# CHANGECHAT: AN INTERACTIVE MODEL FOR REMOTE SENSING CHANGE ANALYSIS VIA MULTIMODAL INSTRUCTION TUNING

Pei Deng, Wenqian Zhou, Hanlin Wu

School of Information Science and Technology, Beijing Foreign Studies University, Beijing, China

### **ABSTRACT**

Remote sensing (RS) change analysis is vital for monitoring Earth's dynamic processes by detecting alterations in images over time. Traditional change detection excels at identifying pixel-level changes but lacks the ability to contextualize these alterations. While recent advancements in change captioning offer natural language descriptions of changes, they do not support interactive, user-specific queries. To address these limitations, we introduce ChangeChat, the first bitemporal vision-language model (VLM) designed specifically for RS change analysis. ChangeChat utilizes multimodal instruction tuning, allowing it to handle complex queries such as change captioning, category-specific quantification, and change localization. To enhance the model's performance, we developed the ChangeChat-87k dataset, which was generated using a combination of rule-based methods and GPT-assisted techniques. Experiments show that ChangeChat offers a comprehensive, interactive solution for RS change analysis, achieving performance comparable to or even better than state-of-the-art (SOTA) methods on specific tasks, and significantly surpassing the latest general-domain model, GPT-4. Code and pre-trained weights are available at https://github.com/hanlinwu/ChangeChat.

*Index Terms*— Vision-language models, interactive change analysis, change captioning, instruction tuning.

## 1. INTRODUCTION

Remote sensing (RS) change analysis is crucial for monitoring dynamic Earth processes, focusing on detecting alterations in images captured at different times over the same geographical area. This technique is fundamental to applications such as environmental monitoring [1] and disaster management [2].

Change detection, traditionally the cornerstone of RS analysis, has excelled in identifying surface alterations at the pixel level [3, 4]. Despite its accuracy, change detection often lacks the ability to contextualize the changes, leaving out details about the characteristics of the objects involved or their spatial relationships. To bridge this gap, "change captioning" has emerged, translating detected changes into natural language descriptions to provide a richer understand-

ing of the observed changes [5, 6]. However, while enhancing interpretation, change captioning does not support interactive queries or offer user-specific information.

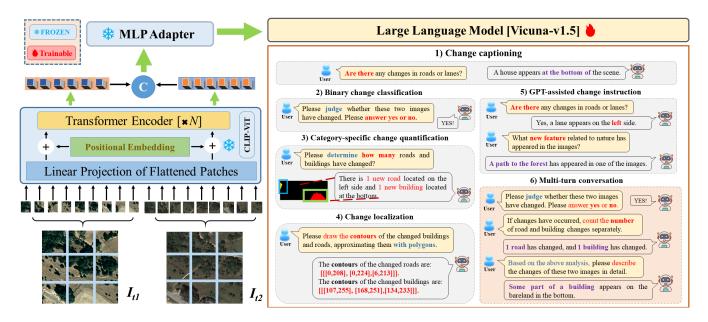
To address these limitations, the interactive Change-Agent model [7] introduced multi-task learning to simultaneously handle both change detection and captioning. However, its reliance on an external large language model (LLM) and lack of an end-to-end framework restrict its functionality within predefined parameters. Recent vision-language models (VLMs) such as GPT-4 [8] and LLaVA [9] show promise across general domains but struggle with the unique characteristics of RS data, often leading to inaccurate or misleading interpretations.

To overcome these challenges, we introduce ChangeChat, the first bitemporal VLM designed for RS change analysis through multimodal instruction tuning. ChangeChat goes beyond predefined workflows, offering a flexible, interactive platform capable of responding to a diverse range of change-related queries. This capability extends to detailed tasks like change captioning, category-specific quantification, and precise change localization.

