Seg the HAB: Language-Guided Geospatial Algae Bloom Reasoning and Segmentation

Patterson Hsieh[1,*] **Jerry Yeh**[2,*] **Mao-Chi He**[2*] **Wen-Han Hsieh**[2,] **Elvis Hsieh**[2,]

UC San Diego[1], UC Berkeley[2]

Abstract

Climate change is intensifying the occurrence of harmful algal bloom (HAB), particularly cyanobacteria, which threaten aquatic ecosystems and human health through oxygen depletion, toxin release, and disruption of marine biodiversity. Traditional monitoring approaches, such as manual water sampling, remain labor-intensive and limited in spatial and temporal coverage. Recent advances in vision-language models (VLMs) for remote sensing have shown potential for scalable AI-driven solutions, yet challenges remain in reasoning over imagery and quantifying bloom severity. In this work, we introduce ALGae Observation and Segmentation (ALGOS), a segmentation-and-reasoning system for HAB monitoring that combines remote sensing image understanding with severity estimation. Our approach integrates GeoSAM-assisted human evaluation for high-quality segmentation mask curation and fine-tunes vision language model on severity prediction using the Cyanobacteria Aggregated Manual Labels (CAML) from NASA. Experiments demonstrate that ALGOS achieves robust performance on both segmentation and severity-level estimation, paving the way toward practical and automated cyanobacterial monitoring systems.

1 Introduction

Harmful algal blooms (HAB) are an escalating global concern driven by climate change. Cyanobacteria-dominated HAB in particular present severe ecological, public health, and economic risks. Numerous studies have shown that HAB create multiple harms. From an ecological perspective, Anderson et al. [1] highlight that HAB cause mass fish mortality and ecosystem disruption. In addition to ecology, public health risks are also severe [3]. Economically, HAB impose billions of dollars in losses annually, specifically the 2017–2018 Florida Red Tide, causing an estimated $2.7 million in losses [13, 14]. These combined effects emphasize the idea that monitoring HAB dynamics is essential for mitigation and policy; however, existing methods, relying on sampling and manual microscopy, are costly, time-consuming, and geographically constrained.

Advances in computer vision and geospatial analysis provide new opportunities for scalable HAB monitoring. However, prior work in AI-based HAB monitoring has only tackled either bloom severity prediction or spatial segmentation, limiting comprehensive monitoring capabilities. Vision-based systems segment bloom patches in local camera imagery [2], yet they neither operate on wide-area remote-sensing data nor estimate severity, limiting the scalability. Dorne et al. [5] estimates severity from Sentinel-2 and ancillary data but provides no explicit spatial delineation. Traditional remote-sensing methods using spectral indices can map blooms but depend on manual thresholds and site-specific tuning [20]. This fragmentation prevents systems from answering queries requiring both spatial and severity reasoning essential for targeted management.

[*]Equal Contribution

Tackling Climate Change with Machine Learning: workshop at NeurIPS 2025.

Motivated by these insights, we introduce ALGOS, a unified vision-language framework that bridges reasoning segmentation with HAB severity assessment in satellite imagery. Following the prior work [15], we leverage the CAML dataset for severity-level reasoning with a novel HAB segmentation dataset curated through GeoSAM-assisted annotation with human evaluation [17]. Our framework extends to HAB-specific remote sensing data, enabling simultaneous spatial localization of bloom extent and severity-level classification through natural language reasoning. Our experimental results demonstrate significant improvements over baseline segmentation models in spatial accuracy and baseline VLM in severity prediction. We believe ALGOS explores a new path for automated HAB monitoring that combines the precision of pixel-level segmentation and the contextual reasoning capabilities necessary for ecological assessment and public health decision-making.

2 Related Work

2.1 Cyanobacteria Detection and Segmentation

Classical remote-sensing methods rely on spectral indices and thresholds, which can break down in optically complex inland waters and often require site-specific tuning [20]. Early computer-vision systems with hand-crafted color features were brittle to illumination and background changes [16]. Modern deep models infer algal presence or chlorophyll-*a* from multispectral imagery and often outperform traditional baselines, but generalization suffers when training data are region-limited [20].

