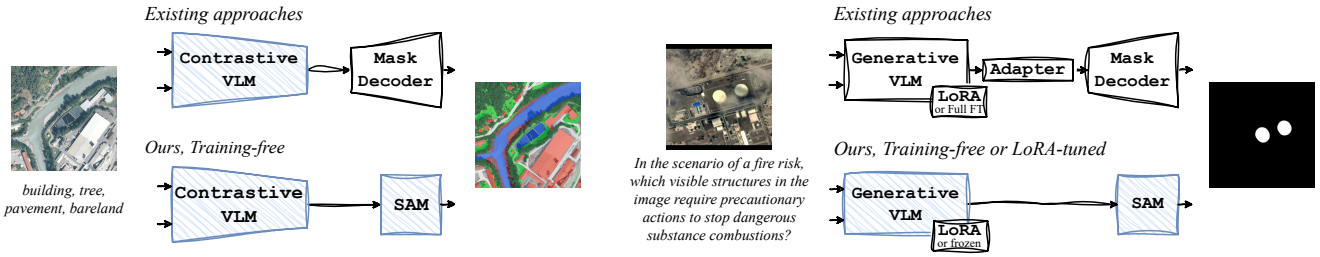
Enabling Training-Free Text-Based Remote Sensing Segmentation

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



(a) Contrastive VLM for text-based RS segmentation

(b) Generative VLM for text-based RS segmentation

Figure 1. Existing methods [32, 34, 54] rely on additional trainable mask decoders and adapters. We propose a training-free methodology that combines VLMs and SAM without introducing new trainable components. Additionally, with LoRA fine-tuning, our method achieves state-of-the-art performance on reasoning segmentation. Blue is for frozen components.

Abstract

Recent advances in Vision Language Models (VLMs) and Vision Foundation Models (VFMs) have opened new opportunities for zero-shot text-guided segmentation of remote sensing imagery. However, most existing approaches still rely on additional trainable components, limiting their generalisation and practical applicability. In this work, we investigate to what extent text-based remote sensing segmentation can be achieved without additional training, by relying solely on existing foundation models. We propose a simple yet effective approach that integrates contrastive and generative VLMs with the Segment Anything Model (SAM), enabling a fully training-free or lightweight LoRA-tuned pipeline. Our contrastive approach employs CLIP as mask selector for SAM's grid-based proposals, achieving state-of-the-art open-vocabulary semantic segmentation (OVSS) in a completely zero-shot setting. In parallel, our generative approach enables reasoning and referring segmentation by generating click prompts for SAM using GPT-5 in a zero-shot setting and a LoRA-tuned Qwen-VL model, with the latter yielding the best results. Extensive experiments across 19 remote sensing benchmarks, including open-vocabulary, referring, and reasoning-based tasks, demonstrate the strong capabilities of our approach. Code will be released here.

1. Introduction

Remote sensing imagery has become a cornerstone of earth observation, supporting critical applications such as land-cover mapping, environmental monitoring, and disaster response [23, 58, 59, 61]. Recent advances in deep learning, in particular for pixel-level segmentation have highly improved the accuracy and scalability of such analyses [17, 87]. However, most methods still follow supervised setups, depending on large-scale, domain-specific annotated datasets for training. The costly and inconsistent process of collecting dense pixel-level annotations continues to limit progress, particularly for fine-grained or rapidly evolving geospatial categories.

Recently, Vision Language Models (VLMs) [1, 3, 49] and Vision Foundation Models (VFMs) [52] have shown impressive zero-shot capabilities on natural images, achieving text-based segmentation without additional supervision. These models offer an appealing direction for remote sensing, where various successful approaches have emerged [4, 27, 32, 34, 53, 54, 75, 78, 83, 84]. Nevertheless, most methods in this domain still rely on additional trainable adapters [34, 36, 83], lightweight heads [32, 78], or token-level bridges [4, 27, 53] to link visual and textual modalities.

Our work is based on the premise that relying exclusively on pretrained foundation models enables a training-

free approach for text-based remote sensing segmentation. This raises a central question: *To what extent can such segmentation be achieved solely through pretrained foundation models, without introducing any additional trainable components?* To explore this, our approach builds on two key elements: VLMs and VFMs. VLMs provide the multi-modal link between textual queries and visual content, while VFMs (such as SAM [52]), offer a generic mechanism for mask generation. We investigate strategies to integrate these components without introducing additional trainable parameters.

Specifically, we propose two complementary pipelines, the first uses *contrastive VLMs*, like CLIP [49], as semantic selectors over SAM's category-agnostic mask proposals for OVSS. The second employs *generative VLMs*, such as GPT-5 [47] and Qwen-VL [3], as SAM prompters with spatial clicks for referring and reasoning-based segmentation scenarios. While the first approach operates in a fully zero-shot manner, the second can be applied either zero-shot or with lightweight LoRA fine-tuning [21]. Figure 1 provides a visual comparison of our proposed approach, and contrasts it with recent text-based remote sensing segmentation methods. In summary, our contributions are as follows:

- We investigate the extent to which text-based remote sensing segmentation can be accomplished by using only existing VLMs and SAM, without introducing additional trainable components.
- We propose two complementary approaches for combining VLMs with SAM: (i) using a contrastive VLM to select masks from SAM's grid-based proposals, and (ii) using a generative VLM to generate click prompts for SAM-based segmentation.
- We demonstrate that contrastive VLM-based pipeline enables fully training-free segmentation, achieving state-of-the-art performance in OVSS of remote sensing imagery.
- We further show that minimal LoRA fine-tuning of the generative VLM-based approach, with SAM kept frozen, yields state-of-the-art results on reasoning and referring segmentation with remote sensing imagery.

2. Related work

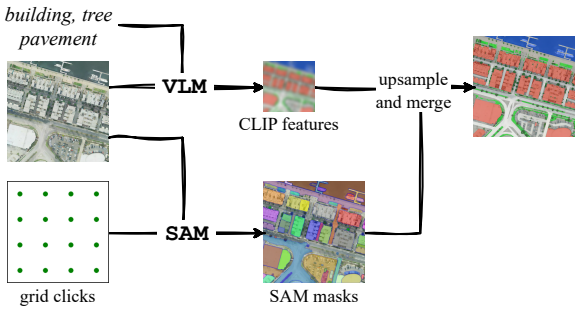
2.1. VLMs for Text-Based Segmentation

Text-based image segmentation aims to segment regions within an image based on natural language descriptions. Recent advances [27, 53, 78] on multi-modal datasets and pretraining strategies for VLMs have revolutionised this field. Consequently, text-based segmentation is increasingly dominated by contrastive and generative VLM-based approaches. These models can localise complex language-guided targets without requiring extensive task-specific supervision.

