VHM: Versatile and Honest Vision Language Model for Remote Sensing Image Analysis

Chao Pang[*1,3], **Xingxing Weng**[*3], **Jiang Wu**[*†2], **Jiayu Li**[3], **Yi Liu**[3], **Jiaxing Sun**[2,6], **Weijia Li**[4], **Shuai Wang**[5], **Litong Feng**[5], **Gui-Song Xia**[‡1,3,6,7], **Conghui He**[‡2,5]

[1]School of Artificial Intelligence, Wuhan University
[2]Shanghai Artificial Intelligence Laboratory
[3]School of Computer Science, Wuhan University
[4]Sun Yat-Sen University
[5]Sensetime Research
[6]State Key Lab. of LIESMARS, Wuhan University
[7]Institue for Math & AI, Wuhan University
{wujiang, sunjiaxing, heconghui}@pjlab.org.cn, {wangshuai4, fenglitong}@sensetime.com
liweij29@mail.sysu.edu.cn, {pangchao, xingxingw, jiayu.li, yi_liu, guisong.xia}@whu.edu.cn

Abstract

This paper develops a Versatile and Honest vision language Model (VHM) for remote sensing image analysis. VHM is built on a large-scale remote sensing image-text dataset with rich-content captions (VersaD), and an honest instruction dataset comprising both factual and deceptive questions (HnstD). Unlike prevailing remote sensing image-text datasets, in which image captions focus on a few prominent objects and their relationships, VersaD captions provide detailed information about image properties, object attributes, and the overall scene. This comprehensive captioning enables VHM to thoroughly understand remote sensing images and perform diverse remote sensing tasks. Moreover, different from existing remote sensing instruction datasets that only include factual questions, HnstD contains additional deceptive questions stemming from the non-existence of objects. This feature prevents VHM from producing affirmative answers to nonsense queries, thereby ensuring its honesty. In our experiments, VHM significantly outperforms various vision language models on common tasks of scene classification, visual question answering, and visual grounding. Additionally, VHM achieves competent performance on several unexplored tasks, such as building vectorizing, multi-label classification and honest question answering.