Segmentation efforts show promise but are typically scoped to localization. Barrientos-Espillco et al. [2] segment CyanoHAB patches using synthetic imagery to mitigate data scarcity, yet do not infer severity. Conversely, multi-source fusion and ensemble models improve severity classification (e.g., Sentinel-2 + climate + terrain), but treat monitoring as point prediction without mapping spatial extent [12]. ALGae Observation and Segmentation addresses these gaps by jointly producing a segmentation mask and a severity estimate with language-driven reasoning to support timely monitoring.

2.2 Geospatial Foundation Models

Geospatial foundation models extend vision–language models (VLMs) to remote sensing by pairing an overhead-image encoder with a language model so the system can describe scenes, answer questions, and follow instructions. Training the visual encoder on satellite or aerial imagery improves transfer. Liu et al. [10] aligns overhead images and text with CLIP-style contrastive learning. Instruction-tuned backbones (e.g., Vicuna) add conversational, task-following ability, but most current VLMs remain text-only, lacking spatial outputs such as maps or pixel-wise masks [4]. To obtain spatially explicit results, recent work augments language models with lightweight segmentation branches so a query can return a pixel-level mask, and adapts generic segmenters to overhead data. For instance, a language-to-mask pathway has been explored in geospatial VLMs [15], while GeoSAM fine-tunes Segment Anything for large, low-contrast satellite imagery [17, 8]. Together, these directions move geospatial AI from text answers toward actionable, per-pixel outputs that practitioners can use for monitoring and geospatial decision.

3 Methods

3.1 HAB Segmentation Dataset Curation

Following the prior work [15], we developed a comprehensive pipeline to generate high-quality pixel-level segmentation masks for HAB detection using the CAML dataset [6]. Our approach addresses the unique challenges of HAB segmentation in Sentinel-2 imagery, where bloom boundaries are often diffuse with subtle spectral signatures.

We extend GeoSAM [17] with an interactive mask generation with human evaluation stage. Unlike the original GeoSAM, which primarily relies on automated point prompt generation, our framework introduces a semi-supervised curation loop. Users provide positive points on bloom areas, negative points on background regions, and a region-of-interest (ROI) box to guide the model's attention. The system then generates candidate masks, which are interactively refined through lightweight morphological filtering and post-processing. This reduces ambiguity in diffuse bloom regions while retaining efficiency.

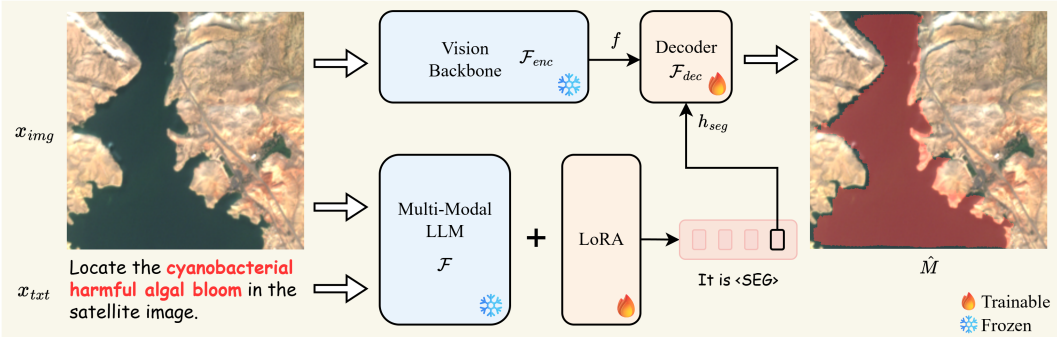


Figure 1: The pipeline of ALGOS. Given the input image and text query, the multimodal LLM (LISAT$_{\text{PRE}}$) generates text output. The last-layer embedding for the <SEG> token is then decoded into the segmentation mask via the decoder. We adopt SAM [8] as our choice of vision backbone.

To ensure the reliability of the resulting dataset, each generated mask undergoes human evaluation. Annotators visually compare the candidate masks against the corresponding Sentinel-2 imagery, retaining only those masks that accurately delineate bloom regions. Masks that fail this validation are discarded or corrected through additional refinement. In this way, the semi-supervised feedback loop ensures that the segmentation masks are consistently aligned with ecological reality. Therefore, the resulting dataset contains high-quality segmentation masks, robust to the heterogeneous quality of Sentinel-2 imagery, from cloud-free high-resolution scenes to partially degraded scenes.

3.2 HAB Reasoning Dataset Curation

To enable severity-level assessment through natural language reasoning, we adopt the synthetic query generation pipeline following Quenum et al. [15] to create HAB-reasoning queries.