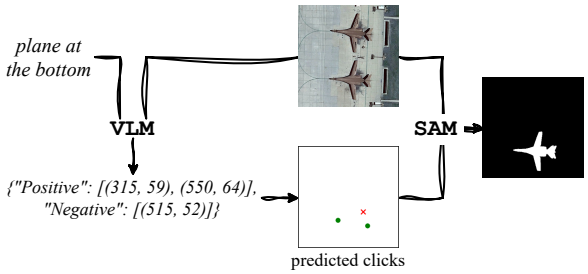
Contrastive VLMs [42, 49] are trained to align image and text representations through contrastive learning on paired data. Starting from CLIP [49], they have shown remarkable zero-shot performance on image classification, which naturally extends to semantic segmentation [41]. These approaches are usually categorised based on the amount of needed supervision. *Training-free* methods attempt to exploit the inherent localisation capabilities of CLIP with minimal modifications. For instance, MaskCLIP [83] proposes to extract the value embedding of the last self-attention block of CLIP's vision encoder for dense prediction tasks. Following this work, other studies [28] generalise the query-key attention to a self-self attention mechanism. This includes, the value-value attention in CLIP-Surgery [37], the query-query and key-key attention in SCLIP [63], and generalised self-self attention combination in GEM [7]. Another stream of work [5, 26, 55, 60, 68] adopts a two-stage method. The first stage generates category-agnostic mask proposals, while the second stage classifies them. *Trainable* methods allow models to be trained on some base classes in a supervised or weakly supervised manner. Typically, some works [19, 42, 46, 50] train a localisation-aware CLIP for dense predictions. Others [13, 16, 31, 40, 74] instead fine-tune a subset of CLIP's pre-trained parameters or add a few trainable ones to adapt it for dense prediction on base classes. Finally, with additional adapters [6, 25, 29, 36, 64] CLIP can be connected with other foundation models (SAM [52] or DINO [9]) to enhance the localisation ability.

Generative VLMs [1, 3, 73] are trained to model the joint or conditional distribution of images and text via auto-regressive generation. These models allow for complex reasoning between image and language modalities. However, none of them directly support image segmentation, so need to be extended with extra components. LISA [27] establishes the paradigm by introducing a `<SEG>` token to connect LLMs with segmentation decoders. Finetuning VLMs with these novel tokens is further explored in PixelLM [53] and more recent works [4, 51, 65, 66, 76, 78, 79, 81]. SAM4MLLM [12] and SegAgent [85] propose the solution without adding novel trainable tokens, namely to use SAM [52] as a tool, conditioned by clicks or boxes, produced by VLM in form of text. Another possible tool for image generation for VLMs can be diffusion models, *e.g.* Qwen-Image [70] for Qwen3 [73] or GPT-Image-1 [48] for GPT-5 [47]. This technically gives VLMs the ability to solve segmentation, since segmentation mask can be imagined as just an image.

Overall, despite the proven capabilities of aforementioned contrastive and generative VLM-based methods on natural images, none of them originally attempted text-based remote sensing segmentation.



(a) Contrastive VLM for text-based RS segmentation



(b) Generative VLM for text-based RS segmentation

Figure 2. Inference schemes of our segmentation approaches with (a) contrastive and (b) generative VLMs.

2.2. Text-Based Remote Sensing Segmentation

Early studies on text-based remote sensing segmentation primarily rely on masked language models such as BERT [15] to encode textual inputs. These models are typically followed by a task-specific conditional convolutional decoder [10, 30, 33, 35, 39, 77, 80, 87] or diffusion-based architectures such as DiffRIS [17]. Despite notable progress, these approaches remain heavily supervised and rely on dedicated model designs for specific datasets or prompt types. Recently, SegEarth-OV [32] marks a shift towards reducing training dependency by introducing a nearly training-free framework. Their method incorporates a frozen CLIP model for image–text alignment, while training only an unsupervised mask decoder. They further explore remote sensing adaptations of CLIP, such as GeoRSCLIP [82] and RemoteCLIP [38], which improve visual grounding in geospatial imagery.

A contemporary line of research transitions to generative VLMs to handle more complex referring and reasoning prompts. For example, GeoGround [84] reformulates segmentation as per-tile binary prediction using a VLM backbone without an explicit decoder. An extended version incorporates multiple auxiliary encoders and a dedicated mask decoder for finer spatial resolution. SegEarth-R1 [34] further advances this line of work by introducing a reasoning-focused remote sensing benchmark. It also proposes a simplified yet trainable mask encoder that is conditioned jointly on the vision and language embeddings.

Typically, existing methods achieve language-conditioned segmentation by adding trainable components on top of VLMs. These components may include mask decoders, adapters, or token-level bridges. In contrast, our work shows that combining only a VLM with SAM is sufficient to achieve state-of-the-art results across open-vocabulary, referring, and reasoning segmentation tasks in remote sensing, without any additional trainable modules.

3. Methodology

In this work, we address the task of *text-based remote sensing image segmentation*. Given an image I and a textual query t, the goal is to produce a segmentation mask M that corresponds to the region in the image described by the text. Our objective is to design a solution that requires minimal training, and ideally operates in a fully zero-shot manner, *i.e.*, without any task-specific training.

To process textual instructions and input images, we adopt VLMs that are already pre-trained on large-scale image–text pairs in a self-supervised manner. These models have demonstrated strong zero-shot generalisation across a wide range of vision–language tasks. However, existing VLMs do not inherently generate segmentation masks, limiting their direct applicability to segmentation tasks. On the other hand, VFMs such as SAM [52] have shown remarkable performance across various segmentation tasks, including semantic, instance, and panoptic segmentation. As illustrated on Figure 2, we propose to combine these powerful pre-trained components, a VLM for understanding text and images, and SAM for mask generation. This combination does not require any additional trainable modules, enabling a fully training-free paradigm for text-based remote sensing segmentation.

VLMs can be broadly categorised into two types. The first type is *contrastive* models, such as CLIP [49], which are trained to align images and text in a shared embedding space. The second type is *generative* models, such as Qwen-VL [3], which are trained to autoregressively generate text conditioned on visual inputs. In the following sections, we describe how we use both types of VLMs in combination with SAM to solve the task of text-based remote sensing segmentation.

3.1. Contrastive VLMs as SAM Mask Selectors

Pipeline. Let $\mathscr{F}$ denote a contrastive VLM and $\mathscr{S}$ the SAM model. Given an input image I and a textual prompt t, the contrastive VLM processes I and t and computes a per-pixel foreground probability map $p(x, y)$ for t. In parallel, SAM produces a set of K category-agnostic mask proposals $\{M_k\}_{k=1}^{K}$ given I and a set of clicks $\mathcal{C}$ in form of a regular 2D grid.