Code & Data — https://github.com/opendatalab/VHM

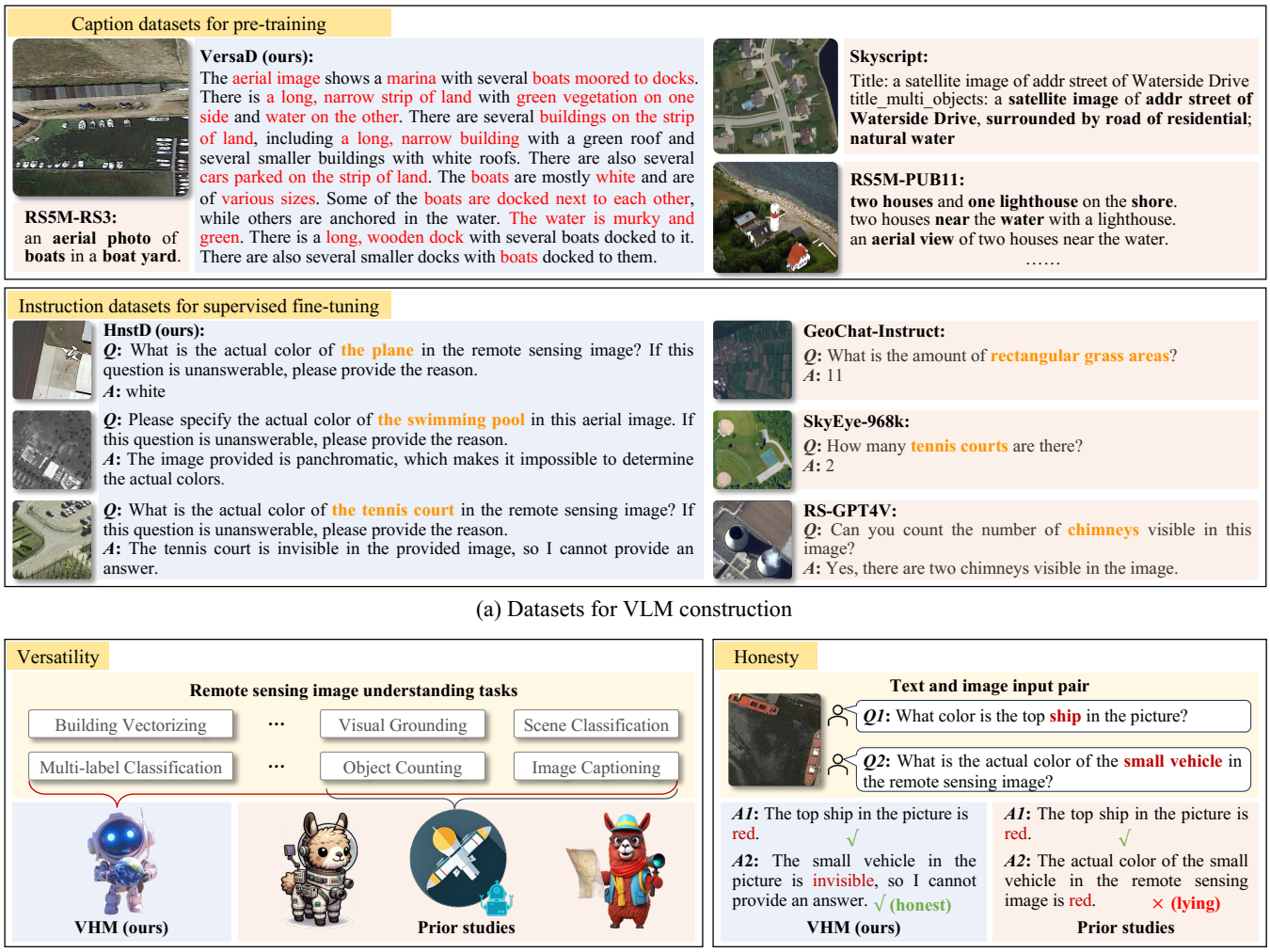
1 Introduction

The remarkable achievements of Vision Language Models (VLMs) in computer vision have sparked a wave of research into tailoring VLMs for remote sensing (RS) (Hu et al. 2023; Kuckreja et al. 2024; Zhang et al. 2024; Muhtar et al. 2024), aiming to enhance RS image analysis in a more intelligent and human-like manner (Li et al. 2024b). Following the typical construction pipeline of first pretraining and then supervised fine tuning (Liu et al. 2024), most studies (Hu et al. 2023; Kuckreja et al. 2024; Luo et al. 2024; Bazi et al. 2024), based on network weights pretrained on large natural image-text datasets, design RS-specific instruction datasets for the fine-tuning process. Given the domain shift between natural and RS images (Wang et al. 2022a), the most recent study (Muhtar et al. 2024) introduces an RS image-text dataset (*i.e.*, LHRS-Align) for additional pretraining. Although existing studies have reported promising results, two main issues still need to be addressed:

(1) They usually pre-train VLMs using image-text pairs with sparse-content captions, which is inadequate for RS. As illustrated in Fig. 1(a), RS images typically contain various objects due to their large field of view. However, most RS image-text pairs (Zhang et al. 2023; Wang et al. 2024) focus on a few prominent objects and their relationships (*e.g.* water and houses). Moreover, even with these prominent objects, they merely mention their presence and neglect crucial details such as their color and shape. This significantly impedes VLMs from achieving a thorough understanding of RS images, thereby limiting their ability to perform diverse RS image analysis tasks.

(2) They usually fine-tune VLMs using instruction datasets that contain only factual questions, which makes VLMs prone to lying. As shown in Fig. 1(a), questions in instruction datasets such as GeoChat-Instruct (Kuckreja et al. 2024) and SkyEye-968k (Zhan, Xiong, and Yuan 2024) involve queries about real objects within images and are accompanied by affirmative answers. Consequently, when

[*]These authors contributed equally.
[†]Project lead.
[‡]Corresponding author.



(a) Datasets for VLM construction

(b) Versatility and honesty of VLM

Figure 1: Illustration of versatility and honesty. In (a), words in red and bold are key pieces of information in the captions. Existing datasets for pretraining VLMs typically contain sparse-content captions, focusing on a few prominent objects and their relationships. In contrast, VersaD captions provide detailed descriptions of image properties, object attributes, and scene context. These rich-content captions contribute to a more thorough understanding of RS images, thereby enhancing VLMs' ability to perform diverse RS tasks. Additionally, instruction datasets for fine-tuning VLMs usually contain only factual questions about existent objects within images (see words in orange in (a)), which can result in VLMs lying to produce affirmative answers to nonsense queries about non-existent objects. In contrast, our HnstD includes both factual and deceptive questions, designed to instill honesty in VLMs.

faced with deceptive questions, VLMs lie to produce affirmative answers. For example, VLMs misrepresented the color of a non-existent object (Fig. 1(b)).

In this paper, we devote ourselves to developing VHM, a more **V**ersatile and **H**onest VLM for RS image analysis, aimed at performing a broader range of downstream tasks and providing truthful answers. Specifically, we construct a large-scale RS image-text dataset called VersaD, featuring captions with rich content. Given the complexity of RS images, it is laborious to generate accurate and rich-content captions for massive images (*e.g.* 1.4M). Motivated by the fact that existing VLMs trained with sparse-content captions, such as LLaVA (Liu et al. 2024), have achieved remarkable performance by pre-training on large-scale, noisy datasets and fine-tuning on smaller, clean data, we explore applying a similar strategy to the case with the rich-content caption. Considering noise tolerance and construction costs, we propose leveraging off-the-shelf models to generate rich-content captions. Notably, in our prompt design, we emphasize the inclusion of metadata (*e.g.* modality, resolution), object attributes (*e.g.* semantic category, material), and scene context (*e.g.* spatial layout, scene category). These contribute to the versatility of our VHM (see Table 2). Furthermore, we customize the answer format of the off-the-

shelf model and impose constraints on uncertain object descriptions to ensure accuracy in its answers.

To enable and verify VHM's honesty, we create an RS-specific honest dataset, dubbed HnstD. The sample in the HnstD dataset consists of an RS image paired with a question and an answer. The question addresses the relative positions of objects and their attributes such as presence, color, and absolute position. Furthermore, each type of question (except those regarding presence) is divided into factual and deceptive categories. For example, asking the color of an existing object versus that of a non-existent one. Based on HnstD, we endow VHM with honesty by using it as an additional instruction dataset for fine-tuning. Additionally, we introduce a new task, honest question answering to compare the honesty of different models.

