RSICCLLM: A Multimodal Large Language Model for Remote Sensing Image Change Captioning

Yelin Wang[1,2,*], Zijia Song[3,*], Shuo Ye[2], Chuanguang Yang[1], Miaoyu Wang[1], Yong Xu[3], Zhulin An[1,†], Yongjun Xu[1], and Zitong Yu[2,4,†]

[1] State Key Laboratory of AI Safety, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China
[2] Great Bay University, Dongguan, China
[3] Harbin Institute of Technology, Shenzhen, China
[4] Dongguan Key Laboratory for Intelligence and Information Technology, Dongguan, China

Abstract. Remote Sensing Image Change Captioning (RSICC) aims to describe changes between bi-temporal remote sensing images and holds significant research and application value. However, most existing methods rely on conventional deep learning architectures, and the limited model capacity constrains performance. Although large-model post-training techniques have achieved great success in general domains, their direct transfer to RSICC remains challenging due to data scarcity and the need for fine-grained change understanding. To address this, we propose RSICCLLM, the first post-training framework for large vision-language models in RSICC. Specifically, we design **a data generation paradigm**, release the instruction dataset RSICI, and establish **a task-specific RSICC benchmark**. We further introduce **Difference-aware Supervised Fine-tuning** to explicitly extract change representations and guide the model in perceiving and understanding temporal differences. In addition, we propose **Dual-Negative Preference Optimization (DNPO)**, which employs two complementary negative-sample construction strategies to construct the preference dataset RSICP and further refine model performance. Extensive experiments validate the superior capability of RSICCLLM, which achieves outstanding results with only 7B parameters, surpassing models of substantially larger scales. The code and dataset will be made publicly available at `https://github.com/keaill/RSICCLLM`.

Keywords: Data generation paradigm · Difference-aware supervised fine-tuning · Dual-Negative preference optimization · Task-specific RSICC benchmark

[*] These authors contributed equally to this work.
[†] Corresponding Authors.

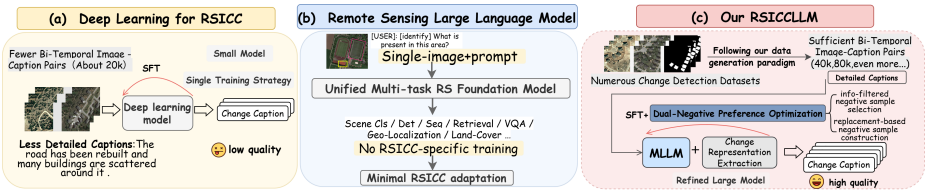

Fig. 1: The differences between the previous approaches and ours. (a) Previous methods are constrained by **limited model capacity** and **data scarcity**, hindering performance in change captioning. (b) Existing remote-sensing foundation models are rarely post-trained for RSICC with instruction or preference alignment, mainly due to limited RSICC data and preference supervision. (c) We introduce a data generation paradigm and train a refined larger model with Supervised Fine-tuning and Dual-Negative Preference Optimization, achieving superior results.

1 Introduction

Remote sensing image change captioning (RSICC) generates descriptions of changes between bi-temporal remote sensing images [18, 30, 59]. Unlike remote sensing image change detection (RSICD), which reports only visual differences [32, 55], RSICC maps visual evidence to semantics, improving interpretability and supporting many important public affairs, such as environmental assessment and disaster response. As an essential component of the remote sensing field, RSICC still faces multiple challenges [51,59], such as accurately perceiving real changes between bi-temporal images while ignoring irrelevant factors like illumination, and capturing subtle, diverse land-cover variations while aligning them with natural language. Therefore, RSICC holds substantial scientific and practical significance.

Recently, RSICC has attracted increasing research attention and made continuous progress. Semantic-CC [58] uses a SAM-based dual-image encoder with a lightweight multi-scale change detection (CD) decoder for pixel-level semantic guidance. Pix4Cap [28] adds a CD branch to produce pseudo-labels and a semantic-fusion augmentation to inject CD features into the captioner. Prompt-CC [31] introduces an image-level change classifier and a multi-prompt strategy that leverages an off-the-shelf caption generator to produce plausible captions without retraining. However, these methods are primarily based on conventional deep learning frameworks, and their relatively small model capacity imposes an upper limit on overall performance.