Severity-Based Query Generation. We fine-grain the WHO recreational guidance thresholds [18] into a five-level defined as follows: Level 1: $x < 2 \times 10^4$, Level 2: $2 \times 10^4 \leq x < 1 \times 10^5$, Level 3: $1 \times 10^5 \leq x < 1 \times 10^6$, Level 4: $1 \times 10^6 \leq x < 1 \times 10^7$, Level 5: $x \geq 1 \times 10^7$ cells/mL. Based on the scale, we curate natural language query templates that require the model to infer algal bloom severity directly from satellite imagery. Each template prompts the model to classify severity on this ordinal five-point scale, ensuring consistent outputs while retaining ecological interpretability across heterogeneous observational conditions.

Multi-modal Alignment. Each reasoning query is paired with the corresponding satellite image and severity label defined above, creating a structured instruction–image–answer triplet (Appendix C). This triplet design enables the model to jointly learn the relationship between visual appearance, ecological context, and ordinal severity categories, thereby supporting robust estimation of harmful algal blooms following [15, 9].

3.3 Vision-Language Model Architecture for HAB Monitoring

Our framework adopts the embedding-as-mask paradigm for HAB-specific applications, integrating domain-adapted visual encoders with language models fine-tuned on HAB-reasoning queries .

Multimodal Integration. Following LISAT's architecture [15], we employ a Vicuna-7B language model [4] as our base LLM, coupled with a Remote-CLIP ViT-L/14 encoder [10] optimized for satellite imagery processing. The visual encoder processes Sentinel-2 multispectral bands, while a learnable linear projection aligns visual features with the language model's embedding space. We expand the vocabulary with a specialized <SEG> token that, when generated, triggers segmentation mask prediction through a SAM decoder head [8].

Training Objectives.

Our model is optimized end-to-end with a joint objective that integrates both text generation and segmentation. The overall training loss $\mathcal{L}$ is formulated as a weighted combination:

$$\mathcal{L} = \omega_{\text{txt}}\,\mathcal{L}_{\text{txt}} + \omega_{\text{mask}}\,\mathcal{L}_{\text{mask}}. \tag{1}$$

Table 1: Segmentation results comparing LISAT, LISA, and ALGOS. cIoU: per-image class-balanced mean IoU; gIoU: dataset-level/global IoU.

Model	cIoU	gIoU
LISAT	$0.1083_{\pm 0.0124}$	$0.1052_{\pm 0.0132}$
LISA 7B	$0.1373_{\pm 0.0182}$	$0.1274_{\pm 0.0160}$
ALGOS	$\mathbf{0.6493}_{\pm 0.0301}$	$\mathbf{0.5969}_{\pm 0.0268}$

Table 2: Severity prediction results comparing the LLaVA baseline and ALGOS.

Model	MSE	RMSE	MAE
LLaVA-7B	3.868	1.967	1.587
ALGOS	**2.984**	**1.727**	**1.365**

Here, the text generation objective $\mathcal{L}_{\text{txt}}$ is defined as the standard autoregressive cross-entropy loss:

$$\mathcal{L}_{\text{txt}} = \mathbf{CE}(\hat{\mathbf{y}}_{\text{txt}}, \mathbf{y}_{\text{txt}}). \tag{2}$$

The segmentation objective $\mathcal{L}_{\text{mask}}$ combines a per-pixel binary cross-entropy (BCE) term and a DICE loss, balanced by ω_{bce} and ω_{dice}:

$$\mathcal{L}_{\text{mask}} = \omega_{\text{bce}}\, \mathbf{BCE}(\hat{\mathbf{M}}, \mathbf{M}) \;+\; \omega_{\text{dice}}\, \mathbf{DICE}(\hat{\mathbf{M}}, \mathbf{M}). \tag{3}$$

Implementation Details. All experiments were conducted on eight NVIDIA DGX A100 GPUs (80GB each). For severity prediction, we fine-tuned LLaVA-7B with LoRA for 50 epochs on the HAB dataset. For segmentation, ALGOS was trained jointly on HAB, FP-Ref-COCO [19], and ReasonSeg [9]. LoRA [7] was applied to the multimodal language model, while the SAM decoder was fully fine-tuned. The learning rate was set to 3×10^{-4}, with other configurations kept consistent with standard practice. For the composite loss, we set the weighting coefficients to $\omega_{\text{txt}} = 1.0$ and $\omega_{\text{mask}} = 1.0$. Within the segmentation objective, the per-pixel binary cross-entropy and DICE terms are weighted as $\omega_{\text{bce}} = 2.0$ and $\omega_{\text{dice}} = 0.5$, respectively. Empirically, this configuration yielded the best overall performance. Training required approximately 6 hours across eight DGX A100 GPUs.