Despite these advances, the absence of specialized multimodal instruction-tuning datasets for RS change analysis initially limited ChangeChat's potential. In response, we developed the ChangeChat-87k dataset, comprising 87,195 instructions tailored specifically for RS change scenarios. For dataset generation, we first utilized the captions and change maps from the LEVIR-MCI dataset to obtain diverse multimodal change-related instructions through a rule-based automated pipeline. Additionally, inspired by recent advancements in instruction tuning [9], we leveraged ChatGPT's in-context learning capabilities to automatically convert multimodal samples into appropriate instruction-following formats. This approach ensures a broad variety of instructions, enhancing both the flexibility and robustness of the dataset for RS change analysis tasks.

In summary, this work makes the following contributions:

1) RS change instruction dataset: We developed the RS change instruction dataset with 87k instruction-response pairs, using a pipeline that combines rule-based methods with GPT-assisted generation to automatically create instruction-response pairs. This dataset encompasses a variety of instruction types, including classification, counting, and descriptive



**Fig. 1**. Overview of the proposed ChangeChat. The left side illustrates the network architecture, while the right side shows examples of various types of change analysis.

tasks. It effectively addresses the shortage of instructionfollowing data in the field of RS change analysis and significantly enhances the performance of the ChangeChat model.

2) ChangeChat: We introduce ChangeChat, the first bitemporal VLM specifically designed for interactive RS change analysis. The proposed model integrates multimodal instruction tuning, creating a general-purpose RS change analysis assistant. It is capable of responding interactively to a broad array of queries, thereby providing a comprehensive solution that extends well beyond the capabilities of traditional change detection and captioning methods.

## 2. METHODOLOGY

In this section, we first provide a detailed explanation of the construction of the instruction-tuning dataset for change analysis. Then, we introduce the architecture of ChangeChat.

### 2.1. RS Change Instruction Dataset Generation

The LEVIR-CC [5] has been a pioneering large-scale dataset for addressing the challenge of change captioning in remote sensing images (RSIs), comprising 10,077 pairs of bitemporal images, each accompanied by five natural language descriptions. However, due to the lack of fine-grained change information in LEVIR-CC, such as the quantity and location of changed objects, it is limited to the single task of change captioning. The subsequently released LEVIR-MCI [7] dataset includes pixel-level change maps, but it is still insufficient for building a comprehensive model capable of in-depth change analysis. To address these limitations, we

develop a new large-scale multimodal RS change instruction dataset tailored for change analysis tasks.

Inspired by the success of GPT-based automatic dataset generation tasks [9], our RS change instruction dataset is constructed using a hybrid approach that combines rule-based methods with ChatGPT's in-context learning capabilities to automatically generate both instructions and responses. The proposed change instruction dataset contains six instruction types: 1) change description, 2) binary change detection, 3) category-specific change quantification, 4) change localization, 5) GPT-assisted change instruction, and 6) multi-turn conversation. Table 1 summarizes the number of instructions for each type.

Change captioning. We first expand the change captioning sample  $(I_{t1}, I_{t2}, C)$  into an instruction-following version in a straightforward manner, Human:  $I_{t1} I_{t2} Q < STOP > Assistant: <math>C < STOP >$ . Here,  $I_{t1}, I_{t2}$  and C represent the bitemporal images and the corresponding change caption, respectively. In this case, we fix the instruction Q as: "Please briefly describe the changes in these two images."

**Binary change classification.** We create instructions that ask ChangeChat to determine whether changes have occurred between the two images, expecting a binary "yes" or "no" answer. The ground truth is obtained based on the change map from the LEVIR-MCI dataset.

Category-specific change quantification. We create instructions that guide ChangeChat to quantify changes within specific categories like counting added buildings or roads. These quantity-related instructions are generated based on templates, and the ground truth for the responses is obtained by analyzing the change map using the OpenCV library.

**Table 1**. Overview of the RS change instruction dataset.

Instruction	Type	Number
Change captioning	Rule-based	34,075
Binary change classification	Rule-based	6,815
Category-specific change quantification	Rule-based	6,815
Change localization	Rule-based	6,815
GPT-assisted instruction	GPT-assisted	26,600
Multi-turn conversation	Rule-based	6,815
Total		87,195

**Change localization.** For precise change localization tasks, we create instructions that ask ChangeChat to delineate the contours of changed buildings and roads. The ground truth for the responses is obtained by extracting contours and approximating them with polygons from the objects in the change map. To adapt to the language model, we represent the polygons using a sequence of vertex coordinates:  $P = [(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)]$ , where  $(x_k, y_k)$  represents the normalized coordinates of the k-th vertex, and n is the total number of vertices.