For each SAM mask M_k, we determine whether it corresponds to the target object by counting the proportion of pixels with $p(x, y) > 0.5$ with the following indicator function,

$$\delta_k = \begin{cases} 1, & \text{if } \frac{1}{|M_k|} \sum_{(x,y)\in M_k} \mathbf{1}[p(x, y) > 0.5] > 0.5 \,, \\ 0, & \text{otherwise.} \end{cases}$$

The final prediction is obtained by merging all relevant masks as follows,

$$M = \bigcup_{k=1}^{K} \{M_k \mid \delta_k = 1\} \,.$$

Extension to multi-class segmentation. Given m text prompts $\{t_i\}_{i=1}^m$, the VLM predicts per-pixel probabilities $p_i(x, y)$ for each class. Each pixel is assigned to the class with the highest probability as follows,

$$\text{class}(x, y) = \arg \max_{i\in\{1,\dots,m\}} p_i(x, y) \,.$$

To mitigate CLIP's global bias, we apply the debiasing technique from [32], which subtracts a scaled <CLS> token from all patch tokens. Each SAM mask M_k is then assigned to the class that dominates within its area,

$$\ell_k = \arg \max_{i\in\{1,\dots,m\}} |\{(x, y) \in M_k \mid \text{class}(x, y) = i\}| \,.$$

The final segmentation for class i is expressed as:

$$M^i = \bigcup_{k=1}^{K} \{M_k \mid \ell_k = i\} \,.$$

Note that masks or pixels not assigned to any class remain background.

3.2. Generative VLMs as SAM prompters

Pipeline. To avoid introducing any trainable components between the VLM and SAM, the only way to condition SAM with language is by expressing its prompts in text. SAM supports two types of lightweight prompts (*clicks* and *bounding boxes*) that can be easily described with text. For consistency with our contrastive-VLM approach, we focus only on click-based prompting.

Let the generative VLM be denoted as $\mathscr{F}$. Given an image I and a textual instruction t, the VLM outputs a set of clicks $\mathcal{C} = \{c_i\}_{i=1}^n$ that indicate the target region

$$\mathcal{C} = \mathscr{F}(I, t) \,.$$

These clicks serve as prompts for SAM ($\mathscr{S}$), which generates the final segmentation mask M:

$$M = \mathscr{S}(I, \text{prompt} = \mathcal{C}) \,.$$

Each click is labeled as either *positive* or *negative*, indicating whether the corresponding location should be included or excluded from the segmentation mask. In textual form, the click set is represented as:

$$\{\text{“Positive”} : [(x_{+,1}, y_{+,1}), (x_{+,2}, y_{+,2}), \dots], \\ \text{“Negative”} : [(x_{-,1}, y_{-,1}), (x_{-,2}, y_{-,2}), \dots]\} \,.$$

If no negative clicks are present, SAM generates the mask using only the positive set.

Training. Preliminary experiments show that this pipeline already achieves reasonable segmentation quality. This is observed in a fully zero-shot manner when $\mathscr{F}$ is a large proprietary VLM. However, to further improve performance and generalisation, we propose to fine-tune a smaller open-source generative VLM for click generation, while keeping SAM completely frozen.

To train the VLM, we simply concatenate the image I, textual instruction t, and the click sequence $\mathcal{C}$ in text form into a single token sequence. The model is then optimised using standard next-token prediction (cross-entropy loss). However, the missing component is the supervision signal, since existing segmentation datasets provide masks M but not click annotations. We address this by automatically converting masks into click sequences, as described below.

Training clicks generation. Existing text-based image segmentation datasets usually provide ground-truth annotations in the form of per-pixel masks. Our goal is to convert these masks M into a sequence of clicks $\mathcal{C}$ without human involvement. To do so, we adopt an iterative strategy inspired by interactive segmentation [2, 57].

Starting from an image and its corresponding ground-truth mask, SAM is prompted with an initial positive click inside the target object to produce a mask. This prediction is then compared with the ground-truth mask to identify under-segmented and over-segmented regions. Additional clicks are placed in these regions, positive clicks in missing areas and negative clicks in unwanted regions. Then, SAM is prompted again to update the mask. This process is repeated until a stopping condition is met (*e.g.*, achieving a target IoU or reaching a maximum number of clicks). The resulting synthetic click sequences $\mathcal{C}$ are then used to fine-tune the generative VLM for click generation. More details about this process are given in Appendix.

3.3. Application to Text-Based Remote Sensing Segmentation

In text-based remote sensing segmentation, text prompts vary significantly in complexity. Existing settings can be grouped into three categories: (i) **OVSS:** each class is described using a short phrase or a couple of keywords (*e.g.*, *road*, *industrial area*). (ii) **Referring segmentation:** each prompt is a full sentence that describes a specific region

or object within the image (*e.g.*, *The vehicle on the upper right*). (iii) **Reasoning segmentation:** the prompt requires multi-step reasoning or implicit understanding, without explicitly naming the target region (*e.g.*, *Which part of the infrastructure is best for rapid patient transport by emergency services?*).

Our contrastive and generative VLM-based pipelines naturally align with these three levels of complexity. Contrastive VLMs perform well with short, unambiguous prompts, making them suitable for OVSS. However, their capability degrades when prompts become longer, descriptive, or require contextual reasoning. Alternatively, generative models are better at understanding complex linguistic instructions and grounding them spatially through click prompts. Therefore, we employ the generative VLM approach for referring and reasoning-based segmentation.

This design choice raises an important question: *why not use the generative approach for all three tasks?* The limitation lies in the nature of how generative VLMs interact with SAM. These models typically produce only a small set of clicks, resulting in a single (or very few) connected mask. This is adequate for referring and reasoning segmentation tasks, where usually only one instance is expected. However, in OVSS, many semantic categories such as *forest*, *water*, or *urban area* might consist of multiple spatially disconnected regions. A single SAM prompt, even with various positive and negative clicks, often fails to capture all relevant areas when SAM is kept frozen. Consequently, OVSS requires combining multiple SAM-generated masks, which our contrastive VLM-based mask selection approach naturally enables. In summary, contrastive VLMs are preferable for OVSS tasks, while generative VLMs are more appropriated for referring and reasoning-based segmentation.

4. Experiments

4.1. Datasets and Implementation Details

OVSS. Following prior works [32], we evaluate our approach on 17 widely used datasets for multi-class and single-class remote sensing semantic segmentation. For multi-class standard semantic segmentation, we use 5 datasets depicting satellite images including OpenEarthMap [71], LoveDA [67], iSAID [69], Potsdam, and Vaihingen [22]. We also consider 3 additional datasets containing UAV images, UAVid [43], UDD5 [11], and VDD [8]. In the context of single-class semantic segmentation, the datasets depict two classes: the foreground class, corresponding to building, road or water respectively, and the background class. We employ 4 datasets for building extraction, WHUAerial [24], WHUSat.II [24], Inria [44], and xBD [20]; 4 for road extraction, CHN6-CUG [86], DeepGlobe [14], Massachusetts [45], and SpaceNet [62]; and 1 for flood detection, WBS-SI [56]. More details about datasets could be found in Appendix.

