Appendix: Enabling Training-Free Text-Based Remote Sensing Segmentation

Jose Sosa, Danila Rukhovich, Anis Kacem, Djamila Aouada
SnT, University of Luxembourg
{jose.sosa,danila.rukhovich,anis.kacem,djamila.aouada}@uni.lu

In this Appendix, we first outline additional implementation details (Sec. 1), then present extra quantitative (Sec. 2) and qualitative results (Sec. 3), and conclude with further information about the datasets (Sec. 4).

1. Implementation Details

Click generation procedure. To fine-tune our generative VLM approach for referring and reasoning segmentation tasks, we use the training splits from the RRSIS-D and EarthReason datasets. Our objective is to train the VLM to output click positions in textual form, which are subsequently used to prompt SAM. For this purpose, we utilise the input images and their corresponding ground truth masks. Each mask M is automatically converted into a sequence of clicks $\mathcal{C}$ without human intervention.

Inspired by interactive segmentation methods [1, 16], we adopt an iterative click generation strategy. Formally, at iteration i, given the current click sequence $\mathcal{C}_{i-1}$, we compute an intermediate predicted mask using SAM:

$$M_{i-1} = \mathscr{S}(I, \text{prompt} = \mathcal{C}_{i-1}).$$

The discrepancy between M_{i-1} and the ground-truth mask M reveals both under-segmented and over-segmented regions. We define two binary error maps:

$$E_+ = M - M_{i-1}, \quad E_- = M_{i-1} - M,$$

where E_+ contains pixels that should be included (false negatives), and E_- contains pixels that should be excluded (false positives).

Next, we compute a distance transform over the union $E_+ \cup E_-$, which yields a probability distribution emphasizing pixels far from already-correct regions. A new click c_i is then sampled from this distribution:

$$c_i \sim \text{DistanceTransform}(E_+ \cup E_-).$$

If $c_i \in E_+$, it is labeled as a *positive* click; if $c_i \in E_-$, it is labeled as a *negative* click. The click set is updated as:

$$\mathcal{C}_i = \mathcal{C}_{i-1} \cup \{c_i\}.$$

This process is repeated until a stopping condition is met (*e.g.*, achieving a target IoU or reaching a maximum number of clicks) as depicted in algorithm 1. The resulting synthetic click sequences $\mathcal{C}$ are then used to finetune the generative VLM for click generation. Figure 1 illustrates this process.

Contrastive VLM inference. We use the same CLIP (ViT-B/16) model as in [9], initialised with the official weights provided by OpenAI. For the text encoder, we adopt the OpenAI ImageNet prompt template, e.g., "a photo of a *class name*," as input. Following [8], we also rename some of the official classes listed on Table 3 for OVSS. For CLIP, input images are resized such that the long side is 448 pixels on main paper experiments, and slide inference is performed using a 224×224 window with a stride of 112. The input to SAM retains the original image dimensions. To avoid memory issues with extremely large images, we cap the maximum image size at 1024 pixels. Images larger than this are split into 1024×1024 non-overlapping patches, processed individually by SAM, and the resulting mask predictions are then merged.

Generative VLM inference. Examples of prompts used for the generative VLMs (Qwen3-VL, GPT-5, and GPT-Image-1) are provided in Figure 3. The prompts remain fixed across all experiments, with only the input image and the question component varying. Note that Qwen3-VL and GPT-5 produce outputs in textual form, whereas GPT-Image-1 directly generates the corresponding segmentation mask.

Inference time. As shown in Table 2, our contrastive VLM pipeline is faster than SegEarth-OV [8] while achieving comparable accuracy with a small SAM grid (10×10). Note that larger grids improve IoU at the cost of speed. The generative VLM pipeline is slightly slower than SegEarth-R1 [9], mainly due to the larger 2B-parameter Qwen3-VL model compared to the 1.3B-parameter Phi model, but yields a $+2\%$ IoU improvement. Notably, both SegEarth-OV and SegEarth-R1 rely on mask decoders trained on remote sensing data, whereas our approach does not.

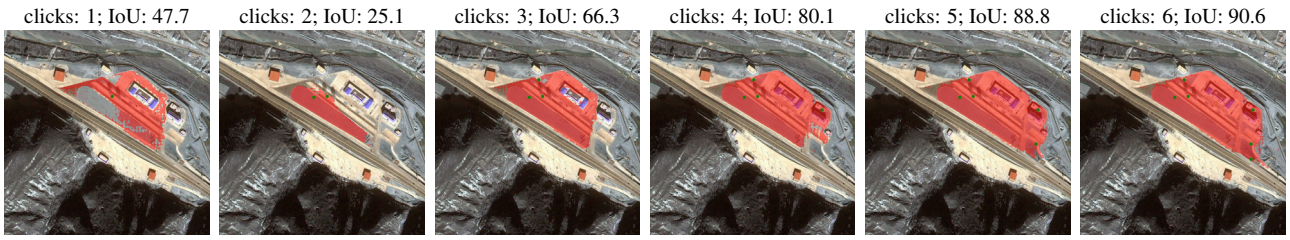

Figure 1. Visualisations from click generation procedure. Masks (red) are produced by SAM prompted by clicks (green). Reported IoU is compared to groung truth mask.

