-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 1 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 1:

This paper proposes FarmSeg-VL, the first large-scale image-text benchmark dataset for farmland segmentation, which fills the gap of the lack of high-quality farmland multimodal data in the field of remote sensing. The research has significant innovation and application value, the experimental design is systematic, the results are analyzed in detail, the data are open and transparent, and it meets the publication criteria of journals. However, some of the methodological details, scope of application, and writing expressions need to be further optimized.

Response: Thank you for your positive and constructive comments. We sincerely appreciate your recognition of the novelty, application value, and completeness of our work, including the contribution of the FarmSeg-VL dataset and the experimental design. In response to your suggestions regarding methodological details, scope of application, and writing clarity, we have carefully revised the manuscript. Specifically, we have (1) provided further clarification on the methodological design and key parameters, (2) expanded the discussion on the dataset's applicability across different agricultural conditions, and (3) refined the language throughout the manuscript to improve readability and precision. We hope that these revisions address your concerns and further enhance the quality of the paper.

Point 1：

The specific application of the model in annotation, such as parameter setting and manual correction ratio, needs to be further explained. In addition, it is necessary to quantify the improvement of annotation efficiency, such as the comparison of time consumption with traditional manual annotation.

Response 1:

Thank you for your comment. In response to the specific application of the model in annotation, we provide the following additional explanation. As shown in Fig. 5 of the manuscript, for mask annotation, we integrated the Segment Anything Model (SAM) to assist in generating farmland masks by creating AI-generated polygons. For text annotation, the tool automatically extracts longitude, latitude, and acquisition month information from the image filenames and populates the corresponding fields.

Additionally, we designed a farmland keyword selection widget to construct standardized image description texts. The resulting mask and text information are stored in separate JSON files. Regarding the parameter settings, we have used the default parameters of SAM without any additional adjustments. Concerning the manual correction ratio, in most cases, the generated masks require no manual modification, with only approximately 10% of the images necessitating minor boundary adjustments.

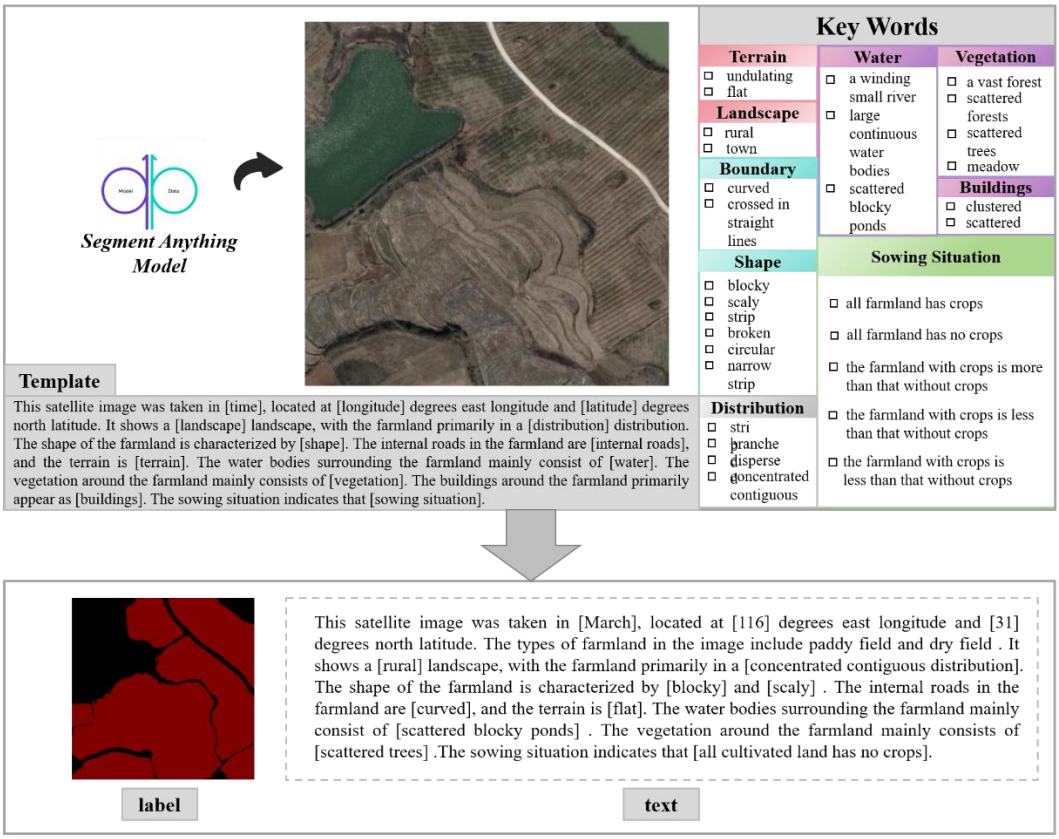

Fig. 1. Farmland semi-automated annotation framework.

