Towards Comprehensive Interactive Change Understanding in Remote Sensing: A Large-scale Dataset and Dual-granularity Enhanced VLM

Junxiao Xue, Quan Deng, Xuecheng Wu, Kelu Yao, Xinyi Yin, Fei Yu
Wei Zhou, Yanfei Zhong, Yang Liu, and Dingkang Yang

***Abstract*—Remote sensing change understanding (RSCU) is essential for analyzing remote sensing images and understanding how human activities affect the environment. However, existing datasets lack deep understanding and interactions in the diverse change captioning, counting, and localization tasks. To tackle these gaps, we construct ChangeIMTI, a new large-scale interactive multi-task instruction dataset that encompasses four complementary tasks including change captioning, binary change classification, change counting, and change localization. Building upon this new dataset, we further design a novel vision-guided vision-language model (ChangeVG) with dual-granularity awareness for bi-temporal remote sensing images (*i.e.*, two remote sensing images of the same area at different times). The introduced vision-guided module is a dual-branch architecture that synergistically combines fine-grained spatial feature extraction with high-level semantic summarization. These enriched representations further serve as the auxiliary prompts to guide large vision-language models (VLMs) (*e.g.*, Qwen2.5-VL-7B) during instruction tuning, thereby facilitating the hierarchical cross-modal learning. We extensively conduct experiments across four tasks to demonstrate the superiority of our approach. Remarkably, on the change captioning task, our method outperforms the strongest method Semantic-CC by 1.39 points on the comprehensive S_m^* metric, which integrates the semantic similarity and descriptive accuracy to provide an overall evaluation of change caption. Moreover, we also perform a series of ablation studies to examine the critical components of our method. The source code and associated data for this work are publicly available at Github.**

***Index Terms*—Remote sensing change understanding, Dataset, Large vision-language models, Change captioning.**

This research was supported by the Key R&D Program of Zhejiang under Grant No. 2024C01036 and the Zhejiang Provincial Natural Science Foundation of China under Grant No. LQ23F030009.

Junxiao Xue and Kelu Yao are currently with the Research Center for Space Computing System, Zhejiang Lab, Hangzhou, 311100, China (E-mail: xuejx@zhejianglab.cn);

Quan Deng is with the Hangzhou Institute for Advanced Study, University of Chinese Academy of Sciences, Hangzhou, 310024, China. (E-mail: dengquan23@mails.ucas.ac.cn);

Xuecheng Wu is currently with the School of Computer Science and Technology, Xi'an Jiaotong University, Xi'an, 710049, China. (E-mail: wuxc3@stu.xjtu.edu.cn);

Xinyi Yin is with the School of Cyber Science and Engineering, Zhengzhou University, Zhengzhou, 450002, China (E-mail: yinxinyi@stu.zzu.edu.cn);

Fei Yu is with Liaoning University of Technology, Jinzhou, 123099, China (E-mail: yufei_hitcs@163.com);

Wei Zhou is with the School of Computer Science and Informatics, Cardiff University, Cardiff, CF24 4AG, United Kingdom. (E-mail: zhouw26@cardiff.ac.uk);

Yanfei Zhong is with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing (LIESMARS), Wuhan University, Wuhan, 430072, China. (E-mail: zhongyanfei@whu.edu.cn);

Yang Liu is with the College of Electronic and Information Engineering, Tongji University, Shanghai, 201804, China. (E-mails: yang_liu@ieee.org).

Dingkang Yang is with the College of Intelligent Robotics and Advanced Manufacturing, Fudan University & Fysics AI, Shanghai, 200433, China (E-mail: dkyang20@fudan.edu.cn);

Junxiao Xue, Quan Deng, and Xuecheng Wu deserve equal contributions.

Work done during Quan Deng's research internship at Zhejiang Lab.

Corresponding authors: Dingkang Yang & Yang Liu.

I. Introduction

In recent years, remote sensing change understanding (RSCU) techniques have emerged as powerful tools for monitoring land cover dynamics [1], urban expansion [2], environmental degradation [3], and disaster management [4]. A key focus of these techniques has been change detection, which aims to identify differences between bi-temporal (captured at two different time points) remote sensing images.

Despite significant progress in remote sensing change detection in recent years, particularly in pixel-level and region-level analysis using widely adopted deep learning architectures such as Siamese networks, encoder-decoder frameworks, and semantic segmentation models [5], [6], [7], existing methods remain primarily focused on identifying the spatial location and extent of changes. However, they are generally unable to generate contextual or semantic descriptions of the detected changes effectively [8], [9]. Moreover, these approaches generally provide limited support for interactive exploration, making it difficult for users to query or interpret changes in a flexible, task-driven manner. Consequently, there is a growing need to move beyond localization-oriented change detection toward comprehensive change understanding and natural language captioning. Describing changes through natural language can facilitate human-understandable interpretation and enable higher-level reasoning in remote sensing analysis.

With the rapid developments of VLMs, these models have demonstrated exceptional capabilities in joint vision-language understanding and reasoning. They have been widely applied in various tasks involving natural images, such as image captioning and visual question answering (VQA) [10], [11]. Due to their ability to align visual information with natural language representations effectively, VLMs are emerging as promising tools for addressing semantic understanding challenges in remote sensing images. Recent studies have begun to explore the use of VLMs for change captioning in remote sensing, generating natural language descriptions of differences between image pairs. However, most efforts remain limited to static caption generation and have not fully leveraged the interactive and multi-task reasoning capabilities of VLMs [12]. Such as change counting, spatial localization, and comprehensive captioning remain insufficiently addressed. Therefore, a more unified and comprehensive approach is required, one that moves beyond simple captioning and supports interactive, question driven understanding of changes in the remote sensing images.

In this work, we construct a new large-scale multi-task instruction-tuning dataset for remote sensing change understanding, named ChangeIMTI. The dataset not only includes the task of change captioning, but also extends to a variety of VQA tasks, such as determining whether a change has occurred, counting the number of changed objects, and localizing the regions of change. These tasks are designed to comprehensively enhance the model's semantic understanding and interactive reasoning capabilities in the context of remote sensing images. By jointly training on these complementary tasks, the model is able to develop more robust and generalizable multimodal representations. For instance, binary change classification improves the model's awareness of changes, while change counting enhances sensitivity to object-level differences. The resulting multi-task synergy not only improves individual task performance but also leads to better overall generalization across diverse remote sensing scenarios.

With this dataset, we propose a unified framework that incorporates a vision-guided module to enhance fine-grained spatial and semantic feature extraction. After receiving bi-temporal remote sensing image pairs, the vision-guided module processes them through two parallel branches: one branch is designed to extract fine-grained visual cues such as spatial location and object count, which serve as detailed prompts to improve the model's sensitivity to subtle changes; the other branch focuses on capturing coarse-grained semantic information, such as overall scene-level differences relevant to change captioning. These visual features are subsequently integrated into a VLM, which is fine-tuned on the ChangeIMTI. This design facilitates more interpretable and interactive remote sensing change understanding by tightly coupling detailed visual perception with high-level semantic reasoning.

In conclusion, the main contributions of this paper can be summarized as follows:

- We construct a new large-scale multi-task instruction-tuning dataset denoted ChangeIMTI for RSCU. It covers change captioning, binary change classification, change counting, and change localization tasks, providing a unified testbed for applying VLMs in the remote sensing change domain.
- We propose a unified VLM-based framework for remote sensing change understanding, equipped with a vision-guided module that achieves dual-granularity change perception through simultaneous modeling of fine-grained and global representations. The framework supports both change captioning and VQA, thereby enhancing semantic interpretability and interactive reasoning.
- We have conducted extensive evaluations across multiple downstream remote sensing change understanding tasks and achieve impressive model performance, demonstrating the effectiveness and generalizability of our introduced approach.

II. Related Work

A. Remote Sensing Change Datasets

Remote sensing change datasets can generally be categorized into three types based on the form of change annotation they provide, each serving different research purposes and application scenarios. The first type includes only binary change masks (*e.g.*, LEVIR-CD [13], WHU-CD [14]), which are primarily designed for pixel-level change detection, offering high-resolution satellite or aerial images annotated with binary labels indicating whether a change has occurred. However, these datasets typically lack semantic information about the nature or category of the change, limiting their utility for higher-level interpretation or reasoning tasks. The second type focuses on textual descriptions of scene changes, without providing explicit pixel-level masks. Datasets such as RSICCFormer [15] include natural language annotations that describe how a scene has changed over time, capturing semantic and contextual information. However, they lack fine-grained mask annotations, limiting their effectiveness in tasks that require precise spatial localization of changes. The third type offers both binary change masks and corresponding textual descriptions, enabling multi-level and multi-modal change understanding. A representative example is LEVIR-MCI [8]. Although the dataset incorporates both fine-grained spatial annotations and high-level semantic descriptions, it was not specifically designed or optimized for use with VLMs, limiting its applicability in this emerging research area.

B. Remote Sensing Change Understanding

RSCU requires not only generating captions for change images but also capturing fine-grained differences. A key subtask, Remote Sensing Change Captioning (RSCC), aims to generate natural language descriptions of semantic differences between bi-temporal remote sensing images [8]. Early methods followed image captioning pipelines with CNN encoders and RNN/Transformer decoders, often introducing fusion strategies to highlight change regions [16], [17], [18], [19], [20]. but struggled with subtle variations. More recently, the emergence of VLMs has introduced new paradigms for addressing RSCC tasks. The approach leverages powerful pretrained models such as CLIP for visual representation and GPT-style LLMs for text generation, enabling improved semantic reasoning and generalization, such as Liu *et al.* [21]. However, such two-stage paradigms remain relatively underexplored. Moreover, despite these advancements, existing RSCC models continue to face significant challenges, including the accurate description of small or subtle changes, fine-grained spatial localization, and effective contextual reasoning, particularly in complex or cluttered remote sensing.

In summary, although RSCC has achieved notable progress through the use of deep learning techniques and the availability of curated benchmark datasets, existing approaches remain limited in their ability to deliver human-aligned and semantically rich descriptions. And current research still lacks sufficient exploration into interactive methods for fine-grained change understanding, leaving a critical gap in fully leveraging these models for detailed change analysis. The integration of advanced VLMs presents a promising avenue for addressing these limitations. These models offer the potential to enhance contextual reasoning, align fine-grained visual changes with language outputs, and support more interactive and interpretable remote sensing applications.

C. Vision-language Models in Remote Sensing

VLMs have demonstrated strong potential in enhancing image understanding by integrating visual and linguistic information [22], [23], [24], particularly in remote sensing [25], [26], [27]. Current studies mainly follow two directions: (1) pretraining on large-scale remote sensing image datasets using self-supervised learning, and (2) fine-tuning existing VLMs with small but high-quality remote sensing datasets.