Method	Building extraction				Road Extraction				Flood Detection	Avg.
	WHU-A	WHU-S	Inria	xBD-pre	CHN6	DG	MA	SpaceNet	WBS-SI	
Trained on remote sensing data										
SegEarth-OV [8]	49.9	–	48.9	**43.1**	32.8	**20.1**	17.2	29.1	**57.9**	37.4
Zero-shot methods										
Ours	**58.7**	–	**53.8**	39.1	**33.5**	15.0	**20.3**	**29.5**	56.3	**38.3**

Table 1. Additional results of our contrastive VLM-based approach for text-based remote sensing segmentation on the OVSS task using images of size 896×896. Avg. denotes the average across all datasets. Best results are highlighted in **bold**.

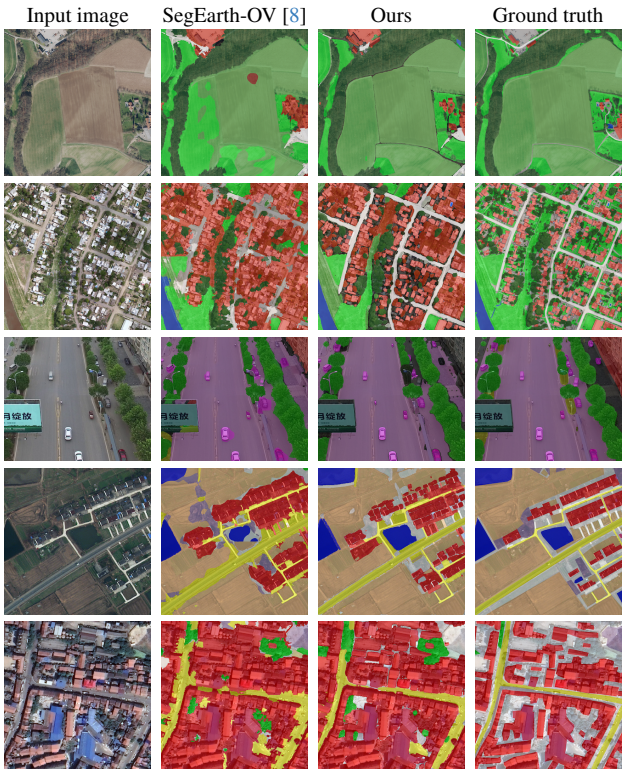

Figure 2. Qualitative comparison with baseline SegEarth-OV on OVSS datasets.

2. Quantitative Results

Table 1 provides additional OVSS results for the single-class extraction datasets using CLIP with an input resolution of 896×896. All other experimental settings follow the configuration described in the main experiments. Notably, our model achieves higher average performance than the SegEarth-OV [8] baseline on this setting, despite not being trained on remote sensing data.

3. Qualitative Results

Figure 2 presents additional visualisations of our contrastive VLM-based approach on OVSS. We showcase examples from a subset of multi-class datasets [11, 18, 21] and visually compare our predictions with SegEarth-OV [8] and the corresponding ground-truth annotations. Overall, our method produces more precise and better-defined segmentation masks in several categories compared to [8]. This improvement is particularly noticeable for roads (rows 2, 4, and 5) and cars (row 3). In row 2, our approach handles crowded scenes effectively, correctly segmenting roads, buildings, and vegetation. However, small regions between densely packed buildings remain challenging and still cause occasional confusion.

Figure 4 shows additional visualisations of our generative VLM-based approach for reasoning-based segmentation. We visually compare our predicted masks with those produced by the recent SegEarth-R1 [9] and GPT-Image-1 [14]. In addition, Figure 5 shows more qualitative results for reasoning-based segmentation on EarthReason, while Figure 6 depicts visualisations for referring segmentation