Using the proposed datasets, we develop a novel VLM for RS image analysis, VHM, by implementing a two-stage training strategy and exploring the integration of multi-level visual representations. Comparing VHM with recent RS-specific VLMs, it is able to perform more downstream tasks such as building vectorizing and multi-label classification. For typical RS image understanding tasks, VHM achieves state-of-the-art performance across multiple RS datasets. Additionally, VHM offers insights into the honesty of VLMs, a crucial aspect in RS applications.

In summary, our paper made the following contributions:

- We construct VersaD, a large-scale RS image-text dataset featuring captions with rich content. This dataset facilitates VLMs in achieving a thorough understanding of RS images, enhancing their versatility.
- We create HnstD, an RS-specific honest dataset comprising questions with factual and deceptive categories. By utilizing it as an additional instruction dataset, VLMs are endowed with honesty.
- Building upon our datasets, we develop VMH, a versatile and honest VLM tailored for RS image analysis. VHM explores the integration of multi-level visual representations, showcasing its capability to perform numerous downstream tasks and achieve superior performance across multiple common RS image understanding tasks.

2 Versatile and Honest Datasets

2.1 VersaD

VersaD construction. We construct VersaD with the goal of incorporating a wide range of RS visual knowledge into VLMs. To achieve this, we collect several open-source RS datasets acquired from various geographic locations with different imaging sensors and conditions, ensuring diversity in ground objects and richness in images. The datasets include MillionAID (Long et al. 2021), CrowdAI (CrowdAI 2018), fMoW (Christie et al. 2018), CVUSA (Workman, Souvenir, and Jacobs 2015), CVACT (Liu and Li 2019), and LoveDA (Wang et al. 2021). For dataset statistics, please refer to Appendix B.1. VersaD includes nearly 1.4 million RS images with spatial resolutions ranging from 0.08 to 153 meters per pixel. Considering noise tolerance and construction costs, we choose an off-the-shelf model, namely Gemini-Vision (Team et al. 2023), to generate captions for these 1.4 million images. To make the caption rich in content, we carefully design prompts, as detailed in Appendix B.1. The prompt emphasizes information about image properties, object attributes, and scene context. Additionally, it constrains the description of uncertain objects and the answer format to improve accuracy in Gemini-Vision's answers. Ultimately, we produce VersaD, a large-scale RS image-text dataset featuring captions with rich content. Several examples of VersaD are provided in Appendix B.2.

VersaD quality assessment. Since automatic caption generation may be inaccurate, we randomly sample 315 image-text pairs for quality assessment through manual inspection. Specifically, image captions in VersaD have rich content and multiple sentences. Therefore, we divide the captions sentence by sentence, resulting in 2002 sentences. Each sentence may contain multiple key pieces of information simultaneously, such as the relative position between objects and their attributes. Thus, we further divide sentences based on pieces of information and then combine the corresponding images to check each piece, categorizing the sentences into three groups: completely accurate, completely inaccurate, and partially accurate. Completely accurate means that all pieces of information within the sentence are entirely without error, whereas the opposite would be completely inaccurate. Partially accurate refers to a sentence that contains both accurate and inaccurate pieces of information. Statistically, 73%, 10%, and 17% of the sentences from the 315 image-text pairs are completely accurate, completely inaccurate, and partially accurate, respectively. Among these partially accurate sentences, 55% of the information pieces are accurate. More details about the manual inspection can be found in Appendix B.3.

Consequently, the overall accuracy of VersaD stands at 82.3%, surpassing that of CC3M (79%) (Sharma et al. 2018), which serves as the pretraining dataset for LLaVA. Prior studies on the construction of RS image-text datasets usually present strategies or rules to filter out image-text pairs due to their sparse content and succinct sentences. However, our captions may include all three categories of sentences, posing challenges for filtering. Experimentally, we observe that VLMs pre-trained on VersaD outperform those trained on datasets characterized by sparse-content and accurate captions, such as SkyScript with 96.1% accuracy (Wang et al. 2024), as shown in Table 8. Thus, we infer that rich content in captions can make up for the presence of noise.

VersaD-Instruct construction. Following LLaVA (Liu et al. 2024), we create VersaD-Instruct for fine-tuning VLMs. To ensure that the generated conversations contain crucial information, such as the location and quantity of objects within the images, three object detection datasets: DOTA-v2 (Xia et al. 2018), Fair1M (Sun et al. 2021) and DIOR (Li et al. 2020), are used as the image sources for VersaD-Instruct. We randomly sample 30K RS images and then generate rich-content captions for them using Gemini-Vision. Based on these rich-content captions and bounding-box annotations, we prompt language-only Gemini to generate multi-turn conversation and reasoning data. This process

results in VersaD-Instruct, comprising 30K RS images, with 26K images dedicated to conversation and 4K to complex reasoning. For details about the prompts and in-context examples, as well as examples of VersaD-Instruct, please refer to Appendix B.4.