**GPT-assisted change instruction.** We leveraged Chat-GPT's in-context learning capabilities to generate a wider variety of instruction-following data. We began by providing it with a system message to guide its responses. Then, we manually designed a few seed examples for each type of task to help it understand the desired output structure. Specifically, it generated two types of conversational data: (1) question-answer pairs based on five given captions describing changes, and (2) more fine-grained queries incorporating extracted contour and quantification information, enabling detailed questions about quantity and relative localization.

**Multi-turn conversation.** We designed multi-turn dialogues to encourage ChangeChat to perform change analysis using a chain-of-thought (CoT) approach. The instructions are presented in increasing difficulty, beginning with simple binary change classification, followed by change object and quantity identification, and progressing to the complex and detailed change captioning task.

## 2.2. Instruction Tuning for ChangeChat

# 2.2.1. ChangeChat Architecture

ChangeChat follows an architecture inspired by LLaVA [10] but with significant adaptations to suit our change analysis tasks. The model is built upon three key components: i) a vision tower based on CLIP-ViT [11], ii) a cross-modal adaptor, and iii) an LLM based on Vicuna-v1.5 [12].

Our ChangeChat differs significantly from LLava in three aspects. First, LLava can only accept a single image as visual input and lacks the capability to conduct change analysis on multi-temporal images. Second, we enable ChangeChat to output spatial locations represented by coordinates through a carefully constructed instruction dataset, thus bridging the

gap in language models' ability to handle visual localization tasks. Finally, LLava is oriented towards general domains and performs poorly in reasoning about RSIs. Next, we will provide a detailed introduction to each component.

**Vision tower.** We used the pre-trained CLIP-ViT [11] model as the visual feature extractor, which divides each  $256 \times 256$  image into a  $14 \times 14$  grid of patches and encodes each patch into a 1024-dimensional token.

**Cross-modal adaptor.** Following [10], we employ a multi-layer perceptron (MLP) with one hidden layer to align the visual and language modalities. Specifically, the 1024-dimensional tokens from the vision tower are mapped to 4096 dimensions to serve as the input embeddings for the LLM.

**LLM.** We use the LLM Vicuna-v1.5 [12] as the brain of our ChangeChat. We improved the Vicuna model by integrating visual embeddings from two different temporal phases, enhancing its capabilities in change analysis tasks. By finetuning the LLM with low-rank adaptation (LoRA) [13], we optimize key matrix elements to enhance speed while preserving the model's linguistic capabilities. This approach integrates a broader range of contextual knowledge into change analysis tasks, enhancing our model's capabilities in change detection, counting, and localization.

# 2.2.2. Training Details

We initialize the model's weights using the pre-trained CLIP-ViT [11], a pre-trained MLP from [10], and Vicuna-v1.5 [12] as the LLM. The LLM is efficiently fine-tuned using LoRA [13], while the parameters of the visual tower and the MLP adapter are frozen. The LoRA fine-tuning is implemented with a rank r of 64 and a scaling factor  $\alpha$  of 128. The training leverages mixed-precision techniques to accelerate the process and reduce memory usage, while dynamic loss scaling is employed to maintain numerical stability.

We train ChangeChat for one epoch on the ChangeChat-87k dataset using the AdamW optimizer with a cosine learning rate scheduler. The learning rate is set to  $2\times 10^{-4}$ , with a warmup ratio of 3%. The training was completed using a single NVIDIA L20 GPU with 48GB of memory. To achieve an effective global batch size of 96, we use a batch size of 16 and accumulate gradients over 6 steps.

