MAgSeg: Segmentation of Agricultural Landscapes in High-Resolution Satellite Imagery using Multimodal Large Language Models

Piyush Tiwary[1,4,*], **Utkarsh Ahuja**[2], **Depanshu Sani**[1], **Aishwarya Jayagopal**[1], **Sagar Gubbi**[1], **Subhashini Venugopalan**[3], **Alok Talekar**[1] and **Vaibhav Rajan**[1]
[1]Google DeepMind, India
[2]Google LLC, India
[3]Google Research, USA
[4]Indian Institute of Science, Bangalore

Abstract

Agricultural landscape segmentation in the Global South is challenging as it is characterized by fragmented plots, high intra-class variance, and a scarcity of labeled training data. Recent advances in segmentation have been made by Multimodal Large Language Models (MLLMs). However, current approaches encounter critical context length bottlenecks and a domain alignment gap in understanding satellite features. We address these limitations through MAgSeg, a novel, decoder-free MLLM segmentation approach. MAgSeg is an architecturally efficient approach that enables standard MLLMs to perform segmentation of complex smallholder agricultural landscapes from high-resolution satellite imagery, without requiring auxiliary vision decoders. We introduce a novel instruction tuning data format designed to enable scalable fine-tuning and post-training on high resolution satellite imagery, which enables MAgSeg to learn from the global context of the image while generating text tokens for only a patch within the image. Extensive evaluations on datasets spanning three countries in the Global South demonstrate that MAgSeg significantly outperforms state-of-the-art MLLM baselines, offering a scalable solution to map smallholder agricultural environments.

1 Introduction

Agricultural Landscape Segmentation involves identifying the polygonal boundaries and constituent areas of agricultural features, such as crop fields, trees and water bodies, using satellite imagery. It is a foundational step for large-scale mapping and analysis of agricultural environments [FAO *et al.*, 2023; OECD, 2023], which, in turn, support advanced, field-level analytics for multiple downstream precision agriculture applications such as crop type and yield management, policy making and resource allocation [Rudel *et al.*, 2009; Kang and Özdoğan, 2019]. Effective solutions in these applications can have significant impact on mitigating climate change and managing global food security [FAO *et al.*, 2023; OECD, 2023]. This task is especially challenging for *smallholder farms* in the Global South, which are characterized by an intricate mosaic of very small fields ($<$ 2 ha) [Ministry of Agriculture and Farmers Welfare, 2024], diverse land uses, and interspersion with natural vegetation and water resources [Lesiv *et al.*, 2019; Samberg *et al.*, 2016; Rada and Fuglie, 2019], and are responsible for more than half of the global agricultural production [Samberg *et al.*, 2016; Sylvester and others, 2015].

The nature of smallholder systems pose unique technical challenges in landscape segmentation. These include high variance within the *field* class (due to diverse crops, shapes, and sizes), and low variance between fields and other classes like forests; and large distribution shifts with varying landscape characteristics across different regions, climates, and seasons [Wang *et al.*, 2022; Kerner *et al.*, 2023]. Accurate delineation of small-sized fields and other features necessitate the use of very-high-resolution (VHR) satellite imagery (e.g., WorldView-3 data at 0.31 to 2 m/pixel). However, adapting standard image segmentation models, primarily developed for natural images, for this task remains challenging due to the high scale variation, densely packed small-scale objects and computational challenges arising from high dimensional inputs [Zhu *et al.*, 2017; Rolf *et al.*, 2024]. Moreover, nearly all previous approaches use specialized supervised learning architectures and rely on labelled data. Obtaining such labeled data is labor-intensive, costly, and difficult to scale across the diverse geographies of the Global South.

Multimodal Large Language Models (MLLMs) harness the inherent reasoning capabilities and broad world knowledge of LLMs, and have been recently used for various segmentation tasks with natural images [Zhou *et al.*, 2024]. MLLMs offer the potential for scalability and accessibility with relatively lesser labeled data requirements to address the smallholder landscape segmentation problem. However, existing MLLM-based segmentation approaches perform poorly on VHR satellite images, particularly on this challenging task. The dominant 'embedding-as-mask' paradigm, exemplified by LISA [Lai *et al.*, 2024] and GSVA [Xia *et al.*, 2024], relies on *auxiliary decoder heads* to generate pixel-level masks. This design introduces parameter overhead and creates an alignment bottleneck between

[*]Work done as a Student Researcher at Google DeepMind.

the LLM's text-centric latent space and the pixel decoder. The other 'text-as-mask' paradigm is a *decoder-free* approach proposed by Text4Seg [Lan *et al.*, 2025]. It encodes masks as text sequences and seamlessly integrates segmentation into the standard autoregressive generation process of MLLMs.

However, this approach also fails when applied to VHR satellite imagery for two reasons. Encoding a full VHR mask (e.g., 615×615 px) would require about $\mathcal{O}(3.6 \times 10^5)$ tokens, far exceeding the context windows of many MLLMs. Text4Seg relies on severe downsampling (e.g., to 32×32 px) to keep token counts of the encoded masks low. While acceptable for natural images, this destroys critical high-frequency details in satellite data required for boundary delineation. The second reason is the domain alignment gap: general-purpose MLLMs lack intrinsic understanding of overhead satellite imagery. E.g., while they can identify 'trees' in a photograph, they struggle to distinguish similar features like 'crops' versus 'pasture' in satellite views without explicit pixel-level grounding [Quenum *et al.*, 2025].

In this work, we present a decoder-free MLLM-based Agricultural Landscape Segmentation approach (MAgSeg) that systematically overcomes these limitations. We create an instruction tuning dataset with a novel format, which is designed to enable efficient fine-tuning and aims at retaining both local or global context in VHR masks. This resolves the token length problem through a patch-based strategy: MAgSeg receives the full VHR image as context but is tasked with generating a text-based mask for only a specific sub-patch. Supervised fine tuning on this data helps, but is insufficient since it is optimized for next-token prediction (for the text-based mask) and not spatial accuracy. We address this domain alignment gap via Group Relative Policy Optimization (GRPO) using the mean-DICE score as a direct reward signal. This enables MAgSeg to prioritize pixel-level accuracy over mere textual fluency on the encoded masks.

We use another metric, *overhead*, to measure the parameter efficiency of MLLM-based segmentation models which quantifies the additional *learnable* parameters (e.g., in pixel decoders and segmentation heads) beyond the base LLM backbone; and serves as a proxy for computational cost and memory footprint during inference. MAgSeg, which has zero overhead, shows dramatic gains in segmentation performance over previous zero-overhead decoder-free approaches in our experiments. Moreover, MAgSeg also significantly outperforms previous decoder-based methods which achieve better performance, relative to previous zero-overhead methods, at the cost of high overhead.

