A large-scale image-text dataset benchmark for farmland segmentation

Chao Tao, Dandan Zhong, Weiliang Mu, Zhuofei Du, and Haiyang Wu
School of Geosciences and Info-Physics, Central South University, Changsha 410083, China
Correspondence to: Haiyang Wu (245001024@csu.edu.cn)

Abstract. Understanding and mastering the spatiotemporal characteristics of farmland is essential for accurate farmland segmentation. The traditional deep learning paradigm that solely relies on labeled data has limitations in representing the spatial relationships between farmland elements and the surrounding environment. It struggles to effectively model the dynamic temporal evolution and spatial heterogeneity of farmland. Language, as a structured knowledge carrier, can explicitly express the spatiotemporal characteristics of farmland, such as its shape, distribution, and surrounding environmental information. Therefore, a language-driven learning paradigm can effectively alleviate the challenges posed by the spatiotemporal heterogeneity of farmland. However, in the field of remote sensing imagery of farmland, there is currently no comprehensive benchmark dataset to support this research direction. To fill this gap, we introduced language-based descriptions of farmland and developed FarmSeg-VL dataset—the first fine-grained image-text dataset designed for spatiotemporal farmland segmentation. Firstly, this article proposed a semi-automatic annotation method that can accurately assign caption to each image, ensuring high data quality and semantic richness while improving the efficiency of dataset construction. Secondly, the FarmSeg-VL exhibits significant spatiotemporal characteristics. In terms of the temporal dimension, it covers all four seasons. In terms of the spatial dimension, it covers eight typical agricultural regions across China, with a total area of approximately 4,300 km². In addition, in terms of captions, FarmSeg-VL covers rich spatiotemporal characteristics of farmland, including its inherent properties, phenological characteristics, spatial distribution, topographic and geomorphic features, and the distribution of surrounding environments. Finally, we perform a performance analysis of the vision language model and a deep learning model that relies only on labels trained on FarmSeg-VL. Models trained on the vision language model outperform deep learning models that rely only on labels by 10%-20%, demonstrating its potential as a standard benchmark for farmland segmentation. The FarmSeg-VL dataset will be publicly released at https://doi.org/10.5281/zenodo.15860191(Tao et al., 2025a).

1 Introduction

Farmland has been the foundation of agricultural food security, and accurately monitoring farmland has been crucial for implementing policies such as farmland improvement, enhanced supervision, and planning and control (Sishodia et al., 2020).

Currently, the intelligent interpretation of remote sensing images for farmland based on deep learning has become a primary method for farmland monitoring(Li et al., 2023; Tu et al., 2024) .

However, existing farmland remote sensing image segmentation methods mainly follow a label-driven deep learning paradigm, which faces significant bottlenecks in both data and model. Specifically, in terms of datasets, although existing benchmark datasets have contributed to the advancement of farmland segmentation technology to some extent, they rely solely on label-driven deep learning paradigm, which has two main limitations: First, a single label can only drive the model to learn shallow visual features of farmland, which fails to reveal the underlying driving mechanisms affecting the spatial distribution and temporal evolution of farmland. Additionally, it is difficult to represent the spatial-temporal heterogeneity in complex agricultural environments. Specifically, the surface cover of farmland shows seasonal differences in complete coverage, partial coverage, and no coverage with the growth cycle of crops, while diverse terrain leads to significant geographical differentiation in the spatial distribution of farmland and its associations with surrounding features such as water bodies, buildings, and vegetation. However, existing datasets cannot represent this kind of spatial-temporal heterogeneity, making it difficult for models to establish the inherent relationships between farmland and its surrounding environment. In terms of model, although technologies such as convolutional neural networks (CNNs), graph convolutional networks (GCNs), and Transformer have significantly enhanced feature representation capabilities, the existing label-driven paradigm inherently has clear theoretical flaws. First, the existing label-driven paradigm to excessively rely on visual cues and neglect the logical connections between farmland and its surrounding environment in complex farmland scenarios. Second, the label struggles to reflect the evolution of farmland across seasons and growth stages, severely limiting the model's generalization ability in spatiotemporal dynamic scenarios. Therefore, there is an urgent need to break through the theoretical framework of the traditional label-driven deep learning paradigm and explore a new paradigm capable of uncovering the deep semantic logic of farmland.

With the emergence of vision language models (VLMs) and their expanding applications across various fields, studies (Devlin et al., 2019; Wu et al., 2025b, a) have shown that language can reveal deeper semantic clues behind visual information. These VLMs typically follow a general construction process: first, feature representations are extracted from images through a visual encoder, a process aimed at capturing key visual representation in the images. For example, in the LLaVA model (Liu et al., 2023), the image representations generated by the fixed visual encoder lay the foundation for subsequent processing. Next, to establish a connection between vision and language, the model needs to map the extracted visual features to the space of the language model, enabling visual representation to be translated into natural language descriptions or understood. The LLaVA model precisely utilizes this method, mapping image representations to the prompt space of large language models, helping the model understand the relationship between visual representation and linguistic expressions, thereby achieving efficient downstream tasks. Furthermore, to enable the model to handle complex tasks, integrating visual perception with language understanding becomes a key step. LISA (Lai et al., 2023) is a typical example, it not only combines visual perception capabilities but also incorporates in-depth language understanding abilities, allowing it to perform reasoning-based tasks such as segmentation tasks. This multimodal information processing capability is one of the important characteristics of VLMs, enabling them to consider visual context while understanding and generating language. These breakthroughs make up for the

shortcomings of relying solely on label-guided models to handle complex spatiotemporal heterogeneous farmland scenes, making it possible to mine the complex semantic information in farmland remote sensing images and then model the deep inherent logical relationship between farmland and its surroundings. Specifically, language can guide models to capture farmland features across multiple dimensions, including shape and boundaries, phenological characteristics that reflect seasonal changes and crop growth states, spatial layout based on latitude and longitude, and geographical features such as terrain and landscape morphology. Additionally, language can describe the relative positional relationships between farmland and surrounding features such as water bodies, buildings, and vegetation. By integrating these rich semantic cues, VLMs can better understand and interpret the complexity of farmland.

However, in remote sensing, many existing image-text datasets struggle to provide detailed captions and precise annotations for specific land features like farmland. As a result, they often fall short of meeting the requirements for high-accuracy farmland segmentation. For example, the first large-scale remote sensing image-text pair dataset RS5M (Zhang et al., 2024) and the SkyScript dataset (Wang et al., 2024), which contains millions of image-text combinations, although large in scale, provide a relatively rough description of farmland and fail to deeply describe the specific characteristics of the farmland. In addition, although the manually annotated dataset RSICap (Hu et al., 2023) provides scene-level semantic descriptions, it lacks a refined depiction of the characteristics of the farmland itself, making it difficult to meet the model's need for deep semantic information extraction of the farmland. In contrast to the methods mentioned above, ChatEarthNet (Yuan et al., 2024) seeks to enhance the richness of semantic captions for land cover types by employing detailed prompt strategies and leveraging semantic segmentation labels from ChatGPT and the WorldCover project. However, due to the inherent randomness of automatically generated captions, these captions tend to emphasize the spatial location of farmland within the image while often lacking detailed information about its inherent attributes. Although these datasets have contributed significantly to advancing image-text understanding in remote sensing, most focus on general remote sensing tasks, with only a small portion dedicated to farmland captions. Moreover, these captions are often neither comprehensive nor in-depth. Existing datasets have not fully reflected the complexity of farmland and its changing characteristics over time and space. This is particularly evident in high-precision farmland segmentation tasks, where there is a lack of deep analysis of farmland characteristics and how they behave in different scenarios.

To address the above issues, this paper constructs the FarmSeg-VL dataset, a dedicated image-text dataset focused on farmland segmentation, which fully reflects the spatiotemporal characteristics of farmland. FarmSeg-VL covers eight typical agricultural regions in China and includes data samples from four seasons, filling the gap of spatial and temporal imbalance in existing datasets. With its extensive geographical coverage and seasonal variations, this dataset ensures effective support for the learning of various forms of farmland.

The contributions of this paper are as follows:

1) This study constructed the first farmland image-text benchmark dataset, filling the gap in remote sensing image-text datasets for the farmland-dedicate domain. This dataset includes various types of farmland, and covers a wide spatial and

temporal range, providing a high-value data foundation for the application research of vision language models in the field of farmland segmentation.

2) We summarize 11 key elements for describing farmland's inherent properties and its surrounding environment, offering a comprehensive framework for characterizing farmland from multiple perspectives. Additionally, a text template for describing farmland images was designed, providing an important reference for constructing a language dataset focused on farmland.

3) This study developed a semi-automated annotation method based on the caption templates constructed in this paper. We utilize the semi-automated annotation approach to generate mask and rich captions, significantly reducing labor time while enhancing the authenticity and reliability of the annotations.

4) Extensive experiments have demonstrated that the model trained on the image-text farmland dataset proposed in this paper significantly improves farmland segmentation performance and exhibits strong transferability, providing a performance baseline for vision language models in farmland segmentation.

2 Review of Existing Remote Sensing Datasets for Farmland Segmentation

2.1 Non image-text dataset

Traditional remote sensing dataset for farmland segmentation are mainly annotated with single-label, which can be divided into two categories: dedicated dataset and non-dedicated dataset. The detailed information is provided in Table 1. Non-dedicated datasets, such as the scene level dataset BigEarthNet (Sumbul et al., 2019), are not very suitable for pixel level farmland segmentation. Piexel-level dataset, such as WorldCover (ESA) (Zanaga et al., 2022), DynamicWorld (DyWorld) (Brown et al., 2022), and LandCover (Karra et al., 2021) , primarily focus on large-scale mapping and macro-level analysis, making them less suitable for fine-grained farmland segmentation. Moreover, Evlab-SS (Wang et al., 2017) focuses on pixel-level classification, but the proportion of farmland pixels is relatively low, and it remains limited in data scale and coverage area. Although GID (Tong et al., 2020), DeepGlobe-LandCover (Demir et al., 2018), and LoveDA(Wang et al., 2022a) cover large farmland areas with relatively high pixel proportions, the farmland samples lack diversity. For example, the farmland forms in DeepGlobe-LandCover and LoveDA are mostly regular and contiguous, lacking diversity in farmland representation. While these non-dedicated datasets provide large amounts of data for farmland segmentation, their annotations are relatively coarse. Specifically, in pixel-level farmland segmentation, they struggle to fully cover the complex shapes, distribution patterns, and finer details, such as crop growth stages.

In contrast, dedicated datasets such as GFSAD30 (Phalke and Özdoğan, 2018), WEIMIN (Hou et al., 2023), VACD (Li et al., 2024), and FGFD (Li et al., 2025) are specifically designed for farmland segmentation. These datasets offer high-precision farmland annotation and cover a broader range of farmland forms, crop distributions, and other relevant information. The GFSAD30 dataset has a spatial resolution of 30m, making it suitable for large-scale farmland monitoring, but not for fine-grained farmland segmentation. By contrary， WEIMIN and VACD offer higher resolutions, however, since WEIMIN only

covers Hebei and VACD only covers Guangdong in China, the diversity of farmland samples is limited. The FGFD dataset includes farmland samples from multiple geographic regions. However, it does not account for the phenological characteristics of farmland, limiting its ability to capture seasonal variations and crop growth stages. Although these dedicated datasets offer high annotation accuracy and support fine-grained regional monitoring, their reliance solely on labels to represent farmland's visual characteristics across different spatiotemporal conditions overlooks its inherent complexity and diversity. As a result, they struggle to capture the subtle differences and dynamic changes in farmland driven by seasonal variations and environmental factors.

Table 1. Detailed information on non image-text dataset of farmland.

Type	Dataset	Category	Spatial Resolution	Image size	Farmland Proportion	Region
Non-dedicated datasets	Evlab-SS	11	0.1-2	4500×4500	8.77	/
	GID	15	4	56×56,112×112,224×224	30.66	China
	DGLC	7	0.5	2448×2448	57.74	Thailand, Indonesia, India
	LoveDA	7	0.3	1024×1024	26.79	Nanjing,Changzhou,Wuhan,China
	Bigearthnet	43	10-60	120×120	12.41	/
Dedicated datasets	GFSAD30	3	30	/	/	Europe,Middle East,Russia and Asia
	VACD	2	0.5	512×512	/	Guangdong,China
	WEIMIN	2	0.5-2	512×512	/	Hebei,China
	FGFD	2	0.3	512×512	/	Heilongjiang,Hebei,Shanxi,Guizhou,Hubei,Jiangxi,Xizang,China

2.2 Image-Text Datasets

Existing remote sensing image-text paired datasets, such as UCM-Captions (Qu et al., 2016), RSICD (Lu et al., 2018), RS5M, NWPU-Captions (Cheng et al., 2022), RSICap, SkyScript, and ChatEarthNet, have been widely used in remote sensing research (see Table 2, where CGM denotes Caption Generation Method). However, these datasets are primarily designed for tasks such as image captioning, scene classification, or image-text retrieval, with limited applicability to farmland segmentation. This limitation stems from their insufficient in-depth semantic representations of farmland morphological characteristics, spatial distribution patterns, and contextual relationships with surrounding features. Consequently, these datasets cannot meet the requirements for fine-grained semantic understanding essential for high-precision farmland segmentation.

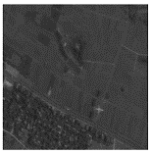
Specifically, most of these datasets focus on high-level descriptions of images, such as scene level or object level characteristics, rather than the detailed semantic annotations needed for fine-grained tasks like farmland segmentation. For example, in SkyScript, the image caption "land use of farmland" provides only broad classification information without offering specific details about farmland characteristics, such as shape, boundaries, crop growth stages, or surrounding environmental features. Similarly, the RS5M dataset provides only brief titles for images, primarily indicating the image source and land cover categories, without offering detailed descriptions of farmland. Additionally, while some datasets use automated methods to generate large-scale image-text pairs, these automatically generated datasets often suffer from inconsistent quality. The generated text frequently lacks detail and contains redundant information, reducing its effectiveness for fine-grained

farmland analysis. For example, in ChatEarthNet, image captions divide each image into four sections—top, bottom, left, and right—focusing on the proportions of primary and secondary land cover types in each section rather than providing a dedicated description of farmland. Manually annotated datasets, such as UCM-Captions, RSICD, and NWPU-Captions, provide five captions per farmland image. However, these descriptions are often repetitive and lack specificity. For example, in UCM-Captions, farmland is described simply as "There is a piece of farmland," while the remaining four descriptions merely rephrase this sentence without adding meaningful details. In RSICD, captions are limited to color and location, such as "green" or "between two forests." NWPU-Captions expands on this slightly by incorporating shape descriptions, like "rectangular," but still lacks deeper insights into farmland characteristics. Although RSICap includes descriptions related to image quality, its farmland annotations remain focused on landscape features and surrounding environments, overlooking inherent farmland attributes. This limited descriptive approach fails to capture farmland's spatiotemporal complexity, making it hard for precise farmland semantic segmentation.

Although these image-text datasets have achieved certain results in large-scale pre-training tasks, their application in the semantic segmentation of farmland remote sensing images is greatly limited due to the lack of pixel-level annotation for semantic segmentation and in-depth description of specific tasks such as farmland segmentation. Therefore, to better support farmland segmentation, the dataset needs to be enhanced by including more fine-grained semantic annotations and comprehensively covering the complex features of farmland.

Table 2. Detailed information on the image-text dataset.

Dataset	Example	CGM	Number	Farmland-related Descriptions
UCM-Captions		manual annotation	2100 images, with 5 captions per image	1.There is a piece of farmland. 2.There is a piece of farmland. 3.It is a piece of farmland. 4.It is a piece of farmland. 5.Here is a piece of farmland.
RSICD		manual annotation	10921 images, each with 5 captions	1.the cream colored and aqua farmland is between two forest. 2.the green and white farmland is between two forest. 3.the cream colored and aqua farmland is between two forest. 4.this farmland with light green parts and bald ones mixes up with those deep green woods. 5.some pieces of green farmlands are together.
RS5M		filtering publicly available image-text paired datasets	5 million images with captions	a satellite image of a farm with a green field
NWPU-Captions		manual annotation	31500 images, 5 captions per image	1.Some rectangular farmlands of different colors are neatly lay out on the ground. 2.Many neatly arranged dark green, light green and tan mixed rectangular farmlands of different sizes. 3.There are some green and uncovered rectangular farmland. 4.There are rectangular farmlands of varying sizes. 5.There are some green rectangular farmlands distributed neatly.