# 3. EXPERIMENTAL RESULTS

## 3.1. Dataset and Evaluation Metrics

We evaluate the model performance using 1929 samples from the LEVIR-CC test set. For the change captioning task, we employ BLEU-1[14] for precision between generated hypotheses and reference sentences, METEOR [15] for alignment consideration of synonyms and word order, ROUGE-L [16] for recall based on the longest common subsequence, and CIDEr [17] for similarity with multiple references. Additionally, we assessed accuracy and recall for

binary change classification, and used mean absolute error (MAE) for category-specific change quantification.

# **3.2.** Comparison with SOTA Change Captioning Models

In this section, we evaluate the performance of ChangeChat across various change analysis tasks. For the change captioning task, we compare it with SOTA methods, including Capt-Dual-Att [18], DUDA [18], MCCFormer-S [19], MCCFormer-D [19], RSICCFormer [5], Prompt-CC [6].

To achieve more robust change captioning results, we propose a CoT reasoning approach, guiding ChangeChat through step-by-step reasoning in a multi-turn dialogue. Specifically, we first prompt ChangeChat to determine if any changes have occurred, then guide it to assess specific changes in roads and buildings. Finally, it generates the change caption based on the gathered information.

The results are shown in Table 2, ChangChat achieves comparable or even better results than SOTA models. Notably, all the comparison models are limited to the single task of change captioning, while our proposal is a more general, multi-task-oriented change analysis system.

**Table 2**. Comparison with SOTA methods on the change captioning task. The best is in **bold**, the second best is underlined.

Method	BLEU-1	METEOR	ROUGE-L	CIDEr-D
Capt-Dual-Att [18]	78.17	35.23	71.60	127.51
DUDA [18]	79.64	35.76	71.47	128.24
MCCFormer-S [19]	79.38	36.88	71.06	127.90
MCCFormer-D [19]	76.23	35.18	68.43	121.10
RSICCFormer [5]	81.96	38.16	72.57	132.00
PSNet [20]	81.97	37.92	73.10	132.87
Prompt-CC [6]	83.66	38.82	73.72	136.44
ChangeChat(Ours)	83.14	38.73	74.01	136.56

### 3.3. Evaluation on Diverse Change Analysis Tasks

For more diverse change analysis tasks, due to the lack of prior research, we take the GPT-4 [8] as the baseline.

**Binary change classification.** We design an instruction for ChangeChat and GPT-4 to determine whether a change has occurred between two RSIs, prompting it to respond with only *yes* or *no* using the phrase "*Please answer yes or no*". As shown in Table 3, ChangeChat outperforms GPT-4 in both accuracy and recall.

Category-specific change quantification. We focus on quantifying changes in roads and buildings, converting number-related outputs from the language model into Arabic numerals using predefined rules to calculate MAE with the ground truth. Table 3 demonstrates that our method outperforms SOTA large VLMs in this task.

**Change localization** To improve reasoning efficiency, we use highly simplified polygons to approximate the general location of the changed objects. We present two examples of change localization in Fig. 2. The results show that

**Table 3.** Comparison on binary change classification and change quantification tasks.

Model	Binary Classification		Change Quantification	
	Acc.	Recall	MAE (road)	MAE (building)
GPT-4 ChangeChat	84.81% <b>93.21%</b>	86.62% <b>92.53</b> %	0.62 <b>0.33</b>	2.91 <b>2.67</b>

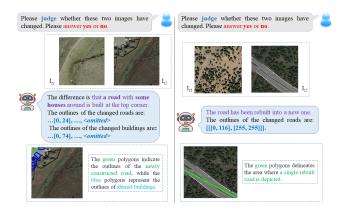


Fig. 2. Two examples of change localization are provided, with the generated coordinates visualized.

ChangeChat can effectively identify and localize the changed objects.

## 3.4. Discussion on CoT reasoning

To validate the effectiveness of our CoT reasoning strategy, we compare the change captioning results with and without using CoT reasoning, as shown in Table 4. The results indicate that guiding ChangeChat through increasingly complex instructions indeed improves its performance.