For pretraining, RingMo [28] applied masked image modeling on 2 million remote sensing images to reduce the domain gap with natural images. Other efforts explored CNN–ViT combinations [29] and extended to SAR or spatio-temporal modalities. RemoteCLIP [25] pioneered vision-language alignment in this domain, achieving strong zero-shot transfer on 16 tasks. GRAFT [30] aligned satellite and ground images without textual supervision. On the fine-tuning side, RSGPT [26] instruction-tuned InstructBLIP with curated caption datasets, while RSPrompter [31] and TTP [32] enhanced SAM for remote sensing instance segmentation and change detection.

Despite progress, remote sensing VLMs still face several challenges. Compared to natural images, remote sensing data exhibits greater scale variability, more complex spatial-temporal structures, and sparse semantic cues, making cross-modal alignment particularly difficult, while high computational costs limit deployment. Future research should explore lightweight fine-tuning, self-supervised cross-modal alignment, and domain adaptation to enhance the generalizability, interactivity, and reasoning capacity of VLMs in remote sensing applications.

III. ChangeIMTI

In this section, we first provide a detailed description of the constructed ChangeIMTI dataset, including task formulation, data sources, annotation format, and instruction generation strategy. Subsequently, we provide a detailed description of the construction process for each category of task data. Finally, we present the statistical characteristics of the dataset.

LEVIR-CC [15] represents a large-scale dataset specifically designed for remote sensing change captioning tasks, containing 10,077 pairs of 256×256 satellite images collected from 20 regions in Texas, USA, with temporal spans ranging from 5 to 15 years. The dataset maintains a balanced composition with 5,038 pairs depicting actual land-cover changes and 5,039 pairs showing no changes, each meticulously annotated with five reference captions to capture description variability. On top of this dataset, LEVIR-MCI [8] provides pixel-level change annotations, making it suitable for change detection tasks. While these datasets are valuable for basic change captioning and detection, they are insufficient to fully exploit the capabilities of VLMs, especially in terms of fine-grained semantic reasoning and diverse VQA. To address this limitation, we construct a large-scale, multi-task instruction-tuning dataset based on LEVIR-CC and LEVIR-MCI. The proposed dataset supports a range of tasks including change captioning, binary change classification, change counting, and change localization, aiming to enhance the generalization and interaction capabilities of VLMs in RSCU. As shown in Fig. 1, an example from our dataset illustrates task performance under changed and unchanged conditions.

TABLE I
STATISTICS OF THE CHANGEIMTI INSTRUCTION TUNING DATASET, DETAILING THE NUMBER OF INSTANCES ACROSS DIFFERENT TASK TYPES.

Task Type	Data Number
Change Caption	24,444
Binary Change Classification	8,148
Change Counting	16,296
Change Localization	16,296
Multi-turn Conversation	24,444
Total	89,628

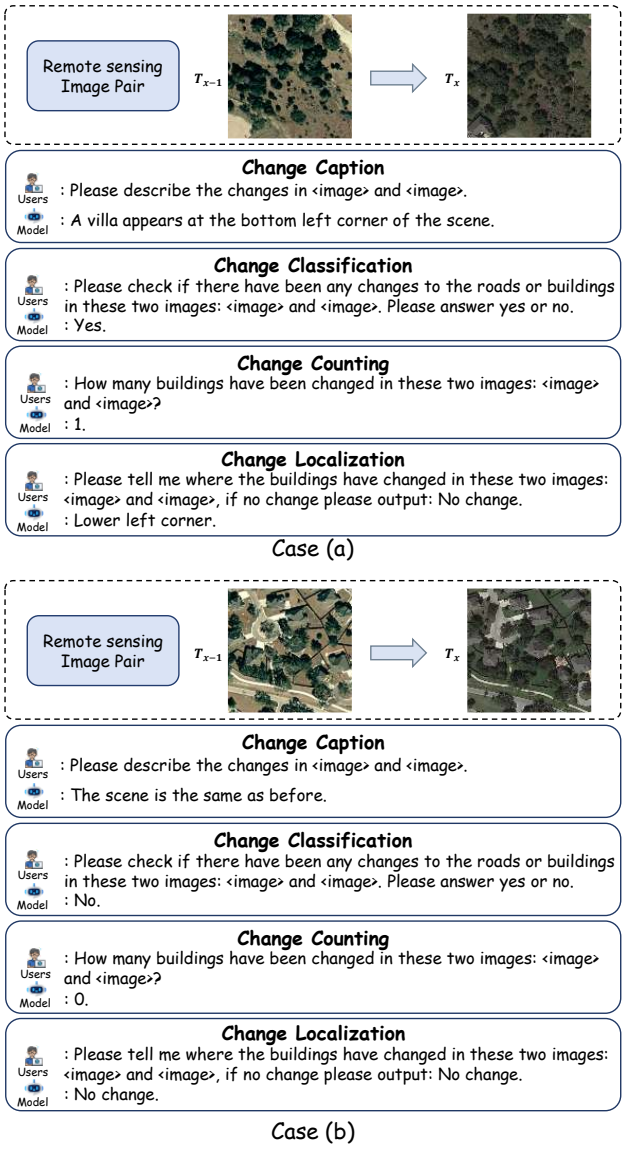

Fig. 1. Examples from the ChangeIMTI dataset illustrating two scenarios. The T_{x-1} and T_x represent images from two different moments in time, with T_{x-1} being the image from the earlier time and T_x being the one from the later time. In Case (a), the image pair contains visible changes. In Case (b), the image pair remains unchanged. The four subtasks (change captioning, binary change classification, change counting, and change localization) demonstrate how changes are described, detected, quantified, and localized, respectively.

A. Change Captioning

Change captioning as the core task, requiring the model to generate natural language descriptions of differences between bi-temporal remote sensing images. To enhance linguistic diversity and improve generalization, we construct five instruction–response pairs for each image pair using the five reference captions from LEVIR-CC. Each instruction is framed to encourage the VLMs to interpret and articulate the semantic differences between the two images.

The task emphasizes the model's ability to understand complex scene changes (*e.g.*, new buildings, demolished roads) and translate them in natural language. Specifically, the user's instruction requires the model to describe the change. The model's output should be a natural-language narrative that delineates this change in detail, including the specific category of the change, the number and locations of the affected elements, and any other pertinent characteristics relevant to the revision. For unchanged image pairs, we construct a single instruction where the expected model response is a negation, such as "the scene is the same as before." helping prevent overfitting toward always generating change descriptions. One example is displayed as follows:

Instruction Template:
User: "Please describe the changes in <image> and <image>."
VLM: [Natural language caption]

B. Binary Change Classification

To introduce decision-making capability, we formulate a binary classification task that determines whether a change has occurred between two images. The task simplifies the complex problem of change detection into a straightforward yes/no decision, making it more manageable for computational models. Specifically, the user's input is a query asking whether any changes have occurred, and the model is expected to respond with only "yes" or "no."

Instruction Template:
User: "Please check if there have been any changes to the roads or buildings in these two images:<image> and <image>. Please answer yes or no."
VLM: "Yes" / "No"

C. Change Counting

Change counting introduces object-level granularity by requiring the model to estimate the number of discrete change regions. To obtain the reference counts, we preprocess the binary change masks using OpenCV's contour detection algorithm, which identifies and counts connected components (*i.e.*, contiguous areas of change). The process ensures that the detected changes are precisely segmented and counted. The task encourages the VLMs to develop instance awareness of changes, going beyond binary classification to quantify scene modifications and identify the specific number of altered regions, thereby enhancing the model's ability to analyze complex scene dynamics. Specifically, the user's query will ask for the number of changes to roads or buildings, and the model's response should be an Arabic numeral providing that count.

Instruction Template:
User: "How many roads[or buildings] have been changed in these two images:<image> and <image>?"
VLM: [Number of detected change regions]

D. Change Localization

To evaluate spatial reasoning, we construct a change localization task that maps each identified change to a coarse-grained spatial region. The task helps the model better understand spatial context by associating changes with specific areas in the image. Specifically, we compute the centroid of each change mask and map it to one of nine predefined regions in a 3×3 grid: top left corner, left, lower left corner, top, center, lower, top right corner, right, and lower right corner. For each image pair, we generate instructions querying the location of changes in specific semantic categories, such as buildings or roads. Specifically, when the user asks about the locations of changes to roads or buildings, the model should respond with one or more directional indicators. The localization process provides additional spatial context, allowing the model to not only detect changes but also infer their positions within the overall scene. The fine-grained localization guides the model's spatial reasoning abilities.

Instruction Template:
User: "Please tell me where the roads[or buildings] have changed in these two images:<image> and <image>, if no change please output: No change."
VLM: [top left corner, left, lower left corner, top, center, lower, top right corner, right, lower right corner](Select a few of them)

E. Multi-turn Dialogue

To simulate real-world user interactions, we synthesize multi-turn conversations that sequentially and seamlessly combine the above tasks into a coherent dialogue. The design follows a structured progression: starting with general change inquiries, the conversations incrementally deepen into specific queries about captioning, classification, counting, and location of changes. This hierarchical approach begins with basic questions, progresses to quantitative analysis, locational identification, and culminates in descriptive analysis. To further align with the stochastic nature of human inquiry patterns, we introduce variability by integrating multiple tasks within a single dialogue in a randomized manner. This dual strategy of combining systematic task progression with randomized task interleaving significantly enhances the model's multi-step reasoning ability, robustness to instruction order variations, and consistency in longer conversational contexts. Such synthetic dialogues not only strengthen the model's analytical capabilities but also lay the groundwork for interactive AI agents in various remote sensing analytics.

F. The Statistics of ChangeIMTI

In summary, our constructed dataset comprises five components, *i.e.*, change captioning, binary change classification, change counting, change localization, as well as multi-turn conversation. The change caption task provides textual descriptions of the detected changes, while binary change classification determines whether changes have occurred. change counting and change localization offer fine-grained information about the changes, and multi-turn dialogue enables interactive capabilities. As displayed in Table I above, it totally contains 89,628 data samples, making it the currently largest instruction-tuning dataset for change detection tasks.

IV. Methodology

As illustrated in Fig. 2, our model consists of two main components: a novel vision-guided module and a VLM backbone. The vision-guided module is designed to extract auxiliary visual cues with dual-granularity semantics, capturing both fine-grained details such as object count and spatial locations and holistic scene-level descriptions in caption format. In contrast to the visual features extracted by ViT, which primarily encode basic image patterns, our module delivers richer and more semantically meaningful representations. These dual-granularity signals are then integrated into the VLM, which jointly processes the bi-temporal remote sensing images and the injected prompts to generate context-aware responses.