Dataset	Example	CGM	Number	Farmland-related Descriptions
RSICap		manual annotation	2585 image-text pair	This is a low-resolution panchromatic satellite image showing a village and farmland. At the bottom of the image, there is a village with dense buildings, and above the village is a large area of farmland, divided into sections by some dirt roads. There is also a body of water in the middle of the farmland. In the image, you can also see an airplane, which was probably captured by the satellite when it was flying over the farmland.
SkyScript		linking remote sensing images with semantics in OSM through geographic coordinates	2.6 million image-text pairs, each image corresponding to a title describing a single object and a title describing multiple objects	Single-object text: landuse of farmland, crop of cotton Multi-object text: landuse of farmland with crop of cotton
ChatEarth Net		automatically generate GPT through effective prompts	ChatGPT3.5 generates 163488 image-text pairs, ChatGPT-4v generates 10000 image-text pairs containing captions	The image primarily consists of crop fields, which are most dominant across all sections. In the top left, there is a significant expanse of crop fields, with a small area of grass and developed land. Moving to the top right, crop fields continue to dominate, followed by a smaller developed area and grassy patches. In the bottom left, the landscape is mostly covered by crop fields, followed by a few trees and a small amount of grass. The bottom right also exhibits a large area of crop fields, accompanied by a small developed area and a small portion of grass. In the middle section, crop fields are again the main feature, with a small number of trees and a tiny developed area. Overall, the image depicts a landscape predominantly characterized by crop cultivation, with minor presence of developed areas, trees, and grass.

3 FarmSeg-VL: A large-scale image-text dataset benchmark for farmland segmentation

3.1 Construction of FarmSeg-VL

The construction process of the FarmSeg-VL is shown in Fig. 1, which is mainly divided into three parts: remote sensing image acquisition and processing, caption construction, and semi-automatic annotation. In the part1, we collected high-resolution images (with a resolution of 0.5m-2m) from various typical agricultural regions in China across four seasons to ensure the dataset covers farmland with diverse spatiotemporal features. In the part2, the study synthesized the spatiotemporal characteristics of farmland and summarized 11 key factors related to its inherent properties and the distribution of surrounding environments. These factors were then used to generate detailed captions, covering aspects such as farmland shape, terrain, sowing situation, and the distribution of surrounding water bodies, vegetation, and buildings. In the part3, a semi-automated manual annotation method was employed to generate corresponding binary masks and a segment of caption for each remote sensing image sample, thus completing the dataset construction.

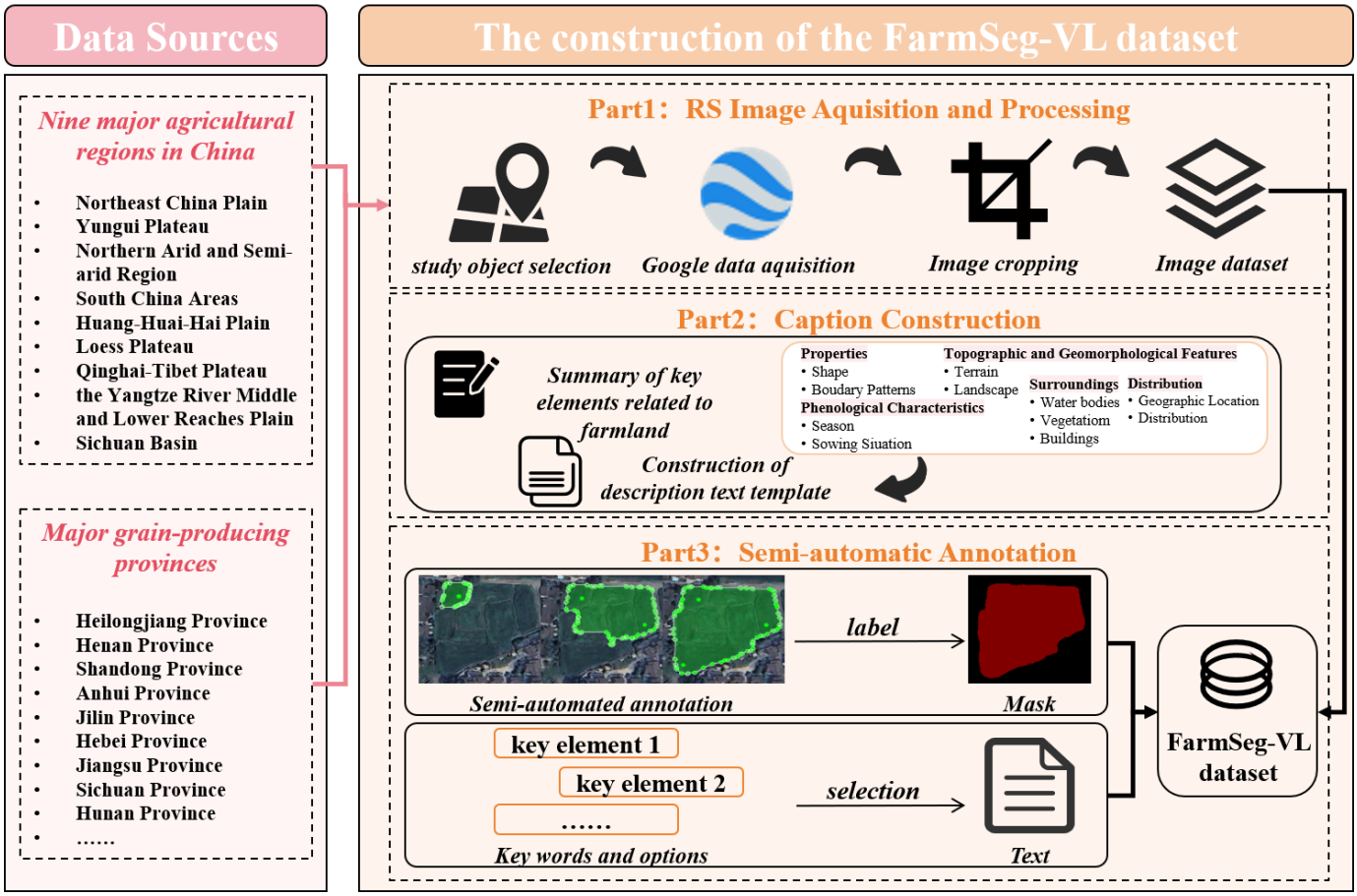


Fig. 1. Dataset construction.

The FarmSeg-VL dataset, as shown in Fig. 2, consists of three key components: image, mask, and text. Specifically, FarmSeg-VL includes image data from eight major agricultural regions across four seasons, and the image features include diversity under different imaging conditions. The caption focuses on five attributes of farmland remote sensing images with a total of eleven key features: inherent properties (such as shape and boundary pattern), phenological characteristics (such as season and sowing situation), spatial distribution (such as distribution and geographic location information), topographic and geomorphic features (such as terrain and landscape), and distribution of surrounding environments (such as buildings, water bodies, and vegetation).

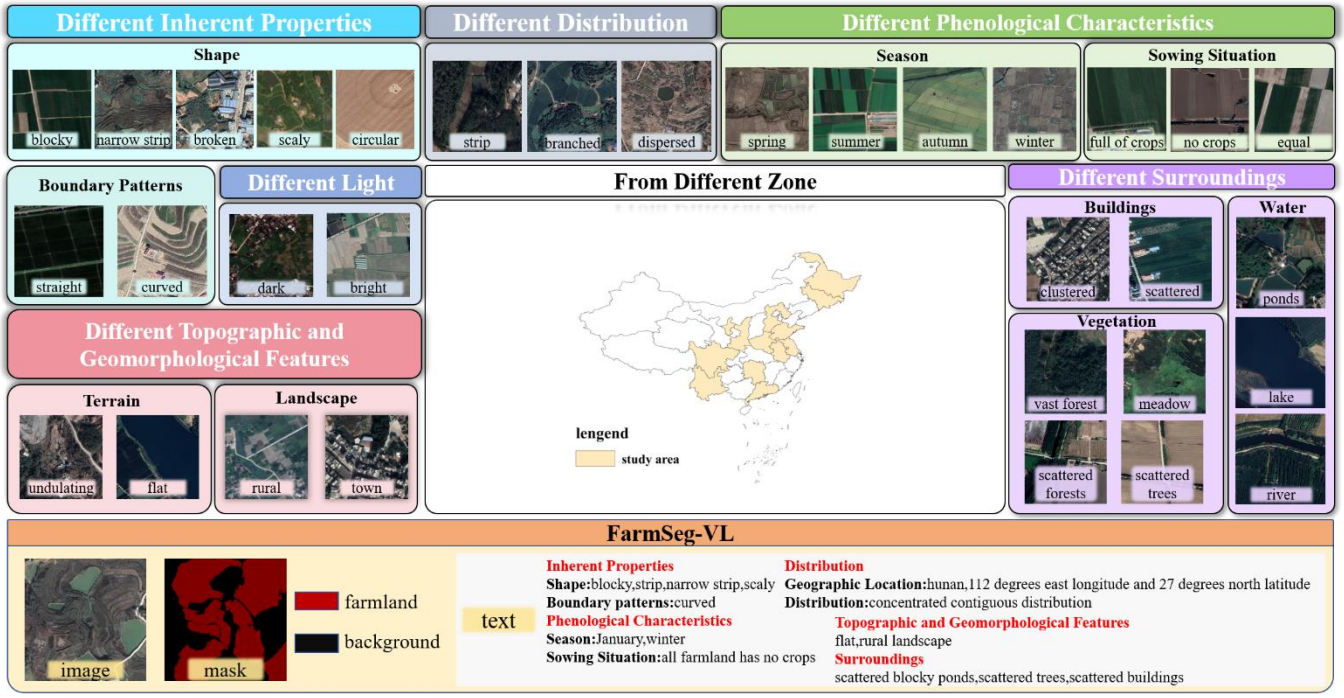


Fig. 2. Attribute annotation and spatiotemporal distribution of FarmSeg-VL.

1) RS Image Acquisition and Processing

China's vast territory, diverse landforms, and complex climate result in significant regional variations in agricultural conditions, leading to highly heterogeneous texture features and distribution patterns of farmland in remote sensing imagery. As a result, farmland exhibits significant spatiotemporal dynamics and fragmented distribution characteristics, presenting diverse spatial patterns due to these regional differences. For example, the land in the Northeast China Plain is flat and fertile, and the farmland has the characteristics of concentrated distribution and regular shape, while the Yungui Plateau in China has complex terrain and diverse climate, and the farmland has the characteristics of dispersed distribution and fragmented shape. The farmland appearance and characteristics of these agricultural areas are unique, which poses different challenges and opportunities for farmland segmentation. This study selected representative agricultural regions based on the spatial distribution and morphological characteristics of farmland. Specifically, based on the spatial aggregation and morphological regularity of farmland, the Northeast China Plain and Huang-Huai-Hai Plain were selected as typical regions characterized by concentrated and regular-shaped farmland. For areas with sloped farmland distribution, the Northern Arid and Semi-Arid Region and the Loess Plateau were chosen as study areas. At the same time, in view of the particularity of farmland morphology, such as narrow and long, striped, and sporadic and fragmented, the South China Areas, Sichuan Basin, Yungui Plateau, and Yangtze River Middle and Lower Reaches Plain were selected as research areas. The study covers 13 provincial-level administrative regions, including Heilongjiang, Jilin, Ningxia, Hebei, Henan, Shandong, Shaanxi, Anhui, Hunan, Jiangsu, Guangdong,

Sichuan, and Yunnan. These regions provide broad spatial coverage, highlight distinct regional characteristics, and are highly representative and typical of China's diverse agricultural landscapes.

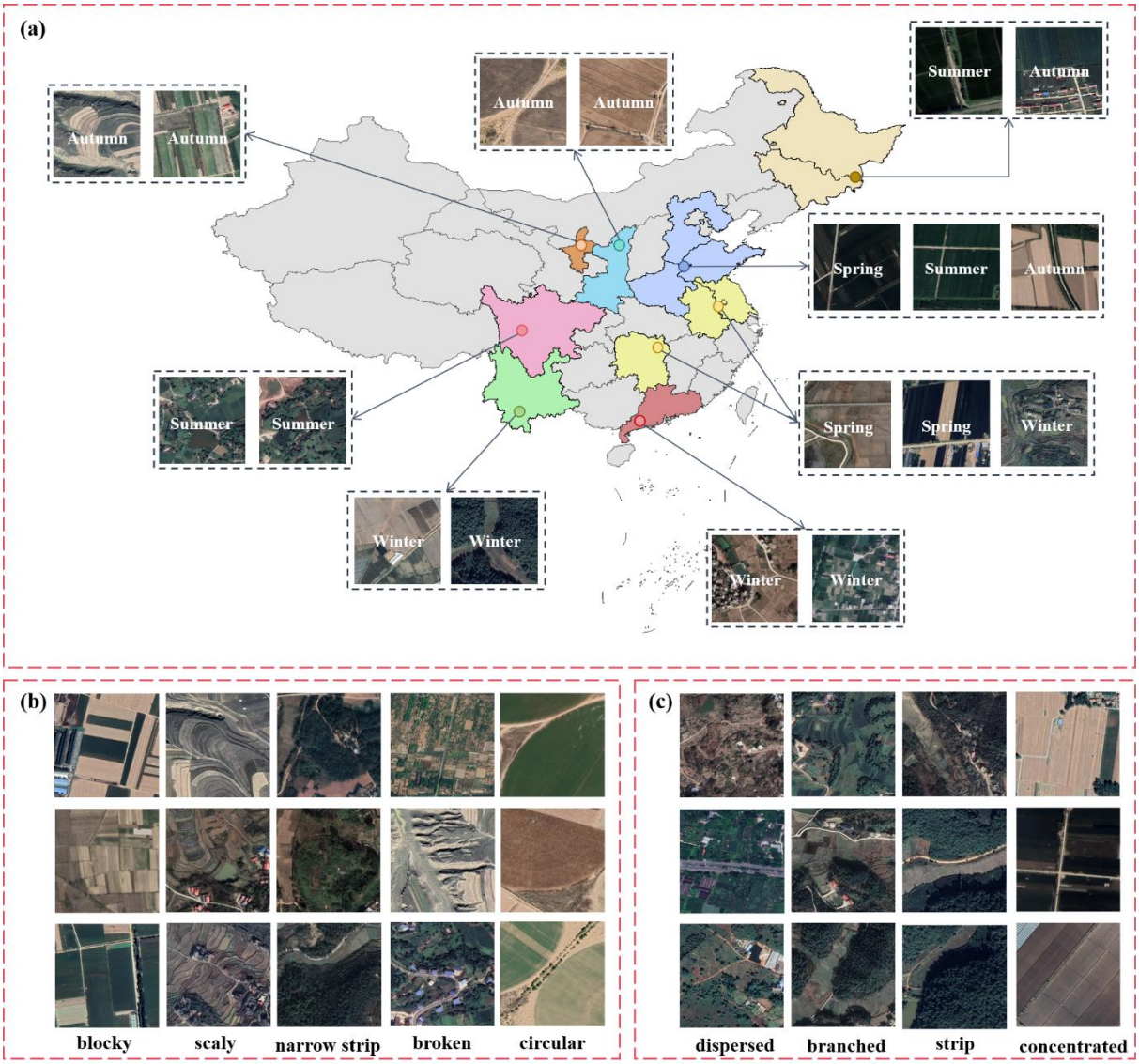


Fig. 3. Demonstration of the diversity of data samples. (a) Farmland samples from different agricultural regions. (b) Farmland samples with different shapes(c) Farmland samples with varying distribution patterns.