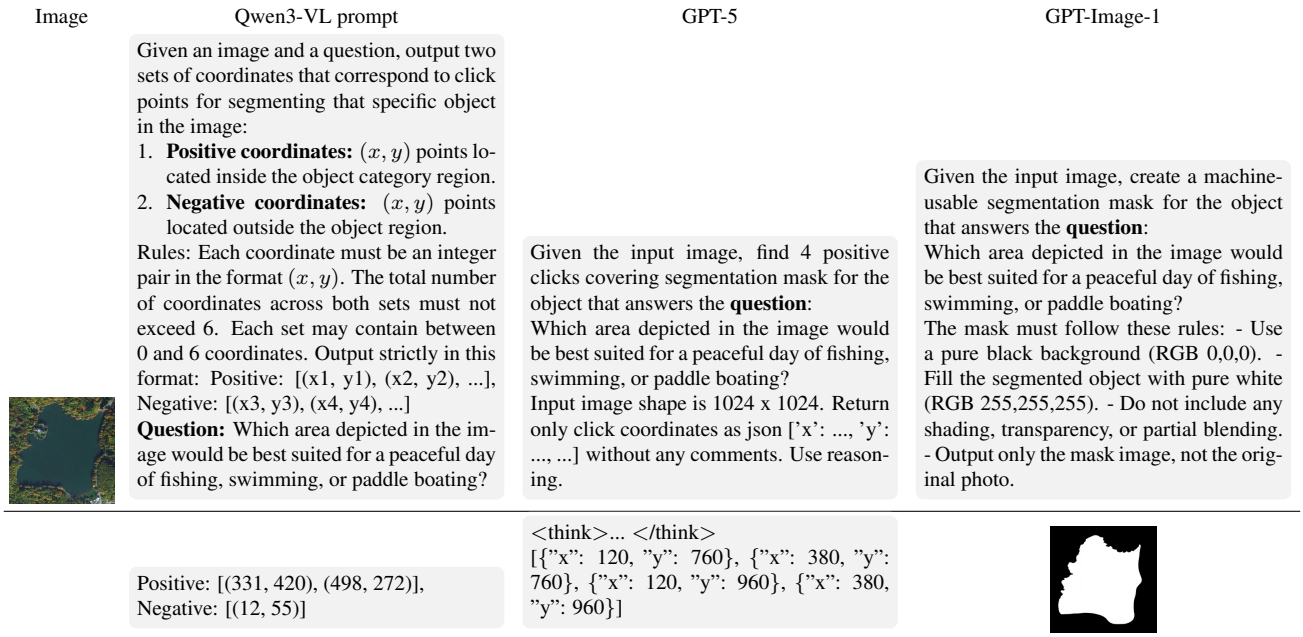

Figure 3. Top row: input image and corresponding prompts for the different generative VLM settings. Bottom row: outputs produced by each model.

using RRSIS-D dataset.

Failure cases. Figure 7 highlights difficult scenarios. In these cases, the generative VLM-based model produces clicks that lead to segmentation masks with low IoU in both referring and reasoning tasks. We observe three recurring failure modes in the reasoning examples. First, the model might produce regions that are plausible but not the correct answer according to the question, as seen in rows one and three. Second, some questions require additional contextual understanding, as in row two. Here, the model selects an area that is semantically reasonable for answering the question (venues for recreational activities). However, the ground-truth annotation indicates a different region. Finally, we observe cases in which the model selects suitable click locations as in row 4. The target region, however, involves multiple, poorly delimited areas, preventing the segmentation model (SAM) from producing a consistent mask. In referring segmentation scenarios, some errors arise from ambiguous descriptions or annotations, such as the example shown in row 5.

4. Datasets

OpenEarthMap [21] provides globally distributed satellite and aerial imagery with a spatial resolution ranging from 0.25 to $0.5m$. It comprises 9 classes including background class. We follow [8] setup and evaluate on its validation set, excluding xBD data.

LoveDA [18] contains $0.3m$ resolution images sourced from Google Earth, covering both urban and rural scenes. It includes 6 foreground categories and 1 background class. We use the validation set for evaluation.

iSAID [19] contains $655,451$ annotated object instances across 15 categories in $2,806$ high-resolution images. The dataset features large scale variation, dense object distributions, and imbalanced category frequencies, reflecting real-world aerial conditions. All images are identical to those in DOTA-v1.0 [20], primarily collected from Google Earth, JL-1, and GF-2 satellites. We evaluate on its validation set, which consist on $11,644$ image patches.

Potsdam and **Vaihingen** [6] datasets are designed for urban semantic segmentation appearing in the 2D Semantic Labeling Contest. Their spatial resolutions are $5cm$ and $9cm$, respectively, each containing 6 classes. We use their validation sets for evaluation.

UAVid [11] consists of 30 video sequences captured in 4K resolution from oblique urban views. We treat individual frames as independent images and follow merging process of some categories as in [8]. The final dataset contains 5 foreground classes and 1 background class. We evaluate it using its test set.

UDD5 [3] is captured by a DJI Phantom 4 UAV flying at altitudes varying from 60 to $100m$. It contains 4 foreground categories and 1 background class. We use its validation set for evaluation.

VDD [2] is collected with a DJI Mavic Air II drone, comprising 400 RGB images with a resolution of 4000×3000