2.2 HnstD

HnstD construction. HnstD is an additional instruction dataset developed to make VLMs honest for RS image analysis. Each sample in HnstD consists of an RS image paired with a single-turn conversation. Different from existing RS instruction datasets (Kuckreja et al. 2024; Hu et al. 2023; Zhan, Xiong, and Yuan 2024) that contain only factual questions, HnstD includes both factual and deceptive questions. This design aims to prevent VLMs from producing affirmative answers to unreasonable queries posed by users. Based on DOTA-v2 (Xia et al. 2018) and Fair1M (Sun et al. 2021), HnstD is designed with four recognition tasks: the relative position between objects, their presence, color, and absolute position. As illustrated in Table. 1, except for the presence task, all other tasks have both factual and deceptive questions. Particularly, deceptive questions about object color arise from either the non-existence of objects or their presence in panchromatic images, while those concerning relative and absolute positions stem from the non-existence of objects. In terms of question format, we use yes or no questions for the presence task, open-ended questions for the color task, and single-choice questions with five candidate answers for the relative position and absolute position tasks. Detailed explanations of each task's construction can be found in Appendix C.1, with examples of HnstD provided in Appendix C.2. HnstD comprises over 45K question-answer pairs in total. Every question and answer in HnstD is manually checked to ensure reliability.

Task	QFmt	#Train Sample (Fact. / Dec.)	#Test Sample (Fact. / Dec.)
Presence	Y/N	8,000 / -	242 / -
Color	OE	8,000 / 4,000+1,000	200 / 300+100
Absolute Position	SC	8,000 / 4,000	100 / 300
Relative Position	SC	8,000 / 4,000	100 / 300

Table 1: Information of HnstD dataset. *QFmt* refers to the question format, covering Y/N (Yes-or-No), OE (Open-End), and SC (Single-Choice). *Fact.*, and *Dec.* stand for factual question, and deceptive question, respectively. For the color task, 4,000/300 and 1,000/100 are the number of deceptive questions stemming from the non-existence of objects and their presence in panchromatic images.

Honest question answering. Based on HnstD, we introduce a new task, namely honest question answering, to compare the honesty of VLMs. For quantitative evaluation, we utilize the matching strategy to calculate the accuracy of the presence, relative position, and absolute position tasks. Since the relative position and absolute position tasks have two categories of questions, their accuracy (Acc) is the average accuracy of factual questions (Acc_{fact}) and deceptive questions (Acc_{dec}):

$$Acc = (Acc_{\text{fact}} + Acc_{\text{dec}})/2.0, \tag{1}$$

where Acc_{fact} (Acc_{dec}) is expressed as the ratio of the number of correctly answered questions to the total number of factual (deceptive) questions. For the color task, we employ the matching strategy and ChatGPT-3.5 API to evaluate factual and deceptive questions, respectively. For the prompts used with ChatGPT-3.5, please refer to Appendix C.3. Since deceptive questions of the color task have two causes, the accuracy of this task is calculated as follows:

$$Acc = (Acc_{\text{fact}} + (Acc^{\text{ex}}_{\text{dec}} + Acc^{\text{pan}}_{\text{dec}})/2.0)/2.0, \tag{2}$$

where $Acc^{\text{ex}}_{\text{dec}}$ and $Acc^{\text{pan}}_{\text{dec}}$ are the accuracy of deceptive questions stemming from the non-existence of objects and their presence in panchromatic images, respectively.

3 Versatile and Honest VLM

3.1 Model Architecture

Similar to LLaVA (Liu et al. 2024), our VHM consists of three main components: a vision encoder, a projection layer, and a Large Language Model (LLM), as shown in Figure 2. The vision encoder is responsible for compressing RS images into more compact visual representations. The LLM receives both visual and textual information to perform reasoning tasks. Since LLMs are limited to text perception, the projection layer is introduced to bridge the modality gap between natural language and images.

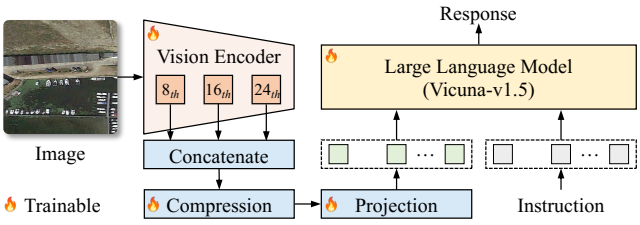


Figure 2: Architecture of the proposed VHM.

In our implementation, we choose the pre-trained CLIP-Large (Radford et al. 2021) with an input resolution of 336×336 and a patch size of 14 for encoding images, as CLIP is trained to align image and text representations. Prior studies (Kuckreja et al. 2024; Li et al. 2024a) usually use the final image feature for subsequent processes. However, several studies (Wang et al. 2022b; Miao et al. 2022) have highlighted that low-level features contain spatial information and contribute to the localization of objects. Therefore, we integrate multi-level features to obtain a comprehensive visual representation. Specifically, image tokens from the 8th, 16th, and 24th Transformer layers are concatenated along the channel dimension and then fed into a compression layer, followed by the projection layer, resulting in a sequence of visual tokens. The compression and projection layers consist of a linear layer and two-layer multi-layer perceptions, respectively. For the LLM, we select Vicuna-v1.5 (Chiang et al. 2023) with 7B parameters due to its excellent instruction-following capabilities in language tasks.