Referring and reasoning segmentation. For referring segmentation, we use the widely adopted RRSIS-D dataset [39]. RRSIS-D includes 17,402 image–description-mask triplets, divided into 12,181 for training, 1,740 for validation, and 3,481 for testing. For the reasoning segmentation task, we adopt the large-scale EarthReason benchmark [34]. This dataset contains 5,434 images, each associated with an average of six questions and corresponding masks. The dataset is split into 2,371, 1,135, and 1,928 images for the training, validation, and test sets, respectively. For both datasets, we report metrics on validation and test sets. Train split is used only for click generation and VLM fine-tuning.

Implementation details. For our contrastive VLM-based approach, we use CLIP-base [49] for image and text encoding, and SAM-L [52] as the mask generator. We sample an uniform grid of 29×29 positive click points for the main experiments. Following [32], all images are resized to 448×448 pixels for CLIP while SAM operates on images at their original resolution. Performance is measured with mIoU for multi-class and foreground IoU for single-class datasets. For the generative VLM-based approach, in the zero-shot setting, we utilise the GPT-Image-1 API [48] and GPT-5 [47] for image and clicks generation, respectively. For the fine-tuning setting, we adopt Qwen3-VL-2B [70] as the backbone model. Training is conducted on four A100 GPUs for 3 epochs, using batch size of 64, LoRA (rank 32), the AdamW optimiser with a learning rate of 2e-4 and a cosine learning rate scheduler. We employ a total of 6 clicks during training. For inference, the same SAM as in the contrastive setup is used. Performance is reported using mIoU for both referring and reasoning tasks; for the reasoning task, the final mask is obtained via average voting over six predictions per image.

4.2. Comparison with Prior Work

Enabling training-free OVSS. We evaluate our contrastive VLM-based approach on multi-class semantic segmentation benchmarks. We compare it with zero-shot natural image baselines and SegEarth-OV [32], which is the closest prior work toward training-free OVSS for remote sensing. As shown in Table 1, our method achieves state-of-the-art zero-shot performance across datasets, demonstrating strong generalisation without task-specific training.

Compared with analogous zero-shot baselines such as CLIP, our model shows a substantial performance gain, highlighting the effectiveness of combining CLIP with a frozen SAM. Against SegEarth-OV, our approach outperforms on 7 of 8 datasets, with the largest gains on UAV imagery (UAVid, UDD5, VDD). Similar to [32], it struggles on iSAID due to fine-grained categories, and on Ope-

Method	OEM	LoveDA	iSAID	Potsdam	Vaihingen	UAVid	UDD5	VDD	Avg.
Trained on remote sensing data									
SegEarth-OV [32]	40.3	36.9	21.7	48.5	40.0	42.5	50.6	45.3	39.2
Oracle	64.4	50.0	36.2	74.3	61.2	59.7	56.5	62.9	58.2
Zero-shot methods									
CLIP [49]	12.0	12.4	7.5	15.6	10.8	10.9	9.5	14.2	11.4
MaskCLIP [83]	25.1	27.8	14.5	33.9	29.9	28.6	32.4	32.9	27.2
SCLIP [63]	29.3	30.4	16.1	39.6	35.9	31.4	38.7	37.9	31.1
GEM [7]	33.9	31.6	17.7	39.1	36.4	33.4	41.2	39.5	32.3
ClearCLIP [28]	31.0	32.4	18.2	42.0	36.2	36.2	41.8	39.3	33.4
Ours	**34.2**	**38.2**	**21.9**	**50.2**	**40.6**	**44.3**	**53.8**	**46.8**	**41.3**

Table 1. Results of our contrastive VLM-based approach for text-based remote sensing segmentation on OVSS task. We evaluate 8 remote sensing multi-class datasets. *Avg.* is for average across all datasets. *Oracle* represents the upper bound, achieved by a fully supervised model [72].

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 [32]	49.2	28.4	44.6	37.0	35.4	17.8	11.5	23.8	60.2	34.2
Zero-shot methods										
CLIP [49]	17.7	3.5	19.6	16.0	7.7	3.9	4.9	7.1	18.6	11.0
MaskCLIP [83]	29.8	14.0	33.4	29.2	28.1	13.2	10.6	20.8	39.8	24.3
SCLIP [63]	33.4	21.0	34.9	25.9	21.1	7.0	7.4	14.9	32.1	22.0
GEM [7]	24.4	13.6	28.5	20.8	13.4	4.7	5.1	11.9	39.5	18.0
ClearCLIP [28]	36.6	20.8	39.0	30.1	25.5	5.7	6.4	16.3	44.9	25.0
Ours	**58.8**	**26.1**	**48.0**	**34.4**	**36.4**	**15.9**	**12.2**	**26.1**	**58.3**	**35.1**

Table 2. Results of our contrastive VLM-based approach for text-based remote sensing segmentation on OVSS task. We evaluate 9 remote sensing single-class datasets across building extraction, road extraction, and flood detection. *Avg.* is for average across all datasets.

nEarthMap minor class confusions slightly reduce performance. Notably, unlike SegEarth-OV, which requires training auxiliary components on remote sensing data (SimFeatUp [18]), our approach is entirely training-free.

On 9 single-class extraction datasets (Table 2), our approach achieves state-of-the-art results among zero-shot baselines and even surpasses [32] on 5. Compared directly to [32], our method ranks first on half of the building extraction datasets and second on the rest. For road extraction, it outperforms [32] on 3 datasets, though with smaller margins, likely due to the challenge of zero-shot localisation of thin, complex structures.

Enabling training-free referring and reasoning segmentation. We evaluate our generative-VLM based approach for referring and reasoning segmentation tasks on the validation and test splits of RRSIS-D [39] and EarthReason [34]. The upper part of Table 3 shows the results of our training-free approach. In this setup, a proprietary generative VLM is prompted to output click positions, which are fed into SAM (second row). This yields better segmentation results than directly prompting the same VLM to output segmentation masks (first row). However, these promising results still fall short of the current state of the art. This limitation likely arises from the difficulty of current VLMs to perform challenging tasks, such as spatial reasoning and referring segmentation on remote sensing imagery.