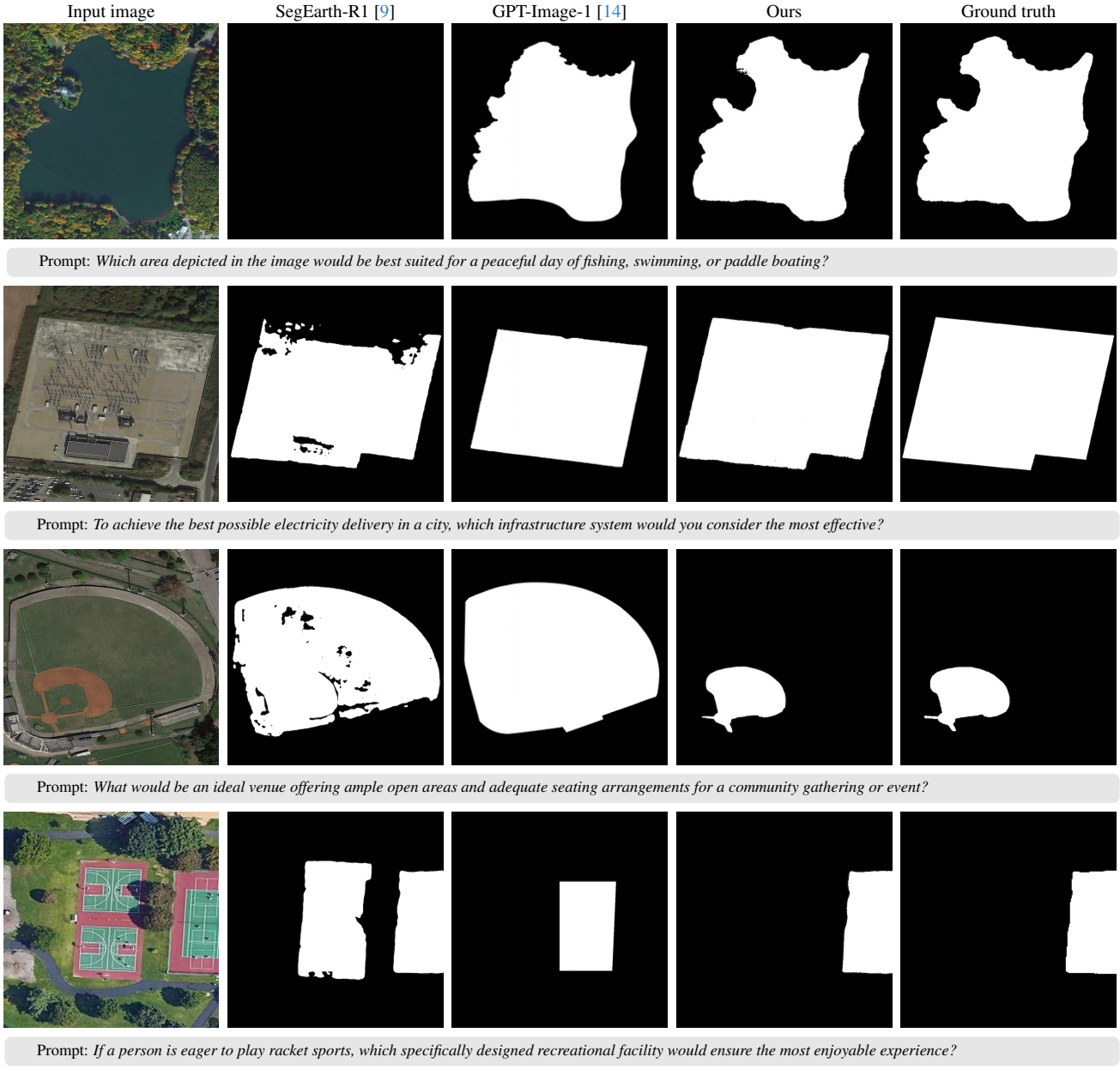

Figure 4. Qualitative comparison with baseline methods on EarthReason dataset.

pixels. The images are taken from altitudes between $50m$ and $120m$. It includes 6 foreground classes and 1 background class. We evaluate on its test set.

WHUAerial [7] comprises manually curated aerial and satellite imagery collections for building extraction. The aerial subset contains over 220 k individual building instances extracted from high-resolution imagery $(0.075m)$ covering approximately $450km^2$ of Christchurch, New Zealand. We use its validation set for evaluation.

WHUSat.II [7] includes six adjacent satellite images covering $860km^2$ across East Asia with 0.45 m ground resolution. It contains $34,085$ manually annotated buildings, cropped into $17,388$ non-overlapping tiles. This subset is specifically designed to assess model generalisation across similar building styles from different data sources within the same geographic region. We use its test set for evaluation.

Inria [12] dataset contains diverse urban environments, from dense city centers (*e.g.*, San Francisco) to alpine towns (*e.g.*, Lienz). Each subset contains distinct cities, enabling evaluation of cross-region generalisation under varying geographic, illumination, and seasonal conditions. It covers 810 km^2 with a spatial resolution of $0.3m$. We use the test

Algorithm 1: Iterative Click Generation for Synthetic Training Sequences

Input: Image I, ground-truth mask M, SAM model $\mathscr{S}$, maximum iterations $T = 6$, IoU threshold $\tau = 0.98$

Output: Click sequence $\mathcal{C}$

Initialise click sequence: $\mathcal{C} \leftarrow \emptyset$

for $i \leftarrow 1$ **to** T **do**

```
// Predict intermediate mask
   from current click set
```

$M_{i-1} \leftarrow \mathscr{S}(I, \text{prompt} = \mathcal{C})$

```
// Stop if sufficiently close to
   ground truth
```

if $IoU(M_{i-1}, M) \geq \tau$ **then**

  break

end

```
// Compute false-negative and
   false-positive regions
```

$E_+ \leftarrow M - M_{i-1}$

$E_- \leftarrow M_{i-1} - M$

```
// Sample next click using a
   distance transform over
   errors
```

$D \leftarrow \text{DistanceTransform}(E_+ \cup E_-)$

Sample $c_i \sim D$

```
// Assign click polarity
```

if $c_i \in E_+$ **then**

  Label c_i as positive

else

  Label c_i as negative

end

```
// Update click sequence
```

$\mathcal{C} \leftarrow \mathcal{C} \cup \{c_i\}$

end

return $\mathcal{C}$

Method	VLM	Decoder	Time↓	IoU↑
Contrastive VLM; LoveDA test set				
SegEarth-OV [8]	0.15B	SimFeatUp*	4.1	36.9
Ours	0.15B	SAM (10×10)	**3.6**	36.2
Ours	0.15B	SAM (29×29)	14.0	**38.2**
Generative VLM; EarthReason test set				
SegEarth-R1 [9]	1.3B	Mask2Former*	**0.44**	70.7
Ours	2B	SAM	0.58	**72.7**