Meanwhile, large-scale models and post-training techniques have shown strong success in general domains [16, 50] and are being introduced into remote sensing [23, 44]. For example, RS-GPT and Falcon leverage large-scale multimodal training to achieve strong performance on remote-sensing captioning and various downstream tasks [20, 53]. Compared with conventional deep learning models, large models offer significant gains from increased capacity. However, due to the scarcity of specific datasets for RSICC and the limited ability of existing vision-language models to understand temporal changes, generating accurate and reliable change descriptions remains challenging.

Motivated by these observations, we explore whether large-scale models and post-training techniques can be introduced into the RSICC field to enhance understanding of changes and achieve performance breakthroughs. Based on this idea, we propose RSICCLLM, the first **post-training framework** of large models tailored for RSICC. The differences between the previous approaches and ours are shown in Fig. 1. We first construct **an instruction dataset generation paradigm** for RSICC using large-scale RSICD data. Specifically, we employ Qwen-VL-Max to generate instruction samples by feeding bi-temporal remote sensing images, together with their corresponding binary masks as geometric priors. By prompting the model to describe changes within the masked regions, we obtain a high-quality instruction dataset named **Remote Sensing Image Change Instruction dataset (RSICI)**. Based on this dataset, we establish **a task-specific RSICC benchmark**. We then adopt Qwen2.5-VL-7B as the backbone for subsequent training. To further improve the model's perception and understanding of temporal changes while suppressing irrelevant factors, we propose **Difference-aware Supervised Fine-tuning**, which explicitly extracts change representation and enhance the visual encoder with Central Difference Convolution (CDConv) [54] and Hough Transform modules [4]. Using RSICI, we perform Supervised Fine-tuning (SFT) to obtain RSICCLLM-sft. Subsequently, we propose **Dual-Negative Preference Optimization (DNPO)** to further enhance model performance. We first design two strategies for constructing negative samples and build the **Remote Sensing Image Change Preference dataset (RSICP)**. In addition, we perform preference optimization on RSICCLLM-sft using RSICP and introduce a new loss operator to stabilize the training process, yielding the final model RSICCLLM.

Our main contributions can be summarized as follows:

- We design **a new instruction dataset generation paradigm** in RSICC, building a large-scale instruction dataset RSICI and establish a task-specific **RSICC benchmark**. We also release the first preference dataset RSICP dedicated to this task.
- We propose RSICCLLM, the first post-training framework of large models for RSICC. To improve the model's understanding of temporal changes, we propose **Difference-aware Supervised Fine-tuning.** And **a new preference optimization objective DNPO** is introduced to progressively refine the model's performance.
- Extensive experiments demonstrate the superior performance of our approach. Notably, RSICCLLM achieves state-of-the-art performance across all evaluation metrics with only 7B parameters.

2 Related Works

Remote Sensing Image Change Captioning. RSICC has recently emerged as a new research direction in remote sensing [10,12,15,46]. Early work by Chouaf *et al.* [11] explored RSICC on a private dataset with a VGG-16 encoder and an RNN decoder, which was later extended by Hoxha *et al.* [18] using image- and