Our contributions are summarized as follows:

1. We propose an architecturally efficient, zero-overhead approach, MAgSeg, that enables standard MLLMs to perform segmentation of complex smallholder agricultural landscapes from very-high-resolution (VHR) satellite imagery, without requiring auxiliary vision decoders.
2. MAgSeg overcomes critical context length bottlenecks and domain alignment gaps in previous decoder-free approaches through instruction tuning with a novel format, tasked to generate text tokens for only a patch within a VHR image, and RL-based post-training to refine pixel-level understanding.
3. On challenging benchmark smallholder agriculture datasets spanning three countries in the Global South, MAgSeg significantly outperforms state-of-the-art LLM-based baselines.

2 Background and Related Works

MLLMs for Vision Tasks. MLLMs demonstrate exceptional capabilities in visual reasoning, captioning, and visual question answering. However, applying MLLMs to dense prediction tasks like segmentation remains non-trivial due to the inherent modality gap between language and vision. A common approach, exemplified by LISA [Lai *et al.*, 2024] and GLaMM [Rasheed *et al.*, 2024], involves appending a specialized pixel decoder (e.g., SAM [Kirillov *et al.*, 2023]) to the LLM. The LLM outputs a special token (e.g., `<seg>`) which cues the decoder to generate a mask. While effective, this 'embedding-as-mask' approach increases training complexity, limits model scalability and disconnects the segmentation process from the LLM's native text-generation capabilities. Other approaches like VisionLLM [Wang *et al.*, 2023; Wu *et al.*, 2024] output polygon coordinates directly, but often struggle with complex geometries in satellite imagery.

To address these limitations Text4Seg [Lan *et al.*, 2025] introduced a decoder-free 'text-as-mask' approach by casting segmentation as a sequence-to-sequence problem, which aligns with the autoregressive training pipeline of LLMs and enables scalable, flexible modeling with minimal architectural changes. The key idea is to split the mask into a grid of fixed-size patches, flatten the patches and represent each patch with a *semantic descriptor*, which can be a word, phrase or a more complex textual description. Thus, we obtain a sequence of text tokens whose length can become exceedingly large with both number of patches and more complex descriptions. To compress this description, they apply Run-Length Encoding (RLE) to the sequence, and empirically find that applying RLE row-wise (RRLE), with each row separated by a newline character, to the mask patches works well in practice. This transformed visual instruction data is used to fine tune the MLLM.

However, Text4Seg, which achieves state-of-the-art performance on several segmentation tasks on natural images, fails on the challenging task of smallholder landscape segmentation (as seen in our experiments). These landscapes have both large number of objects to be delineated and high scale variation, e.g., ranging from small to very large fields. Hence, the masks lead to complex and long descriptions with long sequences even after RRLE. The approach of downsampling the masks used by [Lan *et al.*, 2025] leads to loss of critical high-frequency spatial details required for boundary delineation, whereas increasing the number of patches leads to loss of contextual information for each instance, which, in turn, deteriorates segmentation performance. Our approach, MAgSeg, specifically addresses these problems, while retaining the advantages of their decoder-free approach.

Agricultural Landscape Understanding. Segmentation of agricultural landscapes has been extensively studied in remote sensing – see [Zheng *et al.*, 2025] for a recent survey. Most previous works have focussed on field boundary de-

lineation, particularly in the global North, e.g., [Masoud *et al.*, 2020; Waldner and Diakogiannis, 2020; Waldner *et al.*, 2021]. A few recent works have recognized and addressed the challenges of smallholder agriculture by adapting classical computer vision techniques such as panoptic segmentation [Dua *et al.*, 2024], object detection [Mei *et al.*, 2022] and transfer learning based methods [Wang *et al.*, 2022; Kerner *et al.*, 2023]. All these methods rely on supervised learning (mainly CNN-based) architectures.

To harness their robust reasoning capabilities, visual understanding and world knowledge, MLLMs are being actively explored for various remote sensing tasks such as visual question answering and retrieval [Li *et al.*, 2024]. However, their use in segmentation tasks is relatively underexplored. Two recent segmentation models, LISAt [Quenum *et al.*, 2025] and FSVLM [Wu *et al.*, 2025] follow the same embedding-as-mask architecture of LISA [Lai *et al.*, 2024]: a CLIP-encoded image [Radford *et al.*, 2021] and a text embedding are passed to a MLLM (both use LLaVA [Liu *et al.*, 2023b]); segmentation is triggered by a dedicated `<seg>` token within the LLM's vocabulary; upon generation, the token's embedding is mapped to a SAM-based decoder through an MLP; the decoder subsequently combines this embedding with the base image features to reconstruct the pixel-level mask. The model is trained end-to-end with loss functions to optimize text generation and segmentation. FSVLM was specifically evaluated for farmland segmentation while LISAt demonstrated improvements over many geo-spatial foundation models on visual question answering, visual grounding and image captioning tasks. In our experiments, we demonstrate significant improvements over this state-of-the-art approach on smallholder landscape segmentation datasets.

3 Our Approach: MAgSeg

Let $\mathcal{D} = \{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^{N}$ denote a (training) dataset, where $\mathbf{x}_i \in \mathbb{R}^{h_\mathbf{x} \times w_\mathbf{x} \times 3}$ represents a high-resolution satellite RGB image with height h_x, width w_x, 3 channels, and $\mathbf{y}_i$ is the corresponding multi-class instance map. Our aim is to train a base MLLM, π_θ, with parameters θ, using $\mathcal{D}$, for the task of instance segmentation.

We first transform $\mathcal{D}$ into an instruction-tuning dataset with a novel format that preserves both local and global context in high-resolution images. Since supervised fine-tuning (SFT) alone lacks the fine-grained spatial reasoning needed for pixel-perfect segmentation of complex satellite imagery, we employ a Reinforcement Learning (RL) based post-training step to maximize pixel-level semantic alignment. At inference, the trained model generates semantic masks that are post-processed into instance masks. Figure 1 provides an overview of MAgSeg.

3.1 Confronting High Resolution: Patch-based Instruction Tuning with Global Context

Given a dataset sample $(\mathbf{x}_i, \mathbf{y}_i)$, we first derive the semantic segmentation map, $\mathbf{s}_i$ from the instance map $\mathbf{y}_i$. To facilitate patch-based processing, we extract a random crop from the high-resolution input. We sample top-left coordinates (j, k) uniformly such that $j \sim \mathcal{U}(1, h_\mathbf{x} - h_\mathbf{p})$ and $k \sim \mathcal{U}(1, w_\mathbf{x} - w_\mathbf{p})$. Thus, from $\mathbf{s}_i$, this yields an image patch $\mathbf{p}_i$ of dimensions $h_\mathbf{p} \times w_\mathbf{p}$ and a spatially corresponding semantic mask patch $\mathbf{r}_i$.