Achieving SOTA referring and reasoning segmentation with LoRA-tuned generative VLM. We evaluate the fine-tuned generative VLM-based pipeline on the same datasets and tasks as in the zero-shot setup. As shown in Table 3, this fine-tuning proves highly effective, achieving state-of-the-art performance on both referring (RRSIS-D) and reasoning (EarthReason) segmentation tasks. Unlike previous methods that require full training of LLMs, mask decoders, or additional components, our approach avoids heavy retraining. We fine-tune only a lightweight subset of LLM parameters using LoRA, while keeping the mask generator (SAM) frozen. This design resulted efficient, reducing the number of trainable components compared to prior methods without compromising performance.

Method	LLM	Trained on RS data			RRSIS-D		EarthReason	
		LLM	Mask Decoder	Extra	Val	Test	Val	Test
Zero-shot methods								
GPT-Image-1	GPT-5	✗	✗	✗	20.1	17.2	38.4	41.0
Ours	GPT-5	✗	✗	✗	**25.8**	**24.9**	**46.0**	**47.4**
Classical methods								
RRSIS [77]	BERT-base	✓	✓	✓	60.2	59.4	–	–
LAVT [75]	BERT-base	✓	✓	✓	61.5	61.0	–	–
DiffRIS [17]	CLIP	✓	✓	✓	63.6	62.2	–	–
FIANet [30]	BERT-base	✓	✓	✓	–	64.0	–	–
RMSIN [39]	BERT-base	✓	✓	✓	65.1	64.2	–	–
RSRefSeg [10]	SigLIP-So	✓	✓	✓	–	64.7	–	–
SBANet [33]	BERT-base	✓	✓	✓	66.7	65.5	–	–
BTDNet [80]	BERT-base	✓	✓	✓	66.9	66.0	–	–
Based on generative VLMs								
NExT-Chat [78]	Vicuna-7B	✓	✓	✓	27.0	25.0	–	–
LISA [27]	Vicuna-7B	LoRA	✓	✓	27.8	26.8	61.0	60.9
PixelLM [53]	Vicuna-7B	LoRA	✓	✓	33.9	31.7	57.9	60.0
PSALM [81]	phi-1.5-1.3B	✓	✓	✓	–	–	66.6	68.3
SegEarth-R1 [34]	phi-1.5-1.3B	✓	✓	✓	67.6	66.4	68.6	70.7
GeoPixel [54]	InternLM2-7B	LoRA	✓	✓	68.0	67.3	–	–
Ours	Qwen3-VL-2B	LoRA	✗	✗	**68.1**	**67.6**	**70.6**	**72.7**

Table 3. Results of our generative VLM-based approach for text-based remote sensing segmentation on reasoning and referring tasks. We evaluate on test and validation sets from RRSIS-D [39] (referring) and EarthReason [34] (reasoning) datasets.

Qualitative results. Figure 3 presents results from our contrastive pipeline for multi-class and single-class OVSS. Due to space limits, we show a subset of datasets: OpenEarthMap [71] and LoveDA [67] for multi-class, and Inria [44] for single-class. Our approach correctly identifies most classes, with minor errors in challenging categories (*e.g.*, trees, dense vegetation) and occasional misclassifications in crowded scenes (*e.g.*, buildings vs. roads). For the generative VLM-based approach, Figure 4 shows examples from EarthReason and RRSIS-D for reasoning and referring segmentation. Each row displays the input image with predicted clicks, predicted masks, and ground truths. As observed in the first-row example, our method localises the correct area even when the main object differs from the question target (*e.g.*, the tennis court in the top-right). Additional examples highlight handling of small objects and complex shapes. However, the approach sometimes struggle with ambiguous descriptions, especially when target involves multiple regions. Moreover, SAM's limitations can lead to inaccurate masks for non-well-delimited areas. Full visualisations and in-depth analysis are in the Appendix.

4.3. Ablation Studies

Effect of SAM scale and grid density on contrastive VLM-based pipeline. We conduct ablation studies on SAM size and grid clicks using OEM, LoveDA, UAVid (multi-class), and CHN6 (single-class). From the results reported in Table 4, we observe that the largest variant of SAM achieves the highest performance across datasets, which we use as the default. For experiments with different grid sizes, performance improves steadily up to a grid size of 20×20, after which it plateaus. Consequently, we adopt a 29×29 grid for the final configuration, as it yields superior metrics across most benchmarks. However, in more computationally constrained setups, it would be possible to adopt smaller SAM without significant degradation on performance.

SAM	# clicks	OEM	LoveDA	UAVid	CHN6
SAM-Tiny	29×29	33.9	37.7	43.8	35.2
SAM-Base	29×29	34.2	38.1	44.2	36.2
SAM-Large	10×10	29.7	36.2	40.8	31.0
SAM-Large	20×20	33.1	38.1	44.1	35.6
SAM-Large	29×29	**34.2**	**38.2**	**44.3**	**36.4**

Table 4. Ablation of SAM scale and grid density on contrastive VLM-based approach.

Effect of generative VLM scale and click configuration. We ablate VLM size and number of clicks as depicted in Table 5, on EarthReason and RRSIS-D for reasoning and re-

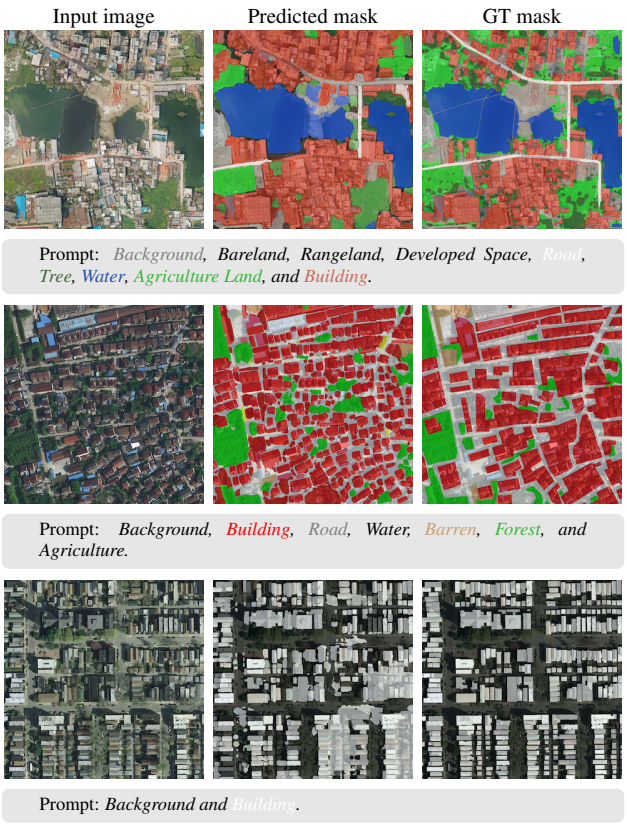


Figure 3. Qualitative results of the training-free contrastive VLM pipeline on multi-class (first and second rows) and single-class (third row) OVSS tasks using remote sensing datasets.

