

## -Detailed Response to Reviewers-

Dear Editors and Reviewers:

Thank you for your letter and for your comments concerning our manuscript entitled "A large-scale image-text dataset benchmark for farmland segmentation". These comments are very helpful for improving the quality of our manuscript. Based on your constructive feedback, we have made some corrections and highlighted the response to Reviewer 2 in red font in the revised manuscript. The responses to these comments of reviewers are as follows.

### **Point-to-Point responses to the Reviewers.**

#### **Reviewer 2:**

Understanding the spatiotemporal characteristics of farmland is essential for accurate farmland segmentation. This study introduced language-based descriptions of farmland and developed FarmSeg-VL dataset, which was the first fine-grained image-text dataset for farmland segmentation. The proposed method is innovative and dataset is of high accuracy, which had great potentials as a standard benchmark for farmland segmentation. However, there are still some problems that deserve to solve before publications.

**Response:** Thank you for your insightful and encouraging comments. We appreciate your recognition of the importance of understanding the spatiotemporal characteristics of farmland, and your acknowledgment of the innovation and potential impact of the FarmSeg-VL dataset as a fine-grained image-text benchmark for farmland segmentation. We are also grateful for your recognition of the accuracy and applicability of our proposed method. In response to your suggestions, we have carefully revised the manuscript to address the remaining issues prior to publication. Specifically, we have refined methodological explanations, clarified the scope of application, and improved the consistency and clarity of language expression throughout the paper. We hope these revisions sufficiently address your concerns and further strengthen the contribution of our work.

#### **Point 1:**

In the introduction part, the author mentioned that the label-driven paradigm had some disadvantages for farmland segmentation. The vision-language models (VLMs) can capture more contextual and background information from imageries. More information about the construction of VLM model can be included.

#### **Response 1:**

Thank you for your valuable comment. We have revised the introduction to include a clearer and more detailed explanation of the construction of VLM model. The detailed modifications are as follows:

#### **1 Introduction**

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.

**Point 2:**

The image-text datasets is also the core of the VLM model, and what is the difference with the traditional label-driven deep learning method.

**Response 2:**

Thank you for your valuable comments. The differences between traditional label-driven deep learning methods and the VLM-based approach are as follows:

(1) The existing label-driven paradigm primarily focuses on learning the shape and texture features of the farmland and its surrounding environment, neglecting the inherent logical relationships between farmland and its environment in complex agricultural scenes. As a result, when the model is tasked with segmenting farmland in complex scenarios, it fails to capture the deep connections between farmland and its environment, leading to insufficient generalization ability. In contrast, language can describe environmental features such as buildings, water resources, and vegetation layout around the farmland. Guided by language, VLMs enhance the inherent logical connections between farmland and surrounding land features.

(2) Labels often fail to fully reflect the evolving characteristics of farmland across different seasons and growth stages, leading to insufficient generalization ability of traditional label-driven deep learning methods in spatiotemporal dynamic scenarios. In contrast, language, by describing information such as seasonal changes and spatial location of the farmland, provides the model with rich prior knowledge. As a result, VLMs are able to capture the phenological features of farmland as it changes with the seasons, climate fluctuations, and crop growth cycles, while also learning the morphological differences caused by geographical variations. This enables a better

understanding of the spatiotemporal dynamic changes of farmland, effectively alleviating the segmentation challenges posed by temporal and spatial variations.

**Point 3:**

In table 1, the abbreviation should be explained, such as SR for spatial resolution.

**Response 3:**

Thank you for your valuable comments. Based on your suggestions, we have made the necessary abbreviations modifications in Table 1. "SR" has been changed to "Spatial Resolution," and "FP" now represents "Farmland Proportion." The detailed modifications are as follows:

| 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, Xinjiang, China |

**Point 4:**

In section 2.2, the author mentioned that the most of image-text datasets just described scene level or object level characteristics instead of specific like farmland segmentation. I wonder the inherent difference among these textural descriptions, and authors can take some detailed examples.