The data samples diversity are shown in Fig. 3. Specifically, we utilized Bigemap software to acquire high-resolution Google satellite imagery covering China, including the eight major agricultural regions previously mentioned. The spatial resolution of images ranges from 0.5m to 2m. Additionally, the software enables us to obtain the shooting time of the image. The total coverage area spans approximately 4300 km², ensuring that the dataset covers a broad geographic region and reflects the diverse characteristics of farmland. The images underwent a series of detailed pre-processing steps, including calibration and

cropping. During image calibration, we corrected geometric distortions caused by the shooting angle and Earth's curvature, ensuring spatial consistency across all images. In the cropping process, irrelevant areas were removed, focusing solely on extracting farmland regions. Additionally, to enhance the dataset's quality, we manually filtered out images affected by cloud or fog cover, stitching artifacts, or overall poor quality, ensuring only high-quality samples remained for analysis. In order to achieve an optimal balance between retaining the detailed features of high-resolution images and improving the efficiency of model training, this study adopted a standardized preprocessing process: all images that passed the quality screening were uniformly normalized, and a standardized cropping strategy of 512×512 pixels was applied. The size selection was based on the following two considerations. First, to preserve spatial resolution and detail, the 512×512 cropping unit can effectively balance the complete expression of local ground features (such as farmland boundaries and vegetation textures) and the efficient allocation of computing resources. Second, to preserve the integrity of spectral information, the cropped images strictly retain the three visible light bands—red, green, and blue—to ensure the effective transmission of spectral features in the model. This normalization processing scheme significantly improves the efficiency of batch data processing during model training by unifying the input data dimensions, while avoiding feature learning bias caused by image size differences. After completing these pre-processing steps, a total of 22,605 image samples were selected. These samples span various seasons, regions, cropping statuses, and feature diverse farmland distributions and shapes, ensuring the comprehensiveness and diversity of the dataset. This provides a rich and varied training dataset for the subsequent farmland segmentation.

2) Caption Construction

For the caption construction of each farmland sample, this study summarizes 11 key elements for describing farmland: shape, boundary morphology, shooting time, sowing conditions, the macro-level distribution of farmland, geographic location information, topographical features, landscape, the distribution of buildings, water bodies, and vegetation. The spatiotemporal characteristics of farmland result from the interaction of multiple factors. (Wang et al., 2022b)Temporally, the variations in crop growth stages lead to distinct visual texture differences in farmland across different seasons.(Zhu et al., 2022) Spatially, farmland exhibits significant spatial differentiation, with different regions affected by factors such as topography, terrain, and water-heat conditions, resulting in noticeable variations in farmland morphology and layout. (Pan and Zhang, 2022) Therefore, this study considers the issue at multiple spatial scales. At the macro-regional scale, typical farmland images were collected from various agricultural regions across China. These regions are not only located in different latitudes and longitudes, but also have different terrains and topography. For instance, farmland in the Northeast China Plain is flat and typically follows a concentrated distribution pattern with regular shapes, which is reflected in descriptions such as “the farmland primarily exhibits concentrated contiguous distribution” and “the shape of the farmland is characterized by blocky.” In contrast, the terrain of South China is predominantly hilly and mountainous, leading to a more dispersed farmland distribution and irregular shapes, which is described in the text as “with the farmland primarily in a dispersed distribution” and “the terrain is undulating.” Similarly, farmland in regions like the Loess Plateau and the arid and Semi-Arid Northern Areas often displays terraced or sloping patterns. At the same time, the spatial coupling relationships between farmland and surrounding features, such as water

bodies and buildings, are key factors influencing the distribution and accuracy of farmland identification. (Duan et al., 2022; Zheng et al., 2022) The relationship between the farmland and surrounding environmental features is expressed, for example, as "the water bodies surrounding the farmland mainly consist of scattered blocky ponds," and "the vegetation around the farmland mainly consists of scattered trees and scattered forests. " Similarly, the segmentation of farmland relies on boundary and texture information, the shape of the farmland and the boundary morphology, is also crucial for accurate identification of farmland. (Xie et al., 2023)

In summary, as shown in Fig. 4, this study categorizes farmland-related attributes into five major aspects: inherent properties, phenological characteristics, spatial distribution, topographic and geomorphic features and distribution of surrounding environments. The inherent properties include the shape of the farmland and the boundary patterns. Phenological characteristics encompass season and the sowing situation of the farmland. The spatial distribution of farmland not only reflects the geographic location information, but also includes the macro-level distribution of farmland in the image, such as concentrated contiguous distribution or dispersed distribution. Farmland shape is a very intuitive and important feature in visual interpretation, closely related to other factors such as terrain, topography, and landscape features, including blocky, striped, or broken. Farmland boundary pattern refers to the spatial shape characteristics of the farmland boundary, primarily manifested in whether its contour lines are relatively straight or exhibit a curved form.

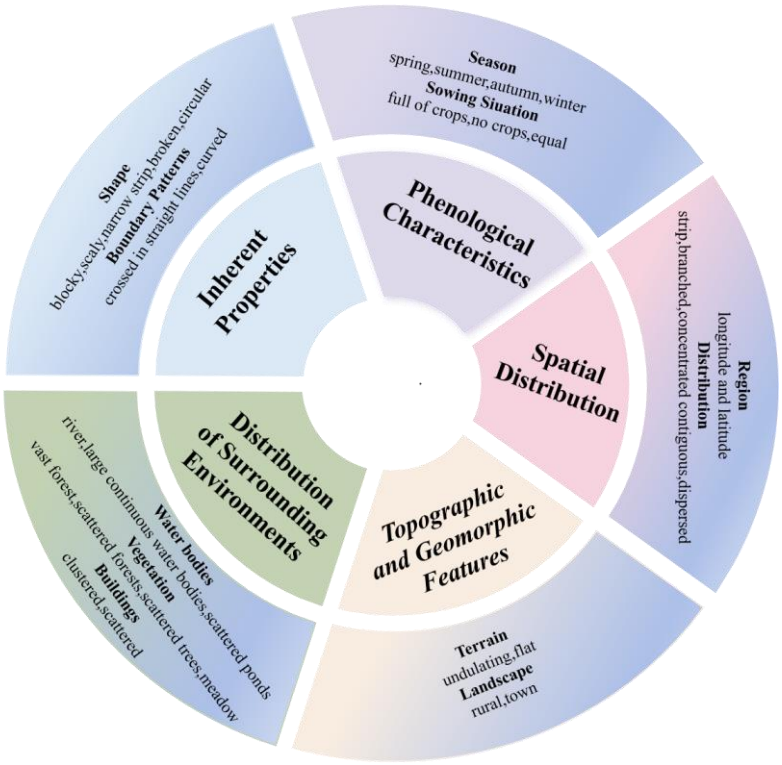


Fig. 4. Farmland description keywords.

3) Semi-Automated Annotation

Currently, there are two main approaches for constructing remote sensing image-text datasets: one involves automatically generating textual annotations using large language models, while the other relies on manual visual annotation by humans. However, both methods face significant challenges in meeting the high-precision requirements of farmland segmentation. Relying solely on automatic annotations generated by large language models has clear limitations. This approach often struggles to capture the nuanced and accurate correspondence between images and text. The granularity of captions is often insufficient, resulting in suboptimal accuracy and completeness in the annotation process. While manual annotation can ensure high-quality data, it has significant drawbacks. This approach requires domain experts to invest substantial time and effort, draining valuable resources and leading to extremely low efficiency. To address these challenges, this study proposes and develops a semi-automatic farmland image-text annotation framework. It is important to highlight that this semi-automatic annotation framework differs from previous methods. In addition to enabling text annotation, it also generates high-quality masks, offering more effective data support for farmland segmentation.

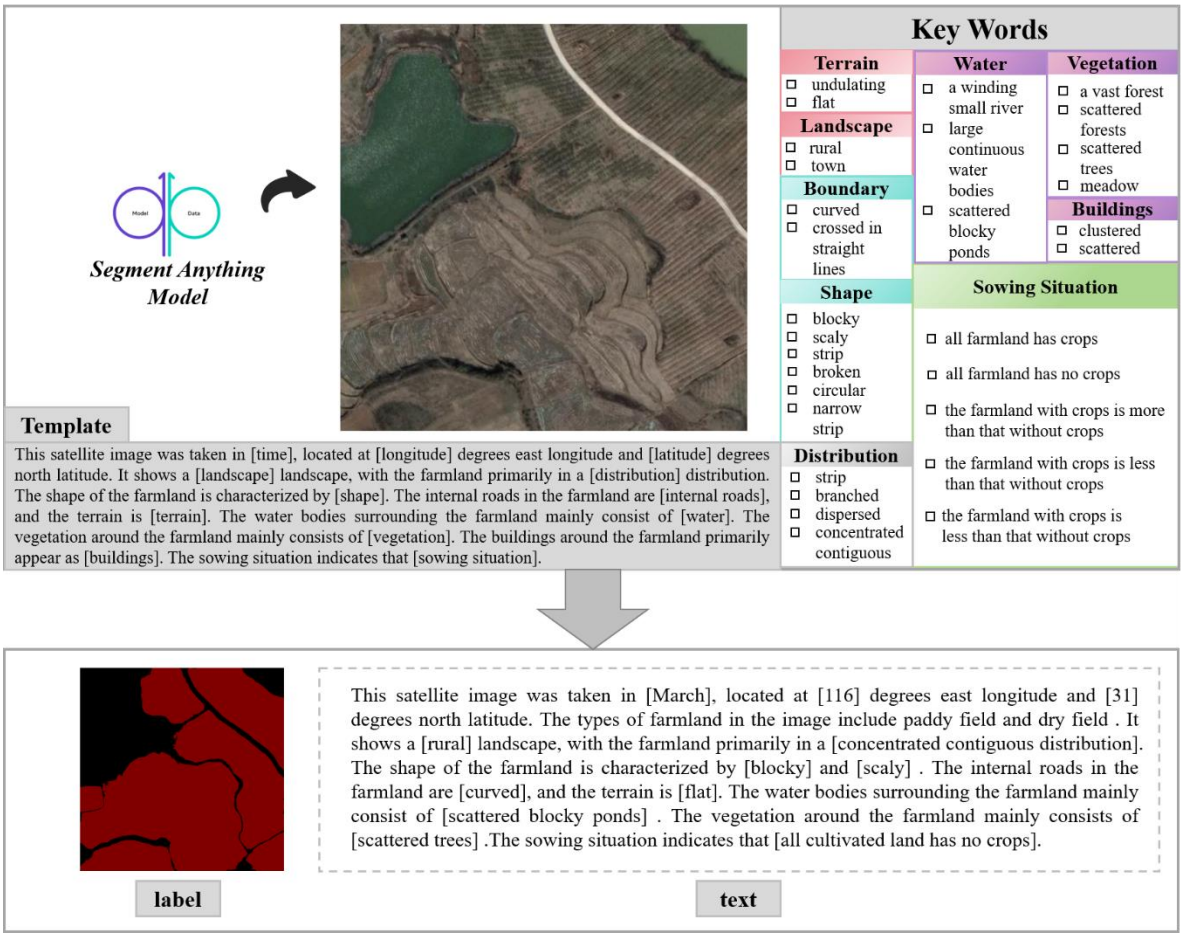


Fig. 5. Farmland semi-automated annotation framework.

The semi-automated annotation framework is illustrated in Fig. 5. Specifically, based on keywords related to farmland descriptions, this study first developed a set of farmland caption templates, providing a standardized reference for annotating image samples. To enable semi-automatic text annotation, this study integrated the constructed farmland caption templates and corresponding keywords into the open-source annotation software labelme. In this way, when annotating the remote sensing images of farmland, semi-automatic text annotation can be completed by visually observing the visual features of the remote sensing images and combining them with manually selected summarized keywords. In particular, the shooting month and longitude and latitude data of the farmland remote sensing images are automatically extracted from the original data. In addition, due to the limitation of cropping size, some images may not contain any land object categories other than farmland. Therefore, when annotating the surrounding environmental attributes using the semi-automated framework, this study requires that the presence of relevant land cover types be verified first, to ensure the accuracy of the captions. Finally, in order to quickly and accurately obtain high-quality farmland masks, this paper connects the Segment Anything Model (SAM) to labelme and performs semi-automatic mask annotation on the image to obtain the image label. Through semi-automatic annotation, humans only need to correct and verify part of the results, which significantly reduces the manpower and time costs compared to traditional fully manual annotation methods. At the same time, the semi-automated process combines the consistency of algorithms with the precision of manual verification, effectively minimizing subjective errors that can occur in manual annotation and thereby enhancing the accuracy and reliability of the labels.

3.2 The Spatiotemporal Characteristics Analysis of FarmSeg-VL Based on Multidimensional Statistics

FarmSeg-VL, as the first large-scale farmland image-text dataset covering multiple regions and seasons in China, is valuable for reflecting the dynamic characteristics of geographical zoning differences, crop growth cycle variations, and tillage practices. This section uses multidimensional statistical methods to analyze the ability of FarmSeg-VL to collaboratively represent spatial breadth and temporal continuity, providing a theoretical basis for evaluating its applicability in cross-regional and cross-seasonal farmland segmentation.

Fig. 6 reveals the spatiotemporal characteristics of FarmSeg-VL from both spatial and temporal perspectives. In terms of the spatial dimension, the sample distribution of agricultural areas in Fig.6(a) shows that FarmSeg-VL fully covers eight agricultural areas, ranging from the Northeast Plain to the Southwest Mountains. Notably, the sample count in the Yangtze River Middle and Lower Reaches Plain is significantly more than in other regions, accurately reflecting the geographical characteristics of the area, which is marked by a high degree of farmland fragmentation and notable terrain complexity. In terms of the temporal dimension, the seasonal distribution in Fig.6 (b) shows that samples in the northern agricultural regions are concentrated in summer and autumn, while the southern agricultural regions exhibit a more balanced distribution throughout the year. This pattern is closely aligned with the differences in crop growth cycles driven by latitude gradients in China. In addition, Fig.6 (c) and (d) illustrate the distribution patterns and shape characteristics of farmland across eight agricultural regions, highlighting the variations between them. Among these, the agricultural areas in the Yangtze River Middle and Lower Reaches Plain exhibit the greatest diversity, featuring four distinct distribution patterns and six different shape

characteristics of farmland. In the Northeast China Plain and the Huang-Huai-Hai Plain, farmland is primarily distributed in concentrated areas, with a predominantly blocky form. In other agricultural regions, there is a clear correlation between the distribution patterns and the shape characteristics of farmland. The diversity and richness of farmland samples across different agricultural regions fully reflect the spatiotemporal variability captured by FarmSeg-VL, underscoring its advantages in farmland segmentation.

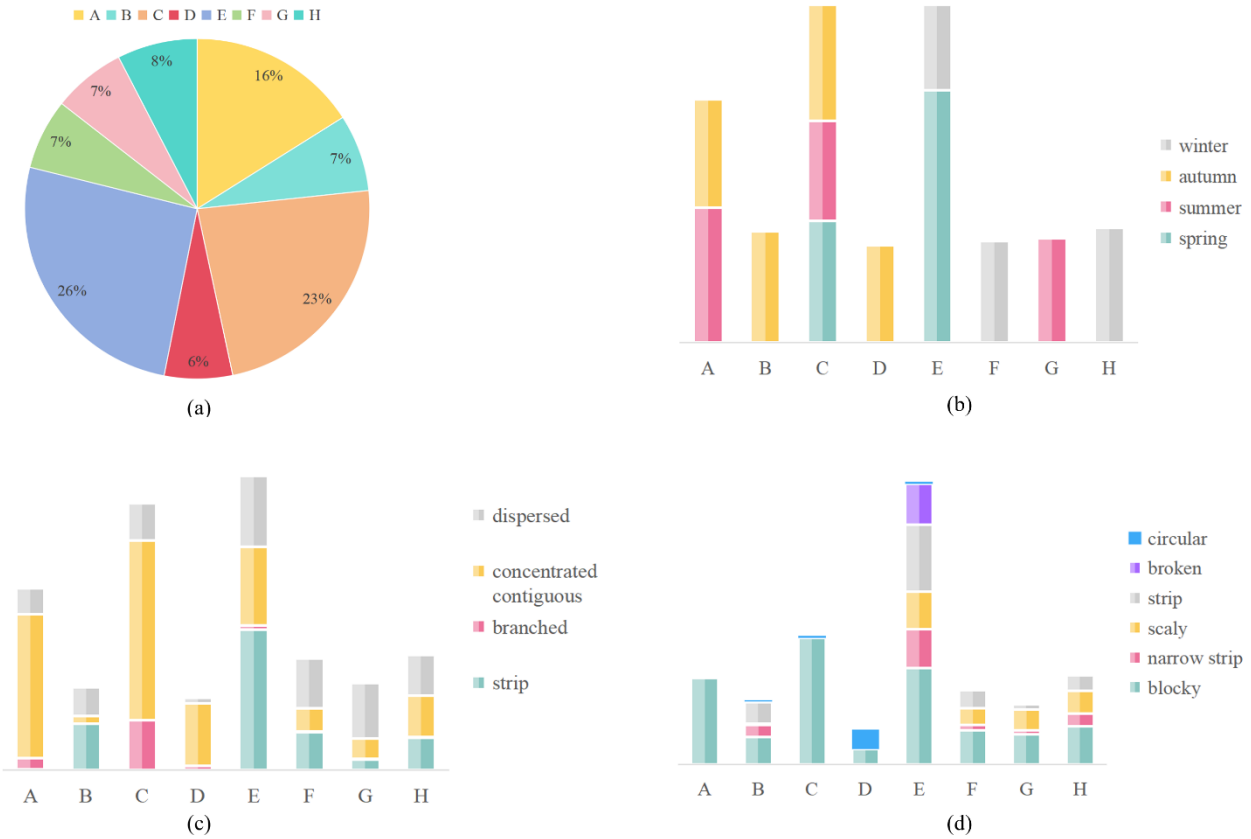


Fig. 6. Diversity of data samples. (a) Sample distribution ratio across different agricultural regions. (b) Sample distribution ratio for different seasons in each agricultural region. (c) Sample distribution ratio based on different farmland distribution patterns in each agricultural region. (d) Sample distribution ratio for different farmland shapes in each agricultural region. Where A represents the Northeast China Plain, B represents the Northern Arid and Semi-arid Region, C represents the Huang-Huai-Hai Plain, D represents the Loess Plateau, E represents the Yangtze River Middle and Lower Reaches Plain, F represents South China Areas, G represents the Sichuan Basin, and H represents the Yungui Plateau.