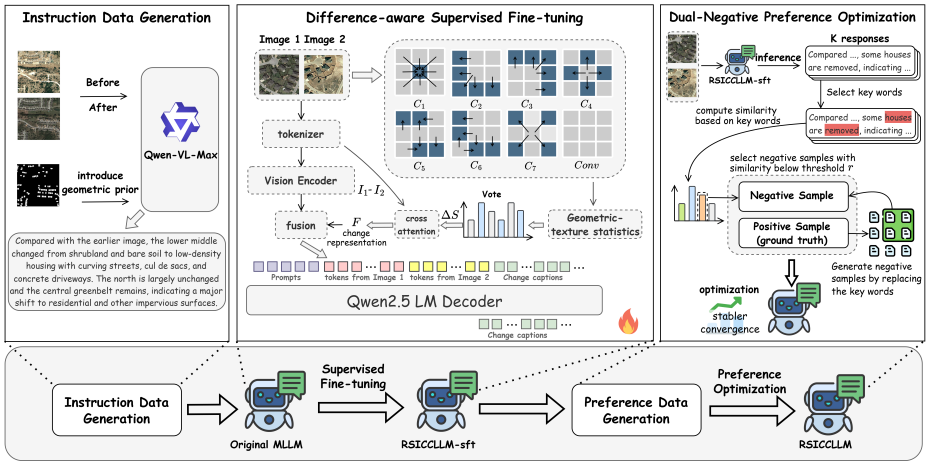

Fig. 2: Overview of the proposed RSICCLLM framework. Left: Instruction Data Generation by Qwen-VL-Max [1] with the binary mask as geometric prior. **Middle**: Difference-aware Supervised Fine-tuning with Central Difference Convolution [54], asymmetric operators, and Soft-Hough voting [4] for explicitly extracting change representation. **Right**: Dual-Negative Preference Optimization with two strategies to construct preference pairs, yielding the final RSICCLLM.

feature-level fusion encoders. Liu *et al.* [30] established the LEVIR-CC benchmark and proposed a CNN–transformer framework to emphasize changed regions. Subsequent studies further improved RSICC with progressive scale-aware modeling [29] and Mamba-based spatio-temporal feature extraction [6]. All the aforementioned methods are deep-learning-based, and their progress is severely constrained by limited datasets and small model capacity. In contrast, our approach addresses these limitations by introducing a data generation paradigm and a large-scale training framework.

Multimodal Reasoning. Recent work shifts multimodal reasoning from supervised learning to reinforcement-based fine-tuning with explicit reasoning [13, 36, 47, 52]. FAST and TON reduce unnecessary computation via difficulty-aware control or a learned "think-or-not" policy [45, 49]. Chain-of-Focus improves region localization with RL-based adaptive search/zooming, and MiCo leverages cross-image contrastive learning with verifiable QA-free rewards [8, 56]. UniVG-R1 advances visual referring through CoT initialization and GRPO-style optimization with reweighting [3]. Kimi-VL and G1 further highlight efficient MoE-style VLMs and RL gains after perception-first cold starts [7, 40], while Qwen-AD advocates a look-before-you-decide strategy [26]. These advances motivate a domain-specific RSICC framework for further breakthroughs.

3 Methodology

3.1 Overview of Framework

We propose a three-stage RSICC framework, RSICCLLM (Fig. 2): **1) Instruction Data Generation Stage.** We use Qwen-VL-Max to synthesize instruction samples from bi-temporal images, conditioned on a binary change mask as

a geometric prior, with prompts eliciting explicit land-surface transition descriptions. **2) Difference-aware Supervised Fine-tuning Stage.** We enhance the vision encoder by injecting manually extracted change representations into its forward pass, and then perform SFT with the generated instructions to obtain RSICCLLM-sft. **3) Dual-Negative Preference Optimization Stage.** To enable preference training, we construct preference datasets using two complementary strategies, which are info-filtered negative sample selection and replacement-based negative sample construction. The resulting datasets are merged and used to further fine-tune RSICCLLM-sft via preference optimization, yielding RSICCLLM with stronger discriminative and descriptive abilities.

3.2 Instruction Data Generation

In the data generation stage, we integrate the LEVIR-CD [5] and SYSU-CD [39] datasets to automatically generate approximately 35k samples, among which 32k are used for training and 3k as ***in-domain evaluation set***, covering building changes as well as diverse man-made and natural changes. Meanwhile, we further integrate S2Looking [37] and CDD [24] to automatically generate approximately 5k samples as ***out-of-domain test set***, which mainly consist of rural scenes, off-nadir building changes, and seasonally varying imagery, providing a more robust assessment for out-of-domain performance. Each sample consists of a pre-change image, a post-change image, and a fine-grained textual description of the differences. For the Remote Sensing Image Change Instruction (RSICI) dataset, the instruction data generation process is shown in Algorithm 1. The data construction details and ablations are in **Appendix A**.

3.3 Difference-aware Supervised Fine-tuning

We perform Supervised Fine-tuning (SFT) on Qwen2.5-VL-7B for change captioning, jointly modeling temporal dynamics and time-invariant spatial geometry to improve sensitivity to *what* changed and *where* it changed.