A. Vision-guided Module

To enhance the instruction-following capability of our VLMs in remote sensing change captioning, we introduce a vision-guided module that processes bi-temporal remote sensing images to extract dual-granularity visual cues, encompassing both fine-grained details and global-level semantics. As shown in Fig. 2, the module comprises three components: a feature extractor, a fine-grained recognition branch, and a global summary branch.

1) Feature Extractor: The feature extraction module is designed to capture rich, multi-scale semantic representations from bi-temporal remote sensing images, serving as the visual backbone for both the fine-grained recognition branch and the global summary branch. Specifically, the two input images acquired at different time points are independently processed by identical encoder networks with shared parameters, based on the SegFormer-B1 encoders [33]. The parameter-sharing strategy ensures consistent feature alignment across temporal dimensions while producing hierarchical visual features at multiple levels of abstraction.

To formally denote the extracted features, let I_{t1} and I_{t2} represent the input images at two different time points. The shared encoder $\mathcal{F}_\theta$ (SegFormer-B1) extracts a set of multi-level features from each image:

$$\{f_1^{(1)}, f_2^{(1)}, f_3^{(1)}, f_4^{(1)}\} = \mathcal{F}_\theta(I_{t1}), \tag{1}$$

$$\{f_1^{(2)}, f_2^{(2)}, f_3^{(2)}, f_4^{(2)}\} = \mathcal{F}_\theta(I_{t2}), \tag{2}$$

where $f_l^{(i)}$ denotes the feature map at level $l \in \{1, 2, 3, 4\}$ extracted from image I_{ti}. These features encode spatial and semantic cues at increasing levels of abstraction.

The extracted features from both images are subsequently forwarded to the downstream fine-grained recognition branch and global summary branch, which are responsible for detailed instance reasoning and caption generation, respectively.

2) Fine-Grained Recognition Branch: The fine-grained recognition branch enhances the model's capacity to detect and interpret localized changes between bi-temporal remote sensing images. Based on multi-level features extracted by the shared SegFormer-B1 encoders, this branch performs progressive fusion and decoding to generate a high-resolution change mask and fine-grained change attributes.

The multi-scale feature maps from time t_1 and t_2 be represented as:

$$f_t^{(1)} = \left\{ f_1^{(1)}, f_2^{(1)}, f_3^{(1)}, f_4^{(1)} \right\}, \tag{3}$$

$$f_t^{(2)} = \left\{ f_1^{(2)}, f_2^{(2)}, f_3^{(2)}, f_4^{(2)} \right\}. \tag{4}$$

For each level i, a similarity-guided fusion module combines features using cosine similarity and convolution:

$$\alpha_i = \cos(f_i^{(1)}, f_i^{(2)}) + \text{Conv}_{3\times3}(f_i^{(2)} - f_i^{(1)}), \tag{5}$$

$$\tilde{f}_i = \text{Conv}_{3\times3}\left(\text{Concat}[f_i^{(1)}, \alpha_i, f_i^{(2)}]\right), \tag{6}$$

$$f_i^{\text{fused}} = \text{Conv}_{1\times1}\left(\text{ReLU}\left(\text{BN}(\tilde{f}_i)\right)\right). \tag{7}$$

The top-level fused features f_4^{fused} are further processed by a series of stacked BI3 layers [8] to enhance their semantic representation, which can be formulated as:

$$f_4^{\text{refined}} = \text{BI3Layer}_2\left(\text{BI3Layer}_1\left(f_4^{\text{fused}}\right)\right). \tag{8}$$

Next, fused features from all levels are progressively decoded through deconvolution layers. The decoding process can be represented as:

$$d_4 = \text{DeConv}(f_4^{\text{refined}}), \tag{9}$$

$$d_i = \text{DeConv}(f_i^{\text{fused}} \oplus d_{i+1}), \quad \text{for } i = 3, 2, 1, \tag{10}$$

where $\oplus$ denotes channel-wise concatenation.

The final decoded feature map d_1 is passed through a 1×1 convolution with sigmoid activation to produce the binary change mask:

$$M = \sigma\left(\text{Conv}_{1\times1}(d_1)\right). \tag{11}$$

We utilize OpenCV-based post-processing techniques to analyze the predicted binary change mask M. Through contour detection and connected component analysis, we extract the number and spatial locations of changed instances, specifically focusing on road and building changes. These attributes serve as auxiliary cues for downstream reasoning tasks such as change counting and localization.

3) Global Summary Branch: The global summary branch is designed to provide an initial semantic understanding of the scene-level changes between bi-temporal remote sensing images. In contrast to the fine-grained recognition branch, which focuses on localized and detailed change patterns, this branch generates a coarse-grained summary that serves as a visual prompt to guide the subsequent reasoning process of the VLMs. By capturing the overall transformation trends in

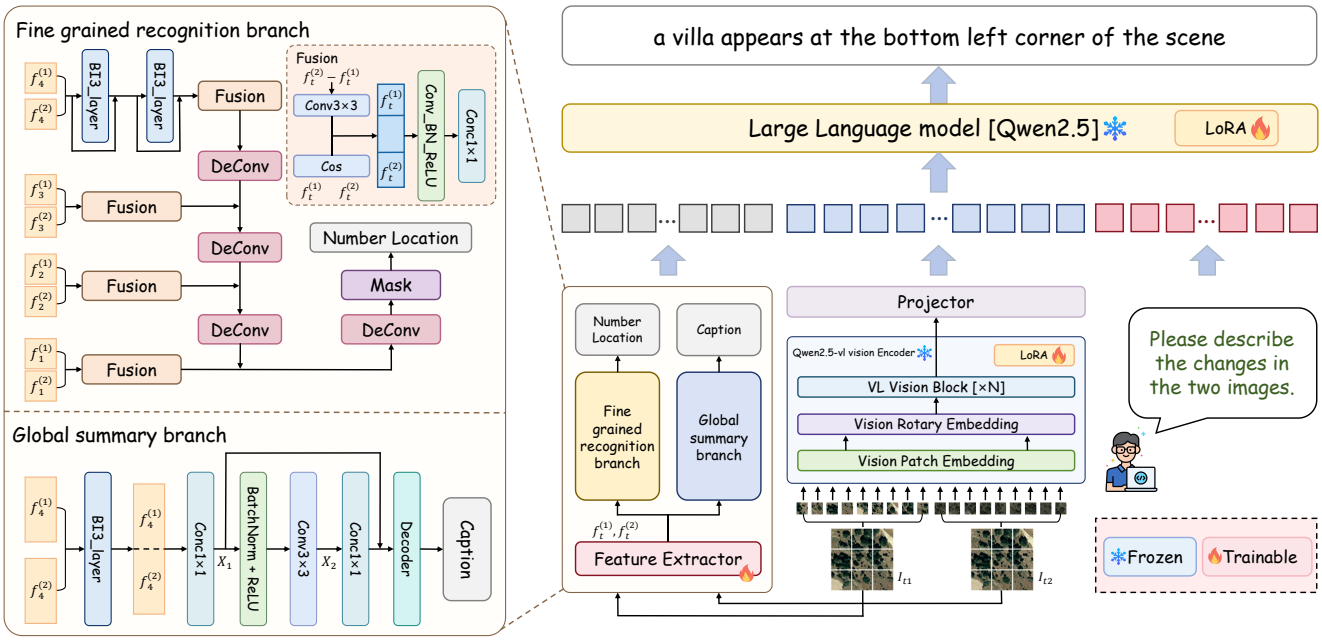

Fig. 2. Overview of ChangeVG. The right panel illustrates the overall architecture of our model, which takes the input image and the user query, processes them through Qwen2.5-VL-7B and a vision-guided module, and finally generates a response to the query. The left panel presents the details of the vision-guided module, which consists of a global summary branch and a fine-grained recognition branch. These branches extract coarse-grained caption information and fine-grained details, respectively.

the input image pair, the global summary branch offers high-level contextual information that facilitates more coherent and accurate language generation in downstream tasks such as change captioning.

Specifically, we feed the high-level semantic features $f_4^{(1)}$ and $f_4^{(2)}$, extracted from the two temporal images, into the BI3 module to enhance and fuse the global contextual representations across time. The resulting bi-temporal feature is then passed through a convolutional transform layer, which transforms the visual features from the image domain into a semantic embedding space suitable for language modeling. Finally, these embeddings are decoded by a Transformer-based language decoder to produce an initial coarse-grained change description, which serves as a prompt for subsequent reasoning by the VLMs. The transform layer $Trans(\cdot)$ can be formally represented as follows:

$$\mathrm{Trans}(f^{(1)}, f^{(2)}) = X_1 + \mathrm{conv}_{1\times1}(X_2), \tag{12}$$

$$X_1 = \mathrm{conv}_{1\times1}\left(\mathrm{concat}\left([f^{(1)}, f^{(2)}]\right)\right), \tag{13}$$

$$X_2 = \mathrm{ConvBNReLU}_{3\times3}\left(\mathrm{conv}_{1\times1}(X_1)\right). \tag{14}$$

The transformed representation is then fed into a Transformer-based language decoder to produce a coarse-grained natural language description of the detected changes. The caption summarizes the global transformation context and serves as a visual prompt to guide subsequent fine-grained reasoning in the downstream VLMs.

B. Instruction Tuning for ChangeVG

To enable the model to effectively understand and respond to queries based on bi-temporal remote sensing images, we fine-tune the open-source VLMs Qwen2.5-VL-7B [22] on our constructed *ChangeIMTI* dataset. Qwen2.5-VL-7B is a vision-language model built upon the Qwen2.5-7B language backbone, equipped with a visual encoder for image understanding and a multi-layer cross-modal fusion module to align image and text representations.

Our instruction tuning aligns the model's generative capabilities with multi-task RSCU. Each training sample is formatted as an instruction–response pair:

$$\mathcal{D} = \{(I_{t1}, I_{t2}, q, r)\}, \tag{15}$$

where I_{t1} and I_{t2} represent the bi-temporal remote sensing images, q is the user query (*e.g.*, "Please describe the changes."), and r is the target response generated by human annotators.

To better exploit both fine-grained and global summary information, we design a prompt. Specifically, we leverage a vision-guided module to extract coarse-grained scene-level captions and fine-grained change indicators, including object counts and spatial localization cues. These outputs are translated into natural language and injected into the final prompt as structured contextual priors. The fusion module then integrates these visual priors with the user query and corresponding image pair to construct a comprehensive multimodal prompt display in Fig. 3. The enriched prompt serves as input to

TABLE II
COMPARE WITH SOTA METHODS ON THE CHANGE CAPTION TASK. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD. THE SECOND-BEST RESULTS ARE INDICATED WITH UNDERLINES.