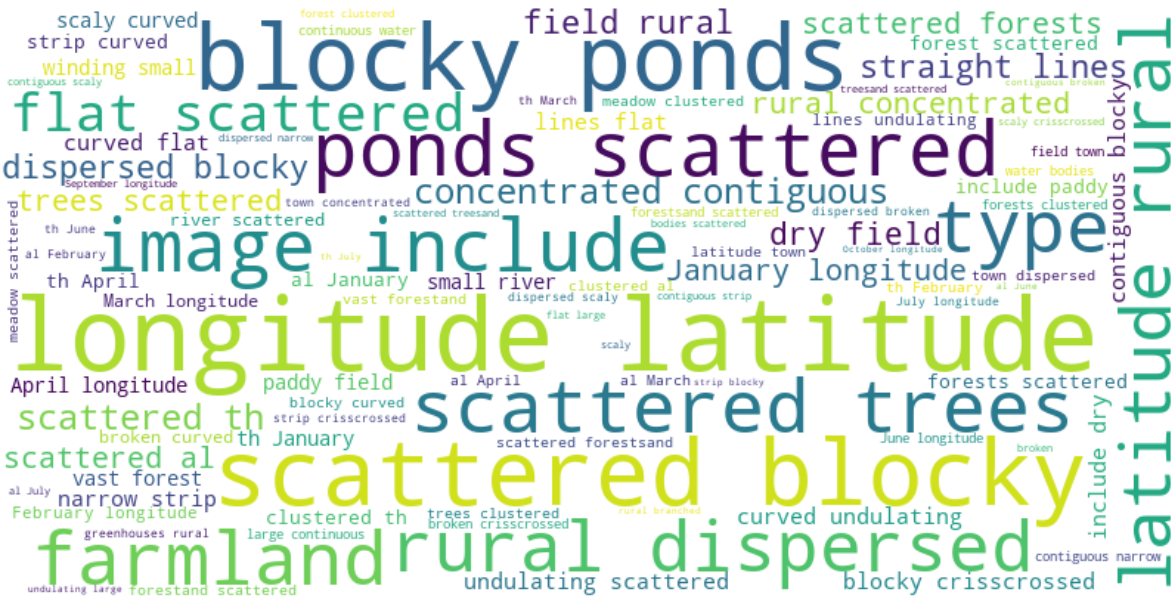


Fig. 7. Word cloud of farmland captions.

To further reveal the spatiotemporal characteristics of FarmSeg-VL, we extracted keywords from its caption and generated a word cloud of farmland-related attributes. As shown in Fig. 7, the spatiotemporal characteristics of FarmSeg-VL are further illustrated through the keyword cloud. High-frequency spatial attributes (e.g., " latitude" and " longitude") show strong semantic associations with temporal attributes (e.g., "January"), indicating that the captions in the FarmSeg-VL dataset effectively link temporal and spatial concepts. The spatial differentiation of morphological descriptors such as "concentrated contiguous" and "dispersed" aligns closely with the statistical results shown in Fig. 6(c) and Fig. 6(d), indicating that text annotations can effectively reflect and convey the geographical patterns of farmland morphology. Notably, the prominent presence of non-farmland attributes such as "ponds" and "forests" among the keywords suggests that FarmSeg-VL reflect not only the characteristics of farmland itself but also emphasizes the logical connections between farmland and its surrounding environment. In summary, the composite captions in FarmSeg-VL at both temporal and spatial levels not only reflect the fundamental characteristics of farmland but also reveal the external driving factors behind its spatiotemporal evolution.

3.3 Why is FarmSeg-VL More Suitable as a Dataset Benchmark for Farmland Segmentation?

Comprehensive spatiotemporal coverage with rich seasonal and regional diversity. The FarmSeg-VL offers extensive coverage across both temporal and spatial dimensions, spanning all four seasons—spring, summer, autumn, and winter—while also including eight typical agricultural regions of China. The dataset reflects the seasonal differences in agricultural landscapes, as well as the unique geographic features of each region, such as variations in farmland characteristics and surrounding environments. These factors enhance the diversity of the dataset.

Rich semantic captions capturing comprehensive farmland attributes. Unlike traditional datasets with simple image annotations, FarmSeg-VL incorporates detailed language captions summarizing the spatiotemporal features of farmland images. Specifically, it covers 11 key descriptive points, including farmland inherent properties, phenological characteristics, spatial distribution, topographic and geomorphic features, and the distribution of surroundings. The rich semantic captions significantly enhance the model's accuracy in farmland segmentation.

Comprehensive seasonal-regional coverage enhances model robustness. Seasonal and climatic variations significantly influence farmland morphology and distribution. Unlike traditional datasets, which typically focus on a single season and limit model adaptability, the FarmSeg-VL spans all four seasons, enabling models to better capture seasonal dynamics and varying crop growth conditions. Additionally, FarmSeg-VL covers diverse agricultural regions across China, reflecting distinct differences in farmland characteristics due to climate and geographic variation. The dataset's extensive seasonal and regional coverage enhances the model’s robustness, ensuring accurate and efficient farmland segmentation under diverse seasonal and climatic conditions.

4 Experiments

This chapter outlines the experimental setup in Section 4.1. Section 4.2 evaluates the effectiveness of the FarmSeg-VL for farmland segmentation by comparing a model fine-tuned on FarmSeg-VL with a vision language model (VLM) trained on a general image-text dataset. This comparison aims to verify whether a dedicated farmland image-text dataset can enhance model performance in farmland segmentation. In Section 4.3, we assess segmentation performance across different agricultural regions, comparing VLMs trained on FarmSeg-VL with the deep learning models that rely solely on labels, including U-Net, DeepLabV3, FCN, and SegFormer. We also analyze the generalization capability of models trained on FarmSeg-VL in diverse agricultural landscapes and their adaptability to spatiotemporal heterogeneity. Section 4.4 investigates the transferability of VLMs trained on FarmSeg-VL through comparative experiments with traditional models on public datasets, evaluating their cross-dataset generalization and cross-domain potential. Finally, Section 4.5 compares FarmSeg-VL with existing farmland datasets in the context of farmland segmentation applications.

4.1 Experimental Setup

Dataset Partitioning. To avoid the influence of sample similarity between the training, testing, and validation sets on the reliable evaluation of the model's generalization ability and domain transferability, this paper selects samples from different agricultural regions for each set. This approach helps reduce spatial homogeneity and ensures a more robust assessment of the model's performance. The dataset is divided into training, validation, and test sets in a 7:2:1 ratio. Specifically, the training set comprises 15,821 samples, the validation set contains 4,512 samples, and the test set includes 2,272 samples. The distribution of test set samples across different agricultural regions is as follows: 363 samples from the Northeast China Plain, 531 samples from the Huang-Huai-Hai Plain, 146 samples from the Northern Arid and Semi-Arid Region, 16 samples from the Loess

Plateau, 587 samples from the Yangtze River Middle and Lower Reaches Plain, 152 samples from South China, 156 samples from the Sichuan Basin, and 171 samples from the Yungui Plateau.

Evaluation Metrics. To assess model performance, this study uses four widely adopted metrics in farmland segmentation: Mean Accuracy (mACC), Mean Intersection over Union (mIoU), Mean Dice Coefficient (mDice), and Recall. Specifically, mACC represents the average pixel classification accuracy across all categories, while mIoU quantifies the mean ratio of intersection over union, a standard metric in semantic segmentation. mDice measures the similarity between predicted and ground-truth segmentation results, and Recall evaluates the proportion of correctly identified positive samples, reflecting the model's ability to capture relevant farmland regions.

4.2 Fine-Tuning General VLMs with FarmSeg-VL: Bridging Domain Gaps and Enhancing Semantic Comprehension for Farmland Segmentation

In order to verify the advantages of the model trained on FarmSeg-VL in farmland segmentation compared to models trained on general image-text datasets. This study systematically evaluates the impact of FarmSeg-VL based fine-tuning on farmland segmentation accuracy across three mainstream vision language segmentation models: LISA (Lai et al., 2023), PixelLM(Ren et al., 2024), and LaSagna(Wei et al., 2024). Among them, LISA is a model that integrates a large language model (LLM) with segmentation mask generation capabilities, enabling reasoning-driven segmentation based on complex textual prompts. LaSagnA extends LISA's architecture by adopting a unified sequence format to handle more complex queries while enhancing perceptual ability through the incorporation of semantic segmentation. This design demonstrates superior performance in processing intricate prompts and improving reasoning capability. PixelLM, in contrast, is a multimodal model specialized for pixel-level reasoning. It addresses the challenge of generating pixel-wise masks for multiple objects by introducing a lightweight pixel decoder and a segmentation codebook, which improves both efficiency and granularity in segmentation tasks.

The experimental results are shown in Table 3. It can be clearly seen that in farmland segmentation, after fine-tuning the model using the FarmSeg-VL, the performance of the model has been significantly improved, with an improvement of nearly 30% to 40%. Specifically, across all methods, the fine-tuned models consistently achieve higher mIoU scores compared to their non-fine-tuned counterparts, highlighting the effectiveness of FarmSeg-VL in improving segmentation accuracy. This result demonstrates that fine-tuning significantly enhances the model's ability to capture and accurately segment relevant features. Notably, the PixelLM model does not produce results in its non-fine-tuned state, as it has not been exposed to farmland-related semantic information during pretraining and is therefore incapable of generating effective predictions without fine-tuning. However, after being trained on the FarmSeg-VL, PixelLM becomes capable of accurately predicting farmland, with performance approaching that of the other two VLMs. This further underscores the importance of fine-tuning with a domain-dedicated dataset to enhance model performance for specialized tasks. To more intuitively analyze the experimental results, this study visualized the segmentation outcomes. As shown in Fig. 8, models that have not undergone fine-tuning tend to misclassify large areas of buildings and forests as farmland. This suggests that non-fine-tuned models struggle to accurately

capture inherent properties of farmland, leading to high uncertainty and significant errors in segmentation results, as well as a lack of stability and consistency.

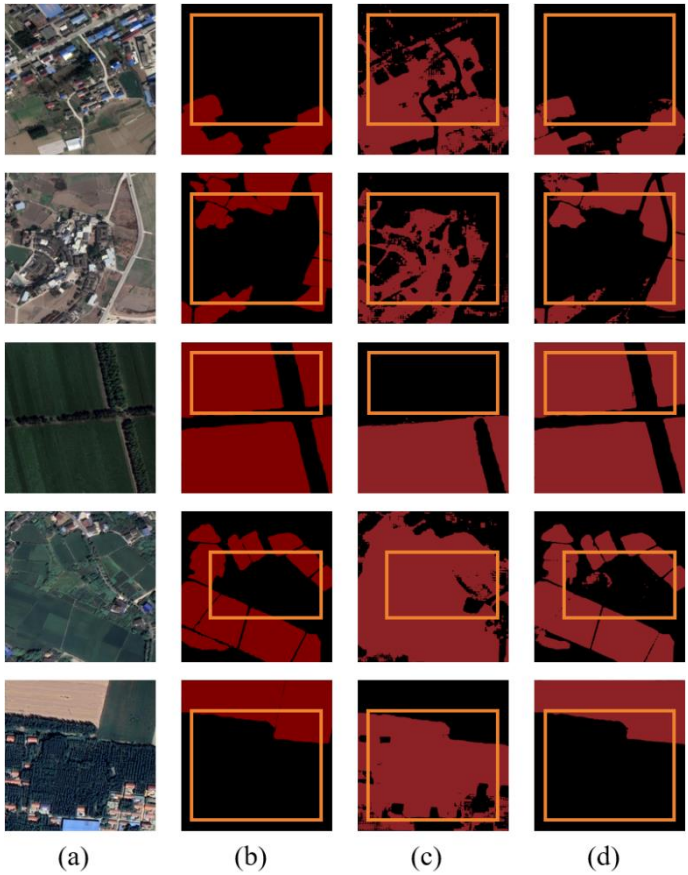


Fig. 8. Visualization of partial experimental results fine-tuned on the FarmSeg-VL Dataset. (a) Original image. (b) Ground truth. (c) Test results without fine-tuning. (d) Test results after fine-tuning.

Table 3. Comparison of fine-tuning results on the FarmSeg-VL dataset.

Method	No Fine Tuning(%)				Fine Tuning(%)			
	mIoU	mACC	mDice	Recall	mIoU	mACC	mDice	Recall
LISA	46.50	58.42	58.39	58.76	87.71	93.47	93.45	93.46
PixelLM	/	/	/	/	83.65	91.13	91.09	91.16
LaSagna	32.31	52.00	47.16	56.51	86.95	93.03	93.02	93.00

In summary, the FarmSeg-VL offers more precise domain-dedicated knowledge for farmland segmentation, allowing models to better capture fine-grained features of farmland. Specifically, FarmSeg-VL contains high-quality farmland annotations that cover multiple semantic dimensions, such as farmland shape, distribution, and sowing situation. This comprehensive information significantly improves the model's ability to understand and segment farmland features with

greater accuracy. Compared to general datasets, FarmSeg-VL effectively reduces cross-domain discrepancies, allowing the model to focus on farmland features, thereby further enhancing the accuracy of farmland segmentation.

4.3 Comparing Model Performance Trained on FarmSeg-VL in Different Agricultural Regions

To explore the application effect of models trained on the FarmSeg-VL in different agricultural regions, this section divides the test set into various agricultural regions, including the Northeast China Plain, Huang-Huai-Hai Plain, Northern Arid and Semi-Arid Region, Loess Plateau, Yangtze River Middle and Lower Reaches Plain, South China, Sichuan Basin, and Yungui Plateau. These regions will be tested using both vision-language models (PixelLM, LaSagna, LISA) and the deep learning models that rely solely on labels (U-Net, DeepLabV3, FCN, SegFormer). Notably, these models that rely solely on labels do not incorporate any language modality, they are trained and tested exclusively using original farmland image and ground truth.

Figure C1 to C8 display the testing accuracy of the model in different agricultural regions. From the overall results, both the deep learning models that rely solely on labels and VLMs demonstrated strong testing accuracy in the agricultural regions of the Northeast China Plain and the Huang-Huai-Hai Plain. However, in the remaining six agricultural regions, the performance differences between the two model types became more pronounced. The primary reason for these differences lies in the varying complexity of the spatial structure of farmland across different agricultural regions. In the Northeast China Plain and Huang-Huai-Hai Plain, the terrain is relatively flat, and the farmland is distributed in a more regular and contiguous manner. As a result, both models exhibit strong segmentation performance in these relatively simple scenarios. In other agricultural regions, particularly in South China Areas, the farmland generally exhibits scattered and fragmented characteristics. Additionally, it shares a high degree of textural similarity with surrounding non-farmland features, such as forests and water bodies, which makes it difficult for the model to segment farmland. By incorporating language, VLMs can effectively comprehend the spatial distribution of farmland and its surrounding environment, thereby alleviating the segmentation challenges caused by spatial differentiation and demonstrating advantages in these different agricultural regions.

