<u>27.63</u>	<u>29.14</u>	**95.27**	<u>50.68</u>	<u>30.51</u>	<u>15.84</u>	<u>23.18</u>	**66.13**	<u>67.07</u>	<u>68.14</u>	<u>67.11</u>	<u>55.76</u>
SatLAERS (Ours)	2.1B	**83.59**	**31.26**	**32.74**	<u>94.64</u>	**52.88**	**31.88**	**18.76**	**25.32**	<u>65.65</u>	**67.25**	**69.48**	**67.46**	**57.31**

Table 2: Component and objective ablations on SatDISA. Granularity scores follow the aggregation protocol of Table 1; K/N and N_r denote mean visual-token retention and reasoning length.

Variant	Image	Region	Pixel	Situation	Overall	K/N	N_r
Full SatLAERS	**83.59**	**52.88**	**25.32**	**67.46**	**57.31**	0.46	61
w/o multi-scale adaptation	82.08	49.45	23.15	65.40	55.02	0.47	62
w/o budget augmentation	82.56	50.94	24.12	65.50	55.78	0.54	70
Random token selection	79.12	43.93	20.94	62.17	51.54	0.46	60
Fixed-budget routing	82.44	50.83	24.41	65.80	55.87	0.50	63
Standard LoRA	82.74	51.30	24.65	65.11	55.95	0.46	64
w/o Stage III GRPO	83.01	52.02	24.98	66.19	56.55	0.62	96

Table 3: Computation allocation across semantic granularities.

Level	K/N	N_r
Image	0.29	19
Region	0.45	32
Pixel	0.68	27
Situation	0.49	106

5.3 Ablations and Analysis

Component and objective ablations. Table 2 evaluates the major components of SatLAERS under the same four-granularity aggregation used in the main comparison. Removing multi-scale adaptation reduces the overall score from 57.31 to 55.02, with larger degradation on region- and pixel-level understanding, highlighting the importance of multi-scale spatial evidence. Removing granularity-aware budget augmentation yields 55.78 overall and increases visual-token consumption. Random token selection causes the largest degradation, reducing the overall score to 51.54, particularly on spatially demanding region and pixel tasks. Fixed-budget routing recovers performance to 55.87 but remains below adaptive budgeting, while replacing sparse low-rank experts with standard LoRA Hu et al. (2021) reaches 55.95. Removing Stage III GRPO Shao et al. (2024) preserves a relatively high score of 56.55 but increases K/N from 0.46 to 0.62 and N_r from 61 to 96, showing that reinforcement learning primarily improves computation-aware allocation while maintaining task performance.

Granularity-dependent computation allocation. Table 3 reveals distinct computation patterns across semantic granularities. Image-level tasks use the smallest visual budget ($K/N = 0.29$), whereas region- and pixel-level tasks retain more visual evidence with ratios of 0.45 and 0.68, reflecting their stronger spatial requirements. Situational tasks retain a moderate visual budget (0.49) but require substantially longer reasoning sequences ($N_r = 106$), indicating greater language-side reasoning demand.

Table 4: Adaptation efficiency and performance on SatDISA. Trainable is the maximum stage-specific trainable fraction.

Model	Train.(%)	Stg.	GPU-h	Overall
Qwen3-VL-2B (tuned)	7.9	1	68.7	53.93
TinyRS-R1 (tuned)	6.5	1	81.3	55.76
SatLAERS	**2.575**	3	**49.3**	**57.31**

Table 5: Onboard inference efficiency on Jetson Orin NX Super 16 GB.

Model	TTFT(ms)	E2E(s)	RAM(GB)
Qwen3-VL-2B	315.3	11.27	4.389
TinyRS-R1	492.9	64.46	7.304
SatLAERS	555.0	16.99	4.461