Method	BLEU-4	METEOR	ROUGE$_L$	CIDEr-D	S_m^*
DUDA [34]	57.79	37.15	71.04	124.32	72.58
MCCFormer-S [35]	56.68	36.17	69.46	120.39	70.68
MCCFormer-D [35]	56.38	37.29	70.32	124.44	72.11
RSICCFormer [15]	62.77	39.61	74.12	134.12	77.66
PSNet [36]	62.11	38.80	73.60	132.62	76.78
Chg2Cap [37]	64.39	40.03	75.12	136.61	79.04
Prompt-CC [21]	63.54	38.82	73.72	136.44	78.13
Semantic-CC [12]	<u>64.51</u>	<u>40.58</u>	**77.76**	<u>138.51</u>	<u>80.34</u>
ChangeVG (Ours)	**65.08**	**42.06**	<u>76.95</u>	**142.83**	**81.73**

You are an assistant proficient in remote sensing change reasoning. Please answer the user's question based on the following auxiliary information.

Visual auxiliary information:
Caption: xxx
Count: { road: x,
building: x}
Location: { road: {xxx},
building: {xxx}}

Query:
The user's question.

Fig. 3. The example of our prompt. It illustrates the architecture of the final prompt fed into the VLM.

the instruction-tuned LLM, enabling it to reason over both high-level semantics and low-level details during generation. By explicitly encoding the auxiliary visual cues alongside user intent, the fusion mechanism significantly enhances the model's capacity for precise, coherent, and contextually grounded change reasoning. Moreover, in multi-turn dialogues, we merge historical conversation with subsequent prompts to support interactive user engagement.

V. EXPERIMENTS

A. Experimental settings

In this work, all the training and evaluations of our introduced model are conducted on a machine with 8 × NVIDIA 4090 GPUs. For comparisons with state-of-the-art VLMs, we deploy their corresponding APIs. The test set employed in our experiments is derived from the LEVIR-MCI dataset, following the preprocessing procedures described in Section III, which can ensure the consistency between the training and evaluation data format.

B. Evaluation Metrics

To evaluate both the effectiveness and the generalizability of our proposed approach, we conduct experiments not only on the change captioning task but also on three additional tasks: binary change classification, change counting, and change localization.

- Change captioning task aims to describe the differences between two images captured at different times using natural language.
- Binary change classification task aims to determine whether a change has occurred in a specific object or region between two images captured at different times.
- Change counting task aims to estimate the number of changed objects of a specified category between two images taken at different time points.
- Change localization task aims to identify the spatial locations of changed objects within the image pair, indicating where changes have occurred over time.

For the change captioning task, we evaluate the model using BLEU-4 [38], METEOR [39], ROUGE-L [40], and CIDEr-D [41] to assess 4-gram precision, alignment with human judgment (considering synonymy, stemming, and word order), the ability to capture the longest common subsequence, and consensus with human-written descriptions based on TF-IDF weighting of n-grams, respectively. To comprehensively evaluate the model's performance on the change captioning task, we use an aggregate metric S_m^* [42], which is defined as the average of four widely-used evaluation metrics: BLEU-4, METEOR, ROUGE-L, and CIDEr-D. The equation is as follows:

$$S_m^* = \frac{1}{4} \times (BLEU_4 + METEOR + ROUGE_L + CIDEr - D). \tag{16}$$

The metric provides a balanced assessment by integrating lexical precision, semantic similarity, and n-gram consensus between the generated and reference captions.

For the binary change classification task, we evaluate the model using four standard metrics: Accuracy, Precision, Recall, and F1-score. These metrics comprehensively reflect the model's classification performance by measuring its overall correctness (Accuracy), its ability to avoid false positives (Precision), its sensitivity to actual changes (Recall), and the balance between Precision and Recall (F1-score).

For the change counting task, we evaluate the model performance using Mean Absolute Error (MAE), applied separately to the road and building categories. MAE quantifies the average absolute difference between the predicted and ground-truth counts, providing an intuitive measure of the model's counting accuracy for each type of changed object.

For the change localization task, we evaluate the model per-

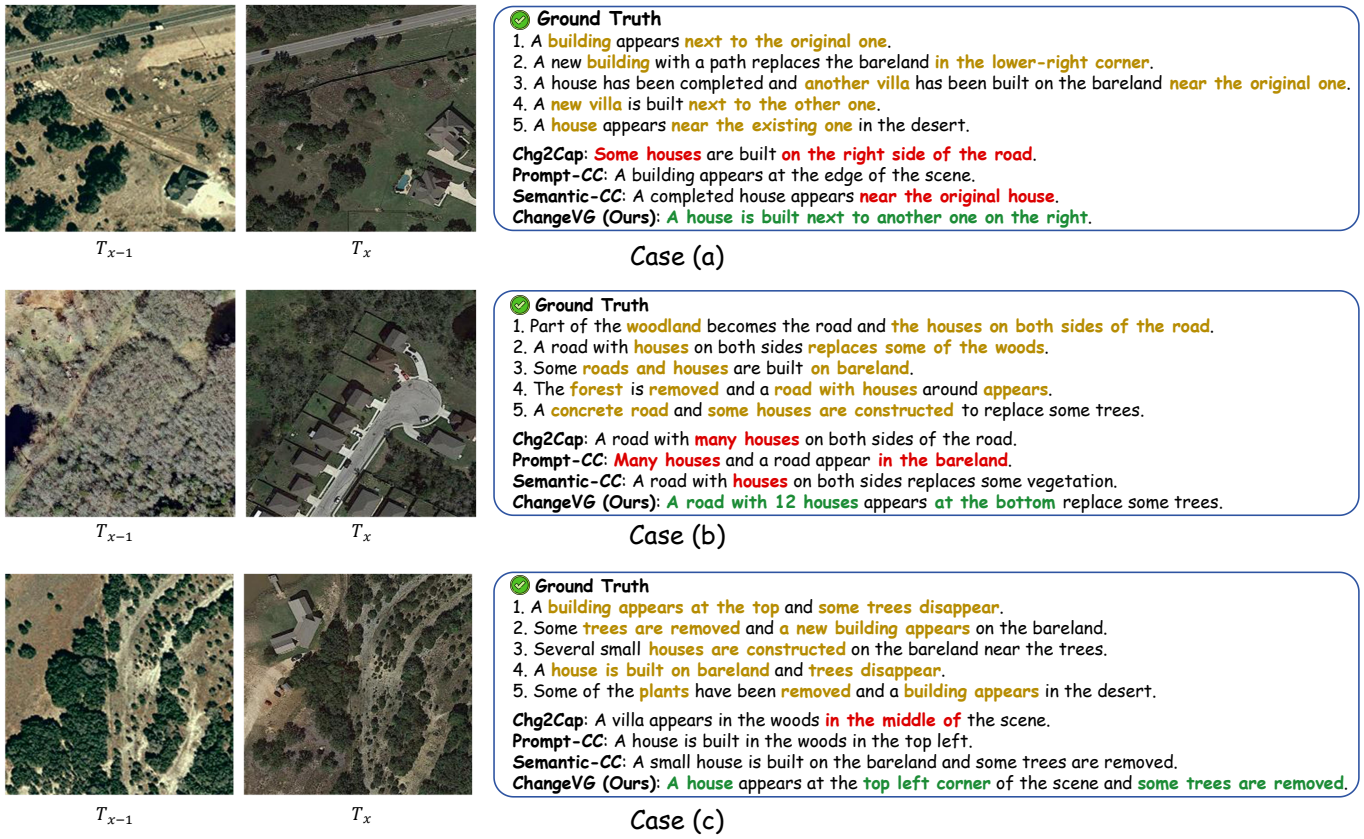

Fig. 4. A direct comparison of our method against existing approaches on the change captioning task, where green indicates correct descriptions and red highlights inaccurate or inappropriate ones.

formance on both road and building categories using example-based accuracy, micro precision, micro recall, micro F1-score, and subset accuracy. These metrics jointly assess the model's ability to accurately and completely localize changes across multiple categories.

C. Main Results

1) Change Captioning: We compare our method with multiple SOTA approaches, including DUDA [34], MCCFormer-S, MCCFormer-D [35], RSICCFormer [15], PSNet [36], Chg2Cap [37], Prompt-CC [21], Semantic-CC [12].

As shown in table II, our model achieves the highest performance across BLEU-4, METEOR, and CIDEr-D metrics, with relative improvements of 1.07%, 3.65%, and 3.12% over the best existing method, respectively. In addition, our approach achieves the best overall score on the aggregated metric S_m^*, outperforming Semantic-CC by 1.39 points. These results demonstrate the advantage of integrating fine-grained and global visual cues into the VLMs, leading to more accurate and contextually rich change descriptions.

To more intuitively assess the effectiveness of our proposed method, Fig. 4 presents a visual comparison using three representative image samples, highlighting the differences between our approach and existing methods. As observed in the figure, our method outperforms other approaches in both quantity and location recognition, accurately identifying the number of changed objects as well as their spatial positions.

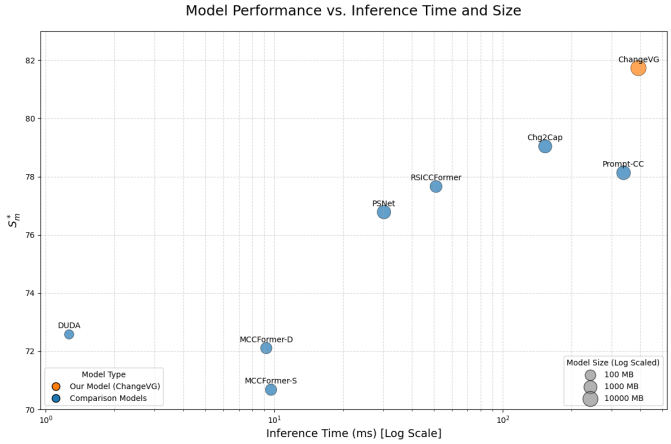

Fig. 5. Comparison of model size, inference time, and model performance across different models. The x-axis represents inference time, the y-axis denotes model performance, and the circle size corresponds to model size.

We further investigate the trade-offs among model size, inference time, and model performance. As illustrated in Figure 5, this plot compares different models across three dimensions: the x-axis represents inference time on a logarithmic scale, the y-axis denotes model performance, and the size of the circles corresponds to model scale. Orange circles represent our model (ChangeVG), while blue circles denote comparative methods. As shown in Table III, our approach

TABLE III
COMPARISON OF MODEL SIZE, INFERENCE TIME, AND MODEL PERFORMANCE ACROSS DIFFERENT MODELS.

Method	Model Size(MB)	Inference time(ms)	S_m^*	Improvement to Ours
DUDA [34]	35.59	1.266	72.58	-11.20%
MCCFormer-S [35]	186.25	9.678	70.68	-13.52%
MCCFormer-D [35]	186.25	9.225	72.11	-11.77%
RSICCFormer [15]	353.979	51.021	77.66	-4.98%
PSNet [36]	1666.85	30.198	76.78	-6.01%
Chg2Cap [37]	1311.33	153.284	79.04	-3.29%
Prompt-CC [21]	2315.34	337.496	78.13	-4.40%
ChangeVG (Ours)	16285.25	392.121	81.73	-

TABLE IV
PERFORMANCE COMPARISONS WITH VLMS ON THE BINARY CHANGE CLASSIFICATION TASK.