**Response 4:**

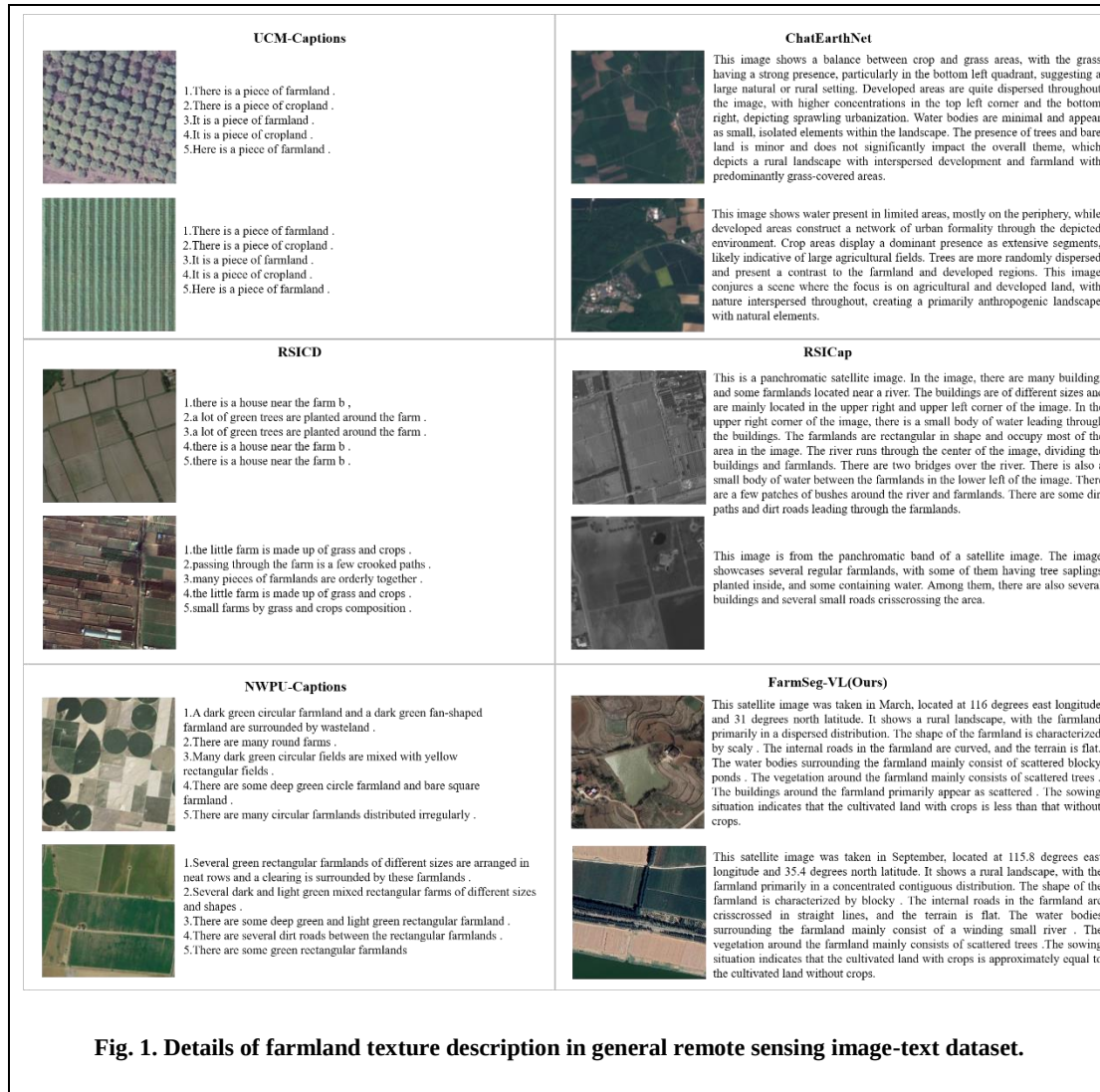
Thank you for your valuable comments. To further illustrate the limitations of farmland texture descriptions in existing image-text datasets and the differences with our dataset, we have added several specific examples and analyzed the inherent differences in terms of granularity and task suitability. The relevant supplements have been placed in the appendix, the detailed modifications are as follows:

## Appendix

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

As shown in Fig.9, 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.



## Point 5:

China has a vast territory and different regions have different agricultural conditions. The construction of FarmSeg-VL dataset and related textural descriptions should consider this.