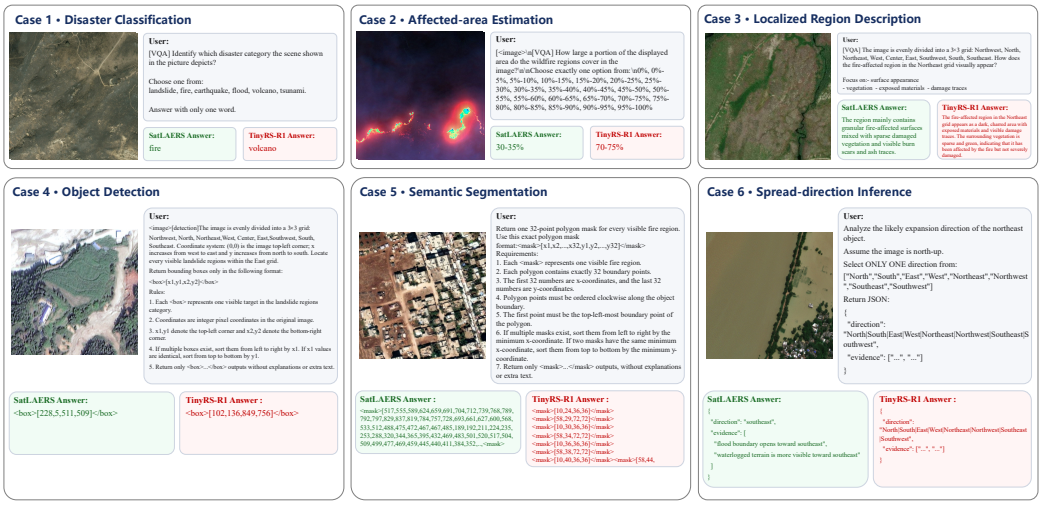

Figure 3: Qualitative comparison of SatLAERS and TinyRS-R1 on multi-granularity disaster understanding tasks.

5.4 Adaptation and Onboard Efficiency

Tables 4 and 5 evaluate adaptation and deployment efficiency. SatLAERS updates only 2.575% of its parameters and completes the three-stage adaptation in 49.3 GPU-hours, while achieving the highest overall performance among efficient models on SatDISA (57.31). On Jetson Orin NX Super 16 GB, SatLAERS requires 4.461 GB peak RAM and 16.99 s end-to-end latency. Compared with TinyRS-R1 Köksal & Alatan (2025), SatLAERS improves the overall score by 1.55 points while substantially reducing deployment memory and latency, demonstrating a more favorable accuracy-efficiency trade-off under onboard constraints. Qwen3-VL Bai et al. (2025) achieves lower raw inference latency, reflecting the inherent trade-off between computational cost and multimodal understanding capability.

5.5 Qualitative Results

Figure 3 presents qualitative comparisons between SatLAERS and TinyRS-R1 across multi-granularity disaster understanding tasks. The results highlight the ability of SatLAERS to preserve task-relevant visual evidence for recognition, grounding, and disaster-oriented reasoning.

6 Conclusion

We presented SatLAERS, a granularity-aware adaptive RS-VLM for resource-efficient first-look disaster understanding. SatLAERS integrates adaptive visual allocation, sparse language adaptation, and budget-aware optimization to dynamically allocate computation across heterogeneous tasks. Together with the SatDISA benchmark, SatLAERS achieves 57.31 overall while updating only 2.575% of parameters, demonstrating effective and efficient disaster-oriented multimodal understanding for onboard deployment.

AI USE STATEMENT

Generative AI tools were used solely to aid and polish the writing, including improving grammar, clarity, and readability. They were not used for research ideation or execution, literature retrieval, substantive content generation, synthetic data generation, or mathematical proofs. All AI-assisted edits were reviewed by the authors, who take full responsibility for the final content of this work.

REFERENCES

Wei Ao, Zhi Gao, Boan Chen, Xinyu Gong, Yuyang Zhai, Ziyao Li, and Qiao Wang. Towards efficient multi-hazard detection from single satellite imagery via spectral-semantic graph reasoning. *Remote Sensing of Environment*, 345:115566, 2026.

Aydin Ayanzadeh, Pranshu Dixit, Sadia Kamal, and Milton Halem. Wildfirevlm: Ai-powered analysis for early wildfire detection and risk assessment using satellite imagery. *ArXiv*, abs/2602.13305, 2026. URL https://api.semanticscholar.org/CorpusID:285616125.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *ArXiv*, abs/2101.03961, 2021. URL https://api.semanticscholar.org/CorpusID:231573431.

Gianluca Giuffrida, Lorenzo Diana, Francesco de Gioia, Gionata Benelli, Gabriele Meoni, Massimiliano Donati, and Luca Fanucci. Cloudscout: A deep neural network for on-board cloud detection on hyperspectral images. *Remote. Sens.*, 12:2205, 2020. URL https://api.semanticscholar.org/CorpusID:220872731.