Method	VQA	VG	IC	SC	OC	IM	IR	OR	GM	BV	MC	HQA	Others
RSGPT (Hu et al. 2023)	✓		✓										
GeoChat (Kuckreja et al. 2024)	✓	✓	✓	✓									RC
SkyEyeGPT (Zhan, Xiong, and Yuan 2024)	✓	✓	✓	✓	✓								VC, REG
EarthGPT (Zhang et al. 2024)	✓	✓	✓	✓	✓								RC, OD
LHRS-Bot (Muhtar et al. 2024)	✓	✓	✓	✓	✓	✓	✓	✓					
VHM (ours)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	

Table 2: Capability comparisons of RS-specific VLMs. The *VQA*, *VG*, *IC*, *SC*, *OC*, *IM*, *IR*, *OR*, *GM*, *BV*, *MC*, *HQA*, *RC*, *VC*, *OD* and *REG* are short for Visual Question Answering, Visual Grounding, Image Captioning, Scene Classification, Object Counting, Image Modality, Image Resolution, Object Recognition, Geometric Measurement, Building Vectorizing, Multi-label Classification and Honest Question Answering, Region Caption, Video Caption, Object Detection and Referring Expression Generation, respectively. Examples of our model's versatility refer to Appendix E.2.

3.2 Training Strategy

We employ a two-stage strategy for training VHM. Specifically, starting with the pre-trained weights of LLaVA, we optimize all components of VHM using the large-scale VersaD dataset to incorporate RS visual knowledge into the model. Utilizing 16 NVIDIA A100-80G GPUs, we pre-train VHM with a batch size of 256 for 1 epoch (approximately 5400 iterations). We apply an initial learning rate of 2e-5, following a cosine scheduler learning rate strategy.

Subsequently, we use three instruction datasets, VersaD-Instruct, VariousRS-Instruct, and HnstD, for supervised fine-tuning (SFT) of the compression layer, projection layer, and LLM of VHM. VersaD-Instruct contains multi-turn conversation and complex reasoning data, aiming to enable VHM to engage in dialogue with users. VariousRS-Instruct comprises RS images paired with single-turn conversations, enabling VHM to perform specific RS image analysis tasks. Detailed information about the VariousRS-Instruct dataset can be found in Appendix D. Since visual question answering, visual grounding, and scene classification are common capabilities of existing VLMs, we, following comparative methods, build training sets for fine-tuning to ensure fairness. For the unique capabilities of our VHM, such as building vectorizing and multi-label classification, we create corresponding test sets to evaluate its performance quantitatively. HnstD is used to endow VHM with honesty. During the SFT process, only the vision encoder is frozen to maintain generalization. Based on 8 NVIDIA A100-80G GPUs, we set a batch size of 128 for 1 epoch (nearly 1400 interactions). The learning rate setting remains the same as previously mentioned.

4 Experiments

4.1 Datasets

We employ multiple RS datasets to conduct comparison experiments across tasks supported by existing VLMs. These include five scene classification datasets (NWPU, METER-ML, SIRI-WHU, AID, WHU-RS19), two visual question answering datasets (RSVQA-LR and RSVQA-HR), and one visual grounding dataset (DIOR-RSVG). To evaluate VHM-specific capabilities, such as honest question answering and building vectorizing, we use the test sets of HnstD and VariousRS-Instruct. Evaluation metrics for each task are detailed in Appendix E.1.

4.2 Evaluation on Versatility

VHM-specific capabilities. In Table. 2, we list the capabilities exhibited by VLMs tailored for RS image analysis. Notably, VHM can perform more tasks, such as building vectorizing and multi-label classification, which are crucial for natural resource monitoring. Since the tasks related to VHM-specific capabilities involve open-ended questions that are not supported by competitors, we only evaluate VHM's performance quantitatively using the test set from VariousRS-instruct. As shown in Table. 3, VHM excels in image attribute recognition, achieving a 95% accuracy on the image modality task and a mean absolute error of 0.24 on the image resolution task. For building vectorizing and zero-shot multi-label classification tasks, VHM demonstrates competent performance. However, it faces challenges in accurately counting objects and measuring geometric properties, evidenced by higher mean absolute errors of 6.75 and 12.82, respectively. Overall, these results confirm the potential of VLMs to facilitate more RS image analysis tasks.

Task	Metric	Score
Object Counting	mean absolute error↓	6.75
Image Modality	accuracy↑	95.00%
Image Resolution	mean absolute error↓	0.24
Geometric Measurement	mean absolute error↓	12.82
Building Vectorizing	complexity-aware IoU↑	71.25%
Multi-label Classification	F_1-measure↑	51.87%

Table 3: Performance of VHM on specific tasks.

VLM-common capabilities. To further verify the versatility of our model, we compare VHM with various VLMs on common RS image analysis tasks, as shown in Table. 4 to 7. From Table. 4, it is evident that VLMs pre-trained on large-scale RS image-text datasets outperform generic VLMs by a large margin, confirming the importance of accounting for domain shift between natural and RS images. While LHRS-Bot performs excellently, VHM is able to obtain a further boost in both fully supervised and zero-shot

settings[1]. The improvements on the NWPU and SIRI-WHU datasets (10.6% and 8.22%) are remarkable, demonstrating VHM's effective acquisition of RS visual knowledge and strong generalization ability.

Method	NWPU	METER-ML	SIRI-WHU	AID	WHU-RS19	Avg.
LLaVA-1.5	34.96	21.73	17.71	31.10	54.55	32.01
MiniGPTv2	28.15	14.29	35.46	32.96	64.80	35.13
Qwen-VL-Chat	42.73	38.77	54.58	55.30	72.25	52.73
Gemini-Vision	66.89	23.36	60.46	66.43	68.70	57.16
LHRS-Bot	83.94	69.81	62.66	91.26	93.17	80.17
VHM (ours)	**94.54**	**72.74**	**70.88**	**91.70**	**95.80**	**85.13**

Table 4: Performance of VLMs on various scene classification datasets.