To visually illustrate the performance differences among various models in farmland segmentation tasks, Fig. C9 to C16 present the segmentation results for each agricultural region. From this, it can be observed that in agricultural regions such as the Northeast China Plain and the Huang-Huai-Hai Plain, although the overall accuracy is high, the deep learning models that rely solely on labels still exhibit certain limitations. For example, this type of model is prone to misjudgment when encountering terrain features that resemble farmland, such as ponds and grasslands, and often exhibits issues such as boundary blurring and discontinuity in the segmentation of farmland. In South China Areas, the highly fragmented nature of farmland, with its scattered or narrow distribution, the segmentation challenge is further acerbated. The deep learning models that rely solely on labels struggle to effectively identify such atypical farmland, leading to a significant decrease in segmentation accuracy. In contrast, VLMs have demonstrated notable advantages in the aforementioned agricultural regions. By incorporating farmland-related key words—such as “concentrated buildings” and “narrow strips”, VLMs enhance their comprehension of both the inherent properties of farmland and the contextual information of its surrounding environment. This enriched understanding contributes to improved completeness and accuracy in farmland segmentation. In addition, this

advantage is not limited to the aforementioned agricultural regions but is also consistently performance in the segmentation results in the other five regions. This further validates the generalization capability and robustness of the VLMs in diverse agricultural landscapes.

In summary, compared to the deep learning models that rely solely on labels, VLMs that incorporate caption demonstrate significant advantages in farmland segmentation across all agricultural regions. Language information effectively compensates for the limitations of the deep learning models that rely solely on labels in complex scenarios, enhancing the model's understanding of farmland morphology and the relationship between farmland and surrounding land cover, thereby significantly improving farmland segmentation accuracy.

4.4 Cross-Domain Performance Evaluation of Models Trained on FarmSeg-VL

In order to evaluate the performance of models trained on the FarmSeg-VL dataset in cross-domain tasks, this paper conducted relevant experiments. Specifically, this section presents transfer tests using VLMs (PixelLM, LaSagna, LISA) and the deep learning models that rely solely on labels (U-Net, DeepLabV3, FCN, SegFormer) trained on the FarmSeg-VL across multiple public datasets. The test datasets include DeepGlobe Land Cover (DGLC), LoveDA, and the Fine-Grained Farmland Dataset (FGFD). Specifically, the DGLC dataset covers regions in Thailand, Indonesia, and India, while the LoveDA includes areas in Nanjing, Changzhou, and Wuhan in China. The FGFD farmland dataset encompasses regions such as Heilongjiang, Hebei, Shaanxi, Guizhou, Hubei, Jiangxi, and Tibet in China. The specific details are provided in Table 1. Specifically, to maintain consistency with the FarmSeg-VL test set and ensure the data is more suitable for the model, we performed data preprocessing on the DGLC and LoveDA. This preprocessing primarily involved cropping the images to a size of 512×512 and merging non-farmland pixel labels, among other steps.

Table 4. Farmland segmentation results of different methods on FGFD.

Evaluation	The deep learning models that rely solely on labels				Vision-Language Model		
Metrics(%)	U-Net	Deeplabv3	FCN	SegFormer	PixelLM	LaSagna	LISA
mACC	72.38	74.76	76.52	76.40	78.59	80.70	**83.33**
mIoU	57.48	60.11	62.43	62.34	64.68	66.83	**70.58**
mDice	72.71	74.94	76.74	76.66	78.55	80.00	**82.65**
Recall	72.38	74.76	76.52	76.40	78.98	80.84	**83.87**

Table 5. Farmland segmentation results of different methods on LoveDA.

Evaluation	The deep learning models that rely solely on labels				Vision-Language Model		
Metrics(%)	U-Net	Deeplabv3	FCN	SegFormer	PixelLM	LaSagna	LISA
mACC	70.83	77.05	73.65	73.78	78.79	80.45	**81.76**
mIoU	47.77	63.85	61.41	60.57	60.75	64.03	**65.74**
mDice	64.65	77.47	75.22	74.73	74.78	77.54	**78.82**
Recall	70.83	77.05	73.65	73.78	77.73	78.87	**80.75**

Table 6. Farmland segmentation results of different methods on DGLC.

Evaluation Metrics(%)	The deep learning models that rely solely on labels				Vision-Language Model		
	U-Net	Deeplabv3	FCN	SegFormer	PixelLM	LaSagna	LISA
mACC	64.60	71.73	69.10	70.32	66.13	71.69	**72.23**
mIoU	48.73	55.68	50.17	52.15	49.38	55.78	**56.36**
mDice	64.71	71.41	66.81	68.55	66.11	71.59	**72.06**
Recall	64.60	71.73	69.10	70.32	69.14	72.22	**72.44**

Tables 4-6 present the experimental results on the FGFD, LoveDA, and DGLC, respectively. Overall, both the deep learning models that rely solely on labels and VLMs exhibit strong cross-domain transfer transferability. This can be attributed to the FarmSeg-VL dataset's broad geographic coverage and diverse seasonal variations, which provide a solid foundation for cross-domain feature learning. Notably, VLMs demonstrate significantly superior cross-domain transfer performance across all three datasets compared to traditional labeled data-dependent deep learning models. This advantage is primarily attributed to the fine-grained captions provided by FarmSeg-VL, which inject transferable semantic prior knowledge into the VLMs. For instance, when caption prompts such as "strip-shaped farmlands in spring" are provided, the models autonomously correlate farmland shape characteristics across different regions under spring conditions. This integration of semantic priors enables VLMs to overcome the representational limitations inherent in single-modality visual features, thereby maintaining enhanced discriminative capabilities in cross-domain scenarios.

Through the cross-domain experiments, this study has drawn two key conclusions: Firstly, models trained on the FarmSeg-VL exhibit significant cross-domain transferability, fully demonstrating the improvement of model generalization performance by the FarmSeg-VL. Secondly, the introduction of captions breaks through the limitations of the deep learning models that rely solely on labels, enabling the model to decouple spatiotemporal heterogeneity interference and effectively improve segmentation accuracy in complex farming scenes.

4.5 Enhanced Model Transferability: Comparative Analysis of FarmSeg-VL and Conventional Farmland Datasets

To verify that the model trained on the FarmSeg-VL outperforms models trained on existing farmland datasets in both segmentation accuracy and generalization, we conducted extensive comparative experiments in this section. First, to ensure the reliability of the experimental results, this study uses the latest dedicated dataset, FGFD, as a benchmark for comparison. Since most existing farmland datasets follow the traditional "Image + Label" format (i.e., a paradigm that solely relies on labeled data), four commonly used the deep learning models that rely solely on labels—U-Net, Deeplabv3, FCN, and SegFormer—are selected to train on the FGFD dataset. For the proposed FarmSeg-VL dataset, three VLMs are selected for comparative experiments. Additionally, to ensure fairness, all trained models are uniformly tested on the LoveDA dataset.

The experimental results, shown in Table 7, reveal that models trained on the FarmSeg-VL dataset using VLMs outperform those trained on the FGFD dataset with the deep learning models that rely solely on labels when tested on the LoveDA dataset. Specifically, the mIoU improved by 10% to 40%, and the mAcc increased by 10% to 30%. This gap indicates that models

trained on the FarmSeg-VL dataset with added language modality have significant transferability in farmland segmentation compared to models trained on the traditional dataset FGFD. Moreover, FarmSeg-VL reflects multiple aspects of farmland characteristics through captions—such as phenological characteristics, spatial distribution, topographic and geomorphic features and distribution of surrounding environments—allowing the model to learn rich and comprehensive information about farmland. With these detailed captions of farmland, models trained on the FarmSeg-VL not only improve the accuracy of farmland segmentation but also enhance the model's ability to handle complex scenes. In summary, the FarmSeg-VL is a large-scale, high-quality image-text dataset of farmland, it has demonstrated great potential in cross-scenario farmland segmentation and provides a strong data foundation for future research in farmland segmentation.

Table 7. Performance of different datasets and methods on the LoveDA dataset.

Evaluation	FGFD				FarmSeg-VL		
Metrics(%)	U-Net	Deeplabv3	FCN	SegFormer	PixelLM	LaSagna	LISA
mACC	63.80	57.93	59.19	67.48	78.79	80.45	**81.76**
mIoU	38.15	29.78	36.62	50.08	60.75	64.03	**65.74**
mDice	55.17	45.33	53.61	66.29	74.78	77.54	**78.82**
Recall	63.80	57.93	59.19	67.48	77.73	78.87	**80.75**

5 Data availability

The FarmSeg-VL dataset is accessible on the Zenodo data repository at https://doi.org/10.5281/zenodo.15860191(Tao et al., 2025a).The FarmSeg-VL dataset consists of image data, labels, and corresponding farmland text descriptions in JSON files.

6 Conclusion

This study constructs FarmSeg-VL, a high-quality image-text dataset specifically designed for farmland segmentation, with key features including high-precision images and masks, extensive spatiotemporal coverage, and refined captions of farmland characteristics. In the dataset construction process, Google imagery with a resolution of 0.5-2 meters was selected as the image data source. Through in-depth analysis of numerous farmland samples, five key attributes were summarized: inherent properties, phenological characteristics, spatial distribution, topographic and geomorphic features and distribution of surrounding environments. These were further refined into 11 specific descriptive dimensions, covering shape, boundary patterns, season, sowing situation, geographic location, distribution, terrain, landscape features, as well as the distribution of water bodies, buildings, and trees in the surrounding environment. Based on the above keywords, a farmland description template was designed, and a semi-automated annotation method was used to generate binary mask labels and their corresponding captions for each image. Ultimately, a dedicated dataset consisting of 22,605 image-text pairs was constructed. To verify the advantages of the FarmSeg-VL in enhancing farmland segmentation accuracy compared to general image-text

datasets, this study first conducted fine-tuning experiments on three leading vision language segmentation models: LISA, PixelLM, and LaSagna. The experimental results demonstrate that the model fine-tuned with the FarmSeg-VL significantly outperforms the model trained with general image-text datasets in segmentation performance. Additionally, this study compared the VLMs trained on the FarmSeg-VL to a traditional deep learning model that relies solely on labels. The results show a 10% to 20% improvement in segmentation accuracy across different agricultural regions and datasets, highlighting that language guidance effectively mitigates the impact of spatiotemporal heterogeneity on farmland segmentation. Finally, the study compared the performance of the traditional deep learning model relies solely on labels trained on the FGFD dataset with the models trained using three VLMs on the FarmSeg-VL dataset. The evaluation on the LoveDA dataset showed an improvement in test accuracy by approximately 15%. Experimental results show that the model trained on FarmSeg-VL significantly improves accuracy and robustness in farmland segmentation. As the first large-scale image-text dataset for farmland segmentation, FarmSeg-VL holds significant academic value and application potential. It is expected to advance research on semantic understanding of farmland in remote sensing imagery, promote the development of more efficient and generalized segmentation models, and better serve the diverse needs of agricultural monitoring.

Appendices

Appendix A More details of farmland texture description in image-text dataset

As shown in Fig. A1, mainstream remote sensing image-text datasets, such as UCM-Captions, NWPU-Captions, RSICD, RSICap, and ChatEarthNet, generally adopt scene-level or object-level descriptions. These datasets often lack detailed characterization of farmland morphology, temporal features, and environmental context, making them insufficient for farmland segmentation tasks that require high-level semantic and structurally rich textual information.

For example, UCM-Captions provides only simple and repetitive descriptions like “There is a piece of farmland,” without any specific texture or spatial information. NWPU-Captions offers slight improvements by adding color and shape descriptions, such as “Many dark green circular fields are mixed with yellow rectangular fields,” but still does not include background context or agricultural semantics. RSICD focuses only on aggregated forms or land cover components, with descriptions like “The little farm is made up of grass and crops,” lacking both temporal cues and environmental context. RSICap provides relatively richer descriptions, for example, “In the image, there are many buildings and some farmlands located near a river,” which reflects spatial relationships between farmland and buildings or water bodies. However, these descriptions are mostly static and fail to capture the dynamic properties of farmland over time. ChatEarthNet, designed primarily for land cover classification, presents slightly more complex descriptions such as “This image shows a balance between crop and grass areas.”, but still lacks detailed information about farmland morphology, terrain, crop types, or surrounding environmental elements. In contrast, the proposed FarmSeg-VL dataset is specifically designed for the farmland segmentation task, placing greater emphasis on fine-grained semantic information closely tied to the spatiotemporal characteristics of farmland. For each remote sensing image, the accompanying textual description includes the image capture time, geographic coordinates, and detailed

references to landform, shape, boundary characteristics, topography, as well as surrounding features such as water bodies, vegetation, and buildings. Additionally, the descriptions incorporate attributes such as cropping patterns and spatial layouts, providing comprehensive semantic support for accurate and context-aware farmland segmentation.

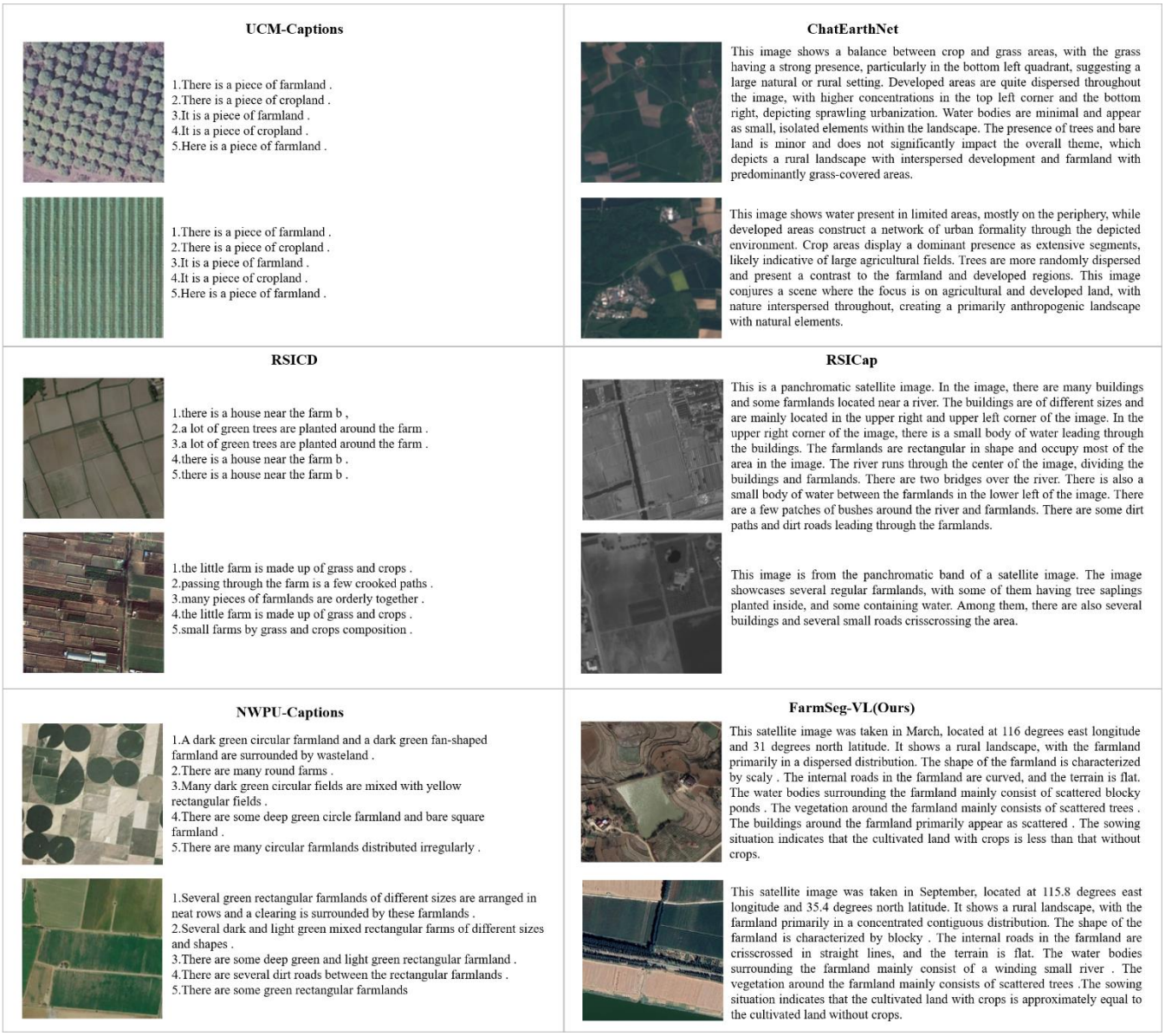


Fig. A1. Details of farmland texture description in general remote sensing image-text dataset.