**Table 4.** Comparison between with and w/o CoT reasoning.

CoT Reasoning	BLEU-1	METEOR	ROUGE-L	CIDEr-D
×	81.16	37.05	73.37	135.46
$\checkmark$	83.14	38.73	74.01	136.56

#### 4. CONCLUSION

In this paper, we introduced ChangeChat, the first bitemporal VLM designed specifically for RS change analysis. Unlike traditional change captioning models, ChangeChat supports interactive, user-specific queries through multimodal instruction tuning. The model, enhanced by the ChangeChat-87k dataset, excels in tasks such as change captioning, category-specific quantification, and change localization. Experiments show that ChangeChat outperforms SOTA methods and significantly exceeds the performance of the latest general-domain large model, GPT-4, demonstrating its effectiveness in comprehensive RS change analysis.

#### 5. REFERENCES

- [1] Zehua Jiao, "The application of remote sensing techniques in ecological environment monitoring," *High-lights in Science, Engineering and Technology*, vol. 81, pp. 449–455, 2024.
- [2] Sultan Al Shafian and Da Hu, "Integrating machine learning and remote sensing in disaster management: A decadal review of post-disaster building damage assessment," *Buildings*, vol. 14, no. 8, pp. 2344, 2024.
- [3] Tengfei Bao, Chenqin Fu, Tao Fang, and Hong Huo, "Ppcnet: A combined patch-level and pixel-level endto-end deep network for high-resolution remote sensing image change detection," *IEEE Geoscience and Remote* Sensing Letters, vol. 17, no. 10, pp. 1797–1801, 2020.
- [4] Zhuo Zheng, Yanfei Zhong, Shiqi Tian, Ailong Ma, and Liangpei Zhang, "Changemask: Deep multi-task encoder-transformer-decoder architecture for semantic change detection," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 183, pp. 228–239, 2022.
- [5] Chenyang Liu, Rui Zhao, Hao Chen, Zhengxia Zou, and Zhenwei 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.
- [6] Chenyang Liu, Rui Zhao, Jianqi Chen, Zipeng Qi, Zhengxia Zou, and Zhenwei Shi, "A decoupling paradigm with prompt learning for remote sensing image change captioning," *IEEE Transactions on Geo*science and Remote Sensing, 2023.
- [7] Chenyang Liu, Keyan Chen, Haotian Zhang, Zipeng Qi, Zhengxia Zou, and Zhenwei Shi, "Change-agent: Towards interactive comprehensive remote sensing change interpretation and analysis," *IEEE Transactions on Geo*science and Remote Sensing, 2024.
- [8] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al., "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.
- [9] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee, "Visual instruction tuning," *Advances in neural information processing systems*, vol. 36, 2024.
- [10] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee, "Improved baselines with visual instruction tuning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 26296–26306.

- [11] Yi Tay, Minh C Phan, Luu Anh Tuan, and Siu Cheung Hui, "Learning to rank question answer pairs with holographic dual lstm architecture," in *Proceedings of the 40th international ACM SIGIR conference on research and development in information retrieval*, 2017, pp. 695–704.
- [12] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al., "Judging Ilm-asa-judge with mt-bench and chatbot arena," *Advances in Neural Information Processing Systems*, vol. 36, 2024.
- [13] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen, "Lora: Low-rank adaptation of large language models," *arXiv preprint arXiv:2106.09685*, 2021.
- [14] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing 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.
- [15] Satanjeev Banerjee and Alon 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.
- [16] Chin-Yew Lin, "Rouge: A package for automatic evaluation of summaries," in *Text summarization branches out*, 2004, pp. 74–81.
- [17] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh, "Cider: Consensus-based image description evaluation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 4566–4575.
- [18] Dong Huk Park, Trevor Darrell, and Anna Rohrbach, "Robust change captioning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 4624–4633.
- [19] Yue Qiu, Shintaro Yamamoto, Kodai Nakashima, Ryota Suzuki, Kenji Iwata, Hirokatsu Kataoka, and Yutaka Satoh, "Describing and localizing multiple changes with transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 1971–1980.
- [20] Chenyang Liu, Jiajun Yang, Zipeng Qi, Zhengxia Zou, and Zhenwei 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.