In Table. 5 and 6, we report the results of different VLMs on the visual question answering task under both fully supervised and zero-shot settings. With the same training setting (one epoch and 10K samples), VHM achieves supervised results comparable to LHRS-Bot's (89.33% vs 89.19%)[2]. In the zero-shot setting, VHM stands out as the top-performing model, exceeding the sub-optimal EarthGPT by 2.6% in average performance. This demonstrates that our VersaD, which contains varied-resolution RS images, contributes to incorporating RS visual knowledge at different scales into VLMs. As a result, VHM excels in handling high-resolution RS images that were not encountered during training.

Method	LR-rural	LR-presence	LR-compare	Avg.
Gemini-Vision	63.00	60.95	70.32	64.76
RSGPT	**94.00**	**91.17**	**91.70**	**92.29**
GeoChat	94.00	91.09	90.33	91.81
SkyEyeGPT	75.00	88.93	88.63	84.19
LHRS-Bot	89.07	88.51	90.00	89.19
VHM (ours)	88.00	90.11	89.89	89.33

Table 5: Performance of VLMs on the visual question answering dataset (RSVQA-LR).

Method	HR-presence	HR-compare	Avg.
Gemini-Vision	63.60	64.60	64.10
LLaVA-1.5	**69.83**	67.29	68.56
MiniGPTv2	40.79	50.91	45.85
Qwen-VL-Chat	66.44	60.41	63.43
GeoChat	58.45	83.19	70.82
EarthGPT	62.77	79.53	71.15
VHM (ours)	64.00	**83.50**	**73.75**

Table 6: Performance of VLMs on the visual question answering dataset (RSVQA-HR).

[1]The instruction datasets of LHRS-Bot and VHM include only the NWPU and METER-ML datasets. All other datasets are absent.

[2]RSGPT is fine-tuned for five epochs on the RSVQA dataset and GeoChat is trained with 50K training samples.

Table. 7 presents the results of the visual grounding task on the DIOR-RSVG dataset, using an intersection over union (IoU) threshold of 0.5 as the evaluation metric. The original size of images in this dataset is 800×800. To adapt to VLMs' input, images are usually downsampled. The higher the downsampling ratio, the smaller the object size and the fewer its visual features, posing challenges for accurately locating objects. Nevertheless, our VHM surpasses the best competitor by 11.59, even with the smallest input size of 336×336. This advantage stems from our design of leveraging spatial information within low-level features.

Method	Input Size	DIOR-RSVG
CogVLM*	490×490	44.58
Qwen-VL-Chat	448×448	31.86
VHM (ours)	336×336	**56.17**

Table 7: Performance of VLMs on the visual grounding dataset (DIOR-RSVG). * indicates use of the cogvlm-grounding-generalist version.

Qualitative results. In Appendix E.2, we present examples of conversations between users and VHM. These examples highlight VHM's ability to provide detailed descriptions of objects, scene, and their attributes, such as color, shape, and layout, showcasing its comprehensive understanding of input RS images. Consequently, VHM effectively performs RS image analysis tasks, including object counting, relative position recognition, and image resolution estimation, while producing honest answers.

4.3 Evaluation on Honesty

Honesty is a significant property for VLMs in RS applications such as national defense security. Based on the HnstD dataset, we compare VHM with various VLMs on the honest question answering task, as shown in Table. 9. For the presence task, which includes only factual questions, VHM outperforms all competitors by a large margin with a gap of 10.35%, thanks to the incorporation of RS visual knowledge. In other tasks, competitors generally achieve higher accuracy on factual questions compared to deceptive questions, indicating a tendency toward dishonesty. Notably, CogVLM (Wang et al. 2023) gives excellent results on deceptive questions of the color task, but at the cost of a significant accuracy drop on corresponding factual questions. In contrast, VHM consistently gets good accuracy on both factual and deceptive questions, particularly in the color and absolute position tasks. While VHM's accuracy on the relative position task is remarkable, there is still room for improvement. To encourage further research on the honesty of VLMs in the RS domain, we will release the HnstD dataset.

4.4 Ablation Studies

Training strategy. We compare the model additionally pretrained on VersaD with one that directly adopts the pretrained weights of LLaVA. Both models are fine-tuned on our VariousRS-Instruct and HnstD. To ensure fairness, we

w/ RS Pretraining	w/ Multi-level Vis. Rep.	Pretraining Dataset	Scene Classification	Visual Question Answering	Visual Grounding	Honest Question Answering
✗	✗	-	73.22	77.19	28.23	61.65
✓	✗	VersaD (ours)	84.71	82.09	51.06	**80.50**
✓	✓	RS5M-5M (Zhang et al. 2023)	83.38	82.22	48.71	75.45
✓	✓	RS5M-1.5M (Zhang et al. 2023)	83.03	82.53	48.08	75.84
✓	✓	SkyScript (Wang et al. 2024)	80.20	**83.32**	16.01	65.38
✓	✓	VersaD (ours)	**84.79**	81.18	**59.52**	78.59

Table 8: Comparisons with different training strategies and pretraining datasets. *Vis. Rep.* stands for Visual Representation. The performance of scene classification and visual question answering is the average accuracy across multiple datasets.