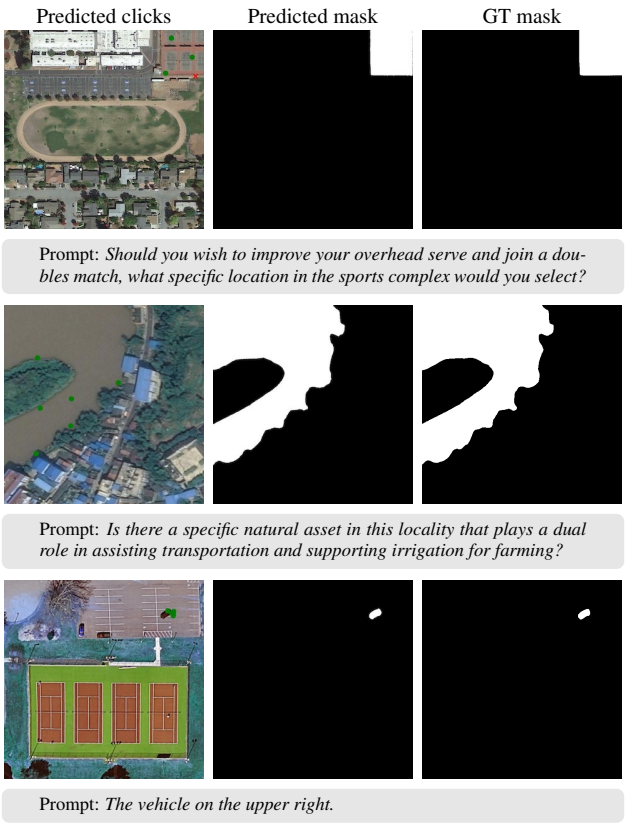


Figure 4. Qualitative results of the LoRA-tuned generative VLM pipeline on reasoning (first and second rows) and referring (third row) tasks using remote sensing datasets.

Method	# clicks	EarthReason		RRSIS-D	
		Val	Test	Val	Test
Qwen2.5-VL-7B	6	67.8	69.3	61.4	60.9
Qwen3-VL-4B	6	70.4	71.7	67.3	67.2
Qwen3-VL-2B	2	63.8	64.9	61.1	61.0
Qwen3-VL-2B	4	67.9	69.8	66.4	66.4
Qwen3-VL-2B	6	**70.6**	**72.7**	**68.1**	**67.6**

Table 5. Ablation of generative VLM scale and click configuration.

ferring segmentation, respectively. According to the results, upgrading from QwenVL-2.5 to QwenVL-3 improves performance, making QwenVL-3 our baseline. Further experiments comparing the 4B ($\sim$80M trainable parameters) and 2B ($\sim$50M trainable parameters) variants of Qwen3-VL show a performance gain for the smaller 2B model, which we then use for the main experiments. In terms of click configuration, performance increases consistently up to six clicks, which improves results by $+6.8/+7.8$ (EarthReason val/test) and $+7.0/+6.6$ (RRSIS-D val/test) over two clicks variant.

5. Conclusion

We introduced a simple yet powerful approach for zero-shot text-based segmentation of remote sensing imagery. Our approach combines contrastive (CLIP) and generative (GPT-5, Qwen-VL) VLMs with the Segment Anything Model (SAM). The resulting two pipelines achieve state-of-the-art results on 19 remote sensing benchmarks, including open-vocabulary, referring, and reasoning segmentation. The contrastive pipeline enables fully training-free OVSS, while the generative pipeline supports both zero-shot inference (via GPT-5) and lightweight LoRA fine-tuning (via Qwen-VL) for more complex linguistic reasoning. Despite the used VLMs being primarily pre-trained on natural images, our results demonstrate that the proposed approach remains effective for earth perception tasks. As foundation models continue to evolve, we anticipate even better zero-shot capabilities and improved alignment between visual and textual representations for more complex, real-world geospatial understanding.

Acknowledgments. We thank Rim Sleimi and Nicla Notarangelo for great discussions. This work was supported by FNR HPC BRIDGES project, with reference HPC BRIDGES/2022/17978225/AI4CC. Experiments were performed on MeluXina, special thanks to LuxProvide team for the support.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 1, 2

[2] 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. 4

[3] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 1, 2, 3

[4] Zechen Bai, Tong He, Haiyang Mei, Pichao Wang, Ziteng Gao, Joya Chen, Zheng Zhang, and Mike Zheng Shou. One token to seg them all: Language instructed reasoning segmentation in videos. *Advances in Neural Information Processing Systems*, 37:6833–6859, 2024. 1, 2

[5] Luca Barsellotti, Roberto Amoroso, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Training-free open-vocabulary segmentation with offline diffusion-augmented prototype generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3689–3698, 2024. 2

[6] Luca Barsellotti, Lorenzo Bianchi, Nicola Messina, Fabio Carrara, Marcella Cornia, Lorenzo Baraldi, Fabrizio Falchi, and Rita Cucchiara. Talking to dino: Bridging self-supervised vision backbones with language for open-vocabulary segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22025–22035, 2025. 2

[7] Walid Bousselham, Felix Petersen, Vittorio Ferrari, and Hilde Kuehne. Grounding everything: Emerging localization properties in vision-language transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3828–3837, 2024. 2, 6

[8] 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. 5

[9] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 2

[10] Keyan Chen, Jiafan Zhang, Chenyang Liu, Zhengxia Zou, and Zhenwei Shi. Rsrefseg: Referring remote sensing image segmentation with foundation models. *arXiv preprint arXiv:2501.06809*, 2025. 3, 7

[11] 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. 5

[12] Yi-Chia Chen, Wei-Hua Li, Cheng Sun, Yu-Chiang Frank Wang, and Chu-Song Chen. Sam4mllm: Enhance multi-modal large language model for referring expression segmentation. In *European Conference on Computer Vision*, pages 323–340. Springer, 2024. 2

[13] Seokju Cho, Heeseong Shin, Sunghwan Hong, Anurag Arnab, Paul Hongsuck Seo, and Seungryong Kim. Cat-seg: Cost aggregation for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4113–4123, 2024. 2

[14] DeepGlobe Consortium. Deepglobe: A challenge to parse the earth through satellite images. http://deepglobe.org/, 2018. Accessed: 2025-11-11. 5

[15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019. 3

[16] Jian Ding, Nan Xue, Gui-Song Xia, and Dengxin Dai. Decoupling zero-shot semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11583–11592, 2022. 2

[17] Zhe Dong, Yuzhe Sun, Tianzhu Liu, and Yanfeng Gu. Diffris: Enhancing referring remote sensing image segmentation with pre-trained text-to-image diffusion models. *arXiv preprint arXiv:2506.18946*, 2025. 1, 3, 7