Table 2. Inference time comparison. * indicates that the component is trained on remote sensing data.

set for evaluation.

xBD [5] is a large-scale, high-resolution satellite imagery benchmark designed for building damage assessment and

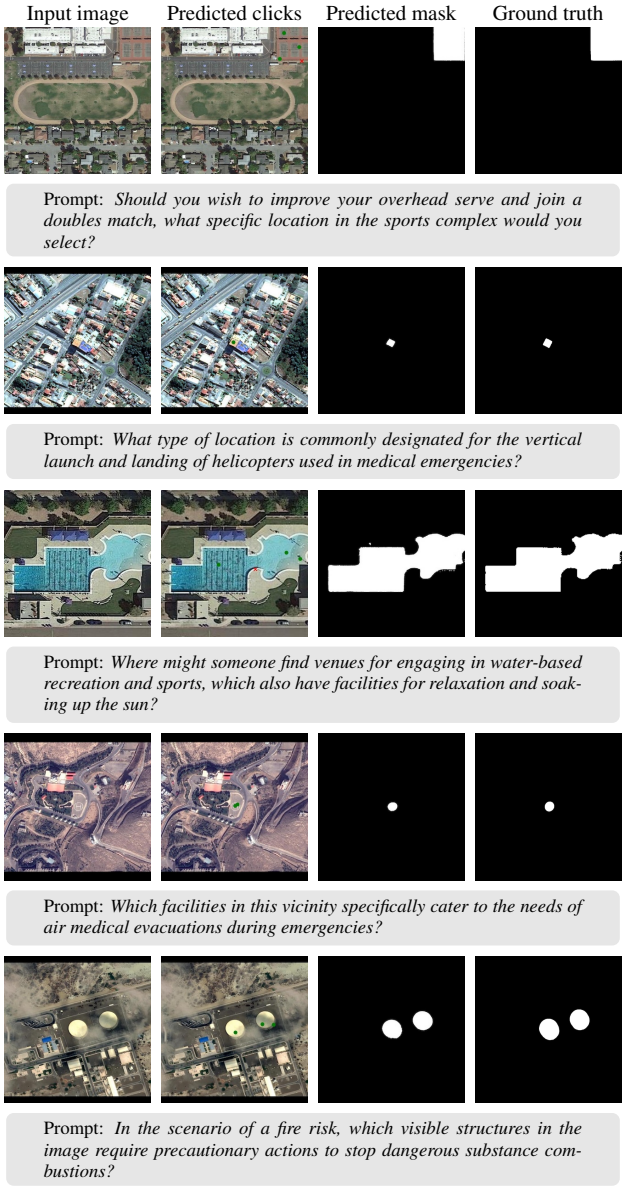

Figure 5. Qualitative results for reasoning segmentation on EarthReason dataset.

change detection in disaster scenarios. It comprises $22,068$ pre and post-disaster image pairs covering 19 natural hazard events, spanning a total area of $45,362km^2$ and including $850,736$ annotated building footprints. Its spatial resolution is $0.8m$. We use the pre-disaster satellite data of test set for evaluation.

CHN6-CUG [23] is a large-scale, manually annotated benchmark for pixel-level road extraction from satellite imagery, collected from Google Earth. It includes imagery from six representative Chinese cities capturing diverse levels of urbanisation and road structures. It contains 4511

Dataset	Source / Sensor	Resolution	Classes (FG+BG)	Split Used
Multi-class Open-Vocabulary Semantic Segmentation				
OpenEarthMap [21]	Satellite & aerial imagery (global)	0.25–0.5 m	8 + 1	Val
LoveDA [18]	Google Earth (urban/rural)	0.3 m	6 + 1	Val
iSAID [19]	Google Earth, JL-1, GF-2	-	15 + 1	Val
Potsdam[6]	Aerial (urban)	5 cm	5 + 1	Val
Vaihingen[6]	Aerial (urban)	9 cm	5 + 1	Val
UAVid [11]	UAV video (4K, slanted view)	-	5 + 1	Test
UDD5 [3]	DJI Phantom 4 UAV	60–100 m altitude	4 + 1	Val
VDD [2]	DJI Mavic Air II UAV	50–120 m altitude	6 + 1	Test
Single-class Open-Vocabulary Semantic Segmentation				
WHUAerial	Aerial imagery (Christchurch, NZ)	0.075 m	1 + 1	Val
WHUSat.II	Satellite imagery (East Asia)	0.45 m	1 + 1	Test
Inria	Aerial imagery (global cities)	0.3 m	1 + 1	Val
xBD	Satellite imagery (multi-disaster)	0.8 m	1 + 1	Test
CHN6-CUG	Google Earth (Chinese cities)	0.5 m	1 + 1	Test
DeepGlobe	Satellite imagery (Asia)	5 m	1 + 1	Val
Massachusetts	Satellite imagery (urban/suburban/rural)	1 m	1 + 1	Test
SpaceNet	Satellite imagery (Las Vegas, Paris, Shanghai, Khartoum)	0.3 m	1 + 1	Test
WBS-SI	Satellite imagery (global)	-	1 + 1	Custom [8]
Referring Segmentation				
RRSIS-D	-	-	-	Val & Test
Reasoning Segmentation				
EarthReason	-	-	-	Val & Test

Table 3. Summary of datasets used in our paper. Rows highlighted in brown correspond to building extraction datasets, green to road extraction, and blue to flood detection datasets. *FG* and *BG* are for foreground and background classes respectively.