The mask patch $\mathbf{r}_i$ is subsequently converted into a text-based RRLE representation, denoted as $\mathbf{t}_i = \texttt{RRLE}(\mathbf{r}_i)$, following the semantic descriptors devised for Text4Seg [Lan *et al.*, 2025] (cf. Algorithm 1 in Appendix A). We then construct an instruction-tuning sample consisting of the prompt input and the target response:

$$\text{Input:}\{\mathcal{I}_{\text{text}}, \mathbf{x}_i, \mathbf{p}_i\} \quad \rightarrow \quad \text{Target:}\mathbf{t}_i$$

Here, $\mathcal{I}_{\text{text}}$ is the textual prompt instructing the LLM to generate the RRLE mask for the provided patch. The full high-resolution image $\mathbf{x}_i$ is passed alongside each patch $\mathbf{p}_i$ to provide global context, while targeting only the patch tokens $\mathbf{t}_i$ keeps us within the MLLM's context window. This lets us apply Text4Seg's decoder-free seq-to-seq framework to HR satellite images without downsampling or losing critical high-frequency details.

To fine-tune π_θ we employ Low-Rank Adaptation (LoRA) [Hu *et al.*, 2022] in conjunction with a standard autoregressive objective. LoRA optimizes a significantly smaller set of parameters, denoted as $\Delta\theta$, by freezing θ and injecting trainable rank decomposition matrices into the linear layer of the MLLM's transformer. This approach ensures scalability, data efficiency as well as generalizability.

Formally, let the ground-truth RRLE sequence be represented as a series of tokens $\mathbf{t}_i = \{w_1, w_2, \ldots, w_T\}$. The trainable parameters $\Delta\theta$ are optimized by minimizing the negative log-likelihood of these target tokens conditioned on the multimodal context-

$$\Delta\theta_{\text{SFT}} = \arg\min_{\Delta\theta} \mathcal{L}_{\text{SFT}}(\Delta\theta) \tag{1}$$

$$\mathcal{L}_{\text{SFT}}(\Delta\theta) = -\sum_{t=1}^{T} \log \pi_{\theta+\Delta\theta}(w_t \mid w_{<t}, \mathbf{x}_i, \mathbf{p}_i, \mathcal{I}_{\text{text}}) \tag{2}$$

where, $\pi_{\theta+\Delta\theta}(\cdot)$ denotes the probability distribution of the next token predicted by the adapted model. Minimizing this objective aligns the model's visual reasoning with the textual RRLE structure, enabling segmentation of the target patch $\mathbf{p}_i$.

After this fine-tuning stage, the MLLM is (a) conditioned to generate masks in the specified RRLE format, and (b) instilled with a foundational understanding of segmentation semantics and class labels, as established in [Lan *et al.*, 2025].

3.2 Mitigating Domain Gap with GRPO

SFT optimizes for *next-token prediction* rather than *spatial accuracy*, often failing to achieve pixel-perfect alignment. To bridge this gap, we employ Group Relative Policy Optimization (GRPO) [Shao *et al.*, 2024] as a post-training step, shifting the objective from likelihood maximization to a pixel-level semantic reward.

For a given input tuple $\{\mathcal{I}_{\text{text}}, \mathbf{x}_i, \mathbf{p}_i\}$, we sample a group of G distinct outputs from the policy model π_θ, denoted as $\{\mathbf{t}_{i,g}\}_{g=1}^{G}$. Each output $\mathbf{t}_{i,g}$ represents a candidate RRLE sequence for the target patch $\mathbf{p}_i$. To evaluate the quality of these candidates, we first decode them into 2D semantic masks:

$$\tilde{\mathbf{s}}_{i,g} = \texttt{decode\_RRLE}(\mathbf{t}_{i,g}) \tag{3}$$

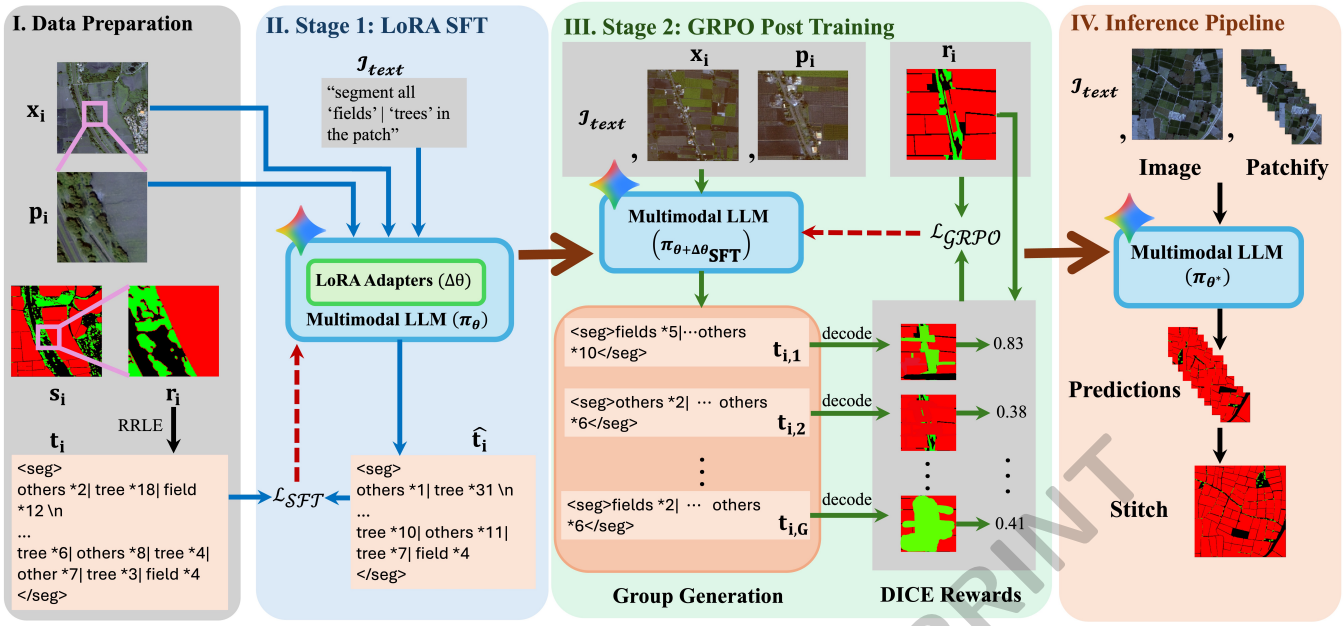

Figure 1: **Overview of MAgSeg. Data Preparation**: from each high-resolution satellite image $\mathbf{x}_i$ and its segmentation map $\mathbf{s}_i$, multiple patches $\mathbf{p}_i$ and their corresponding masks $\mathbf{r}_i$ are extracted, the masks are converted to a text-based RRLE representation $\mathbf{t}_i$ to form the instruction tuning dataset: $\{\mathcal{I}_{\text{text}}, \mathbf{x}_i, \mathbf{p}_i\} \rightarrow \mathbf{t}_i$. **Training**: consists of two stages: (1) LoRA Supervised Finetuning (SFT), where the base multimodal LLM is adapted to generate RRLE-encoded masks using a frozen backbone and trainable LoRA adapters; and (2) GRPO Post-Training, where we sample multiple candidate RLE sequences, decode them into segmentation masks, and use the mean-DICE score as a reward signal to update the policy, ensuring pixel-perfect alignment. **Inference**: uses the trained MLLM to predict on a grid of disjoint patches, which are then stitched and postprocessed to form the final segmentation map.

where, $\tilde{\mathbf{s}}_{i,g}$ is the predicted segmentation map for the g-th sample (cf. Algorithm 2 in Appendix A). To quantify the semantic accuracy, we compute the reward $r_{i,g}$ as the Mean DICE score (mDICE) between the predicted mask $\tilde{\mathbf{s}}_{i,g}$ and the ground truth mask $\mathbf{s}_i$. Let C denote the set of classes (e.g., fields, ponds, trees). The reward is defined as:

$$r_{i,g} = \frac{1}{|C|} \sum_{c \in C} \frac{2 \sum_{x,y} (\tilde{\mathbf{s}}_{i,g}^{(c)} \cap \mathbf{s}_i^{(c)})}{\sum_{x,y} \tilde{\mathbf{s}}_{i,g}^{(c)} + \sum_{x,y} \mathbf{s}_i^{(c)} + \epsilon} \tag{4}$$

where, $\cap$ denotes the intersection of active pixels for class c, and ϵ is a smoothing term. GRPO optimizes the policy by estimating the advantage of each sample relative to the group average, eliminating the need for a separate value function critic. The advantage $A_{i,g}$ for the g-th sample is computed as:

$$A_{i,g} = \frac{r_{i,g} - \mu_r}{\sigma_r} \tag{5}$$

where, μ_r and σ_r are the mean and standard deviation of the rewards within the group of G samples. The GRPO loss objective is formulated as:

$$\mathcal{L}_{\text{GRPO}}(\Delta\theta) = -\frac{1}{G} \sum_{g=1}^{G} \min \left(\rho_{i,g} A_{i,g}, \text{clip}(\rho_{i,g}, 1-\epsilon, 1+\epsilon) A_{i,g} \right) \tag{6}$$

where $\rho_{i,g} = \frac{\pi_{\theta+\Delta\theta}(\mathbf{t}_{i,g})}{\pi_{\theta+\Delta\theta_{\text{SFT}}}(\mathbf{t}_{i,g})}$ is the probability ratio. The final post-training objective is the GRPO loss with KL penalty:

$$\mathcal{L}_{\text{PT}} = \mathcal{L}_{\text{GRPO}}(\Delta\theta) + \beta D_{\text{KL}}(\pi_{\theta+\Delta\theta_{\text{SFT}}} || \pi_{\theta+\Delta\theta}) \tag{7}$$

This objective encourages the model to increase the probability of generating RRLE sequences that yield higher DICE scores, thereby explicitly refining the model's pixel-level understanding via the GRPO loss while at the same time avoids diverging from the reference policy obtained from Stage I, $\pi_{\theta+\Delta\theta_{\text{SFT}}}$.

3.3 Inference

During inference, we patchify the high-resolution input image into a grid of disjoint spatial patches of dimensions $h_{\mathbf{p}} \times w_{\mathbf{p}}$ (cf. Inference pipeline in Figure 1). These patches are collated into a single batch to enable parallel processing, which significantly reduces computational latency compared to sequential execution. Once the individual patch masks are generated, they are spatially concatenated to reconstruct the full-resolution semantic map. To remove stitching artifacts and improve boundary precision, we apply the postprocessing steps recommended in EPOC [Chen *et al.*, 2025]. This includes (i) boundary detection using a SegFormer [Xie *et al.*, 2021] based refinement model and (ii) the Watershed algorithm [Vincent and Soille, 1991] on the refined semantic maps to extract discrete instance segmentation maps.

4 Experiments

4.1 Experiment Settings

Datasets. We conduct our experiments on two datasets which span across 3 countries in the Global South with smallholder agriculture systems. The first is the ALU dataset [Dua *et al.*, 2024] which contains 2009/430/431 images in its

train/val/test splits. This data is from India, annotated with five distinct semantic classes: fields, trees, clouds, ponds, and wells. The second is a publicly available benchmark dataset, AI4SmallFarms [Persello *et al.*, 2023], for agricultural field boundary delineation. The train/val/test splits contain 1962/90/304 images for Vietnam and 1860/605/177 images for Cambodia. In both cases, VHR imagery from Google Maps Satellite View are used. We take images sampled at 0.5 m/pixel GSD, from two satellite sources: Maxar Worldview, Airbus Pleiades.

Baselines. We benchmark our approach against state-of-the-art segmentation MLLM-based methods: LISA [Lai *et al.*, 2024], GSVA [Xia *et al.*, 2024], which are decoder-dependent and the decoder-free Text4Seg [Lan *et al.*, 2025]. We consider two variants of Text4Seg – coarse predictions (C) and refined predictions (R) that uses SAM refinement. Further, we also consider LISAt [Quenum *et al.*, 2025] which extends LISA for segmentation in satellite images. In addition, we include Transformer-based models specialized for segmentation, LAVT [Yang *et al.*, 2022] and GRES [Liu *et al.*, 2023a].

Evaluation Metrics. Following the evaluation protocol in [Dua *et al.*, 2024] for smallholder farms, we report the instance-wise Mean Intersection-over-Union (mIoU), Median IoU, False Positive Rate (FPR), and False Negative Rate (FNR). More details are in Appendix C.