[18] Stephanie Fu, Mark Hamilton, Laura Brandt, Axel Feldman, Zhoutong Zhang, and William T Freeman. Featup: A model-agnostic framework for features at any resolution. *arXiv preprint arXiv:2403.10516*, 2024. 6

[19] Golnaz Ghiasi, Xiuye Gu, Yin Cui, and Tsung-Yi Lin. Scaling open-vocabulary image segmentation with image-level labels. In *European conference on computer vision*, pages 540–557. Springer, 2022. 2

[20] 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

[21] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022. 2

[22] ISPRS Foundation. Isprs benchmark on semantic labeling. https://www.isprs.org/resources/datasets/benchmarks/UrbanSemLab/, 2025. Accessed: 2025-11-11. 5

[23] Johannes Jakubik, Sujit Roy, CE Phillips, Paolo Fraccaro, Denys Godwin, Bianca Zadrozny, Daniela Szwarcman, Carlos Gomes, Gabby Nyirjesy, Blair Edwards, et al. Foundation models for generalist geospatial artificial intelligence. *arXiv preprint arXiv:2310.18660*, 2023. 1

[24] 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. 5

[25] Siyu Jiao, Yunchao Wei, Yaowei Wang, Yao Zhao, and Humphrey Shi. Learning mask-aware clip representations for zero-shot segmentation. *Advances in Neural Information Processing Systems*, 36:35631–35653, 2023. 2

[26] Dahyun Kang and Minsu Cho. In defense of lazy visual grounding for open-vocabulary semantic segmentation. In *European Conference on Computer Vision*, pages 143–164. Springer, 2024. 2

[27] 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*, pages 9579–9589, 2024. 1, 2, 7

[28] Mengcheng Lan, Chaofeng Chen, Yiping Ke, Xinjiang Wang, Litong Feng, and Wayne Zhang. Clearclip: Decomposing clip representations for dense vision-language inference. In *European Conference on Computer Vision*, pages 143–160. Springer, 2024. 2, 6

[29] Mengcheng Lan, Chaofeng Chen, Yiping Ke, Xinjiang Wang, Litong Feng, and Wayne Zhang. Proxyclip: Proxy attention improves clip for open-vocabulary segmentation. In *European Conference on Computer Vision*, pages 70–88. Springer, 2024. 2

[30] Sen Lei, Xinyu Xiao, Tianlin Zhang, Heng-Chao Li, Zhenwei Shi, and Qing Zhu. Exploring fine-grained image-text alignment for referring remote sensing image segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 2024. 3, 7

[31] Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and René Ranftl. Language-driven semantic segmentation. *arXiv preprint arXiv:2201.03546*, 2022. 2

[32] 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, 3, 4, 5, 6

[33] Kun Li, George Vosselman, and Michael Ying Yang. Scale-wise bidirectional alignment network for referring remote sensing image segmentation. *ISPRS Journal of Photogrammetry and Remote Sensing*, 226:350–363, 2025. 3, 7

[34] 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, 3, 5, 6, 7

[35] Rui Li and Xiaowei Zhao. Aeroreformer: Aerial referring transformer for uav-based referring image segmentation. *arXiv preprint arXiv:2502.16680*, 2025. 3

[36] Yongkang Li, Tianheng Cheng, Bin Feng, Wenyu Liu, and Xinggang Wang. Mask-adapter: The devil is in the masks for open-vocabulary segmentation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 14998–15008, 2025. 1, 2

[37] Yi Li, Hualiang Wang, Yiqun Duan, Jiheng Zhang, and Xiaomeng Li. A closer look at the explainability of contrastive language-image pre-training. *Pattern Recognition*, 162:111409, 2025. 2

[38] Fan Liu, Delong Chen, Zhangqingyun Guan, Xiaocong Zhou, Jiale Zhu, Qiaolin Ye, Liyong Fu, and Jun Zhou. Remoteclip: A vision language foundation model for remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–16, 2024. 3

[39] 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. 3, 5, 6, 7

[40] Yong Liu, Sule Bai, Guanbin Li, Yitong Wang, and Yansong Tang. Open-vocabulary segmentation with semantic-assisted calibration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3491–3500, 2024. 2

[41] Timo Lüddecke and Alexander Ecker. Image segmentation using text and image prompts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7086–7096, 2022. 2

[42] Huaishao Luo, Junwei Bao, Youzheng Wu, Xiaodong He, and Tianrui Li. Segclip: Patch aggregation with learnable centers for open-vocabulary semantic segmentation. In *International Conference on Machine Learning*, pages 23033–23044. PMLR, 2023. 2

[43] 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. 5

[44] 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. 5, 7

[45] Volodymyr Mnih. *Machine learning for aerial image labeling*. University of Toronto (Canada), 2013. 5

[46] Jishnu Mukhoti, Tsung-Yu Lin, Omid Poursaeed, Rui Wang, Ashish Shah, Philip HS Torr, and Ser-Nam Lim. Open vocabulary semantic segmentation with patch aligned contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19413–19423, 2023. 2

[47] OpenAI. Introducing gpt-5. https://openai.com/index/introducing-gpt-5/, 2025. Accessed: 2025-11-12. 2, 5

[48] 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, 5

[49] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 1, 2, 3, 5, 6

[50] Kanchana Ranasinghe, Brandon McKinzie, Sachin Ravi, Yinfei Yang, Alexander Toshev, and Jonathon Shlens. Perceptual grouping in contrastive vision-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5571–5584, 2023. 2

[51] Hanoona Rasheed, Muhammad Maaz, Sahal Shaji, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M Anwer, Eric Xing, Ming-Hsuan Yang, and Fahad S Khan. Glamm: Pixel grounding large multimodal model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13009–13018, 2024. 2

[52] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 1, 2, 3, 5

[53] Zhongwei Ren, Zhicheng Huang, Yunchao Wei, Yao Zhao, Dongmei Fu, Jiashi Feng, and Xiaojie Jin. Pixellm: Pixel reasoning with large multimodal model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26374–26383, 2024. 1, 2, 7

[54] Akashah Shabbir, Mohammed Zumri, Mohammed Bennamoun, Fahad S Khan, and Salman Khan. Geopixel: Pixel grounding large multimodal model in remote sensing. *arXiv preprint arXiv:2501.13925*, 2025. 1, 7

[55] Tong Shao, Zhuotao Tian, Hang Zhao, and Jingyong Su. Explore the potential of clip for training-free open vocabulary semantic segmentation. In *European Conference on Computer Vision*, pages 139–156. Springer, 2024. 2