Gianluca Giuffrida, Luca Fanucci, Gabriele Meoni, Matej Batič, *LéonieBuckley*, *AubreyDunne*, *Chris N.vanDijk*, *MarcoEsposito*, *John R.Hefele*, *NathanVercruyssen*, *G* sat-1 mission: The first on-board deep neural network demonstrator for satellite earth observation. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–14, 2022. URL https://api.semanticscholar.org/CorpusID:243793900.

Giorgia Guerrisi, Fabio Del Frate, and Giovanni Schiavon. Artificial intelligence based on-board image compression for the ϕ-sat-2 mission. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 16:8063–8075, 2023. URL https://api.semanticscholar.org/CorpusID:260026520.

Ritwik Gupta, Bryce Goodman, Nirav Patel, Richard Hosfelt, Sandra Sajeev, Eric T. Heim, Jigar Doshi, Keane Lucas, Howie Choset, and Matthew E. Gaston. Creating xbd: A dataset for assessing building damage from satellite imagery. In *CVPR Workshops*, 2019. URL https://api.semanticscholar.org/CorpusID:263890009.

Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv: Learning*, 2016. URL https://api.semanticscholar.org/CorpusID:125617073.

J. Edward Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, She-An Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *ArXiv*, abs/2106.09685, 2021. URL https://api.semanticscholar.org/CorpusID:235458009.

Jeremy Irvin, Emily Liu, Joyce Chen, Ines Dormoy, Jinyoung Kim, Samar Khanna, Zhuo Zheng, and Stefano Ermon. Teochat: A large vision-language assistant for temporal earth observation data. *ArXiv*, abs/2410.06234, 2024. URL https://api.semanticscholar.org/CorpusID:273228085.

Johannes Jakubik, Sujit Roy, Christopher Phillips, Paolo Fraccaro, Denys Godwin, Bianca Zadrozny, Daniel Szwarcman, Carlos Gomes, Gabby Nyirjesy, Blair Edwards, Daiki Kimura, Naomi Simumba, Linsong Chu, S. Karthik Mukkavilli, Devyani Lambhate, Kamal Das, Ranjini Bangalore, Dário A. B. Oliveira, Michal Muszynski, Kumar Ankur, Muthukumaran Ramasubra-

manian, Iksha Gurung, Sam Khallaghi, Han-Xi Li, Michael Cecil, Maryam Ahmadi, Fatemeh Kordi, Hamed Alemohammad, Manil Maskey, Raghu Kiran Ganti, Kommy Weldemariam, and Rahul Ramachandran. Foundation models for generalist geospatial artificial intelligence. *ArXiv*, abs/2310.18660, 2023. URL https://api.semanticscholar.org/CorpusID:264590307.

Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, 2021. URL https://api.semanticscholar.org/CorpusID:231839613.

Aybora Köksal and A. Aydın Alatan. Tinyrs-r1: Compact vision language model for remote sensing. *IEEE Geoscience and Remote Sensing Letters*, 22:1–5, 2025. URL https://api.semanticscholar.org/CorpusID:278739960.

Kartik Kuckreja, Muhammad Sohail Danish, Muzammal Naseer, Abhijit Das, Salman H. Khan, and Fahad Shahbaz Khan. Geochat:grounded large vision-language model for remote sensing. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27831–27840, 2023. URL https://api.semanticscholar.org/CorpusID:265456719.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *ArXiv*, abs/2301.12597, 2023. URL https://api.semanticscholar.org/CorpusID:256390509.

Wentong Li, Yuqian Yuan, Jian Liu, Dongqi Tang, Song Wang, Jie Qin, Jianke Zhu, and Lei Zhang. Tokenpacker: Efficient visual projector for multimodal llm. *International Journal of Computer Vision*, 133:6794 – 6812, 2024. URL https://api.semanticscholar.org/CorpusID:270878717.

Zi-Chuan Lin, Yi-Cheng Liu, Yang Yang, Lvfang Tao, and Deheng Ye. Adaptvision: Efficient vision-language models via adaptive visual acquisition. *ArXiv*, abs/2512.03794, 2025. URL https://api.semanticscholar.org/CorpusID:283466589.

Jiaqi Liu, Lang Sun, Ronghao Fu, and Bo Yang. Towards faithful reasoning in remote sensing: A perceptually-grounded geospatial chain-of-thought for vision-language models. *ArXiv*, abs/2509.22221, 2025. URL https://api.semanticscholar.org/CorpusID:281659100.

Xu Liu and Zhouhui Lian. Rsunivlm: A unified vision-language model for remote sensing via granularity-oriented moe. *Pattern Recognit.*, 179:113717, 2026. URL https://api.semanticscholar.org/CorpusID:287525230.