Method	Presence	Color		Absolute Position		Relative Position	
	Fact.	Fact.	Dec.-Ex / Pan	Fact.	Dec.	Fact.	Dec.
LLaVA-1.5	70.40	66.96	23.33 / 42.00	61.61	12.00	34.71	31.67
CogVLM	74.71	31.25	68.00 / **100.00**	33.93	16.67	29.34	11.00
Qwen-VL-Chat	72.99	47.62	8.33 / 39.00	54.29	29.52	31.79	28.91
VHM (ours)	**85.06**	**81.50**	**93.33** / 93.00	**76.79**	**90.67**	**47.52**	**87.67**

Table 9: Performance of VLMs on the honest question answering task (HnstD). *Fact.* and *Dec.* stand for factual questions and deceptive questions. The suffixes *Ex* and *Pan* refer to deceptive questions stemming from the non-existence of objects and their presence in panchromatic images.

Method	SC	VQA	VG	HQA
Single level	**85.47**	**81.74**	46.96	79.03
Multi level	85.13	81.54	**56.17**	**80.93**

Table 10: Performance with different architectures. *SC*, *VQA*, *VG* and *HQA* are short for Scene Classification, Visual Question Answering, Visual Grounding and Honest Question Answering, respectively.

use the same model architecture for both, as LLaVA only employs single-level visual representation. As shown in Table. 8, the model pretrained with RS data significantly outperforms the baseline across multiple tasks. This confirms the importance of incorporating RS visual knowledge by pertaining with large-scale RS image-text datasets.

Rich-content caption *v.s.* sparse-content caption. Currently, several large-scale RS image-text datasets are available for pretraining, such as RS5M-5M (Zhang et al. 2023) and SkyScript (Wang et al. 2024). We conduct experiments to evaluate the superiority of our VersaD, and the results are presented in Table. 8. Since VersaD-Instruct is created with rich-content captions, all models are fine-tuned only on VariousRS-Instruct and HnstD. We observe that the model pretrained using VersaD significantly outperforms all baselines across various tasks, particularly in object grounding. This is due to VersaD captions containing diverse objects and their attributes, unlike the captions in existing datasets that merely mention prominent objects. Furthermore, an interesting comparison arises between models pretrained on VerasD and SkyScript. Despite SkyScript having a similar number of image-text pairs as VersaD (1.5M for SkyScript vs 1.4M for VersaD) and higher caption accuracy (96.1% vs 82.3%), the model pretrained using VersaD shows an improvement of over 43% on the visual grounding task. Therefore, we infer that rich-content captions are vital for the performance of VLMs and contribute to VLMs being less sensitive to label noise.

Multi level *v.s.* single level. To assess the impact of using multi-level visual representation, we compare it with a model that uses single-level image features from the final Transformer layer of the vision encoder. Both models are optimized on VersaD, VersaD-Instruct, VariousRS-Instruct, and HnstD. The results, presented in Table. 10, show that integrating multi-level visual representation significantly improves the accuracy of object localization, gaining over 9% on the visual grounding task. This clearly demonstrates the necessity of leveraging detailed spatial information within low-level image features.

5 Conclusion

In this paper, we create a large-scale remote sensing image-text dataset with rich-content captions, and an honest instruction dataset comprising factual and deceptive questions, aiming to endow vision language models with versatility and honesty for remote sensing image analysis. Building upon these datasets, we develop VHM by implementing a two-stage training strategy and integrating multi-level visual representations. VHM achieves state-of-the-art performance on various public datasets across multiple common remote sensing tasks. Additionally, it demonstrates the potential of vision language models to facilitate more remote sensing tasks and offers insights into the honesty of models, which is crucial in applications such as national defense security. while VHM shows superiority in several remote sensing image analysis tasks, it currently lacks the capability for pixel-wise perception, preventing it from performing semantic segmentation or change detection of remote sensing images. Addressing this limitation is a key to future research.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (grants No. 62325111 and No. U22B2011), the Shanghai Artificial Intelligence Laboratory, and the National Key R&D Program of China (2021YFB3900503).

References

Bazi, Y.; Bashmal, L.; Al Rahhal, M. M.; Ricci, R.; and Melgani, F. 2024. Rs-llava: A large vision-language model for joint captioning and question answering in remote sensing imagery. *Remote Sens.*, 16(9): 1477.

Chiang, W.-L.; Li, Z.; Lin, Z.; Sheng, Y.; Wu, Z.; Zhang, H.; Zheng, L.; Zhuang, S.; Zhuang, Y.; Gonzalez, J. E.; Stoica, I.; and Xing, E. P. 2023. Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality.

Christie, G.; Fendley, N.; Wilson, J.; and Mukherjee, R. 2018. Functional map of the world. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 6172–6180.

CrowdAI. 2018. Crowdai mapping challenge. https://www.crowdai.org/challenges/mapping-challenge. Accessed on: 2021-02-26.

Hu, Y.; Yuan, J.; Wen, C.; Lu, X.; and Li, X. 2023. Rsgpt: A remote sensing vision language model and benchmark. *arXiv:2307.15266*.

Kuckreja, K.; Danish, M. S.; Naseer, M.; Das, A.; Khan, S.; and Khan, F. S. 2024. Geochat: Grounded large vision-language model for remote sensing. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 27831–27840.