## Response 5:

Thank you for your comment. We have considered the agricultural diversity across different regions of China in the construction of the FarmSeg-VL dataset and related textual descriptions. Specifically, this is addressed in Section 3.1, under 1) RS Image Acquisition and Processing and 2) Caption Construction. We apologize for the unclear description, and we have made appropriate revisions to enhance readability for the readers. The specific changes are as follows:

### 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.

## **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)



**Point 6:**

In figure 2, the authors should an example of 5 types of text characteristics for farmland fragmentation. In future study, more quantitative description can be included.

**Response 6:**

Thank you for your valuable comments. In response to your request to add five examples of farmland fragmentation text features in Figure 2, we have included the corresponding examples in the appendix to further assist readers in understanding farmland shape characteristics. In future studies, from the perspective of the textual descriptions of farmland samples, we may add proportional descriptions, such as farmland area ratio and fragmentation degree, to provide a more comprehensive assessment of farmland. These additional metrics could include the mean patch size of farmland, the proportion of farmland area to total image area, and the number of fragmented patches per hectare, which will further deepen the understanding of farmland fragmentation and its spatial characteristics. From the perspective of the dataset's scale, and coverage, FarmSeg-VL covers 13 provinces in China, spanning approximately 4,300 km<sup>2</sup>. We plan to add data from 2 to 3 other countries (e.g., Africa, Europe, and the United States) to expand our dataset, aiming for global diversification. The detailed modifications are as follows:

**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. 18, beyond those shown in Fig. 10.

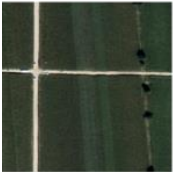
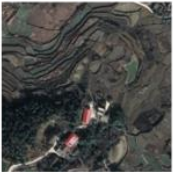
| blocky                                                                              | narrow strip                                                                        | scaly                                                                               | broken                                                                               | circular                                                                              |
|-------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------|
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |

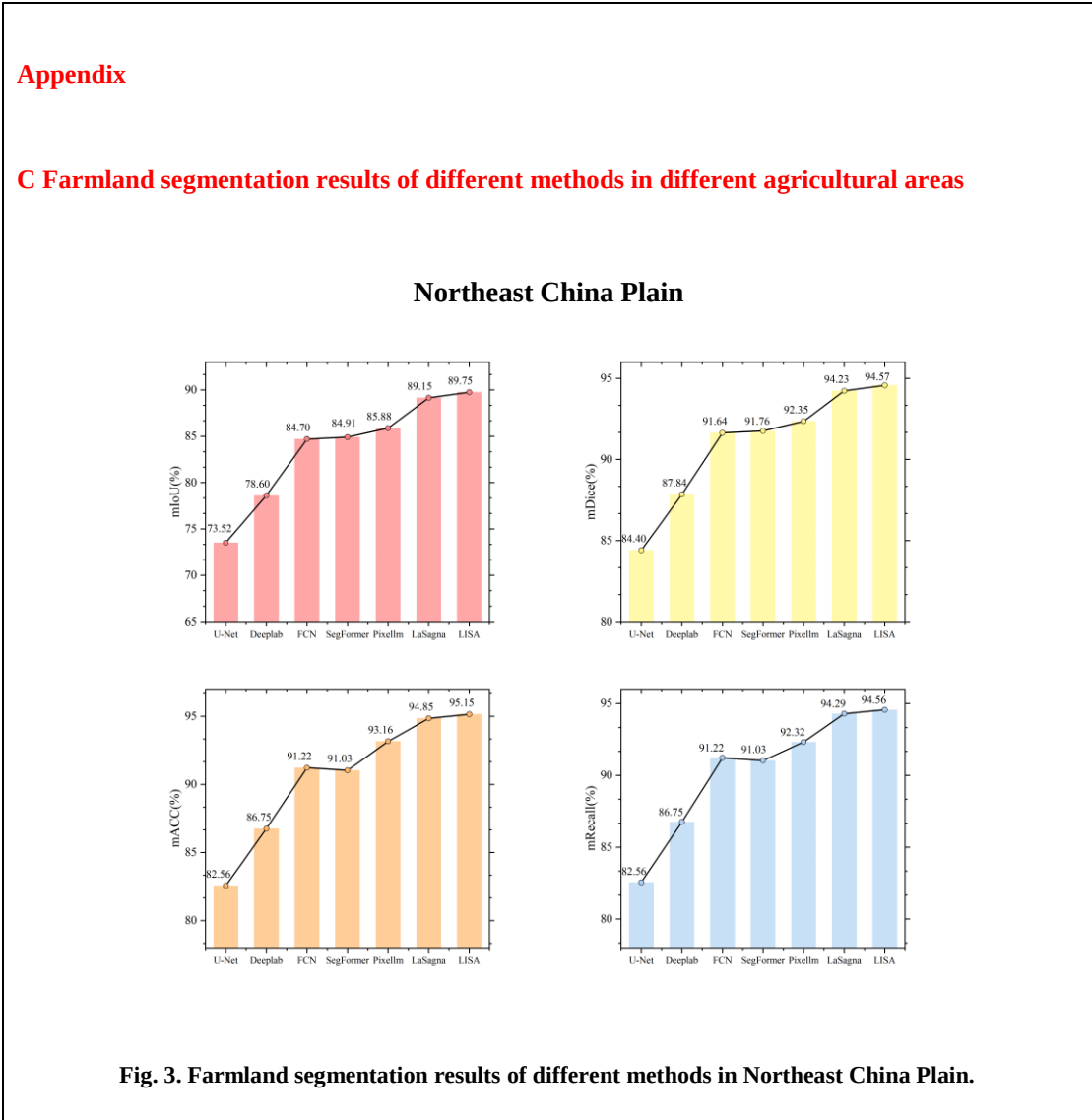
Fig. 2. Example of farmland shape.