Method	Accuracy	Precision	Recall	F1
GLM-4v [43]	0.6580	0.6705	0.6679	0.6692
Qwen2.5-VL-7B [22]	0.5805	0.8888	0.2166	0.3483
Qwen2.5-VL-72B [22]	0.7110	**0.9711**	0.4555	0.6202
GPT-4o [44]	0.7240	0.8558	0.5617	0.6783
Gemini2.5-Flash [45]	0.8036	0.7675	0.8895	0.8240
Claude-Sonnet-4.0 [46]	0.5170	0.8723	0.0791	0.1451
ChangeVG (Ours)	**0.9460**	0.9549	**0.9401**	**0.9474**

TABLE V
PERFORMANCE COMPARISONS OF MAE ON CHANGE COUNTING FOR ROADS AND BUILDINGS ACROSS VLMS.

Method	Road (↓)	Build (↓)
GLM-4v [43]	0.486	4.161
Qwen2.5-VL-7B [22]	0.943	3.875
Qwen2.5-VL-72B [22]	0.408	4.181
GPT-4o [44]	0.348	3.663
Gemini2.5-Flash [45]	1.295	3.909
Claude-Sonnet-4.0 [46]	0.711	4.308
ChangeVG (Ours)	**0.156**	**0.802**

outperforms other methods by up to 10% and by at least 5%, further underscoring its competitiveness in balancing these multidimensional trade-offs.

2) Binary Change Classfication: For this task, we compare our approach with several SOTA VLMs, such as GLM-4v [43], Qwen2.5-VL-7B , Qwen2.5-VL-72B [22], GPT-4o [44], Gemini2.5-Flash [45] and Claude-Sonnet-4.0 [46]. All models are evaluated under the same setting: each receives a pair of bi-temporal images along with a binary instruction, and must classify whether a change is present.

As shown in Table IV, our model can significantly outperform all the advanced baselines across four important evaluation metrics. In particular, our approach achieves an Accuracy of 94.6% and an F1 score of 94.74%, surpassing the best-performing baseline (Gemini2.5-Flash) by +14.26 and +12.34 points respectively. These results demonstrate the effectiveness of our visual-enhanced instruction tuning framework, which improves both precision and recall, and offers robust performance even in challenging change scenarios.

3) Change Counting: As shown in Table V, our model significantly outperforms all baseline VLMs. To be specific, our method achieves an MAE of 0.156 for roads and 0.802 for buildings, outperforming the best-performing baseline (GPT-4o for roads with 0.348 and GPT-4o for buildings with 3.663) by a large margin.

These improvements demonstrate the effectiveness of our design, where the fine-grained recognition branch explicitly segments and localizes change regions, and the visual auxiliary module enriches the model's understanding of spatial details. This fine-grained guidance allows the VLMs to make more accurate numerical predictions during instruction-following, particularly in dense or small-scale change scenarios.

4) Change Localization: As displayed in Tables VI & VII, our method significantly outperforms existing VLMs across all the evaluation metrics. These metrics can jointly reflect both the overall correctness of model predictions (Precision and Recall) and the capability to fully capture multi-label outputs (*i.e.*, subset accuracy).

For the road category, our model achieves a Micro F1 of 0.7720, surpassing the best baseline (Claude-Sonnet-4.0, 0.5543) by 21.77 percentage points. Similarly, for the building, our approach achieves 0.9137, improving upon the second-best result (Gemini2.5-Flash, 0.7192) by a large margin.

The superior performance can be attributed to the fine-grained recognition branch, which explicitly extracts and segments multi-scale spatial features, and the visual auxiliary encoder, which enhances semantic grounding in the image domain. Together, these components provide strong visual priors to the language model, enabling it to localize change regions with high accuracy.

In addition to single-turn evaluations, we design task-oriented multi-turn dialogue settings to assess the model's capability to engage in coherent, step-by-step reasoning across temporally aligned image pairs, while also demonstrating its ability to interact with users. As shown in Fig. 6.

D. Ablation Studies

1) Diffence task and vision-guided: To further investigate the contributions of individual auxiliary tasks and the vision-guided module to the overall performance of our model on the change captioning task, we conduct detailed ablation studies. As shown in Table VIII, we progressively introduce binary change classification, change counting and change localization tasks, as well as the vision-guided module into the instruction tuning process, and evaluate their respective impacts.

TABLE VI
PERFORMANCE COMPARISONS WITH VLM ON THE ROAD CHANGE LOCALIZATION.

	Method	Example-based Accuracy	Micro Precision	Micro Recall	Micro F1	Subset Accuracy
Road	GLM-4v [43]	0.5451	0.3659	0.5881	0.4511	0.5030
	Qwen2.5-VL-7B [22]	0.0088	0.0219	0.0108	0.0145	0.0060
	Qwen2.5-VL-72B [22]	0.4746	0.4771	0.3993	0.4347	0.4720
	GPT-4o [44]	0.6010	0.4724	0.6076	0.5315	0.5608
	Gemini2.5-Flash [45]	0.2712	0.1579	0.6667	0.2554	0.1800
	Claude-Sonnet-4.0 [46]	0.6475	0.4789	0.6578	0.5543	0.6338
	ChangeVG (Ours)	**0.7865**	**0.7850**	**0.7595**	**0.7720**	**0.7330**

TABLE VII
PERFORMANCE COMPARISONS WITH VLMS ON THE BUILDING CHANGE LOCALIZATION.

	Method	Example-based Accuracy	Micro Precision	Micro Recall	Micro F1	Subset Accuracy
Building	GLM-4v [43]	0.6093	0.6295	0.4619	0.5328	0.5310
	Qwen2.5-VL-7B [22]	0.0460	0.2012	0.0483	0.0770	0.0140
	Qwen2.5-VL-72B [22]	0.4974	0.5355	0.2047	0.2962	0.4800
	GPT-4o [44]	0.6652	0.7766	0.4293	0.5530	0.5872
	Gemini2.5-Flash [45]	0.6012	0.6353	0.8285	0.7192	0.3950
	Claude-Sonnet-4.0 [46]	0.5318	0.3828	0.5134	0.4386	0.4889
	ChangeVG (Ours)	**0.8813**	**0.9068**	**0.9208**	**0.9137**	**0.7220**

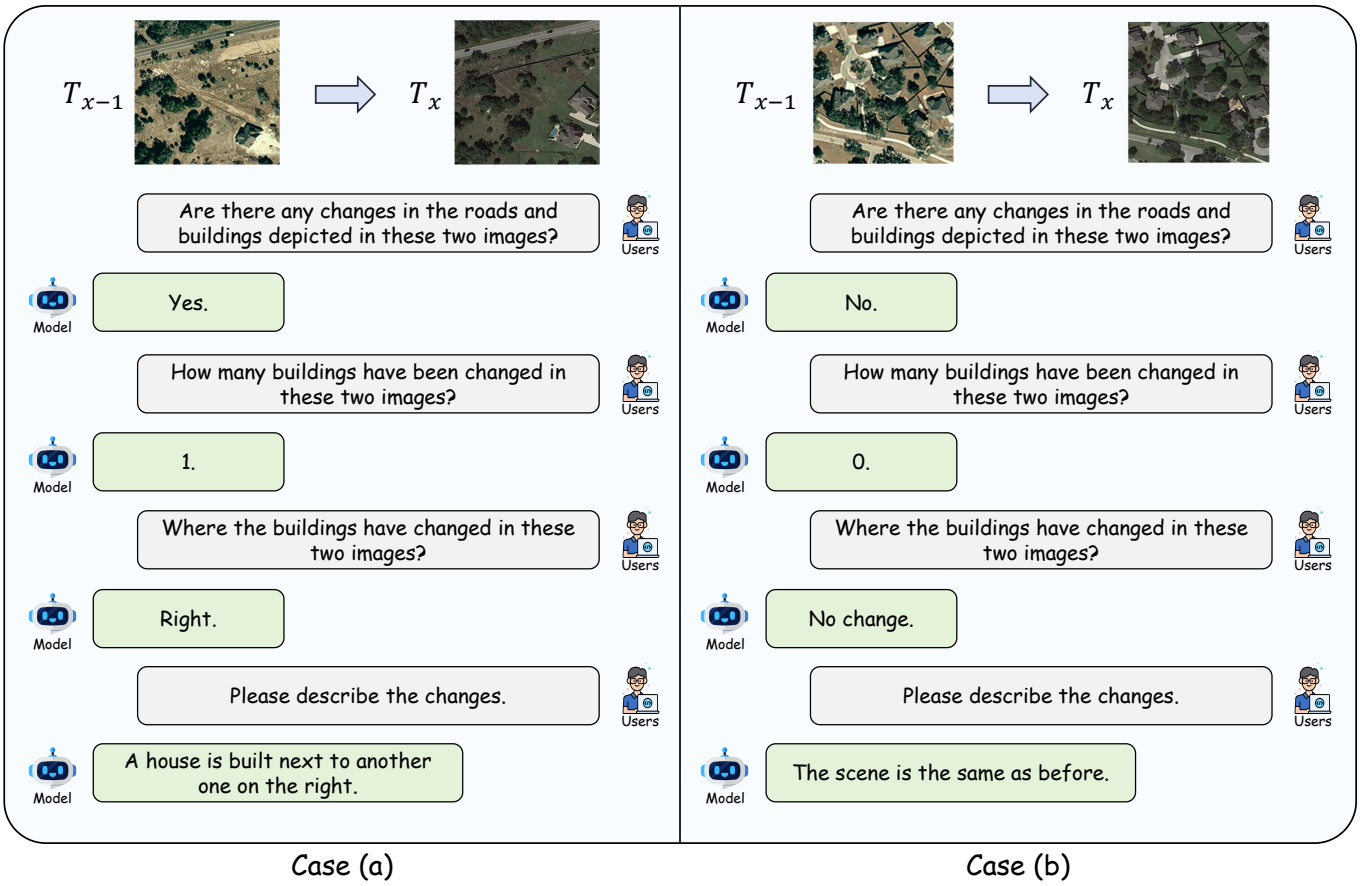

Fig. 6. The conversations example of multi-turn with different user queries. The T_{x-1} and T_x represent images from two different moments in time, with T_{x-1} being the image from the earlier time and T_x being the one from the later time. Case (a) illustrates a scenario with changes, while Case (b) illustrates a scenario without changes.