In addition to standard performance metrics, we use a parameter efficiency metric, denoted as **Overhead**. This metric quantifies the additional learnable parameters required by a method beyond the base LLM backbone. For decoder-dependent baselines, this includes parameters of the external segmentation heads such as SAM-H, pixel decoders, and any projection layers. For decoder-free methods such as ours and Text4Seg, the overhead is zero. Concretely, total number of parameters of a method is denoted as -

$$\Phi_{\text{total}} = \Phi_{\text{base}} + \Phi_{\text{projection}} + \Phi_{\text{decoder}} \tag{8}$$

where, Φ_{base}, $\Phi_{\text{projection}}$ and Φ_{decoder} denote the parameters in the base model, projection layers and decoder heads respectively. Then, the overhead is defined as -

$$\text{Overhead} = \frac{\Phi_{\text{projection}} + \Phi_{\text{decoder}}}{\Phi_{\text{base}}} \tag{9}$$

This metric is vital for two reasons: (1) it highlights the efficacy of the base LLM's native reasoning capabilities without reliance on auxiliary decoder heads, and (2) it serves as a proxy for the computational cost and memory footprint required during inference[1].

Implementation. For MAgSeg, we experiment with MLLMs of two sizes – 4B and 12B – versions of Gemma3 [Team *et al.*, 2025]. We refer to these two variants as MAgSeg (4B) and MAgSeg (12B) respectively. These are comparable to the model sizes used in the MLLM-based baseline methods: LISA (13B), LISAt (7B), GSVA (13B) and Text4Seg (7B). We use $h_{\mathbf{p}} = w_{\mathbf{p}} = 32$ for patch dimension as mentioned in Section 3. Appendix B has other details of MAgSeg and all the baselines.

[1]Here, we refer solely to the additional overhead incurred during training; inference overhead is excluded.

4.2 Results

Segmentation Performance

India. The quantitative comparison on the panoptic segmentation dataset from India is summarized in Table 1. Our proposed method, MAgSeg, advances the state-of-the-art, outperforming all MLLM-based and transformer-based baselines by a substantial margin. Specifically, we observe an improvement of 21 mIoU points over GRES, the closest competitor.

Vietnam and Cambodia. These regions are characterized by dense wetland rice cultivation, offering a different spatial distribution compared to India. As shown in Table 2, MAgSeg consistently achieves state-of-the-art results when trained on these diverse landscapes. The performance gap is distinct in Cambodia split. Here, baseline methods struggle to learn effective segmentation boundaries; notably, even strong decoder-based models like GRES and LISA plateau at approximately 0.12 mean IoU. Text4Seg fails completely (0.04 mean IoU) due to the small, fragmented nature of agricultural fields which are lost during mask downsampling. In contrast, MAgSeg successfully learns the complex landscape features, achieving a mean IoU of 0.43. This result – nearly quadrupling the performance of the nearest baseline – demonstrates that our GRPO-based training objective enables the model to converge on difficult pixel-level tasks where standard cross-entropy or decoder-alignment losses fail. In the Vietnam data, while baseline performance improves (with GRES reaching 0.40 mean IoU), our MAgSeg-12B model maintains better performance with a mean IoU of 0.42 and a better median IoU of 0.35.

We attribute the significantly lower performance of prior methods to their inherent architectural limitations, which are addressed in MAgSeg: (a) *Text4Seg:* This method exhibits poor performance on satellite imagery. As noted in [Lan *et al.*, 2025], Text4Seg relies on downsampling masks (e.g., to 32×32) to maintain tractable token counts. In the context of high-resolution satellite data, this compression causes a severe loss of high-frequency spatial details – such as narrow field boundaries – which cannot be recovered during inference. (b) *Decoder-Based Models (LISA, GSVA, GRES):* While methods like GRES perform better than Text4Seg, they still struggle to match MAgSeg. Specifically, for field segmentation on the ALU dataset, our method outperforms the best decoder-based model (GRES) by 12 and 33 points in mIoU and median IoU, respectively. On the Cambodia dataset, the gains are even larger—31 and 40 points in mIoU and median IoU. These substantial improvements validate our claim. We attribute this to the 'domain alignment gap': the disconnect between the LLM's textual representation and the separate pixel decoder's latent space, which often leads to inconsistent segmentation masks.

Qualitative Analysis

Illustrations in Figure 2 show that MAgSeg produces sharp, instance-aware boundaries for agricultural fields where competitors often output noisy or over-smoothed/undersegmented masks. The visual comparisons corroborate our quantitative findings: Text4Seg outputs suffer from severe blockiness due to the aforementioned downsampling artifacts. Sim-

Class	Metric	LAVT	LISA	LISAt	GSVA	GRES	Text4Seg (C)	Text4Seg (R)	MAgSeg (4B) (Our)	MAgSeg (12B) (Our)
	Overhead	9.95%	9.48%	9.48%	9.35%	15.82%	**0.00%**	**0.00%**	**0.00%**	**0.00%**
Fields	**mean IoU**	0.22	0.22	0.36	0.35	0.37	0.05	0.07	0.58	**0.59**
	median IoU	0.13	0.05	0.30	0.29	0.31	0.02	0.02	0.62	**0.64**
	FNR	8.08	44.25	6.06	7.85	3.48	17.24	20.36	4.71	**3.13**
	FPR	40.66	38.72	52.28	32.58	30.98	32.59	44.99	**20.15**	21.22
Trees	**mean IoU**	0.11	0.00	0.11	0.03	0.13	0.00	0.00	0.20	**0.21**
	median IoU	0.01	0.00	0.02	0.00	0.02	0.00	0.00	0.09	**0.10**
	FNR	45.43	98.63	39.13	87.05	39.83	89.39	91.28	37.26	**34.22**
	FPR	52.19	100.00	54.74	61.47	**37.05**	68.38	76.01	42.08	42.89
Clouds	**mean IoU**	0.19	0.00	0.07	0.03	0.14	0.08	0.09	**0.25**	**0.25**
	median IoU	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.15	**0.17**
	FNR	24.53	100.00	66.04	86.79	47.17	37.74	45.28	19.82	**19.28**
	FPR	56.10	100.00	43.12	91.04	42.47	**29.57**	32.58	41.22	39.08
Ponds	**mean IoU**	0.00	0.00	0.00	0.02	0.01	0.00	0.00	**0.05**	**0.05**
	median IoU	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	FNR	100.00	97.66	100.00	**93.38**	97.90	100.00	100.00	99.56	99.16
	FPR	100.00	100.00	100.00	99.47	**68.42**	100.00	100.00	95.62	96.31
Well	**mean IoU**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	median IoU	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	FNR	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
	FPR	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00

Table 1: Quantitative results on the India (ALU) panoptic segmentation dataset. We report instance-wise mean IoU (↑), median IoU (↑), FPR (↓), FNR (↓), and model overhead (↓).