Li, C.; Wong, C.; Zhang, S.; Usuyama, N.; Liu, H.; Yang, J.; Naumann, T.; Poon, H.; and Gao, J. 2024a. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36.

Li, K.; Wan, G.; Cheng, G.; Meng, L.; and Han, J. 2020. Object detection in optical remote sensing images: A survey and a new benchmark. *ISPRS J. Photogramm. Remote Sens.*, 159: 296–307.

Li, X.; Wen, C.; Hu, Y.; Yuan, Z.; and Zhu, X. X. 2024b. Vision-language models in remote sensing: Current progress and future trends. *IEEE Geosci. Remote Sens. Mag.*, 12: 32–66.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2024. Visual instruction tuning. In *Adv. Neural Inf. Process. Syst.*

Liu, L.; and Li, H. 2019. Lending orientation to neural networks for cross-view geo-localization. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 5624–5633.

Long, Y.; Xia, G.-S.; Li, S.; Yang, W.; Yang, M. Y.; Zhu, X. X.; Zhang, L.; and Li, D. 2021. On creating benchmark dataset for aerial image interpretation: Reviews, guidances, and million-aid. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, 14: 4205–4230.

Luo, J.; Pang, Z.; Zhang, Y.; Wang, T.; Wang, L.; Dang, B.; Lao, J.; Wang, J.; Chen, J.; Tan, Y.; et al. 2024. Skysensegpt: A fine-grained instruction tuning dataset and model for remote sensing vision-language understanding. *arXiv:2406.10100*.

Miao, S.; Xia, M.; Qian, M.; Zhang, Y.; Liu, J.; and Lin, H. 2022. Cloud/shadow segmentation based on multi-level feature enhanced network for remote sensing imagery. *Int. J. Remote Sens.*, 43(15-16): 5940–5960.

Muhtar, D.; Li, Z.; Gu, F.; Zhang, X.; and Xiao, P. 2024. LHRS-Bot: Empowering Remote Sensing with VGI-Enhanced Large Multimodal Language Model. *arXiv:2402.02544*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *Int. Conf. Mach. Learn.*, 8748–8763. PMLR.

Sharma, P.; Ding, N.; Goodman, S.; and Soricut, R. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Annu. Meet. Assoc. Comput. Linguist.*, 2556–2565.

Sun, X.; Wang, P.; Yan, Z.; Xu, F.; Wang, R.; Diao, W.; Chen, J.; Li, J.; Feng, Y.; Xu, T.; Weinmann, M.; Hinz, S.; Wang, C.; and Fu, K. 2021. FAIR1M: A Benchmark Dataset for Fine-grained Object Recognition in High-Resolution Remote Sensing Imagery. arXiv:2103.05569.

Team, G.; Anil, R.; Borgeaud, S.; Wu, Y.; Alayrac, J.-B.; Yu, J.; Soricut, R.; Schalkwyk, J.; Dai, A. M.; Hauth, A.; et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Wang, D.; Zhang, J.; Du, B.; Xia, G.-S.; and Tao, D. 2022a. An empirical study of remote sensing pretraining. *IEEE Trans. Geosci. Remote Sens.*, 61: 1–20.

Wang, G.; Zhang, N.; Liu, W.; Chen, H.; and Xie, Y. 2022b. MFST: A multi-level fusion network for remote sensing scene classification. *IEEE Geoscience and Remote Sensing Letters*, 19: 1–5.

Wang, J.; Zheng, Z.; Ma, A.; Lu, X.; and Zhong, Y. 2021. LoveDA: A remote sensing land-cover dataset for domain adaptive semantic segmentation. *arXiv:2110.08733*.

Wang, W.; Lv, Q.; Yu, W.; Hong, W.; Qi, J.; Wang, Y.; Ji, J.; Yang, Z.; Zhao, L.; Song, X.; et al. 2023. Cogvlm: Visual expert for pretrained language models. *arXiv:2311.03079*.

Wang, Z.; Prabha, R.; Huang, T.; Wu, J.; and Rajagopal, R. 2024. Skyscript: A large and semantically diverse vision-language dataset for remote sensing. In *AAAI Conf. Artif. Intell.*, 5805–5813.

Workman, S.; Souvenir, R.; and Jacobs, N. 2015. Wide-area image geolocalization with aerial reference imagery. In *IEEE Int. Conf. Comput. Vis.*, 3961–3969.

Xia, G.-S.; Bai, X.; Ding, J.; Zhu, Z.; Belongie, S.; Luo, J.; Datcu, M.; Pelillo, M.; and Zhang, L. 2018. DOTA: A large-scale dataset for object detection in aerial images. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 3974–3983.

Zhan, Y.; Xiong, Z.; and Yuan, Y. 2024. Skyeyegpt: Unifying remote sensing vision-language tasks via instruction tuning with large language model. *arXiv:2401.09712*.

Zhang, W.; Cai, M.; Zhang, T.; Zhuang, Y.; and Mao, X. 2024. Earthgpt: A universal multi-modal large language model for multi-sensor image comprehension in remote sensing domain. *IEEE Trans. Geosci. Remote Sens.*, 62: 1–20.

Zhang, Z.; Zhao, T.; Guo, Y.; and Yin, J. 2023. Rs5m: A large scale vision-language dataset for remote sensing vision-language foundation model. *arXiv:2306.11300*.