4 Results

4.1 Setup

Performance metrics. Following Lai et al. [9], Quenum et al. [15], we evaluate segmentation using two IoU variants under a binary setting (algae vs. non-algae). The **cIoU** metric computes the per-image, class-balanced mean IoU; with only one foreground class, this reduces to the average IoU of algae masks across images (assigning a score of 1 when both prediction and ground truth are empty, and 0 otherwise). The **gIoU** metric instead aggregates intersections and unions across all test images before computing the ratio. In the binary case, this measures agreement with the *total* algae extent over the full test set, making it more sensitive to prevalence and large contiguous blooms. For the severity prediction task, we use mean squared error (MSE) as the primary metric, reflecting the ordinal nature of severity levels.

Baselines. Table 1 compares ALGOS with state-of-the-art reasoning segmentation models [9, 15], while Table 2 benchmarks ALGOS against LLaVA [11].

4.2 Results and Observations

ALGOS achieves strong performance across both segmentation and severity prediction tasks. For segmentation, it reaches a cIoU of 0.65 and gIoU of 0.60, far surpassing LISAT (0.11 / 0.10) and LISA-7B (0.14 / 0.13), accurately capturing both per-image bloom regions and large contiguous extents across the dataset (Table 1). For severity prediction, ALGOS substantially reduces error, with mean squared error (MSE) dropping from 3.868 to 2.984, along with corresponding improvements in RMSE and MAE (Table 2). These results show that ALGOS is capable of addressing both spatial segmentation and severity-level estimation, while outperforming baselines in each task.

5 Conclusion

We introduced ALGae Observation and Segmentation, a framework that leverages multimodal language models to reason over heterogeneous data and dynamically segment harmful algal bloom

(HAB) regions. Additionally, we proposed a semi-supervised segmentation pipeline that improves delineation in images where automated methods struggle with unclear bloom boundaries. ALGOS synthesizes spatial and contextual information through a structured reasoning process, ensuring that the segmented regions align with bloom severity levels. Unlike prior work, which has only addressed either severity estimation or localized segmentation in isolation, our approach integrates geospatial foundation models to jointly perform both tasks on wide-area remote sensing imagery. By advancing the ability of geospatial models to reason and adaptively segment polluted areas, ALGOS provides a robust tool for ecological monitoring and policy support, enabling scalable HAB monitoring.

Limitations. Our framework has been evaluated on a limited geographic and seasonal scope based on the CAML dataset, and its generalization to diverse aquatic environments requires larger-scale, cross-region benchmarks. In addition, the reliance on curated datasets highlights the need for continuous data integration that can adapt to evolving ecological conditions. In future work, we will address both limitations by extending our evaluations and data pipelines to support scalable deployment.

References

[1] Donald M. Anderson, Elie Fensin, Christopher J. Gobler, Ann E. Hoeglund, Katharine A. Hubbard, David M. Kulis, Jan H. Landsberg, Kathleen A. Lefebvre, Pieter Provoost, Michael L. Richlen, Jennifer L. Smith, Andrew R. Solow, and Vera L. Trainer. Marine harmful algal blooms (habs) in the united states: History, current status and future trends. *Harmful Algae*, 102:101975, 2021. doi: 10.1016/j.hal.2020.101975.

[2] Fredy Barrientos-Espillco, Esther Gascó, Clara I. López-González, María José Gómez-Silva, and Gonzalo Pajares. Semantic segmentation based on deep learning for the detection of cyanobacterial harmful algal blooms (cyanohabs) using synthetic images. *Applied Soft Computing*, 141:110315, 2023. doi: 10.1016/j.asoc.2023.110315.

[3] Centers for Disease Control and Prevention (CDC). Harmful algal bloom-associated illnesses, 2024. URL `https://www.cdc.gov/harmful-algal-blooms/signs-symptoms/index.html`. Accessed: 2025-08-18.

[4] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. March 2023. URL `https://lmsys.org/blog/2023-03-30-vicuna/`.