Given a bi-temporal image pair Image_1 and Image_2, we first perform differentiable spatial registration to remove non-semantic pixel shifts arising from sensor pose or terrain relief. Afterward, we tokenize the two images and compute a difference map as the initial change cue:

$$T = I_1 - I_2 \,, \tag{1}$$

where T encodes the magnitude of local intensity changes and highlights candidate change locations.

To extract time-invariant spatial geometry, we introduce eight geometry-sensitive convolution kernels inspired by Central Difference Convolution (CDConv) [54]. Given bi-temporal remote-sensing images as input, these kernels can extract filtered geometric textures that are insensitive to illumination and weather. Specifically, they include one standard convolution operator C_8; one CDConv operator C_1; four directional asymmetric operators-C_2, C_3, C_5, and

Algorithm 1: Construction of RSICC-INSTRUCTION from LEVIR_CD, SYSU_CD, CDD, S2Looking

Input: Image pairs D (before, after) with binary masks M; Multimodal LLM; retry limit R=2
Output: Curated instruction dataset $\mathcal{D}$

$\mathcal{D} \leftarrow \emptyset$; `// Initialize dataset`
foreach $(I^{(1)}, I^{(2)}, M, \text{meta}) \in D$ **do**
  Construct prompt P using $(I^{(1)}, I^{(2)}, M)$ with geometric constraints and few-shot examples;
  for $r = 1$ **to** R **do**
    Generate caption y using the MLLM with prompt P;
    Extract structured elements $\{object, direction, verb, quantity\}$ from y;
    if *change detected in mask* M **then**
      Check whether **object** and **verb** are correctly mentioned in y;
    if y *mentions words like "mask" or "segmentation"* **then**
      Reject the sample and retry;
    if y *is excessively long* **then**
      Reject the sample and retry;
    if *semantic hallucinations or off-topic content are detected* **then**
      Reject the sample and retry;
    if *all conditions are satisfied* **then**
      Add $(I^{(1)}, I^{(2)}, M, y, \text{domain})$ to $\mathcal{D}$;
      break
    Adjust decoding strategy: change prompt length or examples;
Remove samples with undesirable keywords or formatting issues;
Three domain experts re-check *all* samples in $\mathcal{D}$; remove any sample that fails the quality criteria;
Apply synonym substitution on extracted elements to augment linguistic diversity;
return $\mathcal{D}$;

C_6; and two symmetric operators-C_4 and C_7. Their formats are shown in Fig. 2 and definitions are as follows:

$$C_i^t(r_x, r_y) = \sum_{(\Delta r_x, \Delta r_y) \in \mathcal{R}} \text{Image}_t(r_x + \Delta r_x, r_y + \Delta r_y) \cdot \left(\sum_{(\Delta r_x, \Delta r_y) \in \mathcal{R}} \delta_i(\Delta r_x, \Delta r_y) \omega(\Delta r_x, \Delta r_y) \right), \tag{2}$$

where $\mathcal{R} = \{(-1,-1), (-1,0), ..., (1,0), (1,1)\}$ denotes the local receptive field of the trainable 3×3 vanilla convolution kernel ω, and $\text{Image}_t \in \mathbb{R}^{C \times H \times W}$ is the input feature map. $t \in \{1, 2\}$ indicates the time index of images, and i is the index of operators. (r_x, r_y) represents the current pixel location, and $\delta_i(\Delta r_x, \Delta r_y)$ is a gate function to determine whether each convolution kernel element $\omega(\Delta r_x, \Delta r_y)$ participates in the computation. We employ Hough Transform [4], and project each response map C_i^t $(t \in \{1, 2\})$ into the Hough parameter space (ρ, θ) to

accumulate geometric evidence via voting:

$$H_i^t(\rho,\theta) \;=\; \mathrm{HoughTransform}\big(C_i^t\big), t \in \{1,2\}, \tag{3}$$

where i denotes the index of operators. After voting, we obtain the final bi-temporal texture statistic:

$$H^t \;\triangleq\; \big\{\, H_i^1(\rho,\theta),\; H_i^2(\rho,\theta) \,\big\}, \tag{4}$$