We first consider the single-task baseline that trains only on the change captioning objective. This configuration achieves a CIDEr-D of 138.08 and an S_m^* score of 78.67. Interestingly, when introducing the binary change classification task alone, the model's performance slightly decreases across multiple metrics (*e.g.*, CIDEr-D drops to 137.52, METEOR drops to

TABLE VIII
THE EFFECTS OF AUXILIARY TASKS AND VISION-GUIDED ON CHANGE CAPTIONING PERFORMANCE.

Caption	Count + Location	Binary	Vision-guided	BLEU-4	METEOR	ROUGE$_L$	CIDEr-D	S_m^*
✓	×	×	×	61.30	40.69	74.61	138.08	78.67
✓	×	✓	×	61.00	40.15	74.46	137.52	78.28
✓	✓	×	×	61.79	40.69	74.93	139.23	79.16
✓	✓	✓	×	62.98	41.11	75.79	140.48	80.09
✓	✓	✓	✓	**65.08**	**42.06**	**76.95**	**142.83**	**81.73**

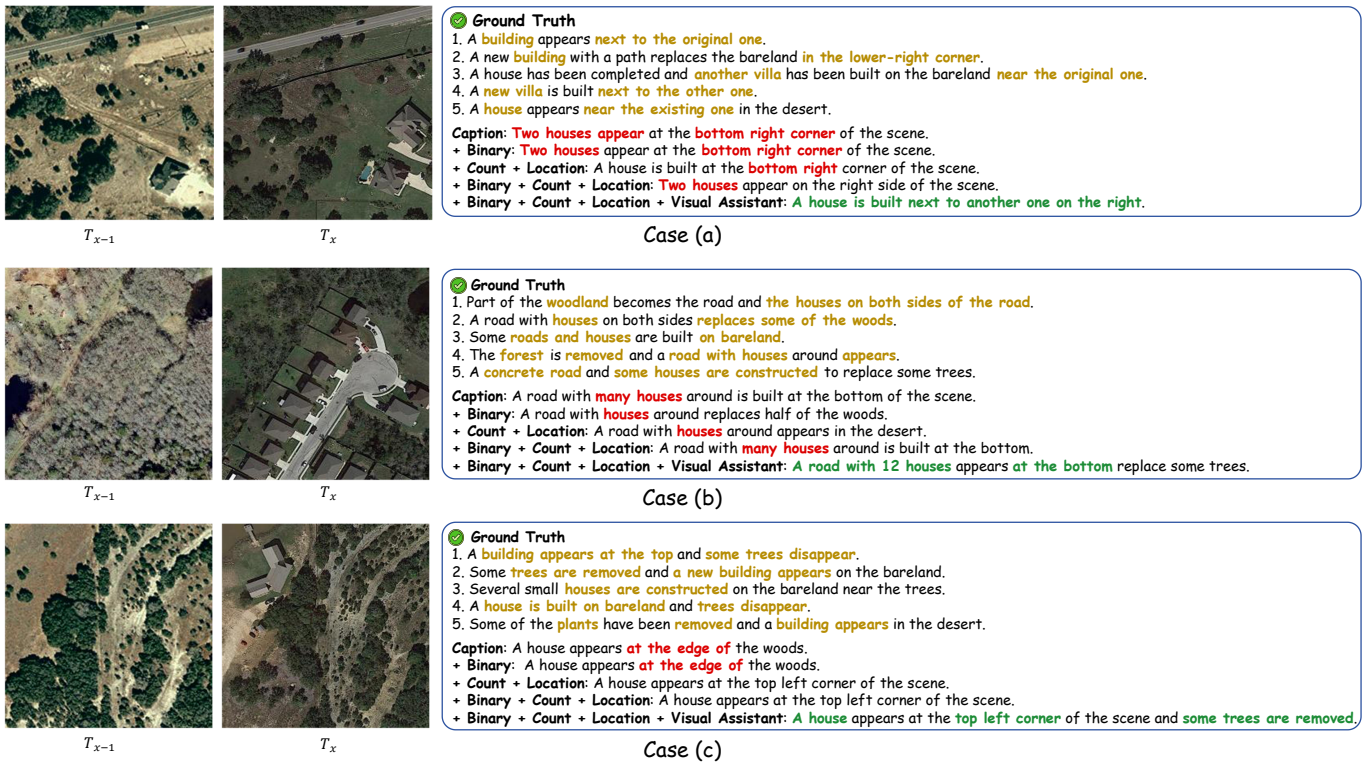

Fig. 7. Example from the change captioning task used for the ablation study. Texts in green indicate correct descriptions, yellow highlights represent partially accurate or ambiguous expressions, and red indicates incorrect outputs.

TABLE IX
THE PRACTICALITY OF THE VISION-GUIDED MODULE ON OTHER VLMS.

Method	BLEU-4	METEOR	ROUGE$_L$	CIDEr-D	S_m^*
InternVL3-8B [47]	65.89	42.09	76.49	142.50	81.74
InternVL3-8B + visual	63.17	42.69	76.95	145.49	82.08
GLM-4.1v-9B [48]	61.77	41.43	75.45	140.01	79.67
GLM-4.1v-9B + visual	64.33	41.81	76.27	141.04	80.86

40.15). This suggests that binary classification, being a coarse-grained task, may not provide sufficiently rich semantic signals to support fine-grained language generation and might even introduce conflicting gradients during optimization.

In contrast, integrating the change counting and change localization tasks alone yields clear performance improvements (CIDEr-D rises to 139.23, S_m^* to 79.16), demonstrating that spatially and numerically grounded auxiliary tasks offer more aligned and informative supervision for descriptive caption generation. When all three auxiliary tasks are jointly trained, the model benefits further, achieving a CIDEr-D of 140.48 and an S_m^* score of 80.09, indicating that the combination of coarse and fine-grained tasks enhances the model's overall semantic understanding.

In the end, the incorporation of the proposed vision-guided module brings the most significant performance gains. Equipped with this module, the model attains the best results across all metrics (*e.g.*, CIDEr-D of 142.83 and S_m^* of 81.73), confirming that integrating structured visual guidance can effectively enhance the model's ability to perceive and

TABLE X
THE EFFECTS OF FINE-GRAINED RECOGNITION BRANCH AND GLOBAL SUMMARY BRANCH ON THE PERFORMANCE OF CHANGE CAPTIONING. W/O REFERS TO WITHOUT.

Method	BLEU-4	METEOR	ROUGE$_L$	CIDEr-D	S_m^*
w/o vision-guided	62.98	41.11	75.79	140.48	80.09
w/o fine-grained recognition	63.92	42.64	76.86	142.02	81.36
w/o global summary	62.12	42.11	76.37	142.41	80.75
ChangeVG(all)	**65.08**	**42.06**	**76.95**	**142.83**	**81.73**

TABLE XI
THE EFFECTS OF FINE-GRAINED RECOGNITION BRANCH AND GLOBAL SUMMARY BRANCH ON BINARY CHANGE CLASSIFICATION PERFORMANCE.

Method	Accuracy	Precision	Recall	F1
w/o vision-guided	0.9220	0.9246	0.9209	0.9228
w/o Fine-Grained	0.9420	0.9618	0.9190	0.9399
w/o Global	0.9440	0.9620	0.9231	0.9421
ChangeVG (Ours)	**0.9460**	0.9549	**0.9401**	**0.9474**

TABLE XII
THE EFFECTS OF FINE-GRAINED RECOGNITION BRANCH AND GLOBAL SUMMARY BRANCH ON CHANGE COUNTING PERFORMANCE.

Method	Road (↓)	Build (↓)
w/o vision-guided	0.164	0.860
w/o Fine-Grained	0.161	0.855
w/o Global	0.158	0.811
ChangeVG (Ours)	**0.156**	**0.802**

describe nuanced bi-temporal changes. These results collectively validate the effectiveness of our multi-task and vision-augmented training strategy in producing accurate and context-aware change captions. As shown in Fig. 7, we present the visualization results of the captioning task under different ablation settings. It can be clearly observed that without the vision-guided module, the generated captions exhibit noticeable inaccuracies in both quantity estimation and spatial localization of the changes.

2) fine-grained recognition branch and global summary branch: To evaluate the effectiveness of the fine-grained recognition branch and the global summary branch in our Vision-guided Module, we conducted ablation studies on each branch individually. Specifically, we performed experiments by masking the information from the respective branch within the vision-guided module. As shown in table X, both branches contribute to improved performance on the captioning task. Notably, since captioning is inherently a global-level task, the model benefits more from the global summary branch alone than from the fine-grained recognition branch alone. We also evaluated the performance of the two branches on these three additional tasks.

As shown in Table XI, both branches improve the model's performance on binary change classification. However, since this task is relatively simple, the contributions of the two branches are comparable. As shown in Table XII, the two branches exhibit distinct contributions on the change counting task. Since this task is fine-grained in nature, the fine-grained recognition branch provides a greater performance gain compared to the global summary branch. Moreover, combining both branches, which integrate global and local information, leads to further performance improvement. Table XIII and table XIV present the performance of the two branches on the change localization tasks for roads and buildings, respectively. Since these tasks are also fine-grained in nature, the fine-grained recognition branch contributes more significantly to performance improvement than the global summary branch.

3) Vision-guided Module in diffent VLMs: To further verify the applicability of our proposed vision-guided module, we conducted additional comparative experiments on two open-source VLMs, InternVL3-8B [47] and GLM-4.1v-9B [48]. Specifically, we first fine-tuned the baseline models on the ChangeIMTI dataset, and then further fine-tuned them by incorporating the vision-guided module to assess the improvements brought by this module. As shown by table IX, after incorporating the vision-guided module, all metrics on the captioning task improved, demonstrating the module's ability to enhance natural language understanding. For instance, the S_m^* for InternVL3-8B increased from 81.74 to 82.08, and for GLM-4.1v-9B, it increased from 79.67 to 80.86.

E. Zero-shot Evaluation

To further validate the generalization capability of our introduced ChangeVG, we conduct zero-shot experiments on two publicly available remote sensing change understanding datasets that demonstrate substantial overlap with ChangeIMTI: QAG-360k [49] and CDVQA [50].

QAG-360 contains 10 classes, 6,810 image pairs, and 8 tasks, including the "change or not" task. This aligns directly with our binary change classification task, as both aim to determine whether a change occurs between image pairs. CDVQA comprises 30 change categories, 4,662 image pairs, and 8 tasks, including a "change or not" task equivalent to our binary change classification task. To maintain consistency with our dataset's task definition, we evaluate the performance of binary change classification on these two datasets and compare it with several SOTA VLMs.

Table XV and Table XVI present the evaluation results of our model alongside other SOTA VLMs on the QAG-360k and CDVQA datasets. As shown, our model consistently outperforms competing approaches in both accuracy and F1 score, demonstrating its superior generalization capability in the domain of remote sensing change understanding.