Appendix B Examples of 5 types of text features for farmland shapes

To provide readers with a more intuitive understanding of the farmland morphology in the FarmSeg-VL dataset, we present five additional examples of farmland shapes in Fig. B1.

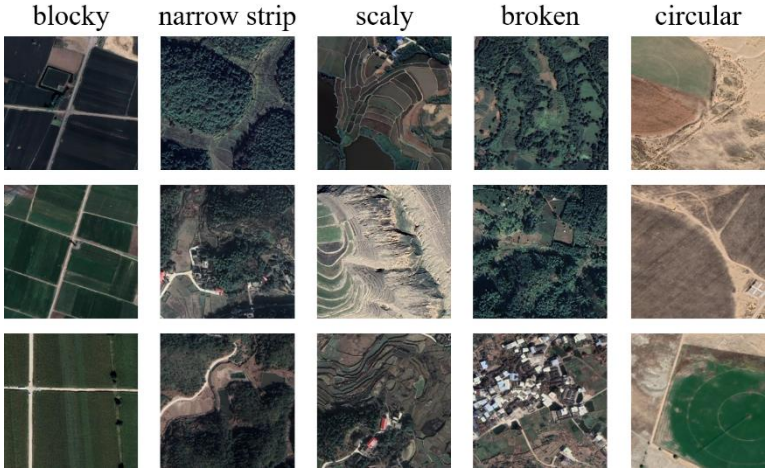


Fig. B1. Example of farmland shape.

Appendix C Farmland segmentation results of different methods in different agricultural areas

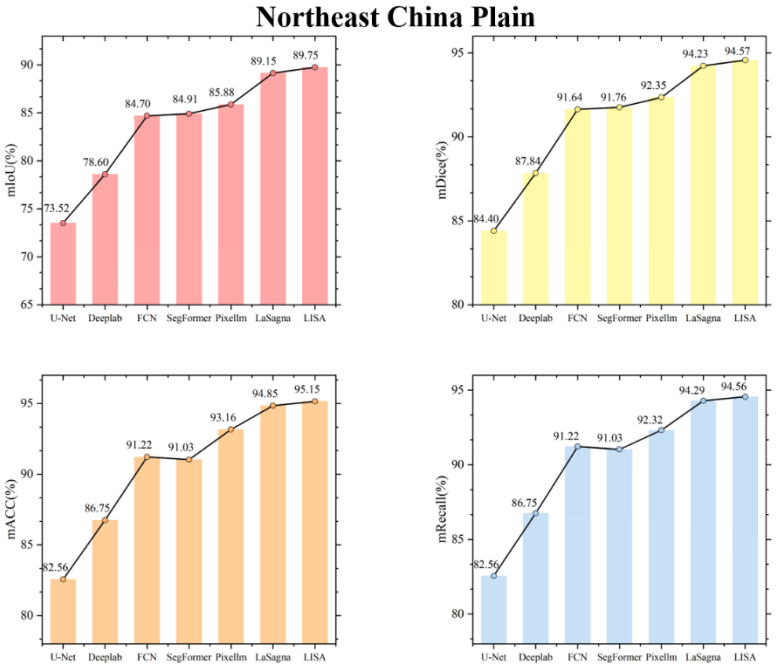


Fig. C1. Farmland segmentation results of different methods in Northeast China Plain.

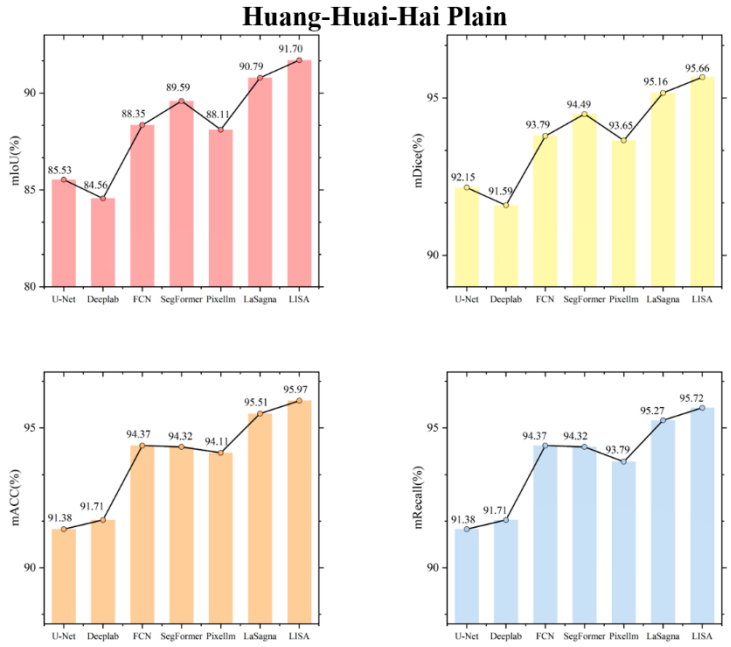


Fig. C2. Farmland segmentation results of different methods in Huang-Huai-Hai Plain.

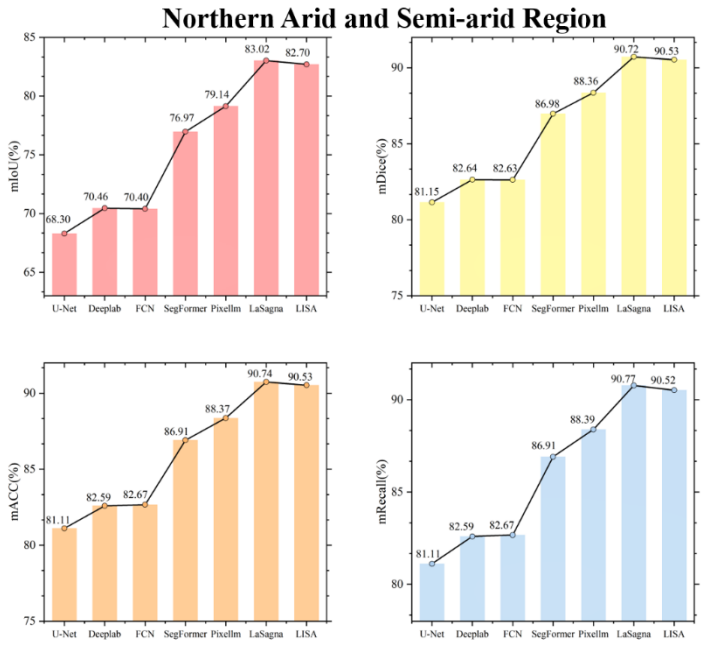


Fig. C3. Farmland segmentation results of different methods in Northern Arid and Semi-arid Region.

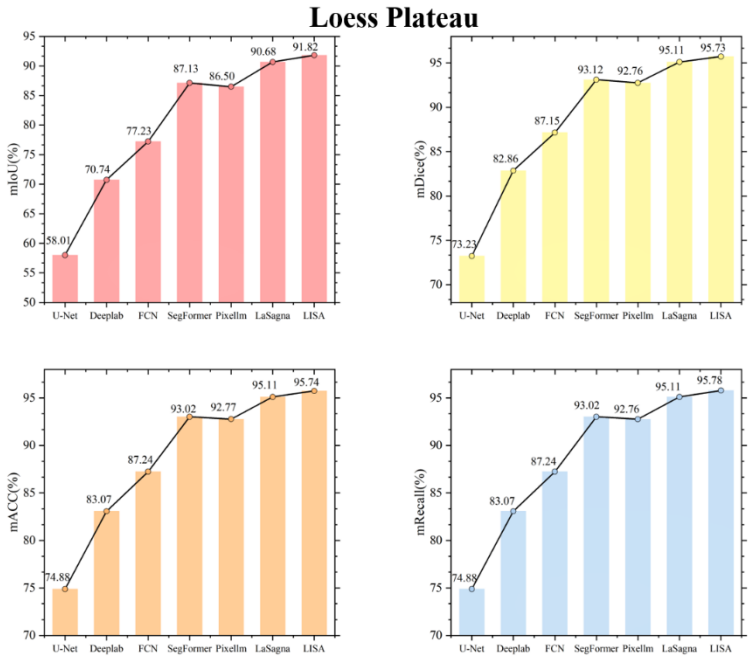


Fig. C4. Farmland segmentation results of different methods in Loess Plateau.

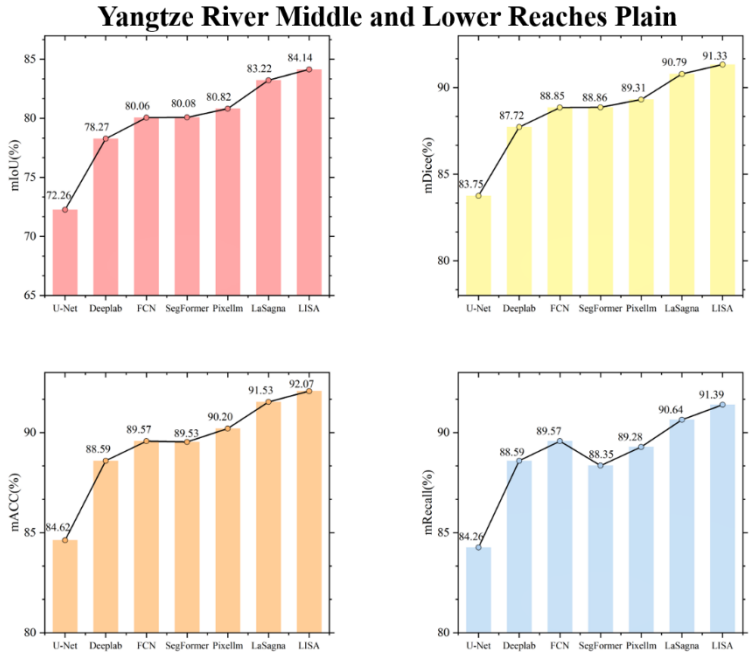


Fig. C5. Farmland segmentation results of different methods in Yangtze River Middle and Lower Reaches Plain.

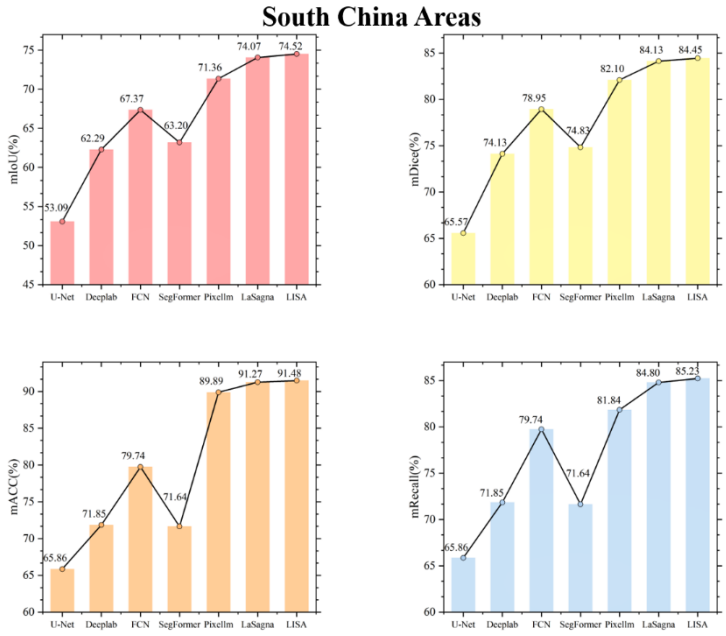


Fig. C6. Farmland segmentation results of different methods in South China Areas.

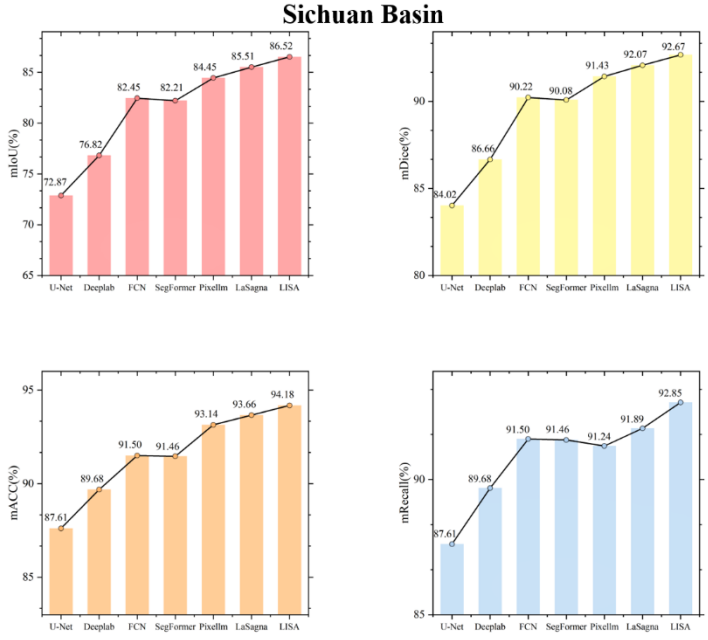


Fig. C7. Farmland segmentation results of different methods in Sichuan Basin.

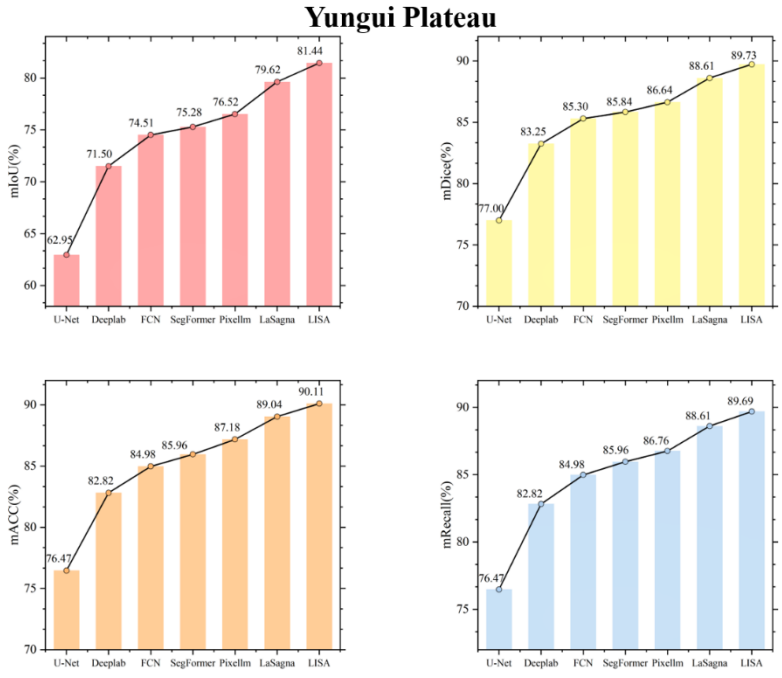


Fig. C8. Farmland segmentation results of different methods in Yungui Plateau.

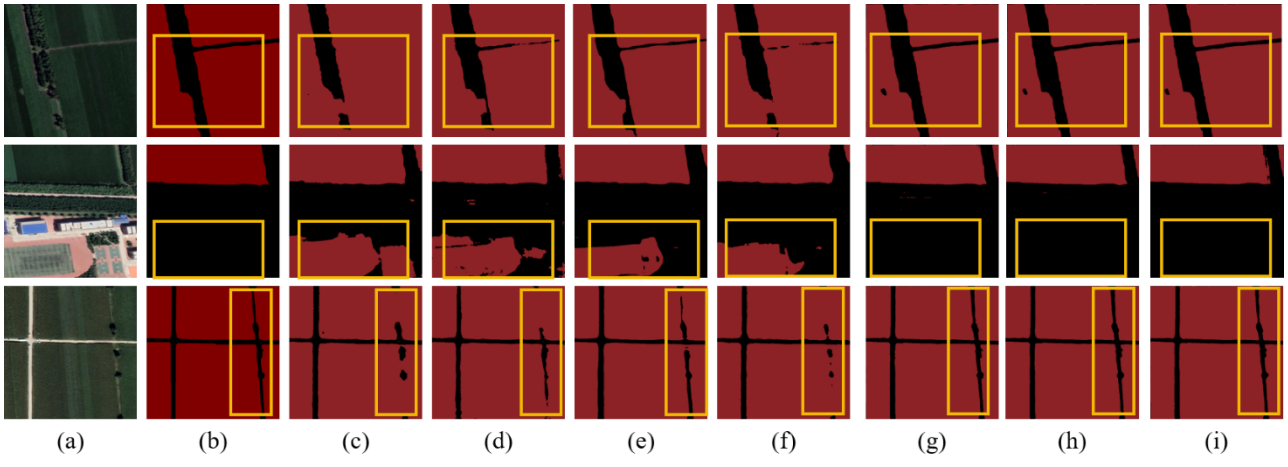


Fig. C9. Farmland segmentation results of different methods in Northeast China Plain. (a)Original image. (b)groundtruth. (c)U-Net. (d)Deeplabv3. (e)FCN. (f)SegFormer. (g)PixelLM. (h)LaSagna. and (i)LISA.