Country	Metric	LAVT	LISA	LISAt	GSVA	GRES	Text4Seg (C)	Text4Seg (R)	MAgSeg (4B) (Our)	MAgSeg (12B) (Our)
Cambodia	**mean IoU**	0.06	0.09	0.12	0.10	0.12	0.04	0.04	0.41	**0.43**
	median IoU	0.02	0.03	0.05	0.04	0.05	0.02	0.01	**0.45**	**0.45**
	FNR	2.11	3.11	3.37	3.60	**0.06**	6.21	5.31	8.69	8.59
	FPR	15.12	33.90	22.75	32.81	**2.56**	19.83	31.79	16.69	14.66
Vietnam	**mean IoU**	0.19	0.22	0.24	0.24	0.40	0.08	0.09	0.40	**0.42**
	median IoU	0.07	0.10	0.13	0.12	0.31	0.03	0.03	0.33	**0.35**
	FNR	9.30	10.01	5.51	8.90	3.67	11.02	11.45	3.99	**3.52**
	FPR	6.28	31.84	34.77	31.51	10.60	21.75	39.69	11.47	**3.86**

Table 2: Quantitative results on the AI4SmallFarms field segmentation dataset. We report instance-wise mean IoU (↑), median IoU (↑), FPR (↓), and FNR (↓) for the Vietnam and Cambodia subsets.

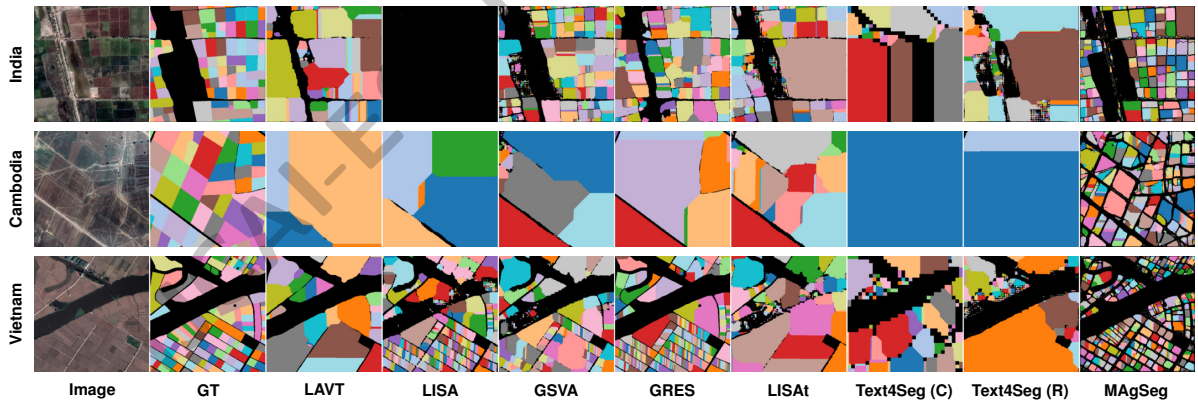

Figure 2: Qualitative results comparing our approach MAgSeg with SOTA baselines. GT: Ground Truth.

ilarly, decoder-based methods like LISA often generate artifacts or fail to capture crisp boundaries, likely due to latent misalignment. In contrast, our proposed method generates aesthetically clean, pixel-precise masks that closely align with the ground truth, effectively capturing the nuanced geometries of agricultural fields. The consistency of these results across both India (ALU) and Southeast Asia (Cambodia and Vietnam) confirms that MAgSeg's superiority is intrinsic to the method, and performs equally well across different landscapes. Figures 5, 6 and 7 in Appendix D.4 show additional illustrations for India, Cambodia and Vietnam.

Country	Metric	LAVT	LISA	LISAt	GSVA	GRES	MAgSeg (4B)
Cambodia	mean IoU	0.09	0.25	0.22	0.21	0.20	**0.40**
	median IoU	0.03	0.15	0.13	0.10	0.10	**0.43**
	FNR	17.87	23.17	22.63	26.73	25.02	**15.69**
	FPR	13.36	23.13	20.10	20.16	21.11	**11.02**
Vietnam	mean IoU	0.13	0.30	0.36	0.29	0.33	**0.41**
	median IoU	0.05	0.19	0.30	0.19	0.22	**0.37**
	FNR	19.98	24.10	**9.05**	26.98	12.57	10.33
	FPR	30.02	50.48	48.41	46.42	47.19	**25.63**

Table 3: Zero-shot evaluation on field boundary segmentation: trained on India and evaluated directly on Cambodia and Vietnam.

	LISA	GSVA	GRES	MAgSeg
Overhead	9.48%	9.35%	15.82%	**0.00%**
Average Time ($s \times 10^{-8}$)	1.69	2.04	1.80	**1.33**

Table 4: Computational efficiency: parameter overhead (↓) and the parameter-normalized training time per token (↓) of MAgSeg and the best baseline methods. Other overhead values are in Table 1.

Zero Shot Generalization

To evaluate the generalization capabilities of MAgSeg across different geographical regions, we conduct a zero-shot evaluation on the AI4SmallFarms dataset. Specifically, models trained solely on the ALU dataset (India) were evaluated directly on the Cambodia and Vietnam test sets without any fine-tuning. This setting is particularly challenging due to differences in agricultural landscapes between the source and target domains. The results are presented in Table 3. We observe that MAgSeg, the 4B version itself, significantly outperforms decoder-dependent baselines in this regime. For Cambodia, our model achieves a mean IoU of 0.40, surpassing the strongest baseline, LISA, which achieves 0.25. Notably, the median IoU for our method is 0.43, whereas baselines struggle to exceed 0.15. A similar trend is observed for Vietnam, where our method achieves a mean IoU of 0.41, compared to 0.36 for LISAt. These results indicate that decoder-based architectures may overfit to the visual statistics of the training domain, leading to poor performance when subjected to domain shifts. In contrast, our generative approach appears to learn more robust, semantic-aware representations that transfer effectively to unseen agricultural landscapes.

Architectural and Training Efficiency

A critical advantage of our approach is highlighted by the *Overhead* metric (Table 1, top row). We observe a clear dichotomy in prior work: decoder-free methods like Text4Seg offer low overhead but poor performance, whereas decoder-based methods achieve better performance at the cost of high overhead. Our method successfully breaks this trade-off, delivering superior segmentation performance with zero decoder overhead. This confirms that proper instruction tuning (via LoRA) and reinforcement learning (via GRPO) are sufficient to ground LLMs in pixel-level tasks without parameter-heavy auxiliary networks.