Dataset	Class names
OpenEarthMap [21]	Bareland, Rangeland, Developed Space, Road, Tree, Water, Agriculture Land, and Building
LoveDA [18]	Building, Road, Water, Barren, Forest, and Agriculture
iSAID [19]	Plane, Ship, Storage Tank, Baseball Diamond, Tennis Court, Basketball Court, Ground Track Field, Harbor, Bridge, Large Vehicle, Small Vehicle, Helicopter, Roundabout, Soccer Ball Field and Swimming Pool
Potsdam [6]	Impervious Surfaces, Buildings, Low Vegetation, Tree, and Car
Vaihingen [6]	Impervious Surfaces, Buildings, Low Vegetation, Tree, and Car
UAVid [11]	Background, Building, Road, Car, Tree, Vegetation, and Human
UDD5 [3]	Vegetation, Building, Road, Vehicle, and Background
VDD [2]	Background, Facade, Road, Vegetation, Vehicle, Roof, and Water
WHUAerial [7]	Background, and Building
WHUSat.II [7]	Background, and Building
Inria [12]	Background, and Building
xBD [5]	Background, and Building
CHN6-CUG [23]	Background, and Road
DeepGlobe [4]	Background, and Road
Massachusetts [13]	Background, and Road
SpaceNet [17]	Background, and Road
WBS-SI [15]	Background, and Water

Table 4. Class names of OVSS datasets.

labeled images with a spatial resolution of 0.5m. We use its test set for evaluation.

DeepGlobe [4] provides high-resolution (0.5 m) satellite imagery sampled from the DigitalGlobe + Vivid collection, covering regions in Thailand, Indonesia, and India. It contains 8,570 RGB images spanning $2,220km^2$ in total. Pixel-level annotations delineate road and background classes, capturing diverse surfaces and urban–rural variations. We use the validation set for evaluation.

Massachusetts [13] is an aerial imagery dataset for road segmentation, designed to address challenges such as occlusions from trees, building shadows, and road texture variations. It contains 1,171 aerial RGB images, covering a total area of $2,600km^2$. Road labels are generated by rasteris-

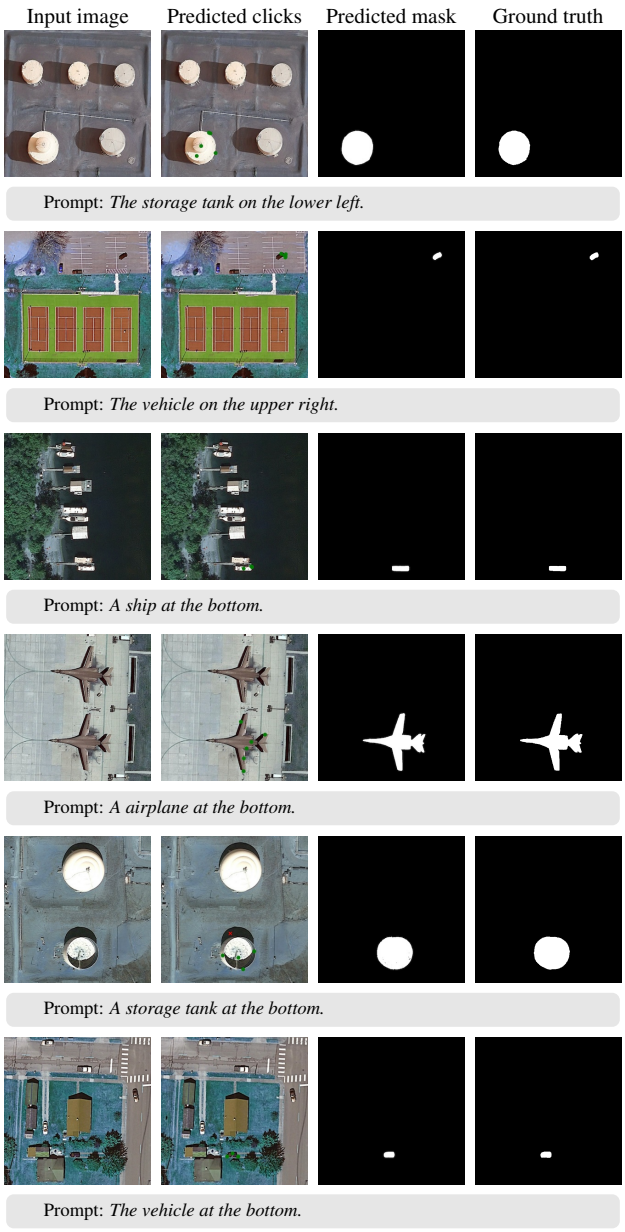

Figure 6. Qualitative results for referring segmentation on RRSIS-D dataset.