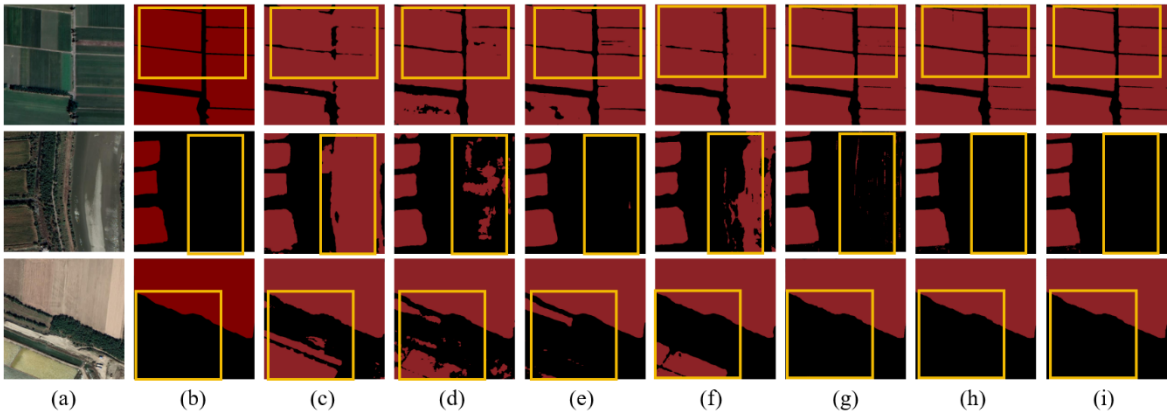


Fig. C10. Farmland segmentation results of different methods in Huang-Huai-Hai Plain . (a)Original image. (b)groundtruth. (c)U-Net. (d)Deeplabv3. (e)FCN. (f)SegFormer. (g)PixelLM. (h)LaSagna. and (i)LISA.

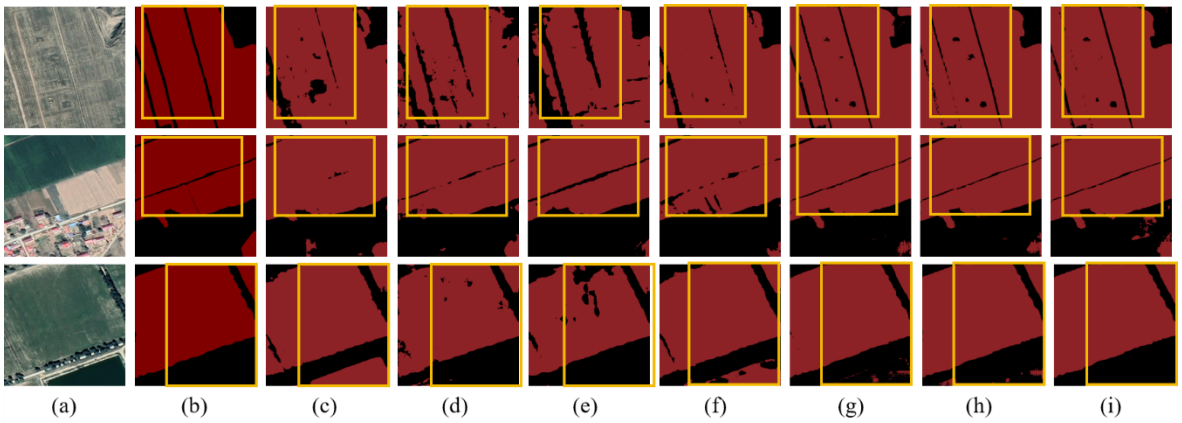


Fig. C11. Farmland segmentation results of different methods in Northern Arid and Semi-arid Region . (a)Original image. (b)groundtruth. (c)U-Net. (d)Deeplabv3. (e)FCN. (f)SegFormer. (g)PixelLM. (h)LaSagna. and (i)LISA.

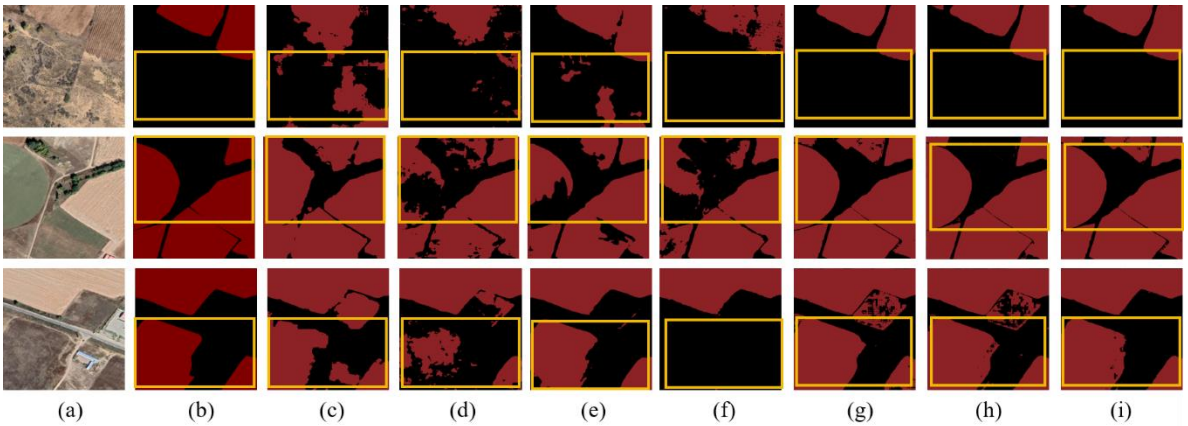


Fig. C12. Farmland segmentation results of different methods in Loess Plateau . (a)Original image. (b)groundtruth. (c)U-Net. (d)Deeplabv3. (e)FCN. (f)SegFormer. (g)PixelLM. (h)LaSagna. and (i)LISA.

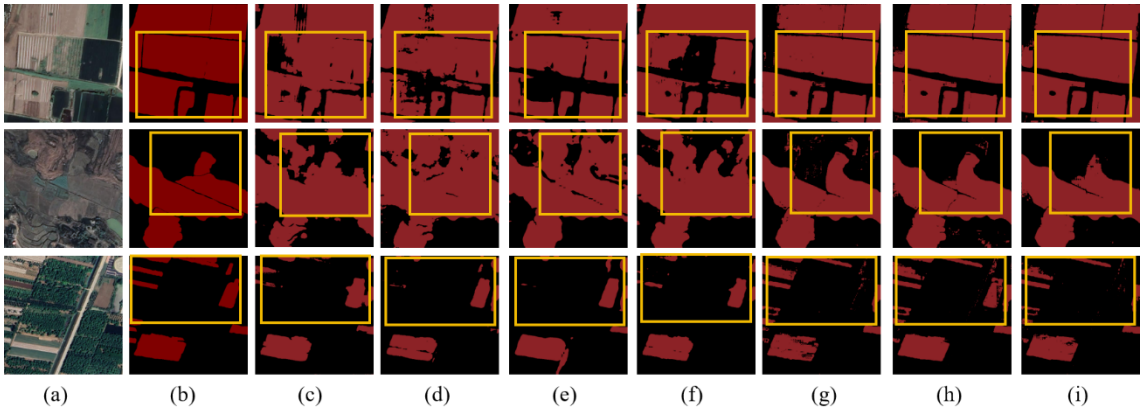


Fig. C13. Farmland segmentation results of different methods in Yangtze River Middle and Lower Reaches Plain . (a)Original image. (b)groundtruth. (c)U-Net. (d)Deeplabv3. (e)FCN. (f)SegFormer. (g)PixelLM. (h)LaSagna. and (i)LISA.

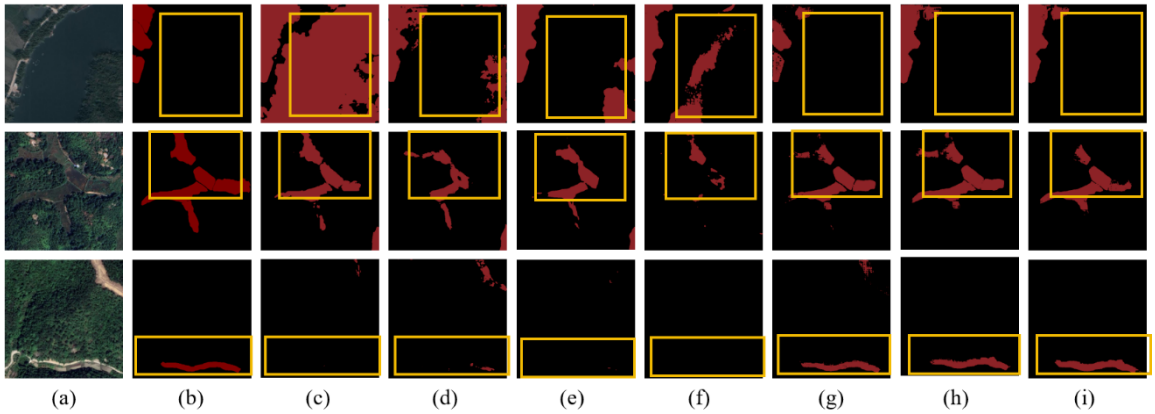


Fig. C14. Farmland segmentation results of different methods in South China Areas . (a)Original image. (b)groundtruth. (c)U-Net. (d)Deeplabv3. (e)FCN. (f)SegFormer. (g)PixelLM. (h)LaSagna. and (i)LISA.

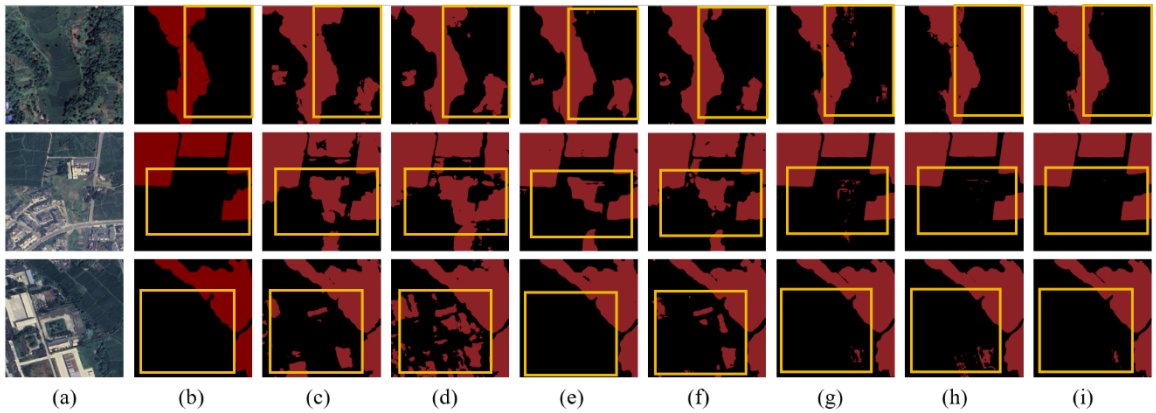


Fig. C15. Farmland segmentation results of different methods in Sichuan Basin . (a)Original image. (b)groundtruth. (c)U-Net. (d)Deeplabv3. (e)FCN. (f)SegFormer. (g)PixelLM. (h)LaSagna. and (i)LISA.

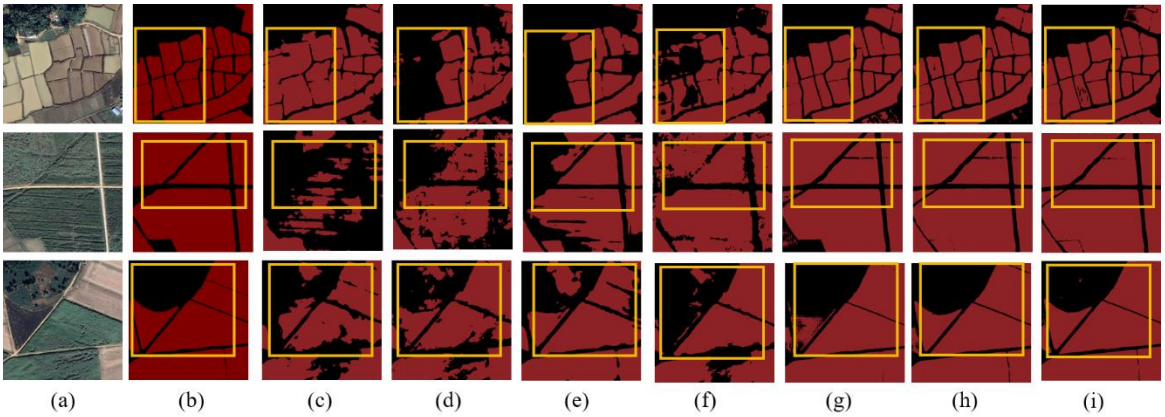


Fig. C16. Farmland segmentation results of different methods in Yungui Plateau . (a)Original image. (b)groundtruth. (c)U-Net. (d)Deeplabv3. (e)FCN. (f)SegFormer. (g)PixelLM. (h)LaSagna. and (i)LISA.

Appendix D Quantitative evaluation of semi-automated annotation efficiency

In order to quantify the annotation efficiency of the semi-automatic annotation framework proposed in this article, comparative experiments were conducted in this section. Specifically, we randomly selected four annotators and annotated the masks and texts on 13 farmland remote sensing images using traditional manual drawing methods and semi-automated annotation methods. Finally, we compared the completion time of the annotations. As shown in Fig. D1, after using the semi-automated annotation method, the average annotation time was significantly reduced, saving approximately 2 minutes per image, and overall efficiency improved by 1.5 times. This result indicates that the annotation tool developed in this article has significantly improved efficiency and usability.

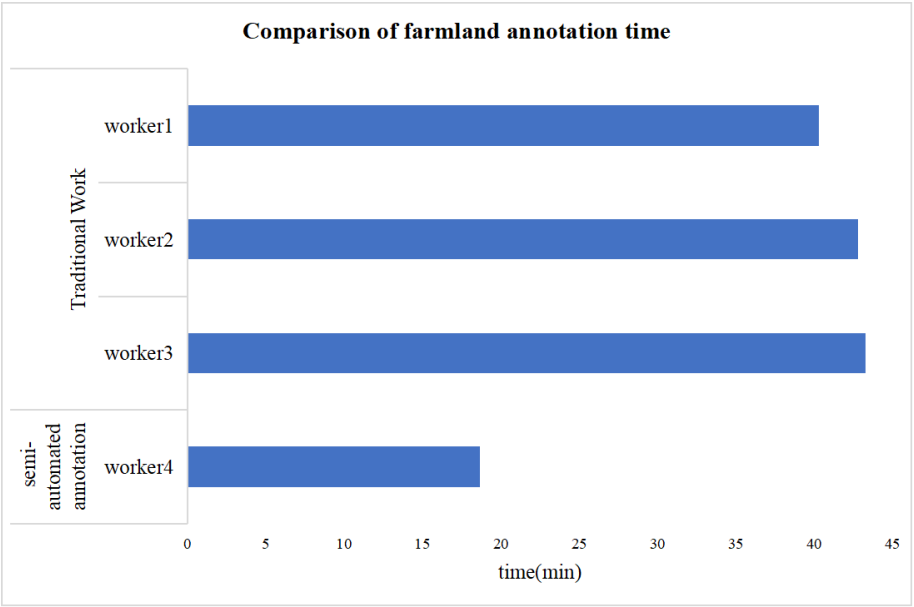


Fig. D1. Comparison of farmland annotation efficiency.

Appendix E Cross-Regional Applicability Assessment of FarmSeg-VL.

To verify the generalization performance of the model trained using FarmSeg-VL on datasets from other countries that have significant differences in climate or cropping patterns compared to FarmSeg-VL, this paper selects a portion of the region in Nordrhein-Westfalen, Germany as the benchmark for testing, test experiments were conducted using the LISA model. Specifically, we selected a subset of data from Nordrhein-Westfalen, Germany, and performed several preprocessing steps, including image downloading, vector boundary processing, and image and label cropping, to adapt it for our farmland segmentation model, the image and label overlay results of the test area are shown in Fig.E1.