We also analyze the computational efficiency of MAgSeg with the best-performing baselines in Table 4. To ensure a fair comparison across different base model sizes, we report a *parameter-normalized training time*, defined as the average training duration per token divided by the total number of model parameters. Our decoder-free approach demonstrates superior efficiency on both metrics. First, regarding architectural complexity, our method incurs zero overhead, whereas decoder-dependent baselines introduce substantial additional parameters, ranging from 9.35% (GSVA) to 15.82% (GRES). Second, this streamlined architecture translates directly to training speed. Our method achieves a normalized training time of 1.33×10^{-8} s, significantly outperforming LISA/LISAt (1.69), GRES (1.80), and GSVA (2.04). This reduction in latency confirms that eliminating auxiliary vision decoders not only simplifies deployment but also enhances the computational throughput of the training process.

Additional Results

We evaluate the impact of GRPO post training and EPOC refinement during inference in MAgSeg through an **ablation study**. Table 1 in Appendix D.1 shows that, while each of them individually improves the mean IoU, the combination of both components yields the best performance. We qualitatively demonstrate the importance of GRPO in suppressing spurious boundaries in figure 4 (Appendix D.3). We also present (in Appendix D.2) the comparative performance on field segmentation **stratified by various criteria**: MAgSeg outperforms the best baselines across 5 different *field size ranges* ($< 100m^2$ to 5 acres; and by a large margin for smaller field sizes $<$ 1 acre), 5 *climatic regions* and 7 *ecological regions*. MAgSeg's robustness to field sizes is particularly important in smallholder landscapes, where plot sizes are highly heterogeneous.

Limitations

Our experiments also highlight limitations of MAgSeg. Performance drops on minority classes (e.g., *wells, ponds*) due to inherent class imbalance, a challenge shared by most segmentation baselines. Additionally, our patch-based inference lacks cross-patch spatial context, which can produce stitching artifacts and geometric discontinuities in final maps (see Figure 8 in Appendix D.5). Future work can address both limitations within our framework.

5 Conclusion

In this work, we presented MAgSeg, the first MLLM framework specifically adapted for smallholder agricultural landscape segmentation from VHR satellite imagery. MAgSeg successfully addresses the critical bottlenecks of prior decoder-free MLLM approaches – token-length limitations and domain alignment inefficiencies, without using parameter-heavy auxiliary decoders. Extensive evaluations on smallholder agriculture datasets spanning 3 countries demonstrate the superiority of our method over previous MLLM-based methods, achieving up to $4\times$ improvement over the nearest competitor while incurring zero parameter overhead. Our evaluation also confirms the robustness of MAgSeg across diverse geographies including in zero-shot settings. Ultimately, this work illustrates that with appropriate grounding strategies, general-purpose foundation models can be effectively adapted for specialized, high-precision agricultural landscape segmentation, paving the way for scalable, accessible, and accurate global agricultural monitoring.

References

[Chen *et al.*, 2025] Delong Chen, Samuel Cahyawijaya, Jianfeng Liu, Baoyuan Wang, and Pascale Fung. Subobject-level image tokenization. In *International Conference on Machine Learning*, 2025.

[Dua *et al.*, 2024] Radhika Dua, Nikita Saxena, Aditi Agarwal, Alex Wilson, Gaurav Singh, Hoang Tran, Ishan Deshpande, Amandeep Kaur, Gaurav Aggarwal, Chandan Nath, Arnab Basu, Vishal Batchu, Sharath Holla, Bindiya Kurle, Olana Missura, Rahul Aggarwal, Shubhika Garg, Nishi Shah, Avneet Singh, Dinesh Tewari, Agata Dondzik, Bharat Adsul, Milind Sohoni, Asim Rama Praveen, Aaryan Dangi, Lisan Kadivar, E Abhishek, Niranjan Sudhansu, Kamlakar Hattekar, Sameer Datar, Musty Krishna Chaithanya, Anumas Ranjith Reddy, Aashish Kumar, Betala Laxmi Tirumala, and Alok Talekar. Agricultural landscape understanding at country-scale. *arXiv*, 2024. https://arxiv.org/abs/2411.05359.

[FAO *et al.*, 2023] FAO, IFAD, UNICEF, WFP, and WHO. The state of food security and nutrition in the world 2023. urbanization, agrifood systems transformation and healthy diets across the rural–urban continuum, 2023. https://doi.org/10.4060/cc3017en.

[Hu *et al.*, 2022] 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.

[Kang and Özdoğan, 2019] Y. Kang and M. Özdoğan. Field-level crop yield mapping with landsat using a hierarchical data assimilation approach. *Remote Sensing of Environment*, 228:144–163, 2019.

[Kerner *et al.*, 2023] Hannah Kerner, Saketh Sundar, and Mathan Satish. Multi-region transfer learning for segmentation of crop field boundaries in satellite images with limited labels. In *Proceedings of the AAAI Workshop on AI to Accelerate Science and Engineering*, 2023.

[Kirillov *et al.*, 2023] 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.

[Lai *et al.*, 2024] 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.

[Lan *et al.*, 2025] Mengcheng Lan, Chaofeng Chen, Yue Zhou, Jiaxing Xu, Yiping Ke, Xinjiang Wang, Litong Feng, and Wayne Zhang. Text4seg: Reimagining image segmentation as text generation. In *The Thirteenth International Conference on Learning Representations*, 2025.

[Lesiv *et al.*, 2019] M. Lesiv, J.C. Laso Bayas, L. See, M. Duerauer, D. Dahlia, N. Durando, R. Hazarika, P. Kumar Sahariah, M. Vakolyuk, and V. Blyshchyk. Estimating the global distribution of field size using crowdsourcing. *Global Change Biology*, 25:174–186, 2019.

[Li *et al.*, 2024] Xiang Li, Congcong Wen, Yuan Hu, Zhenghang Yuan, and Xiao Xiang Zhu. Vision-language models in remote sensing: Current progress and future trends. *IEEE Geoscience and Remote Sensing Magazine*, 12(2):32–66, 2024.

[Liu *et al.*, 2023a] Chang Liu, Henghui Ding, and Xudong Jiang. Gres: Generalized referring expression segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 23592–23601, 2023.

[Liu *et al.*, 2023b] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023.

[Masoud *et al.*, 2020] K.M. Masoud, C. Persello, and V.A. Tolpekin. Delineation of agricultural field boundaries from sentinel-2 images using a novel super-resolution contour detector based on fully convolutional networks. *Remote Sensing*, 12(1):59, 2020.