where $\triangleq$ indicates the voting operation. Crucially, instead of differencing directly in the parameter space, we independently apply the inverse Hough transform to each time step to reconstruct the spatial structural responses:

$$S^{(t)} \;=\; \mathrm{Hough}^{-1}\big(H^t(\rho,\theta)\big), t \in \{1,2\}, \tag{5}$$

where $S^{(t)}$ preserves complete geometric semantics for both bi-temporal images, avoiding the semantic loss caused by early differencing. We obtain a normalized geometric–texture features as follows:

$$\Delta S \;=\; \frac{S^{(2)} - S^{(1)}}{\|S^{(1)}\|_1 + \|S^{(2)}\|_1 + \epsilon}. \tag{6}$$

To better represent the change patterns, we employ cross-attention to fuse the preliminary feature cues T with the geometric–texture features ΔS. Specifically, we treat T as the query vector and ΔS as the key and value vectors. Through learnable projections, both the features T and ΔS are mapped to the same dimension d, and the feature fusion process is computed as follows:

$$F \;=\; \mathrm{softmax}\!\left(\frac{T\Delta S^{\top}}{\sqrt{d}}\right)\Delta S \;\in\; \mathbb{R}^{N\times d}, \tag{7}$$

where N denotes the number of image tokens after tokenization, and F indicates the final change representation. We then enhance the original image embeddings $V_{\mathrm{orig}} \in \mathbb{R}^{N\times d}$ from the vision encoder as follows:

$$V_{\mathrm{enhanced}} \;=\; V_{\mathrm{orig}} + F, \tag{8}$$

where V_{enhanced} focuses more on the change regions within remote sensing images. The enhanced visual representations from the bi-temporal images are then combined with the prompt tokens and fed into the Qwen2.5 LM decoder to generate accurate change captions.

For SFT, we employ autoregressive (AR) objective as loss function. Given a ground truth $y = (y_1, \ldots, y_T)$ corresponding to bi-temporal images, the formula is as follows:

$$\mathcal{L}_{\mathrm{AR}} = -\sum_{t=1}^{T} \log P(y_t \mid y_{<t},\, I_1,\, I_2,\, P;\, \theta)\,, \tag{9}$$

where I_1 and I_2 denote the visual embeddings obtained from the two image inputs, and P represents the prompt token. After concatenating the visual embeddings and the prompt token, they are fed into the decoder with teacher forcing.

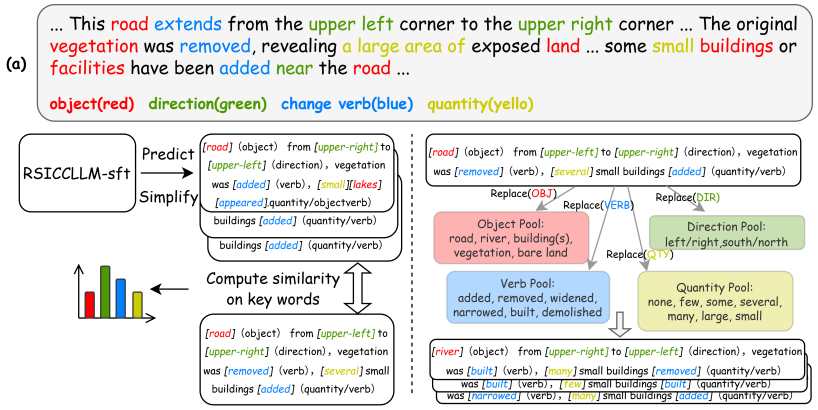

Fig. 3: Two complementary negative sample construction strategies: (a) semantic decomposition of change captions into four key components; (b) info-filtered selection based on keyword similarity; and (c) replacement-based generation via controlled token substitution.

y_t is the t-th token in the ground-truth sequence, and T denotes the sequence length. θ denotes the learnable parameters of the model. After Difference-aware Supervised Fine-tuning, we acquire the RSICCLLM-sft, which possesses preliminary perceptual and analytical capabilities, providing a solid foundation for further refinement through preference optimization.