To quantify the improvement in annotation efficiency, we conducted a comparative experiment. Four operators were randomly selected to annotate 13 remote sensing images using both traditional manual tracing and semi-automatic annotation method . The experimental results, shown in Fig. 19 of the revised manuscript, demonstrate that the average annotation time using semi-automatic annotation method was significantly reduced compared to the traditional approach, resulting in a 1.5 times improvement in overall efficiency. This validates the significant improvement in annotation efficiency and usability of the developed tool. The detailed modifications are as follows:

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. 27, 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.

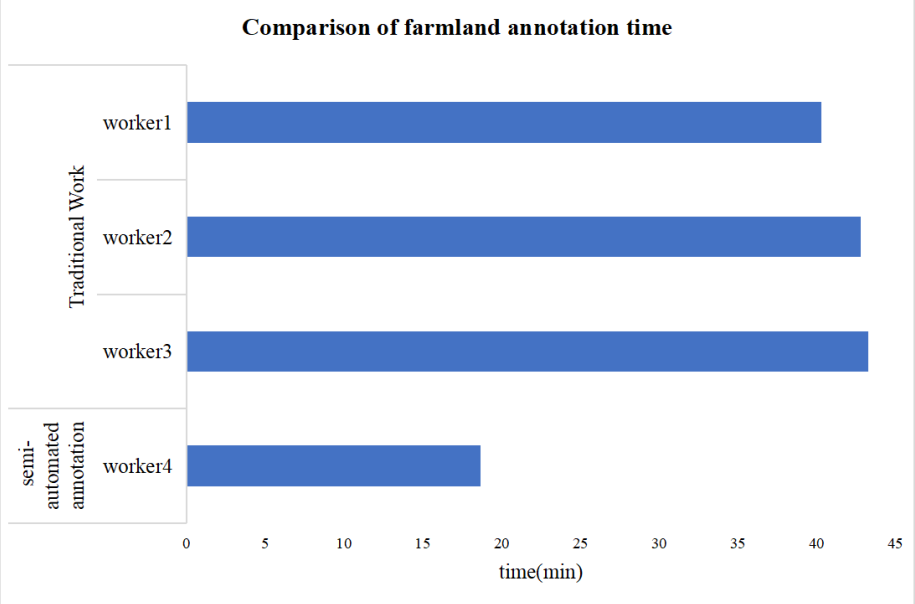

Fig. 27. Comparison of farmland annotation efficiency.

Point 2:

It is suggested to add the selection basis of the 11 key elements, such as whether they have been verified by experts or supported by the literature, in order to enhance the scientific nature of the description framework.

Response 2:

Thank you for your comment. Regarding the selection criteria for the 11 key elements, we have added supporting references from the literature. The detailed modifications are as follows:

3.1 Construction of FarmSeg-VL

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 3：

The current dataset mainly covers the China region. It is suggested that the authors discuss whether the dataset applies to other countries with significant differences in climate or cropping patterns (e.g., Africa, Europe, and the United States). In addition, the authors need to consider whether the global data can be expanded in the future.

Response 3:

Thank you for your comment. Regarding your question, I will answer it from the following two aspects:

(1) From the perspective of the generalizability of the dataset construction process, we believe the process has strong general applicability. China's vast territory spans multiple climate zones, and its geographic and climatic diversity nearly encompasses the main terrain and climate features of other countries. We believe that the descriptive keywords we have designed for farmland can comprehensively cover various cropland morphologies. Although there are differences in climate and cropping practices between China and some other countries, our textual annotation framework is highly flexible. We can adjust the keywords based on the cropland climate of different regions, allowing it to adapt to the farmland characteristics of various regions around the world.

(2) From the perspective of the generalizability of the model trained on FarmSeg-VL, we have further supplemented the relevant experiments in Appendix E of the revised manuscript. The results indicate that, despite FarmSeg-VL being constructed based on remote sensing imagery from China, it still exhibits strong transferability under different climate and cropping pattern conditions, which partially validates the dataset’s applicability in broader agricultural scenarios.

Furthermore, we fully agree with the reviewer's suggestion regarding global expansion. In future work, we plan to further collect and organize remote sensing imagery and farmland annotation information from various countries and regions, continually expanding the spatial coverage of the dataset to support more diverse and globally adaptable agricultural intelligence analysis research. The detailed modifications are as follows:

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.28.