[Mei *et al.*, 2022] Weiye Mei, Haoyu Wang, David Fouhey, Weiqi Zhou, Isabella Hinks, Josh M Gray, Derek Van Berkel, and Meha Jain. Using deep learning and very-high-resolution imagery to map smallholder field boundaries. *Remote Sensing*, 14(13):3046, 2022.

[Ministry of Agriculture and Farmers Welfare, 2024] Ministry of Agriculture and Farmers Welfare. Categorisation of farmers. https://www.pib.gov.in/PressReleasePage.aspx?PRID=2085181, 2024.

[OECD, 2023] OECD. *Agricultural Policy Monitoring and Evaluation 2023: Adapting Agriculture to Climate Change*. OECD Publishing, Paris, 2023. https://doi.org/10.1787/b14de474-en.

[Persello *et al.*, 2023] Claudio Persello, Jeroen Grift, Xinyan Fan, Claudia Paris, Ronny Hänsch, Mila Koeva, and Andrew Nelson. Ai4smallfarms: A dataset for crop field delineation in southeast asian smallholder farms. *IEEE Geoscience and Remote Sensing Letters*, 20:1–5, 2023.

[Quenum *et al.*, 2025] Jerome Quenum, Wen-Han Hsieh, Tsung-Han Wu, Ritwik Gupta, Trevor Darrell, and David M. Chan. LISAt: Language-instructed segmentation assistant for satellite imagery. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025.

[Rada and Fuglie, 2019] N.E. Rada and K.O. Fuglie. New perspectives on farm size and productivity. *Food Policy*, 84:147–152, 2019.

[Radford *et al.*, 2021] 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.

[Rasheed *et al.*, 2024] 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.

[Rolf *et al.*, 2024] Esther Rolf, Konstantin Klemmer, Caleb Robinson, and Hannah Kerner. Mission critical–satellite data is a distinct modality in machine learning. In *International Conference on Learning Representations*, 2024.

[Rudel *et al.*, 2009] Thomas K Rudel, Laura Schneider, Maria Uriarte, Billie Lee Turner, Ruth DeFries, Deborah Lawrence, Jacqueline Geoghegan, Susanna Hecht, Amy Ickowitz, Eric F Lambin, et al. Agricultural intensification and changes in cultivated areas, 1970–2005. *Proceedings of the national academy of sciences*, 106(49):20675–20680, 2009.

[Samberg *et al.*, 2016] L.H. Samberg, J.S. Gerber, N. Ramankutty, M. Herrero, and P.C. West. Subnational distribution of average farm size and smallholder contributions to global food production. *Environmental Research Letters*, 11(12):124010, 2016.

[Shao *et al.*, 2024] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

[Sylvester and others, 2015] Gerard Sylvester et al. Success stories on information and communication technologies for agriculture and rural development, 2015.

[Team *et al.*, 2025] Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

[Vincent and Soille, 1991] Luc Vincent and Pierre Soille. Watersheds in digital spaces: an efficient algorithm based on immersion simulations. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 13(06):583–598, 1991.

[Waldner and Diakogiannis, 2020] François Waldner and Foivos I Diakogiannis. Deep learning on edge: Extracting field boundaries from satellite images with a convolutional neural network. *Remote sensing of environment*, 245:111741, 2020.

[Waldner *et al.*, 2021] François Waldner, Foivos I Diakogiannis, Kathryn Batchelor, Michael Ciccotosto-Camp, Elizabeth Cooper-Williams, Chris Herrmann, Gonzalo Mata, and Andrew Toovey. Detect, consolidate, delineate: Scalable mapping of field boundaries using satellite images. *Remote sensing*, 13(11):2197, 2021.

[Wang *et al.*, 2022] Sherrie Wang, François Waldner, and David B. Lobell. Unlocking large-scale crop field delineation in smallholder farming systems with transfer learning and weak supervision. *Remote Sensing*, 14(22), 2022.

[Wang *et al.*, 2023] Wenhai Wang, Zhe Chen, Xiaokang Chen, Jiannan Wu, Xizhou Zhu, Gang Zeng, Ping Luo, Tong Lu, Jie Zhou, Yu Qiao, et al. Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. *Advances in Neural Information Processing Systems*, 36:61501–61513, 2023.

[Wu *et al.*, 2024] Jiannan Wu, Muyan Zhong, Sen Xing, Zeqiang Lai, Zhaoyang Liu, Zhe Chen, Wenhai Wang, Xizhou Zhu, Lewei Lu, Tong Lu, et al. Visionllm v2: An end-to-end generalist multimodal large language model for hundreds of vision-language tasks. *Advances in Neural Information Processing Systems*, 37:69925–69975, 2024.

[Wu *et al.*, 2025] Haiyang Wu, Zhuofei Du, Dandan Zhong, Yuze Wang, and Chao Tao. Fsvlm: A vision-language model for remote sensing farmland segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 2025.

[Xia *et al.*, 2024] Zhuofan Xia, Dongchen Han, Yizeng Han, Xuran Pan, Shiji Song, and Gao Huang. Gsva: Generalized segmentation via multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3858–3869, June 2024.

[Xie *et al.*, 2021] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*, 34:12077–12090, 2021.

[Yang *et al.*, 2022] Zhao Yang, Jiaqi Wang, Yansong Tang, Kai Chen, Hengshuang Zhao, and Philip H.S. Torr. Lavt: Language-aware vision transformer for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18155–18165, June 2022.

[Zheng *et al.*, 2025] Juepeng Zheng, Zi Ye, Yibin Wen, Jianxi Huang, Zhiwei Zhang, Qingmei Li, Qiong Hu, Baodong Xu, Lingyuan Zhao, and Haohuan Fu. A comprehensive review of agricultural parcel and boundary delineation from remote sensing images: Recent progress and future perspectives. *arXiv preprint arXiv:2508.14558*, 2025.

[Zhou *et al.*, 2024] Tianfei Zhou, Wang Xia, Fei Zhang, Boyu Chang, Wenguan Wang, Ye Yuan, Ender Konukoglu, and Daniel Cremers. Image segmentation in foundation model era: A survey. *arXiv preprint arXiv:2408.12957*, 2024.

[Zhu *et al.*, 2017] Xiao Xiang Zhu, Devis Tuia, Lichao Mou, Gui-Song Xia, Liangpei Zhang, Feng Xu, and Friedrich Fraundorfer. Deep learning in remote sensing: A comprehensive review and list of resources. *IEEE geoscience and remote sensing magazine*, 5(4):8–36, 2017.