**Point 7:**

Table 4-11 can be converted into the figure to improve the readability. What is the LoveDA dataset in section 4.5?

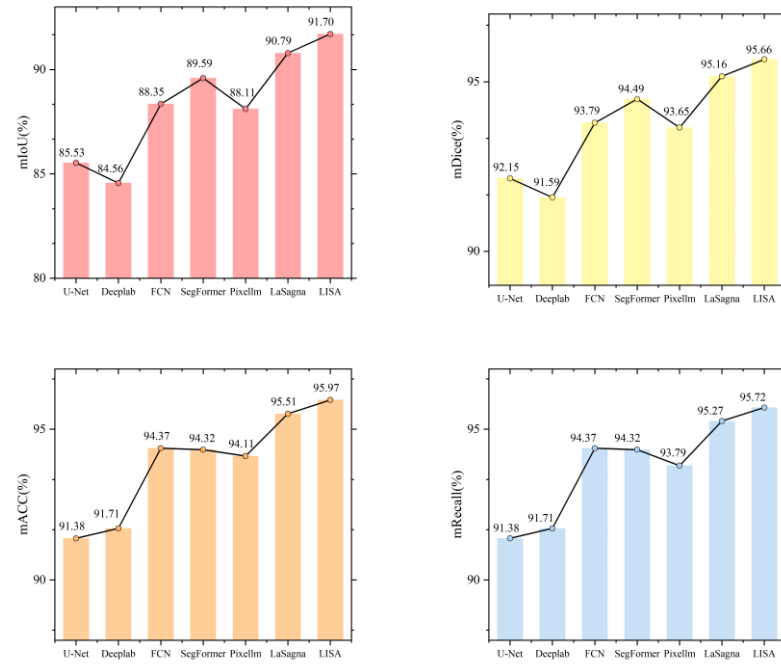
**Response 7:**

We have adopted your suggestion and converted Tables 4-11 into visualized charts (see Fig. 11 to 18 in the revised manuscript), significantly enhancing the intuitiveness and readability of the results. The LoveDA dataset is a publicly available land cover classification dataset that was presented at the NeurIPS 2021 Datasets and Benchmarks Track. Since this paper focuses on farmland segmentation tasks, in the experiment of Section 4.5, we adapted the dataset by merging its original multiple non-farmland categories into a unified background label, retaining only the "farmland" and "non-farmland" categories. This modification aligns its label system with our binary segmentation task. The detailed modifications are as follows:



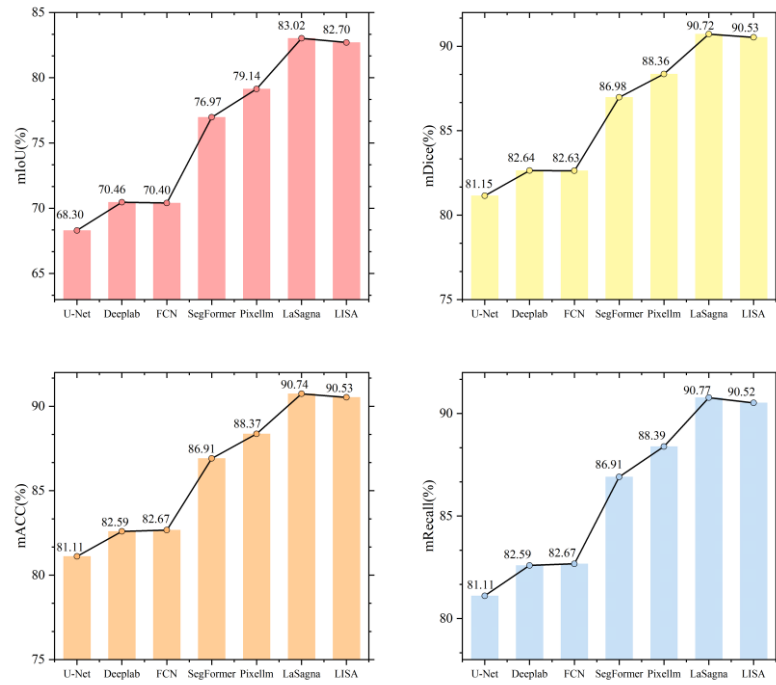


### Huang-Huai-Hai Plain



**Fig. 4. Farmland segmentation results of different methods in Huang-Huai-Hai Plain.**

### Northern Arid and Semi-arid Region



**Fig. 5. Farmland segmentation results of different methods in Northern Arid and Semi-arid Region.**

### Loess Plateau

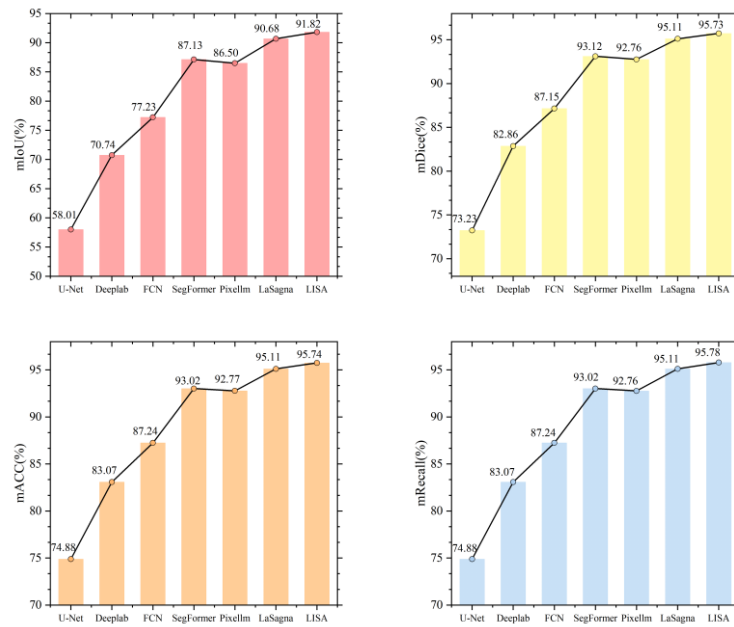


Fig. 6. Farmland segmentation results of different methods in Loess Plateau.