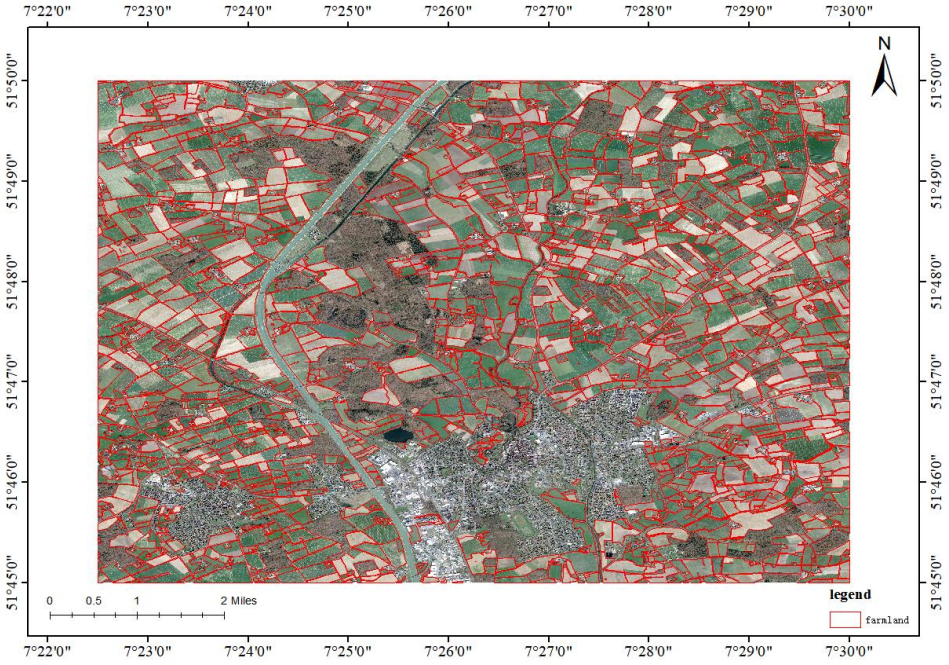

Fig. 2. Example of Fiboa data.

The experimental results are shown in Table 8, 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 Section 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 1. 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**

Point 4:

The text mentions that the data cover four seasons, but it is not clear whether the full growth cycle of different crops is covered. It is suggested that additional clarification be provided.

Response 4:

Thank you for your comment. In the process of data selection, this study has thoroughly considered the seasonal characteristics of farmland and the key stages of the crop growth cycle. Although the data encompasses all four seasons—spring, summer, autumn, and winter—it does not cover the complete growth cycle of all crops. Instead, we selectively focused on key periods when remote sensing imagery exhibited typical texture patterns for different crops in various regions. For example, in the Northeast Plain agricultural region, summer is the peak growth period for major crops, with distinct farmland texture features. Therefore, we primarily collected summer imagery data from the Northeast Plain region, including Heilongjiang and Jilin provinces, to better capture the farmland characteristics during this typical period. By selecting representative temporal imagery, we can effectively enhance the model's ability to recognize farmland spatial distribution and seasonal changes.

Point 5:

It is suggested to add the performance comparison of the model on the training set and the test set, or analyze the influence of data distribution on the robustness of the model through cross-validation.

Response 5:

Thank you for your comment. The training set, test set, and validation set in the FarmSeg-VL dataset were first mixed together, and then randomly divided into three new training, test, and validation sets at a ratio of 7:2:1. The experiments were conducted based on the LISA model. To avoid the influence of random factors, each experiment was repeated three times. The results showed that the model could maintain stable performance under different data partitioning methods, indicating that the model

trained on FarmSeg-VL demonstrates strong robustness and maintains high generalization capability when faced with different data distributions..The detailed modifications are as follows:

Appendix

F The influence of data distribution on the robustness of the model

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 9. 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 9 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 figure, 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.

Point 6:

Tables 4~11 and Figures 9~16 clearly show the situation of each agricultural region, but they occupy more space, I suggest the authors put this part in the supplementary materials.

Response 6:

Thank you for your comment. We have moved Tables 4–11 and Figures 9–16 from the main text to Appendix C of the manuscript, retaining essential descriptions and analyses in the main body to ensure that readers can access the necessary information without compromising the readability of the paper.

Point 7:

Considering the wide international readership, I suggest the authors add some non-Chinese references.

Response 7:

Thank you for your comment. We understand the importance of diverse international references. We would like to clarify that, although some of the cited references were published in Chinese journals, all references are written in English, and no Chinese-language sources are cited in the manuscript. In future research, we will continue to focus on expanding our international perspective to better serve a global audience.