TABLE XIII
THE EFFECTS OF FINE-GRAINED RECOGNITION BRANCH AND GLOBAL SUMMARY BRANCH ON ROAD CHANGE LOCALIZATION.

	Method	Example-based Accuracy	Micro Precision	Micro Recall	Micro F1	Subset Accuracy
Road	w/o vision-guided	0.7697	0.7688	0.7513	0.7600	0.7056
	w/o Fine-Grained	0.7719	0.7711	0.7491	0.7599	0.7076
	w/o Global	0.7843	0.7813	0.7618	0.7714	0.7260
	ChangeVG (Ours)	**0.7865**	**0.7850**	**0.7595**	**0.7720**	**0.7330**

TABLE XIV
THE EFFECTS OF FINE-GRAINED RECOGNITION BRANCH AND GLOBAL SUMMARY BRANCH ON BUILDING CHANGE LOCALIZATION.

	Method	Example-based Accuracy	Micro Precision	Micro Recall	Micro F1	Subset Accuracy
Building	w/o vision-guided	0.8547	0.9105	0.9038	0.9071	0.6580
	w/o Fine-Grained	0.8649	0.9132	0.9158	0.9145	0.6817
	w/o Global	0.8784	0.9181	0.9105	0.9142	0.7119
	ChangeVG (Ours)	**0.8813**	**0.9068**	**0.9208**	**0.9137**	**0.7220**

TABLE XV
PERFORMANCE COMPARISONS WITH VLMS ON BINARY CHANGE CLASSIFICATION IN QAG-360K

Method	Accuracy	Precision	Recall	F1
GLM-4v [43]	0.7677	0.6667	0.7894	0.7229
Qwen2.5-VL-7B [22]	0.6970	0.5714	0.6667	0.6153
Qwen2.5-VL-72B [22]	0.7349	**0.8846**	0.4339	0.5822
GPT-4o [44]	0.6566	0.5322	0.8684	0.6600
Gemini2.5-Flash [45]	0.7677	0.6829	0.7368	0.7089
Claude-Sonnet-4.0 [46]	0.5556	0.4595	**0.8947**	0.6071
ChangeVG (Ours)	**0.8081**	0.7111	0.8421	**0.7711**

TABLE XVI
PERFORMANCE COMPARISONS WITH VLMS ON BINARY CHANGE CLASSIFICATION IN CDVQA

Method	Accuracy	Precision	Recall	F1
GLM-4v [43]	0.7475	0.8113	0.7414	0.7748
Qwen2.5-VL-7B [22]	0.5859	0.8400	0.3621	0.5060
Qwen2.5-VL-72B [22]	0.7727	**0.8889**	0.6667	0.7619
GPT-4o [44]	0.7273	0.7067	**0.9138**	0.7970
Gemini2.5-Flash [45]	0.6465	0.7674	0.5690	0.6535
Claude-Sonnet-4.0 [46]	0.5960	0.6286	0.7586	0.6875
ChangeVG (Ours)	**0.8182**	0.8704	0.8103	**0.8393**

VI. CONCLUSION AND FUTURE WORKS

In this paper, we present a unified framework ChangeVG for RSCU built upon VLMs. To facilitate instruction tuning across diverse tasks, we construct ChangeIMTI, a large-scale interactive multi-task dataset that includes change captioning, binary change classification, object counting, and change localization, to evaluate VLMs in remote sensing. Our framework incorporates a vision-guided module, which enhances the model's ability to capture both fine-grained and global changes across bi-temporal images. Through this integration, the model supports both caption generation and VQA tasks, enabling semantically interpretable and task-aware reasoning. At the same time, our model is capable of interactive question answering with users, providing precise answers about RSCU based on specific questions posed by the user. Extensive experiments conducted on multiple RSCU tasks demonstrate that our approach achieves SOTA performance, validating both the effectiveness of our unified design and its generalizability across tasks. The work highlights the potential of VLMs in remote sensing applications and provides a solid foundation for future research in vision-language geospatial understanding.

Despite promising results, the current ChangeIMTI contains a limited number of object categories, which may constrain the model's ability to generalize to more diverse scenes. In the future work, we plan to expand the dataset to include a broader range of object types, thereby enhancing the model's generalization capability. In addition, incorporating Chain-of-Thought (CoT) pattern into the training samples can be explored to further improve the explainability of model predictions.

REFERENCES

[1] D. R. Panuju, D. J. Paull, and A. L. Griffin, "Change detection techniques based on multispectral images for investigating land cover dynamics," *Remote Sensing*, vol. 12, no. 11, p. 1781, 2020. 1

[2] M. S. Rana and S. Sarkar, "Prediction of urban expansion by using land cover change detection approach," *Heliyon*, vol. 7, no. 11, 2021. 1

[3] M. T. Jabbar and X. Zhou, "Eco-environmental change detection by using remote sensing and gis techniques: a case study basrah province, south part of iraq," *Environmental Earth Sciences*, vol. 64, pp. 1397–1407, 2011. 1

[4] Z. Zheng, Y. Zhong, J. Wang, A. Ma, and L. Zhang, "Building damage assessment for rapid disaster response with a deep object-based semantic change detection framework: From natural disasters to man-made disasters," *Remote Sensing of Environment*, vol. 265, p. 112636, 2021. 1

[5] A. Asokan and J. Anitha, "Change detection techniques for remote sensing applications: A survey," *Earth Science Informatics*, vol. 12, pp. 143–160, 2019. 1

[6] H. Jiang, M. Peng, Y. Zhong, H. Xie, Z. Hao, J. Lin, X. Ma, and X. Hu, "A survey on deep learning-based change detection from high-resolution remote sensing images," *Remote Sensing*, vol. 14, no. 7, p. 1552, 2022. 1

[7] A. Shafique, G. Cao, Z. Khan, M. Asad, and M. Aslam, "Deep learning-based change detection in remote sensing images: A review," *Remote Sensing*, vol. 14, no. 4, p. 871, 2022. 1

[8] C. Liu, K. Chen, H. Zhang, Z. Qi, Z. Zou, and Z. Shi, "Change-agent: Towards interactive comprehensive remote sensing change interpretation and analysis," *IEEE Transactions on Geoscience and Remote Sensing*, 2024. 1, 2, 3, 5

[9] P. Deng, W. Zhou, and H. Wu, "Changechat: An interactive model for remote sensing change analysis via multimodal instruction tuning," *arXiv preprint arXiv:2409.08582*, 2024. 1

[10] Z. Liang, Y. Xu, Y. Hong, P. Shang, Q. Wang, Q. Fu, and K. Liu, "A survey of multimodel large language models," in *Proceedings of the 3rd International Conference on Computer, Artificial Intelligence and Control Engineering*, 2024, pp. 405–409. 1

[11] L. Qin, Q. Chen, Y. Zhou, Z. Chen, Y. Li, L. Liao, M. Li, W. Che, and P. S. Yu, "A survey of multilingual large language models," *Patterns*, vol. 6, no. 1, 2025. 1

[12] Y. Zhu, L. Li, K. Chen, C. Liu, F. Zhou, and Z. Shi, "Semantic-cc: Boosting remote sensing image change captioning via foundational knowledge and semantic guidance," *IEEE Transactions on Geoscience and Remote Sensing*, 2024. 1, 7, 8

[13] H. Chen and Z. Shi, "A spatial-temporal attention-based method and a new dataset for remote sensing image change detection," *Remote sensing*, vol. 12, no. 10, p. 1662, 2020. 2

[14] S. Ji, S. Wei, and M. Lu, "Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set," *IEEE Transactions on geoscience and remote sensing*, vol. 57, no. 1, pp. 574–586, 2018. 2

[15] C. Liu, R. Zhao, H. Chen, Z. Zou, and Z. Shi, "Remote sensing image change captioning with dual-branch transformers: A new method and a large scale dataset," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–20, 2022. 2, 3, 7, 8, 9

[16] C. Liu, R. Zhao, and Z. Shi, "Remote-sensing image captioning based on multilayer aggregated transformer," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022. 2

[17] L. Meng, J. Wang, R. Meng, Y. Yang, and L. Xiao, "A multiscale grouping transformer with clip latents for remote sensing image captioning," *IEEE Transactions on Geoscience and Remote Sensing*, 2024. 2

[18] C. Yang, Z. Li, and L. Zhang, "Bootstrapping interactive image–text alignment for remote sensing image captioning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–12, 2024. 2

[19] G. Hoxha, S. Chouaf, F. Melgani, and Y. Smara, "Change captioning: A new paradigm for multitemporal remote sensing image analysis," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–14, 2022. 2

[20] S. Chouaf, G. Hoxha, Y. Smara, and F. Melgani, "Captioning changes in bi-temporal remote sensing images," in *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*. IEEE, 2021, pp. 2891–2894. 2

[21] C. Liu, R. Zhao, J. Chen, Z. Qi, Z. Zou, and Z. Shi, "A decoupling paradigm with prompt learning for remote sensing image change captioning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–18, 2023. 2, 7, 8, 9

[22] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang *et al.*, "Qwen2. 5-vl technical report," *arXiv preprint arXiv:2502.13923*, 2025. 3, 6, 9, 10, 13

[23] Z. Zhang, X. Zheng, X. Wu, C. Peng, and X. Cao, "Tokenfocus-vqa: Enhancing text-to-image alignment with position-aware focus and multi-perspective aggregations on lvlms," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 1279–1288. 3

[24] Z. Zhang, X. Wu, D. Huang, S. Yan, C. Peng, and X. Cao, "Hkd4vlm: A progressive hybrid knowledge distillation framework for robust multimodal hallucination and factuality detection in vlms," in *Proceedings of the 33rd ACM International Conference on Multimedia*, 2025, pp. 13 881–13 887. 3

[25] F. Liu, D. Chen, Z. Guan, X. Zhou, J. Zhu, Q. Ye, L. Fu, and J. Zhou, "Remoteclip: A vision language foundation model for remote sensing," *IEEE Transactions on Geoscience and Remote Sensing*, 2024. 3

[26] Y. Hu, J. Yuan, C. Wen, X. Lu, Y. Liu, and X. Li, "Rsgpt: A remote sensing vision language model and benchmark," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 224, pp. 272–286, 2025. 3

[27] K. Kuckreja, M. S. Danish, M. Naseer, A. Das, S. Khan, and F. S. Khan, "Geochat: Grounded large vision-language model for remote sensing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 27 831–27 840. 3

[28] X. Sun, P. Wang, W. Lu, Z. Zhu, X. Lu, Q. He, J. Li, X. Rong, Z. Yang, H. Chang *et al.*, "Ringmo: A remote sensing foundation model with masked image modeling," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–22, 2022. 3