### Yangtze River Middle and Lower Reaches Plain

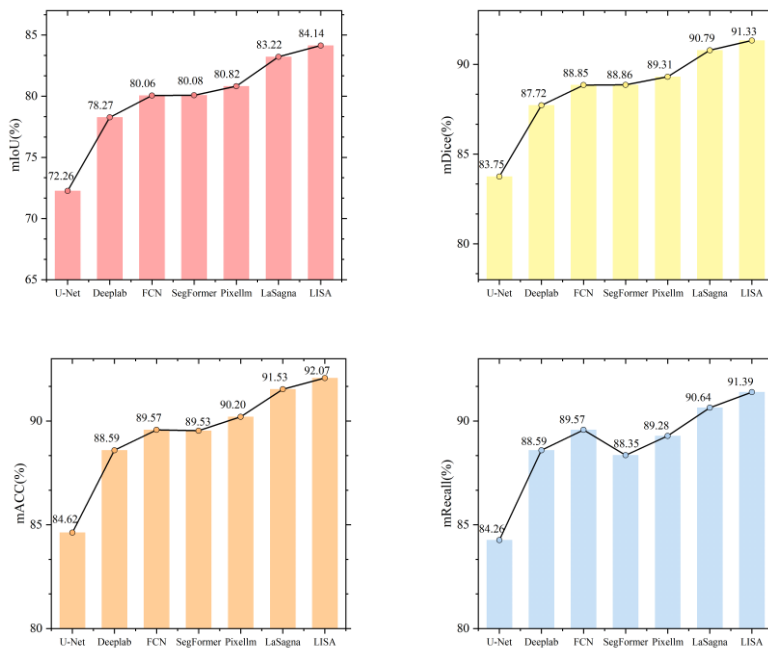


Fig. 7. Farmland segmentation results of different methods in Yangtze River Middle and Lower Reaches Plain.

### South China Areas

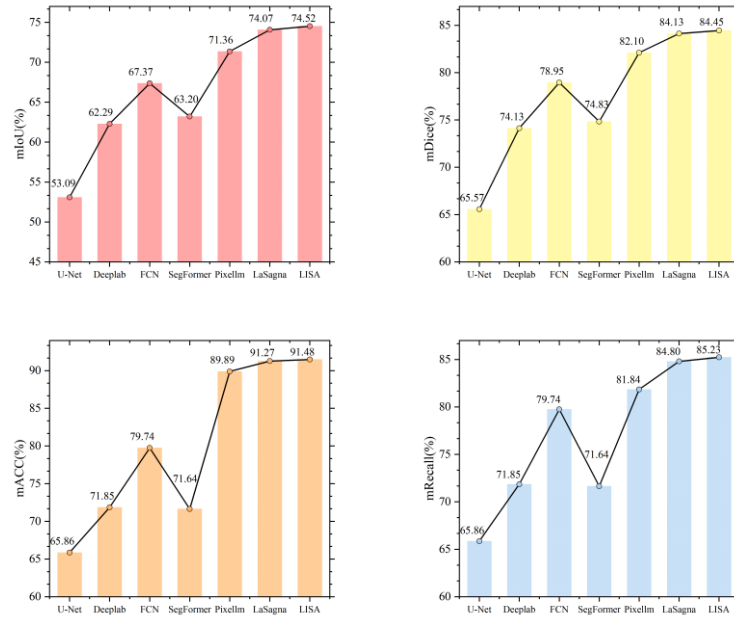


Fig. 8. Farmland segmentation results of different methods in South China Areas.

### Sichuan Basin

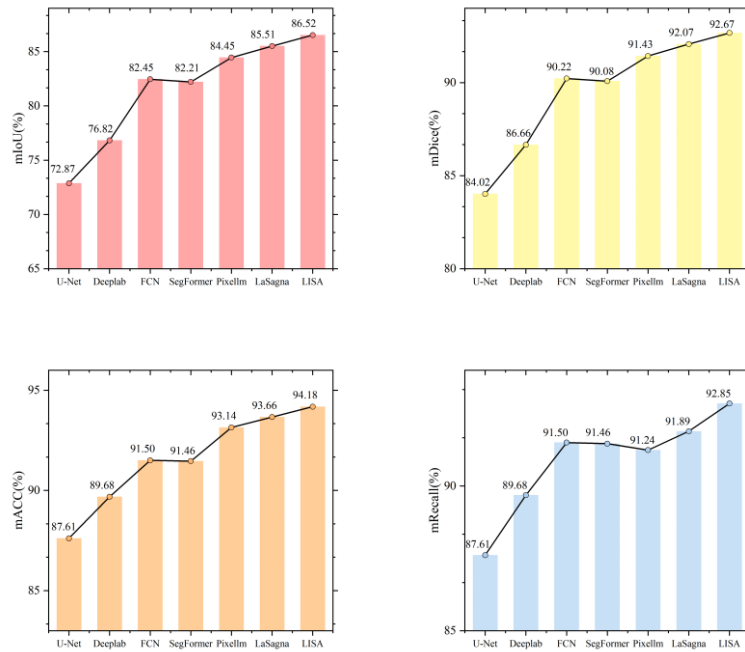


Fig. 9. Farmland segmentation results of different methods in Sichuan Basin.