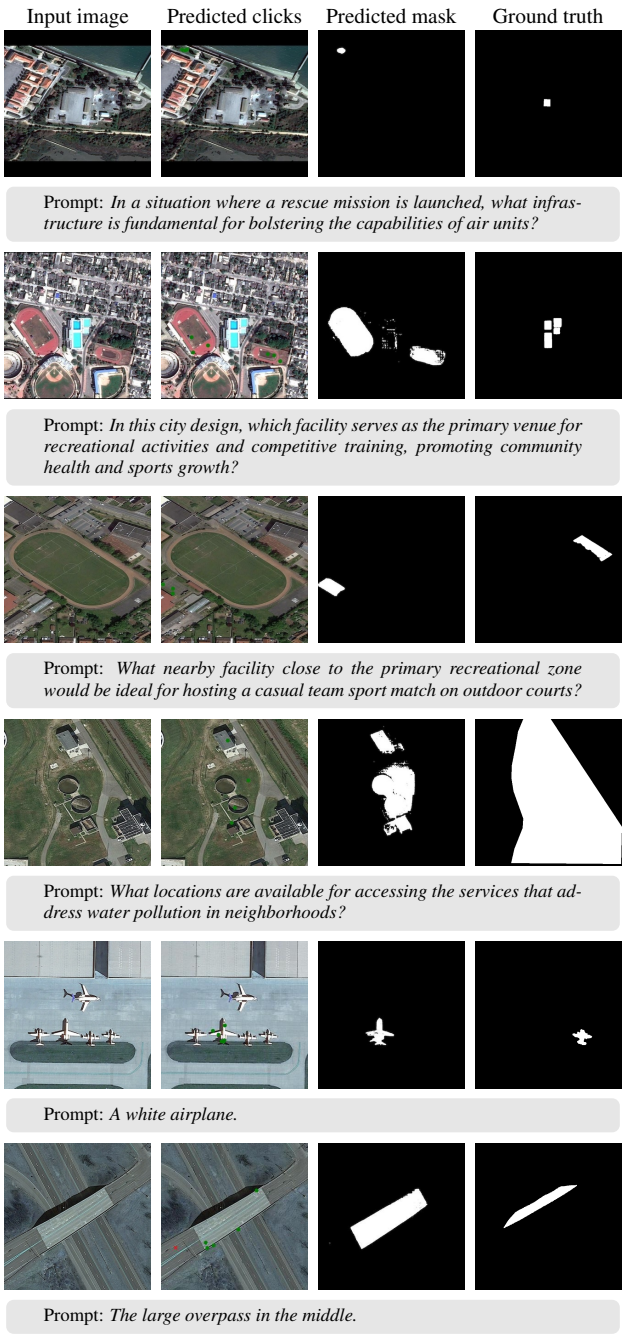

Figure 7. Failure cases on the EarthReason and RRSIS-D datasets for reasoning and referring segmentation, respectively.

ing OpenStreetMap centerlines with a 7-pixel width at 1m spatial resolution. The dataset includes diverse urban, suburban, and rural scenes. We use its test set for evaluation.

SpaceNet [17] contains satellite imagery with a spatial resolution of 0.3m, covering Las Vegas, Paris, Shanghai, and Khartoum. It was introduced for the SpaceNet Road Detection and Routing Challenge, designed to support automated road extraction from very high-resolution imagery. We use the test set for evaluation.

WBS-SI [15] is a satellite imagery dataset designed for water body segmentation. Following the setup in [8], we perform our evaluation using their proposed test split.

EarthReason [9] is the first large-scale benchmark dataset designed for geospatial pixel reasoning. It contains 5,434 high-resolution remote sensing images, each annotated with a manually created segmentation mask targeting specific regions of interest. Alongside these masks, the dataset in-

cludes over 30,000 implicit question-answer pairs that require spatial understanding and reasoning to identify the correct target regions. The dataset is divided into training, validation, and test splits of 2,371, 1,135, and 1,928 images, respectively.

RRSIS-D [10] is a dataset designed for Referring Remote Sensing Image Segmentation (RRSIS), specifically to handle significant variations in spatial resolution and object orientation. The dataset is constructed by converting bounding box annotations from the RSVGD dataset [22] into instance masks. It comprises 20 semantic categories, including aircraft, golf courses, highway service areas, baseball fields, and stadiums. It is further extended with 7 descriptive attributes to enhance the clarity and expressiveness of referring expressions. RRSIS-D exhibits scale variability, with some targets covering minimal pixel areas and others exceeding 400,000 pixels. Overall, the dataset comprises 17,402 image-description-mask triplets, divided into 12,181 for training, 1,740 for validation, and 3,481 for testing.