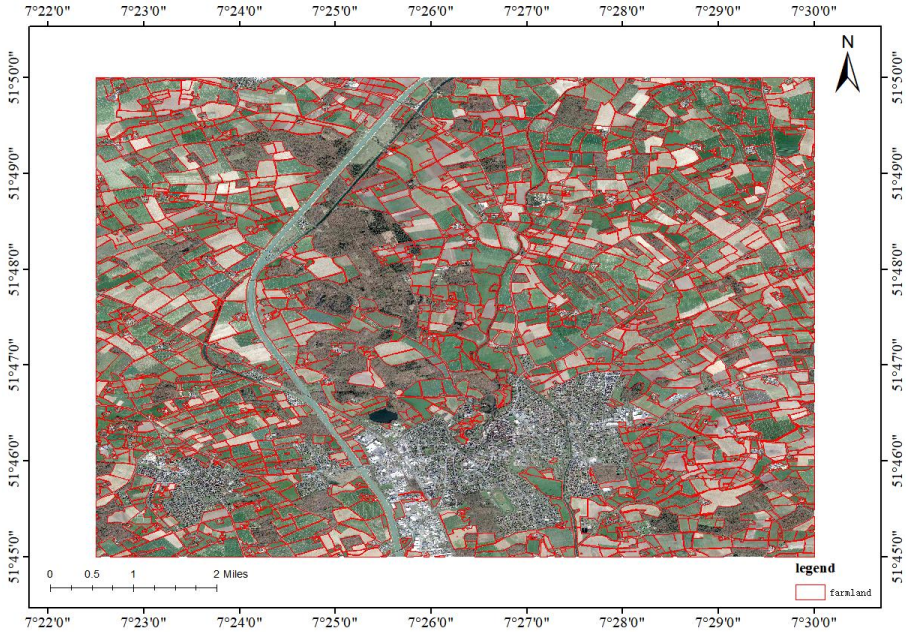


Fig. E1. Example of Fiboa data.

The experimental results are shown in Table E1, where we compare the cross-domain performance of the LISA model trained on the FarmSeg-VL dataset with that of other models evaluated on public datasets in Sect. 4.4. Specifically, the FGFD and LoveDA datasets are from China, while the DGLC dataset covers regions in Thailand, Indonesia, and India. As shown in the table, the LISA model performs well in cross-domain testing, which can be attributed to the extensive geographical coverage and rich seasonal variations of the FarmSeg-VL dataset, providing a solid foundation for cross-domain feature learning. Notably, the LISA model outperforms other models on the Fiboa dataset. This is due to the concentrated, contiguous, and well-defined characteristics of farmland in the Fiboa region, which facilitate the extraction of discriminative features, leading to optimal results in this region. Furthermore, the climatic and cropping system differences between the Fiboa dataset and FarmSeg-VL further validate the applicability and strong generalization capability of the FarmSeg-VL dataset in adapting to the diverse agricultural contexts of different countries. This highlights its potential in global, heterogeneous farmland scenarios.

Table E1. Farmland segmentation results of different methods on fiboa.

Evaluation Metrics(%)	LISA			
	FGFD	LoveDA	DGLC	Fiboa
mACC	83.33	81.76	72.23	**88.05**
mIoU	70.58	65.74	56.36	**78.20**
mDice	82.65	78.82	72.06	**87.73**
Recall	83.87	80.75	72.44	**87.38**

Appendix F Model robustness verification

To evaluate the robustness of the model under different data partitioning conditions, we conducted additional experiments using the LISA model on the FarmSeg-VL dataset. Specifically, we first merged the original training, validation, and test sets, then randomly split the combined dataset into three new training, validation, and test sets following a 7:2:1 ratio. This random splitting procedure was repeated three times to minimize the impact of stochastic variation, and the model was trained and evaluated independently for each split.

Table F1. Farmland segmentation results on different tests.

Evaluation Metrics(%)	Test1	Test2	Test3	Test4
mACC	87.71	87.27	87.33	87.54
mIoU	93.47	93.22	93.26	93.37
mDice	93.45	93.20	93.23	93.36
Recall	93.46	93.20	93.24	93.34

Table F1 shows the results of four different random partitions of the test set. Test1–Test4 represent the results of four different test sets. As shown in the Table F1, the variation in test results across the different test sets is minimal, demonstrating the robustness of the FarmSeg-VL dataset and the model's robustness. This outcome indicates that the balanced distribution and diverse geographical features of the dataset play a crucial role in enhancing the model's stability and generalization capability. Specifically, the FarmSeg-VL dataset is characterized by high-quality image and textual annotations, with a broad distribution that spans different seasons and geographical conditions. This effectively reduces the discrepancies between the datasets, thereby improving the model's robustness to variations in data partitions.

Author contributions

The dataset was conceptualized by CT and HYW. DDZ, WLM, and ZFD carried out the dataset construction, and related experiments. DDZ prepared the initial draft of the manuscript, which was reviewed and revised by all authors.

Competing interests

The contact author has declared that none of the authors has any competing interests.

Disclaimer

Publisher's note: Copernicus Publications remains neutral with regard to jurisdictional claims made in the text, published maps, institutional affiliations, or any other geographical representation in this paper. While Copernicus Publications makes every effort to include appropriate place names, the final responsibility lies with the authors. Regarding the maps used in this paper, please note that Figs. 2 and 3 contain disputed territories.

Financial support

This research was supported in part by the Natural Science Foundation of Hunan for Distinguished Young Scholars under Grant 20221110072 and in part by the National Natural Science Foundation of China under Grant 42471419, Grant 42171376 and Grant 41771458.

References

Brown, C. F., Brumby, S. P., Guzder-Williams, B., Birch, T., Hyde, S. B., Mazzariello, J., Czerwinski, W., Pasquarella, V. J., Haertel, R., Ilyushchenko, S., Schwehr, K., Weisse, M., Stolle, F., Hanson, C., Guinan, O., Moore, R., and Tait, A. M.: Dynamic World, Near real-time global 10 m land use land cover mapping, Sci. Data, 9, 251, https://doi.org/10.1038/s41597-022-01307-4, 2022.

Cheng, Q., Huang, H., Xu, Y., Zhou, Y., Li, H., and Wang, Z.: NWPU-Captions Dataset and MLCA-Net for Remote Sensing Image Captioning, IEEE Trans. Geosci. Remote Sens., 60, 1–19, https://doi.org/10.1109/TGRS.2022.3201474, 2022.

Demir, I., Koperski, K., Lindenbaum, D., Pang, G., Huang, J., Basu, S., Hughes, F., Tuia, D., and Raskar, R.: DeepGlobe 2018: A Challenge to Parse the Earth through Satellite Images, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Salt Lake City, UT, USA, 172–17209, https://doi.org/10.1109/CVPRW.2018.00031, 2018.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, https://doi.org/10.48550/arXiv.1810.04805, 24 May 2019.

Duan, D., Sun, X., Liang, S., Sun, J., Fan, L., Chen, H., Xia, L., Zhao, F., Yang, W., and Yang, P.: Spatiotemporal Patterns of Cultivated Land Quality Integrated with Multi-Source Remote Sensing: A Case Study of Guangzhou, China, Remote Sens., 14, 1250, https://doi.org/10.3390/rs14051250, 2022.

Hou, W., Wang, Y., Su, J., Hou, Y., Zhang, M., and Shang, Y.: Multi-Scale Bilateral Spatial Direction-Aware Network for Cropland Extraction Based on Remote Sensing Images, IEEE Access, 11, 109997–110009, https://doi.org/10.1109/ACCESS.2023.3318000, 2023.

Hu, Y., Yuan, J., Wen, C., Lu, X., and Li, X.: RSGPT: A Remote Sensing Vision Language Model and Benchmark, https://doi.org/10.48550/arXiv.2307.15266, 27 July 2023.

Karra, K., Kontgis, C., Statman-Weil, Z., Mazzariello, J. C., Mathis, M., and Brumby, S. P.: Global land use / land cover with Sentinel 2 and deep learning, in: 2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS, 2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS, Brussels, Belgium, 4704–4707, https://doi.org/10.1109/IGARSS47720.2021.9553499, 2021.

Lai, X., Tian, Z., Chen, Y., Li, Y., Yuan, Y., Liu, S., and Jia, J.: LISA: Reasoning Segmentation via Large Language Model, http://arxiv.org/abs/2308.00692, 3 August 2023.

Li, H., Lin, H., Luo, J., Wang, T., Chen, H., Xu, Q., and Zhang, X.: Fine-Grained Abandoned Cropland Mapping in Southern China Using Pixel Attention Contrastive Learning, IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens., 17, 2283–2295, https://doi.org/10.1109/JSTARS.2023.3338454, 2024.

Li, J., Wei, Y., Wei, T., and He, W.: A Comprehensive Deep-Learning Framework for Fine-Grained Farmland Mapping From High-Resolution Images, IEEE Trans. Geosci. Remote Sens., 63, 1–15, https://doi.org/10.1109/TGRS.2024.3515157, 2025.

Li, M., Long, J., Stein, A., and Wang, X.: Using a semantic edge-aware multi-task neural network to delineate agricultural parcels from remote sensing images, ISPRS J. Photogramm. Remote Sens., 200, 24–40, https://doi.org/10.1016/j.isprsjprs.2023.04.019, 2023.

Liu, H., Li, C., Wu, Q., and Lee, Y. J.: Visual Instruction Tuning, http://arxiv.org/abs/2304.08485, 11 December 2023.

Lu, X., Wang, B., Zheng, X., and Li, X.: Exploring Models and Data for Remote Sensing Image Caption Generation, IEEE Trans. Geosci. Remote Sens., 56, 2183–2195, https://doi.org/10.1109/TGRS.2017.2776321, 2018.

Pan, T. and Zhang, R.: Spatiotemporal Heterogeneity Monitoring of Cropland Evolution and Its Impact on Grain Production Changes in the Southern Sanjiang Plain of Northeast China, Land, 11, 1159, https://doi.org/10.3390/land11081159, 2022.

Phalke, A. R. and Özdoğan, M.: Large area cropland extent mapping with Landsat data and a generalized classifier, Remote Sens. Environ., 219, 180–195, https://doi.org/10.1016/j.rse.2018.09.025, 2018.

Qu, B., Li, X., Tao, D., and Lu, X.: Deep semantic understanding of high resolution remote sensing image, in: 2016 International Conference on Computer, Information and Telecommunication Systems (CITS), 2016 International Conference on Computer, Information and Telecommunication Systems (CITS), Kunming, China, 1–5, https://doi.org/10.1109/CITS.2016.7546397, 2016.

Ren, Z., Huang, Z., Wei, Y., Zhao, Y., Fu, D., Feng, J., and Jin, X.: PixelLM: Pixel Reasoning with Large Multimodal Model, http://arxiv.org/abs/2312.02228, 18 July 2024.

Sishodia, R. P., Ray, R. L., and Singh, S. K.: Applications of Remote Sensing in Precision Agriculture: A Review, Remote Sens., 12, 3136, https://doi.org/10.3390/rs12193136, 2020.

Sumbul, G., Charfuelan, M., Demir, B., and Markl, V.: BigEarthNet: A Large-Scale Benchmark Archive For Remote Sensing Image Understanding, in: IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium, arXiv:1902.06148 [cs], 5901–5904, https://doi.org/10.1109/IGARSS.2019.8900532, 2019.

Tao, C., Zhong, D., Mu, W., Du, Z., and Wu, H.: A large-scale image-text dataset benchmark for farmland segmentation, https://doi.org/10.5281/zenodo.15099885, 2025a.

Tao, C., Zhong, D., Mu, W., Du, Z., and Wu, H.: A large-scale image-text dataset benchmark for farmland segmentation, 2025b.

Tong, X.-Y., Xia, G.-S., Lu, Q., Shen, H., Li, S., You, S., and Zhang, L.: Land-cover classification with high-resolution remote sensing images using transferable deep models, Remote Sens. Environ., 237, 111322, https://doi.org/10.1016/j.rse.2019.111322, 2020.

Tu, Y., Wu, S., Chen, B., Weng, Q., Bai, Y., Yang, J., Yu, L., and Xu, B.: A 30 m annual cropland dataset of China from 1986 to 2021, Earth Syst. Sci. Data, 16, 2297–2316, https://doi.org/10.5194/essd-16-2297-2024, 2024.

Wang, J., Liu, B., and Xu, K.: Semantic segmentation of high-resolution images, Sci. China Inf. Sci., 60, 123101, https://doi.org/10.1007/s11432-017-9252-5, 2017.

Wang, J., Zheng, Z., Ma, A., Lu, X., and Zhong, Y.: LoveDA: A Remote Sensing Land-Cover Dataset for Domain Adaptive Semantic Segmentation, http://arxiv.org/abs/2110.08733, 31 May 2022a.

Wang, J., Zheng, Z., Ma, A., Lu, X., and Zhong, Y.: LoveDA: A Remote Sensing Land-Cover Dataset for Domain Adaptive Semantic Segmentation, n.d.

Wang, Y., Wu, X., and Zhu, H.: Spatio-Temporal Pattern and Spatial Disequilibrium of Cultivated Land Use Efficiency in China: An Empirical Study Based on 342 Prefecture-Level Cities, Land, 11, 1763, https://doi.org/10.3390/land11101763, 2022b.

Wang, Z., Prabha, R., Huang, T., Wu, J., and Rajagopal, R.: SkyScript: A Large and Semantically Diverse Vision-Language Dataset for Remote Sensing, Proc. AAAI Conf. Artif. Intell., 38, 5805–5813, https://doi.org/10.1609/aaai.v38i6.28393, 2024.

Wei, C., Tan, H., Zhong, Y., Yang, Y., and Ma, L.: LaSagnA: Language-based Segmentation Assistant for Complex Queries, https://doi.org/10.48550/arXiv.2404.08506, 12 April 2024.

Wu, H., Mu, W., Zhong, D., Du, Z., Li, H., and Tao, C.: FarmSeg_VLM: A farmland remote sensing image segmentation method considering vision-language alignment, ISPRS J. Photogramm. Remote Sens., 225, 423–439, https://doi.org/10.1016/j.isprsjprs.2025.05.010, 2025a.

Wu, H., Du, Z., Zhong, D., Wang, Y., and Tao, C.: FSVLM: A Vision-Language Model for Remote Sensing Farmland Segmentation, IEEE Trans. Geosci. Remote Sens., 63, 1–13, https://doi.org/10.1109/TGRS.2025.3532960, 2025b.

Xie, D., Xu, H., Xiong, X., Liu, M., Hu, H., Xiong, M., and Liu, L.: Cropland Extraction in Southern China from Very High-Resolution Images Based on Deep Learning, Remote Sens., 15, 2231, https://doi.org/10.3390/rs15092231, 2023.

Yuan, Z., Xiong, Z., Mou, L., and Zhu, X. X.: ChatEarthNet: A Global-Scale Image-Text Dataset Empowering Vision-Language Geo-Foundation Models, https://doi.org/10.5194/essd-2024-140, 26 February 2024.

Zanaga, D., Van De Kerchove, R., Daems, D., De Keersmaecker, W., Brockmann, C., Kirches, G., Wevers, J., Cartus, O., Santoro, M., Fritz, S., Lesiv, M., Herold, M., Tsendbazar, N.-E., Xu, P., Ramoino, F., and Arino, O.: ESA WorldCover 10 m 2021 v200 (v200), https://doi.org/10.5281/ZENODO.7254221, 2022.

Zhang, Z., Zhao, T., Guo, Y., and Yin, J.: RS5M and GeoRSCLIP: A Large Scale Vision-Language Dataset and A Large Vision-Language Model for Remote Sensing, IEEE Trans. Geosci. Remote Sens., 62, 1–1, https://doi.org/10.1109/TGRS.2024.3449154, 2024.

Zheng, Y., Long, H., and Chen, K.: Spatio-temporal patterns and driving mechanism of farmland fragmentation in the Huang-Huai-Hai Plain, J. Geogr. Sci., 32, 1020–1038, https://doi.org/10.1007/s11442-022-1983-8, 2022.

Zhu, Z., Dai, Z., Li, S., and Feng, Y.: Spatiotemporal Evolution of Non-Grain Production of Cultivated Land and Its Underlying Factors in China, Int. J. Environ. Res. Public. Health, 19, 8210, https://doi.org/10.3390/ijerph19138210, 2022.