R. Duncan Luce. Individual choice behavior: A theoretical analysis. 1979. URL https://api.semanticscholar.org/CorpusID:123178801.

Andrés Marafioti, Orr Zohar, Miquel Farré, Merve Noyan, Elie Bakouch, Pedro Cuenca, Cyril Zakka, Loubna Ben Allal, Anton Lozhkov, Nouamane Tazi, et al. Smolvlm: Redefining small and efficient multimodal models. *arXiv preprint arXiv:2504.05299*, 2025.

Robin L. Plackett. The analysis of permutations. *Journal of The Royal Statistical Society Series C-applied Statistics*, 24:193–202, 1975. URL https://api.semanticscholar.org/CorpusID:116534299.

Akashah Shabbir, Mohammed Zumri, Mohammed Bennamoun, Fahad Shahbaz Khan, and Salman Khan. Geopixel: Pixel grounding large multimodal model in remote sensing. *ArXiv*, abs/2501.13925, 2025. URL https://api.semanticscholar.org/CorpusID:275820220.

Zhihong Shao, Peiyi Wang, Qi-Hao Zhu, Runxin Xu, Jun-Mei Song, Xiao Bi, Hao-Wei Zhang, Ming-Chuan Zhang, Y.-K. Li, Yu Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *ArXiv*, abs/2402.03300, 2024. URL https://api.semanticscholar.org/CorpusID:267412607.

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 H. Khan. Earthdial: Turning multi-sensory earth observations to interactive dialogues. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14303–14313, 2024. URL https://api.semanticscholar.org/CorpusID:274859947.

Sara Tehrani, Yong-Hao Xu, Leif Haglund, Amanda Berg, Gulnaz Zhambulova, and Michael Felsberg. Disasterinsight: A building-centric benchmark for evaluating vision–language models in disaster response. 2026. URL https://api.semanticscholar.org/CorpusID:285049964.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

Juan Manuel Delfa Victoria, Taran Cyriac John, and Andrew W. Herson. Navi-orbital: First in-orbit demonstration of a zero-shot vision-language model for autonomous earth observation. *ArXiv*, abs/2606.18271, 2026. URL https://api.semanticscholar.org/CorpusID:289628141.

Junjue Wang, Weihao Xuan, Heli Qi, Zhihao Liu, Kunyi Liu, Yuhan Wu, Hongruixuan Chen, Jian Song, Junshi Xia, Zhuo Zheng, and Naoto Yokoya. Disasterm3: A remote sensing vision-language dataset for disaster damage assessment and response. *ArXiv*, abs/2505.21089, 2025a. URL https://api.semanticscholar.org/CorpusID:278910917.

Peng-Cheng Wang, Zhiquan Wang, Jayoung Lee, Zhuoyan Xu, Ran Xu, Saurabh Bagchi, Yin Li, and Somali Chaterji. Look less, think faster: Joint token-compute adaptation for multimodal llms. *ArXiv*, abs/2607.20357, 2026a. URL https://api.semanticscholar.org/CorpusID:290448660.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025b.

Yong Wang, Jiawei Cui, Chang-Hai Zhai, Xi-Gui Tao, and Yuhao Li. Integrating segmentation and vision-language model for automated and interpretable building damage assessment from satellite imagery. *Adv. Eng. Informatics*, 71:104320, 2026b. URL https://api.semanticscholar.org/CorpusID:284655231.

Jialin Wu, Xia Ben Hu, Yaqing Wang, Bo Pang, and Radu Soricut. Omni-smola: Boosting generalist multimodal models with soft mixture of low-rank experts. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14205–14215, 2023. URL https://api.semanticscholar.org/CorpusID:265609776.

Yulin Xu, Chao-Jun Ouyang, Qing-Song Xu, Dong-Po Wang, Bo Zhao, and Yu Luo. Cas landslide dataset: A large-scale and multisensor dataset for deep learning-based landslide detection. *Scientific Data*, 11, 2024. URL https://api.semanticscholar.org/CorpusID:266743674.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *ArXiv*, abs/1904.09675, 2019. URL https://api.semanticscholar.org/CorpusID:127986044.

Xiaokang Zhang, Weikang Yu, Man-On Pun, and Wenzhong Shi. Cross-domain landslide mapping from large-scale remote sensing images using prototype-guided domain-aware progressive representation learning. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2023. URL https://api.semanticscholar.org/CorpusID:256528190.

A APPENDIX

You may include other additional sections here.