[5] Emily Dorne, Katie Wetstone, Trista Brophy Cerquera, and Shobhana Gupta. Cyanobacteria detection in small, inland water bodies with cyfi. In *Proceedings of the 23rd Python in Science Conference (SciPy 2024)*, SciPy Proceedings, Tacoma, Washington, July 2024. doi: 10.25080/PDHK7238.

[6] S. Gupta, E. Gelbart, R. Gupta, K. Wetstone, and E. Dorne. Cyanobacteria aggregated manual labels dataset, 2024. URL `http://dx.doi.org/10.5067/SeaBASS/CAML/DATA001`.

[7] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021. URL `https://arxiv.org/abs/2106.09685`.

[8] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023.

[9] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[10] Fan Liu, Delong Chen, Zhangqingyun Guan, Xiaocong Zhou, Jiale Zhu, Qiaolin Ye, Liyong Fu, and Junwei Zhou. Remoteclip: A vision language foundation model for remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–13, 2024. doi: 10.1109/TGRS.2024.3374824.

[11] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023. URL https://arxiv.org/abs/2304.08485.

[12] Ioannis Nasios. Ai-driven multi-source data fusion for algal bloom severity classification in small inland water bodies: Leveraging sentinel-2, dem, and noaa climate data. *arXiv preprint*, 2025.

[13] National Oceanic and Atmospheric Administration (NOAA). 2017–2018 florida red tide determined by ocean circulation, 2019. URL https://coastalscience.noaa.gov/news/2017-2018-florida-red-tide-determined-by-ocean-circulation/. Accessed: 2025-08-18.

[14] National Oceanic and Atmospheric Administration (NOAA). Economic impacts of harmful algal blooms. National Centers for Coastal Ocean Science, 2025. URL https://coastalscience.noaa.gov/science-areas/habs/. Accessed: 2025-08-18.

[15] Jerome Quenum, Wen-Han Hsieh, Tsung-Han Wu, Ritwik Gupta, Trevor Darrell, and David M. Chan. Lisat: Language-instructed segmentation assistant for satellite imagery. *arXiv preprint*, 2025.

[16] Arabinda Samantaray, Baijian Yang, J. Eric Dietz, and Byung-Cheol Min. Algae detection using computer vision and deep learning. *arXiv preprint*, 2018.

[17] Raahul Irvin Sultana, Chuang Lia, Hanbing Zhua, Pritam Khanduria, Marco Brocanellib, and Dacheng Zhua. Geosam: Fine-tuning sam with multi-modal prompts for mobility infrastructure segmentation. *arXiv preprint*, 2023.

[18] World Health Organization. Guidelines for safe recreational water environments. volume 1: Coastal and fresh waters. 2003. URL https://iris.who.int/bitstream/handle/10665/42591/9241545801.pdf.

[19] Tsung-Han Wu, Giscard Biamby, David Chan, Lisa Dunlap, Ritwik Gupta, Xudong Wang, Joseph E. Gonzalez, and Trevor Darrell. See, say, and segment: Teaching lmms to overcome false premises, 2023. URL https://arxiv.org/abs/2312.08366.

[20] Liping Yang, Joshua Driscol, Rose Sarigai, Qiaoling Wu, Christopher D. Lippitt, and Madelyn Morgan. Towards synoptic water monitoring systems: A review of ai methods for automating water body detection and water quality monitoring using remote sensing. *Sensors*, 22(6):2416, 2022. doi: 10.3390/s22062416.