[56] shirshmall. Water body segmentation in satellite images. https://www.kaggle.com/datasets/shirshmall/water-body-segmentation-in-satellite-images. Accessed: 2025-11-11. 5

[57] 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. 4

[58] Jose Sosa, Mohamed Aloulou, Danila Rukhovich, Rim Sleimi, Boonyarit Changaival, Anis Kacem, and Djamila Aouada. How effective is pre-training of large masked autoencoders for downstream earth observation tasks? *arXiv preprint arXiv:2409.18536*, 2024. 1

[59] Jose Sosa, Danila Rukhovich, Anis Kacem, and Djamila Aouada. Multimae meets earth observation: Pre-training multi-modal multi-task masked autoencoders for earth observation tasks. *arXiv preprint arXiv:2505.14951*, 2025. 1

[60] Shuyang Sun, Runjia Li, Philip Torr, Xiuye Gu, and Siyang Li. Clip as rnn: Segment countless visual concepts without training endeavor. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13171–13182, 2024. 2

[61] Daniela Szwarcman, Sujit Roy, Paolo Fraccaro, Thorsteinn Eli Gislason, Benedikt Blumenstiel, Rinki Ghosal, Pedro Henrique de Oliveira, Joao Lucas de Sousa Almeida, Rocco Sedona, Yanghui Kang, et al. Prithvi-eo-2.0: A versatile multi-temporal foundation model for earth observation applications. *arXiv preprint arXiv:2412.02732*, 2024. 1

[62] Adam Van Etten, Dave Lindenbaum, and Todd M Bacastow. Spacenet: A remote sensing dataset and challenge series. *arXiv preprint arXiv:1807.01232*, 2018. 5

[63] Feng Wang, Jieru Mei, and Alan Yuille. Sclip: Rethinking self-attention for dense vision-language inference. In *European Conference on Computer Vision*, pages 315–332. Springer, 2024. 2, 6

[64] Haoxiang Wang, Pavan Kumar Anasosalu Vasu, Fartash Faghri, Raviteja Vemulapalli, Mehrdad Farajtabar, Sachin Mehta, Mohammad Rastegari, Oncel Tuzel, and Hadi Pouransari. Sam-clip: Merging vision foundation models towards semantic and spatial understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3635–3647, 2024. 2

[65] Hao Wang, Limeng Qiao, Zequn Jie, Zhijian Huang, Chengjian Feng, Qingfang Zheng, Lin Ma, Xiangyuan Lan, and Xiaodan Liang. X-sam: From segment anything to any segmentation. *arXiv preprint arXiv:2508.04655*, 2025. 2

[66] Junchi Wang and Lei Ke. Llm-seg: Bridging image segmentation and large language model reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1765–1774, 2024. 2

[67] 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. 5, 7

[68] Jinglong Wang, Xiawei Li, Jing Zhang, Qingyuan Xu, Qin Zhou, Qian Yu, Lu Sheng, and Dong Xu. Diffusion model is secretly a training-free open vocabulary semantic segmenter. *IEEE Transactions on Image Processing*, 2025. 2

[69] 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. 5

[70] Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025. 2, 5

[71] 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. 5, 7

[72] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transform-

ers. *Advances in neural information processing systems*, 34: 12077–12090, 2021. 6

[73] Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, et al. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*, 2025. 2

[74] Mengde Xu, Zheng Zhang, Fangyun Wei, Han Hu, and Xiang Bai. San: side adapter network for open-vocabulary semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(12):15546–15561, 2023. 2

[75] Zhao Yang, Jiaqi Wang, Yansong Tang, Kai Chen, Hengshuang Zhao, and Philip HS Torr. Lavt: Language-aware vision transformer for referring image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18155–18165, 2022. 1, 7

[76] Haobo Yuan, Xiangtai Li, Tao Zhang, Zilong Huang, Shilin Xu, Shunping Ji, Yunhai Tong, Lu Qi, Jiashi Feng, and Ming-Hsuan Yang. Sa2va: Marrying sam2 with llava for dense grounded understanding of images and videos. *arXiv preprint arXiv:2501.04001*, 2025. 2

[77] Zhenghang Yuan, Lichao Mou, Yuansheng Hua, and Xiao Xiang Zhu. Rrsis: Referring remote sensing image segmentation. *arXiv preprint arXiv:2306.08625*, 2023. 3, 7

[78] Ao Zhang, Yuan Yao, Wei Ji, Zhiyuan Liu, and Tat-Seng Chua. Next-chat: An lmm for chat, detection and segmentation. *arXiv preprint arXiv:2311.04498*, 2023. 1, 2, 7

[79] Tao Zhang, Xiangtai Li, Hao Fei, Haobo Yuan, Shengqiong Wu, Shunping Ji, Chen Change Loy, and Shuicheng Yan. Omg-llava: Bridging image-level, object-level, pixel-level reasoning and understanding. *Advances in neural information processing systems*, 37:71737–71767, 2024. 2

[80] Tianxiang Zhang, Zhaokun Wen, Bo Kong, Kecheng Liu, Yisi Zhang, Peixian Zhuang, and Jiangyun Li. Referring remote sensing image segmentation via bidirectional alignment guided joint prediction. *arXiv preprint arXiv:2502.08486*, 2025. 3, 7

[81] Zheng Zhang, Yeyao Ma, Enming Zhang, and Xiang Bai. Psalm: Pixelwise segmentation with large multi-modal model. In *European Conference on Computer Vision*, pages 74–91. Springer, 2024. 2, 7

[82] Zilun Zhang, Tiancheng Zhao, Yulong Guo, and Jianwei Yin. Rs5m and georsclip: A large scale vision-language dataset and a large vision-language model for remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 2024. 3

[83] Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from clip. In *European conference on computer vision*, pages 696–712. Springer, 2022. 1, 2, 6

[84] Yue Zhou, Mengcheng Lan, Xiang Li, Litong Feng, Yiping Ke, Xue Jiang, Qingyun Li, Xue Yang, and Wayne Zhang. Geoground: A unified large vision-language model for remote sensing visual grounding. *arXiv preprint arXiv:2411.11904*, 2024. 1, 3

[85] Muzhi Zhu, Yuzhuo Tian, Hao Chen, Chunluan Zhou, Qingpei Guo, Yang Liu, Ming Yang, and Chunhua Shen. Segagent: Exploring pixel understanding capabilities in mllms by imitating human annotator trajectories. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 3686–3696, 2025. 2

[86] 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

[87] Qi Zhu, Jiangwei Lao, Deyi Ji, Junwei Luo, Kang Wu, Yingying Zhang, Lixiang Ru, Jian Wang, Jingdong Chen, Ming Yang, et al. Skysense-o: Towards open-world remote sensing interpretation with vision-centric visual-language modeling. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 14733–14744, 2025. 1, 3