References

[1] Anton Antonov, Andrey Moskalenko, Denis Shepelev, Alexander Krapukhin, Konstantin Soshin, Anton Konushin, and Vlad Shakhuro. Rclicks: Realistic click simulation for benchmarking interactive segmentation. In *Advances in Neural Information Processing Systems*, pages 127673–127710. Curran Associates, Inc., 2024. 1

[2] Wenxiao Cai, Ke Jin, Jinyan Hou, Cong Guo, Letian Wu, and Wankou Yang. Vdd: Varied drone dataset for semantic segmentation. *Journal of Visual Communication and Image Representation*, 109:104429, 2025. 3, 6

[3] Yu Chen, Yao Wang, Peng Lu, Yisong Chen, and Guoping Wang. Large-scale structure from motion with semantic constraints of aerial images. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, pages 347–359. Springer, 2018. 3, 6

[4] DeepGlobe Consortium. Deepglobe: A challenge to parse the earth through satellite images. http://deepglobe.org/, 2018. Accessed: 2025-11-11. 6

[5] Ritwik Gupta, Richard Hosfelt, Sandra Sajeev, Nirav Patel, Bryce Goodman, Jigar Doshi, Eric Heim, Howie Choset, and Matthew Gaston. xbd: A dataset for assessing building damage from satellite imagery. *arXiv preprint arXiv:1911.09296*, 2019. 5, 6

[6] ISPRS Foundation. Isprs benchmark on semantic labeling. https://www.isprs.org/resources/datasets/benchmarks/UrbanSemLab/, 2025. Accessed: 2025-11-11. 3, 6

[7] Shunping Ji, Shiqing Wei, and Meng Lu. Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set. *IEEE Transactions on geoscience and remote sensing*, 57(1):574–586, 2018. 4, 6

[8] Kaiyu Li, Ruixun Liu, Xiangyong Cao, Xueru Bai, Feng Zhou, Deyu Meng, and Zhi Wang. Segearth-ov: Towards training-free open-vocabulary segmentation for remote sensing images. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 10545–10556, 2025. 1, 2, 3, 5, 6, 7

[9] Kaiyu Li, Zepeng Xin, Li Pang, Chao Pang, Yupeng Deng, Jing Yao, Guisong Xia, Deyu Meng, Zhi Wang, and Xiangyong Cao. Segearth-r1: Geospatial pixel reasoning via large language model. *arXiv preprint arXiv:2504.09644*, 2025. 1, 2, 4, 5, 7

[10] Sihan Liu, Yiwei Ma, Xiaoqing Zhang, Haowei Wang, Jiayi Ji, Xiaoshuai Sun, and Rongrong Ji. Rotated multi-scale interaction network for referring remote sensing image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26658–26668, 2024. 8

[11] Ye Lyu, George Vosselman, Gui-Song Xia, Alper Yilmaz, and Michael Ying Yang. Uavid: A semantic segmentation dataset for uav imagery. *ISPRS journal of photogrammetry and remote sensing*, 165:108–119, 2020. 2, 3, 6

[12] Emmanuel Maggiori, Yuliya Tarabalka, Guillaume Charpiat, and Pierre Alliez. Can semantic labeling methods generalize to any city? the inria aerial image labeling benchmark. In *2017 IEEE International geoscience and remote sensing symposium (IGARSS)*, pages 3226–3229. IEEE, 2017. 4, 6

[13] Volodymyr Mnih. *Machine learning for aerial image labeling*. University of Toronto (Canada), 2013. 6

[14] OpenAI. Image generation guide: Gpt-image-1 model. https://platform.openai.com/docs/guides/image-generation?image-generation-model=gpt-image-1, 2025. Accessed: 2025-11-12. 2, 4

[15] shirshmall. Water body segmentation in satellite images. https://www.kaggle.com/datasets/shirshmall/water-body-segmentation-in-satellite-images. Accessed: 2025-11-11. 6, 7

[16] Konstantin Sofiiuk, Ilya A Petrov, and Anton Konushin. Reviving iterative training with mask guidance for interactive segmentation. In *2022 IEEE international conference on image processing (ICIP)*, pages 3141–3145. IEEE, 2022. 1

[17] Adam Van Etten, Dave Lindenbaum, and Todd M Bacastow. Spacenet: A remote sensing dataset and challenge series. *arXiv preprint arXiv:1807.01232*, 2018. 6, 7

[18] Junjue Wang, Zhuo Zheng, Ailong Ma, Xiaoyan Lu, and Yanfei Zhong. Loveda: A remote sensing land-cover dataset for domain adaptive semantic segmentation. *arXiv preprint arXiv:2110.08733*, 2021. 2, 3, 6

[19] Syed Waqas Zamir, Aditya Arora, Akshita Gupta, Salman Khan, Guolei Sun, Fahad Shahbaz Khan, Fan Zhu, Ling Shao, Gui-Song Xia, and Xiang Bai. isaid: A large-scale dataset for instance segmentation in aerial images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 28–37, 2019. 3, 6

[20] Gui-Song Xia, Xiang Bai, Jian Ding, Zhen Zhu, Serge Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, and Liangpei Zhang. Dota: A large-scale dataset for object detection in aerial images. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 3

[21] Junshi Xia, Naoto Yokoya, Bruno Adriano, and Clifford Broni-Bediako. Openearthmap: A benchmark dataset for

global high-resolution land cover mapping. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6254–6264, 2023. 2, 3, 6

[22] Yang Zhan, Zhitong Xiong, and Yuan Yuan. Rsvg: Exploring data and models for visual grounding on remote sensing data. *IEEE Transactions on Geoscience and Remote Sensing*, 61: 1–13, 2023. 8

[23] Qiqi Zhu, Yanan Zhang, Lizeng Wang, Yanfei Zhong, Qingfeng Guan, Xiaoyan Lu, Liangpei Zhang, and Deren Li. A global context-aware and batch-independent network for road extraction from vhr satellite imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 175:353–365, 2021. 5, 6