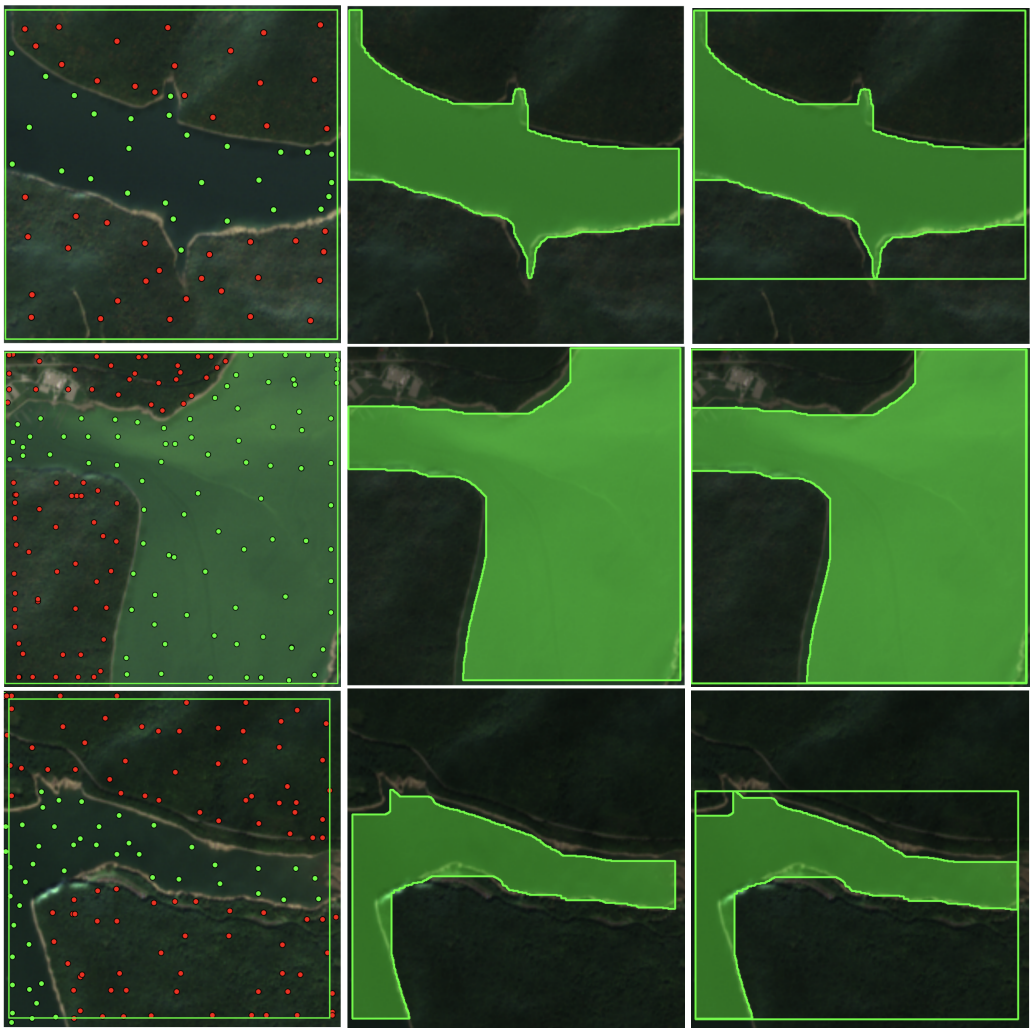
A Qualitative Comparison Across Models

Queries	LISA–7B	LISAT	ALGOS (Ours)	Ground Truth
Locate the cyanobacterial harmful algal bloom in the satellite image.				
Locate all visible harmful algal blooms.				
Find the cyanobacterial harmful algal bloom in the satellite image.				
Segment all visible harmful algal blooms.				
Segment the waterbody affected by cyanobacteria.				
Segment the cyanobacterial harmful algal bloom in the satellite image.				
Segment the algal bloom affected areas.				
Locate the cyanobacterial harmful algal bloom in the satellite image.				

Figure 2: Qualitative comparison of predictions across models.

B Image Segmentation

In this section, we illustrate the interactive segmentation workflow used for data curation (Figure 3). Users first provide positive (green) and negative (red) prompts to guide the model. The model then generates a segmentation mask on Sentinel-2 imagery, from which bounding boxes are extracted to obtain object-level representations for downstream analysis.



(a) User-provided prompts. (b) Generated mask overlay. (c) Bounding boxes extracted.

Figure 3: Interactive segmentation workflow used during data curation.

C Algae Severity Assessment

This sample demonstrates a structured *instruction–image–answer* triplet used for model fine-tuning. Severity levels follow the WHO recreational thresholds, refined into five ordinal categories: 1 = Very low, 2 = Low, 3 = Moderate, 4 = High, 5 = Very high.

Query Prompt

```
<image>
Analyze the provided satellite image of algae-specific conditions.
    Determine the severity level, where:
1 = Very low, 2 = Low, 3 = Moderate, 4 = High, 5 = Very high.
Output only a single digit from 1-5 with no other text.
Example output:3
```

Input Image

Figure 4: Satellite image provided as input.

Ground Truth Label

This example corresponds to **Level 1 (very low)**, with an expected output of: `1.0`