### Yungui Plateau

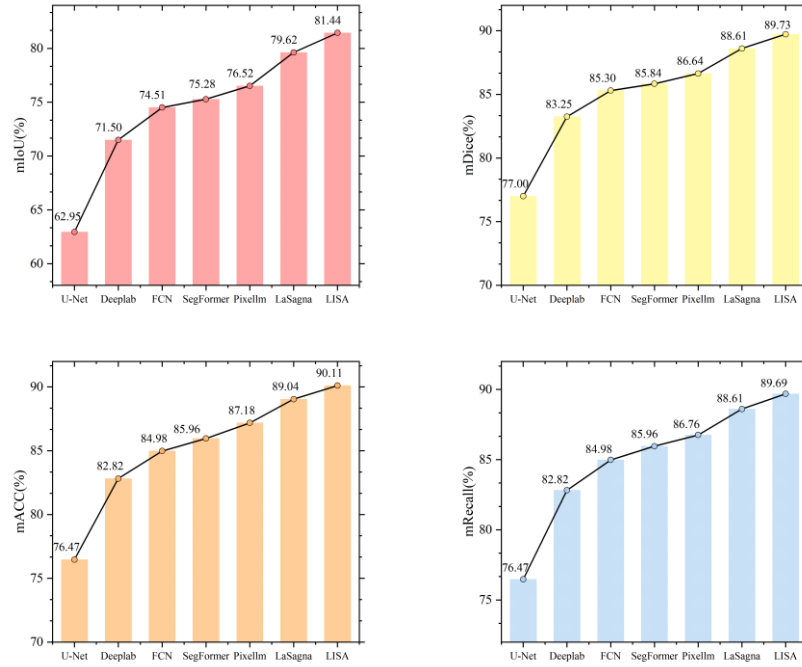


Figure 10. Farmland segmentation results of different methods in Yungui Plateau.

#### Point 8:

In the abstract and conclusion parts, the authors should add more information and quantitative results about the farmland segmentation results in this study.

#### Response 8:

Thank you for your valuable comments. In the abstract and discussion, we have added the quantified results of farmland segmentation. The detailed modifications are as follows:

**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 km<sup>2</sup>. 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.15099885>(Tao et al., 2025).

## 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.

**Point 9:**

I wonder if the author adds some negative samples for the farmland fragmentation? Table show the comparison results solely on labels and VLMs. The deep learning models for these two kinds are different and the training samples for these two kinds are also different. How could these be compared?

**Response 9:**

Thank you for your valuable comments. This study focuses on farmland binary classification, where the image and mask data only include annotations for cropland and non-farmland areas. The non-farmland class, which is labeled as the background, represents the negative samples. In the textual descriptions, we have incorporated rich semantic information about the surrounding environment, which serves as the description for the negative samples. These descriptions include keywords such as "blocky ponds," "meadow," and "scattered buildings," which help the model better understand the relationship between cropland and surrounding land features. Regarding the rationale for model comparison, we understand the reviewers' concerns about comparing different architectures and training sample variations. The core purpose of this comparison is to assess the impact of introducing textual description information on farmland segmentation performance. In the training samples, the images and masks of farmland remain consistent, and only the text description of farmland is added to VLM. Although the model architectures and training data details differ, the comparison is based on one key point: the VLM only adds natural language text during training in addition to the label-based deep learning method, while the visual input and visual encoding remain consistent. During testing, the prompt for the VLM was "Please segment the farmland in the image," which aligns with the goal of the label-based method that relies solely on visual information for inference.

**Point 10:**

There are still some grammatical and lingual problems, and authors should make a thorough revision, such as "annotationand" in line 115.

**Response 10:**

Thank you for your valuable comments. We have carefully reviewed the entire manuscript and made comprehensive revisions to address the grammatical and language issues, including correcting "annotationand" to "annotation and" in line 115.