[29] D. Wang, Q. Zhang, Y. Xu, J. Zhang, B. Du, D. Tao, and L. Zhang, "Advancing plain vision transformer toward remote sensing foundation model," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–15, 2022. 3

[30] U. Mall, C. P. Phoo, M. K. Liu, C. Vondrick, B. Hariharan, and K. Bala, "Remote sensing vision-language foundation models without annotations via ground remote alignment," *arXiv preprint arXiv:2312.06960*, 2023. 3

[31] K. Chen, C. Liu, H. Chen, H. Zhang, W. Li, Z. Zou, and Z. Shi, "Rsprompter: Learning to prompt for remote sensing instance segmentation based on visual foundation model," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–17, 2024. 3

[32] K. Chen, C. Liu, W. Li, Z. Liu, H. Chen, H. Zhang, Z. Zou, and Z. Shi, "Time travelling pixels: Bitemporal features integration with foundation model for remote sensing image change detection," in *IGARSS 2024-2024 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, 2024, pp. 8581–8584. 3

[33] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers," *Advances in neural information processing systems*, vol. 34, pp. 12 077–12 090, 2021. 5

[34] D. H. Park, T. Darrell, and A. Rohrbach, "Robust change captioning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 4624–4633. 7, 8, 9

[35] Y. Qiu, S. Yamamoto, K. Nakashima, R. Suzuki, K. Iwata, H. Kataoka, and Y. Satoh, "Describing and localizing multiple changes with transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 1971–1980. 7, 8, 9

[36] C. Liu, J. Yang, Z. Qi, Z. Zou, and Z. Shi, "Progressive scale-aware network for remote sensing image change captioning," in *IGARSS 2023-2023 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, 2023, pp. 6668–6671. 7, 8, 9

[37] S. Chang and P. Ghamisi, "Changes to captions: An attentive network for remote sensing change captioning," *IEEE Transactions on Image Processing*, vol. 32, pp. 6047–6060, 2023. 7, 8, 9

[38] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 2002, pp. 311–318. 7

[39] S. Banerjee and A. Lavie, "Meteor: An automatic metric for mt evaluation with improved correlation with human judgments," in *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 2005, pp. 65–72. 7

[40] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text summarization branches out*, 2004, pp. 74–81. 7

[41] R. Vedantam, C. Lawrence Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 4566–4575. 7

[42] Z. Zhang, W. Zhang, M. Yan, X. Gao, K. Fu, and X. Sun, "Global visual feature and linguistic state guided attention for remote sensing image captioning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–16, 2021. 7

[43] T. GLM, A. Zeng, B. Xu, B. Wang, C. Zhang, D. Yin, D. Zhang, D. Rojas, G. Feng, H. Zhao *et al.*, "Chatglm: A family of large language models from glm-130b to glm-4 all tools," *arXiv preprint arXiv:2406.12793*, 2024. 9, 10, 13

[44] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023. 9, 10, 13

[45] Google DeepMind. (2025) Gemini-2.5:our most intelligent ai model. Accessed: 2025-06-13. [Online]. Available: https://blog.google/technology/google-deepmind/gemini-model-thinking-updates-march-2025/ 9, 10, 13

[46] claude group. (2025) Introducing claude 4. Accessed: 2025-06-13. [Online]. Available: https://www.anthropic.com/news/claude-4 9, 10, 13

[47] J. Zhu, W. Wang, Z. Chen, Z. Liu, S. Ye, L. Gu, H. Tian, Y. Duan, W. Su, J. Shao *et al.*, "Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models," *arXiv preprint arXiv:2504.10479*, 2025. 11, 12

[48] W. Hong, W. Yu, X. Gu, G. Wang, G. Gan, H. Tang, J. Cheng, J. Qi, J. Ji, L. Pan *et al.*, "Glm-4.1 v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning," *arXiv preprint arXiv:2507.01006*, 2025. 11, 12

[49] K. Li, F. Dong, D. Wang, S. Li, Q. Wang, X. Gao, and T.-S. Chua, "Show me what and where has changed? question answering and grounding for remote sensing change detection," *arXiv preprint arXiv:2410.23828*, 2024. 12

[50] Z. Yuan, L. Mou, Z. Xiong, and X. X. Zhu, "Change detection meets visual question answering," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–13, 2022. 12

Junxiao Xue (Member, IEEE) received the Ph.D. degree in computational mathematics from the School of Mathematical Sciences, Dalian University of Technology, Dalian, China, in 2009. He is currently a researcher at Zhejiang Lab, Hangzhou, China. He has authored more than 80 journal and conference articles in these research areas, including the IEEE Transactions on Knowledge and Data Engineering (TKDE), the IEEE Transactions on Systems, Man, and Cybernetics: Systems (TSMCS), CVPR, etc. His current research interests include Multi-Agents Modeling, Reinforcement Learning, and VLMs.

Quan Deng received his B.E. degree from Xiangtan University, China, in 2021. He is currently pursuing his M.S. degree at the Hangzhou Institute for Advanced Study, University of Chinese Academy of Sciences, Hangzhou, China. His current research interests mainly include multi-modal large language models, scene graph generation and remote sensing.

Xuecheng Wu (Student Member, IEEE) received his B.E. degree (with honors) from Zhengzhou University, Zhengzhou, China, in 2023. He is currently pursuing his M.S. degree at the School of Computer Science and Technology, Xi'an Jiaotong University, Xi'an, China. He has authored over twenty international journal and conference papers in the AI field, including publications at CVPR, MM, EMNLP, ICMR, IEEE TCSS, BDMA, etc. His current research interests mainly include computer vision, multi-modal large language models, large-scale video understanding, as well as AI-generated content detection. Besides, he has served as reviewers for top-tier conferences including CVPR, AAAI, MM, ICCV, WWW, and NeurIPS.

Kelu Yao is currently pursuing a Ph.D. in Artificial Intelligence at Zhejiang University. He is also a Senior Engineer at Zhejiang Lab in Hangzhou, China. He has authored over ten journal and conference papers in the field, including publications at CVPR, AAAI, ICCV, ICML, etc. His current research interests mainly focus on vision-centric multimodal models and interpretable artificial intelligence for computer vision.

Xinyi Yin currently is an undergraduate student at the School of Cyber Science and Engineering, Zhengzhou University, China. Her current esearch interests mainly include computer vision and multi-modal large language models.

Fei Yu (Member, IEEE) received her Ph.D. degree from Harbin Institute of Technology, Harbin, China, in 2021. She was a postdoctoral researcher with Zhejiang Lab from 2022 to 2024. She is now an Associate Professor with Liaoning University of Technology and Zhejiang Lab. Her research interests include semantic communication, visual scene understanding, and 3D vision.

Wei Zhou (Senior Member, IEEE) is currently an Assistant Professor at Cardiff University, United Kingdom. His research interests include multimedia computing, perceptual image and video processing, computational vision, human-centric AI, and machine learning. He has published over 80 papers in leading journals and conferences including IEEE TIP, IEEE TCSVT, IEEE TMM, IEEE TMI, CVPR, ECCV, ACM MM, and MICCAI. He has industrial and research experience at the University of Waterloo (Canada), National Institute of Informatics (Japan), University of Science and Technology of China, Intel, Microsoft Research, and Alibaba Group. Prof. Zhou currently serves as Associate Editor for nine journals including IEEE TNNLS, ACM TOMM, Pattern Recognition, and Neurocomputing, and as Area Editor for Signal Processing: Image Communication. He chairs the Election Committee of the IEEE UK & Ireland Signal Processing Society Chapter and has served as Area Chair for major conferences including ACM MM, IEEE ICME, IEEE ICIP, and BMVC. His honors include the IEEE CASS VSPC Rising Star Honorable Mention, ACM SIGMM China Outstanding Doctoral Dissertation Award, CVPR CLIC Challenge Winner Award, and inclusion in Stanford University's World Top 2% Scientists list. He actively contributes to the research community by organizing special issues, tutorials, challenges, and workshops at venues such as IEEE JBHI, Pattern Recognition, ICCV, and ACM MM.

Yanfei Zhong (Senior Member, IEEE) received the B.S. degree in information engineering and the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, China, in 2002 and 2007, respectively. Since 2010, He has been a Full professor with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing (LIESMARS), Wuhan University, China. He organized the Intelligent Data Extraction, Analysis and Applications of Remote Sensing (RSIDEA) research group. He has published more than 150 research papers in international journals, such as Remote Sensing of Environment, ISPRS Journal of Photogrammetry and Remote Sensing, IEEE Transactions on Geoscience and Remote Sensing. His research interests include hyperspectral remote sensing information processing, high-resolution remote sensing image understanding, and geoscience interpretation for multisource remote sensing data and applications. Dr. Zhong is a fellow of the Institution of Engineering and Technology (IET). He was a recipient of the 2016 Best Paper Theoretical Innovation Award from the International Society for Optics and Photonics (SPIE). He received the second-place prize in the 2013 IEEE GRSS Data Fusion Contest and the Single-View Semantic 3-D Challenge of the 2019 IEEE GRSS Data Fusion Contest. He is serving as an Associate Editor for IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing and International Journal of Remote Sensing.

Yang Liu is an Assistant Professor with the Tongji University, Shanghai, China. He received the B.S. degree (with honors) in Automation from Nanjing University, China, in 2020, and the Ph.D. degree in Computer Science from Fudan University, China, in 2025. From 2023 to 2025, he was a visiting scholar at the University of Toronto, Canada, under a joint Ph.D. program. His research focuses on embodied perception with applications in industry and intelligent transportation systems. Dr. Liu has served as Guest Editor for IEEE TCSS, Area Chair for BMVC 2025 and IEEE ICIP 2025, Workshop/Special Session Co-Chair for IEEE ICASSP, IEEE ICIP, and IEEE WF-IoT, and TPC member for top-tier conferences including CVPR, ICCV, ICLR, NeurIPS, AAAI, and ACM MM. He is an active reviewer for ACM CSUR, PR, INFU, and multiple IEEE Transactions including TPAMI, TIP, TII, TNNLS, TCSVT, TMM, and TIFS.

Dingkang Yang received the B.E. degree (with honors) in Communication Engineering from the joint training program of Yunnan University and the Chinese People's Armed Police (PAP), Kunming, China, in 2020, and the Ph.D. degree in Computer Science from Fudan University, Shanghai, China, in 2025. His research interests include multimodal learning, generative AI, and embodied AI. Dr. Yang has published multiple papers as the first author at the reputable journals and top international conferences, such as IEEE TPAMI, TCSVT, NeurIPS, CVPR, ICCV, ECCV, and AAAI. He previously worked as a Principal Researcher at ByteDance, where his team developed a series of renowned multimodal foundational models, including SAIL-VL and SAIL-Embedding. He currently serves as CTO of Fysics AI, an artificial intelligence company dedicated to developing omni-modal interaction and physical real-world perception.