3.4 Dual-Negative Preference Optimization

Preference learning is a mainstream post-training technique that further improves generation quality beyond SFT by leveraging preference signals, even without explicit annotations. Its effectiveness, however, strongly depends on the quality of the preference dataset. For RSICC, the lack of explicit ground-truth labels makes quality assessment difficult, which in turn hinders the construction of high-quality preference data. To address this issue, we propose Dual-Negative Preference Optimization, which constructs positive-negative sample pairs through two complementary strategies to facilitate preference learning.

We first summarize the characteristics of change captions, as illustrated in Figure 3(a). Among these elements, object, direction, change verb, and quantity carry high information content and directly affect the sentence meaning; thus, they are defined as keywords. In contrast, articles, prepositions, and pronouns mainly serve to enrich sentence fluency and have limited impact on semantics, so they are referred to as non-keywords. The two proposed strategies for constructing the preference dataset are both designed based on this definition.

The first strategy is info-filtered negative sample selection and the process is shown in Fig. 3(b). For each input x containing a pair of bi-temporal images, we use RSICCLLM-sft to generate K candidate captions $\{y_k\}_{k=1}^{K}$. For any candidate caption $y = (w_1, \ldots, w_n)$, we define an indicator function $\phi(w_i) \in \{0, 1\}$ to identify words with high information content. Specifically, we set $\phi(w_i) = 0$ for non-keywords, while assigning $\phi(w_i) = 1$ to keywords with high information density. Accordingly, we obtain a filtered sequence composed only of the keywords

as follows:

$$y^{\text{info}} = \big(w_i \in y \big| \phi(w_i) = 1\big), \tag{10}$$

where y^{info} denotes the filtered sequence for input x, which contains only the extracted keywords. The purpose of introducing $\phi(w_i)$ is to retain only the keywords that represent actual change facts, allowing for more accurate similarity computation. Similarly, for the ground-truth caption y_{GT}, we obtain the corresponding filtered sequence $y_{\text{GT}}^{\text{info}}$. We then calculate the overall similarity between y^{info} and $y_{\text{GT}}^{\text{info}}$ using the averaged BLEU-1, BLEU-2, and ROUGE-1:

$$\begin{aligned} S_{\text{sim}}(y^{\text{info}}, y_{\text{GT}}^{\text{info}}) = \frac{1}{3}(&\text{BLEU-1}(y^{\text{info}}, y_{\text{GT}}^{\text{info}}) \\ &+ \text{BLEU-2}(y^{\text{info}}, y_{\text{GT}}^{\text{info}}) \\ &+ \text{ROUGE-1}(y^{\text{info}}, y_{\text{GT}}^{\text{info}})), \end{aligned} \tag{11}$$

where $S_{\text{sim}}(y^{\text{info}}, y_{\text{GT}}^{\text{info}})$ denotes the overall similarity between a candidate caption y and the ground truth. For each candidate caption y generated from the input x, we compute its similarity score following the above procedure and then select the negative sample set $\mathcal{N}_{\text{sel}}$ for x according to the following criterion:

$$\mathcal{N}_{\text{sel}} = \big\{ y \in \{y_k\}_{k=1}^{K} \big| \tau_1 < S_{\text{sim}}(y^{\text{info}}, y_{\text{GT}}^{\text{info}}) < \tau_2 \big\}, \tag{12}$$

where τ_1 and τ_2 denote the lower and upper bounds of the similarity threshold, respectively. Through this constraint, we can identify an appropriate set of negative samples for each input, ensuring that they are neither too difficult nor too easy to distinguish from the positive examples.

The second strategy is replacement-based negative sample construction, which acquires negative samples by directly modifying the keywords in the ground truth y_{GT}. And the process is shown in Fig. 3(c). Specifically, we define the keyword set of the ground truth as $y_{\text{GT}}^{\text{info}} = \{w \in y_{\text{GT}} \mid \phi(w) = 1\}$ and introduce a replacement mapping $\psi : y_{\text{GT}}^{\text{info}} \to \mathcal{V}$, where $\mathcal{V}$ denotes a contrastive replacement pool built from the corpus. This pool is grouped by object, directions, change verbs, and quantity. When performing replacements, new words are selected only from the same group to preserve part-of-speech and morphological consistency, so that the resulting negative samples remain natural while being semantically incorrect in key aspects. For each ground-truth caption y_{GT}, we randomly perform M replacements to obtain the negative sample set $\mathcal{N}_{\text{con}} = \{\tilde{y}_m\}_{m=1}^{M}$. By combining the datasets from both approaches, we construct the final preference dataset containing 20k image pairs:

$$\mathcal{D}_{\text{pref}} = \left\{ \left(x,\ y_w,\ y_l\right) \ \middle|\ y_l \in \mathcal{N}_{\text{sel}} \cup \mathcal{N}_{\text{con}} \right\}, \tag{13}$$

where y_w denotes the ground truth y_{GT}. Through the synthesized preference dataset, we perform preference optimization with the following loss function:

$$\mathcal{L}(\pi_\theta; \pi_{\text{ref}}) = \mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) + \lambda \mathcal{L}_{\text{S}}(\pi_\theta; \pi_{\text{ref}}), \tag{14}$$

Table 1: Performance comparison of different models on our ***in-domain*** RSICI test set to evaluate their capabilities for remote sensing image change captioning. The best results are highlighted in **bold**.

Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L	SBS
Qwen-VL-Max [1]	27.55	12.12	5.18	2.70	34.20	6.86	20.10	71.79
GLM-4.5V [14]	28.13	12.41	6.67	3.73	31.88	6.73	20.10	69.04
SenseChat-Vision [35]	27.63	12.15	6.24	3.58	31.66	6.47	19.51	68.76
Intern-S1 [21]	26.47	11.97	5.47	3.00	31.98	6.78	19.18	70.89
InternVL-3.5-241B [21]	26.81	12.01	5.32	2.83	32.83	6.80	19.77	72.98
Qwen3-VL-235B [41]	27.15	12.32	5.01	2.82	33.79	6.37	19.42	71.70
Qwen3-VL-32B [41]	26.39	11.66	4.87	2.54	32.44	6.54	18.82	69.57
Qwen2.5-VL-72B [2]	27.54	12.02	4.99	2.65	33.71	6.72	19.38	70.06
TEOChat-7B [22]	8.63	2.74	1.37	0.94	18.89	3.18	11.13	40.97
CCExpert-7B [48]	8.63	2.57	1.34	0.88	17.65	2.94	10.81	40.31
Qwen2.5-VL-7B [2]	7.40	2.35	1.08	0.61	17.27	2.79	10.16	42.06
RSICCLLM-sft	80.13	72.99	65.72	61.68	52.31	27.55	37.51	80.38
RSICCLLM	**81.28**	**74.70**	**66.81**	**62.78**	**53.54**	**28.94**	**38.30**	**82.72**

where π_θ denotes the model continuously optimized during post-training, while π_{ref} refers to RSICCLLM-sft. $\mathcal{L}_{\text{DPO}}(\cdot,\cdot)$ indicates the Direct Preference Optimization(DPO), and the formula is shown as follows:

$$\begin{aligned}\mathcal{L}_{\text{DPO}}(\pi_\theta;\pi_{\text{ref}}) &= -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}_{\text{pref}}}\left[\log p(y_w \succ y_l \mid x)\right] \\ &= -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}_{\text{pref}}}\left[\log\sigma\left(\beta\log\frac{\pi_\theta(y_w \mid x)}{\pi_{\text{ref}}(y_w \mid x)}\right.\right. \\ &\left.\left. - \beta\log\frac{\pi_\theta(y_l \mid x)}{\pi_{\text{ref}}(y_l \mid x)}\right)\right],\end{aligned} \tag{15}$$

where minimizing $\mathcal{L}_{\text{DPO}}$ encourages π_θ to align with the best policy model, thereby increasing the likelihood of producing responses with higher quality. $\mathcal{L}_{\text{S}}(\cdot;\cdot)$, on the other hand, stabilizes training by constraining the divergence between the π_θ and π_{ref}. The formula is shown as follows:

$$\mathcal{L}_{\text{S}}(\pi_\theta;\pi_{\text{ref}}) = \mathbb{E}_{x\sim\mathcal{D}_{\text{pref}}}\left[D_{\text{KL}}\left(\pi_\theta(\cdot \mid x)\,\|\,\pi_{\text{ref}}(\cdot \mid x)\right)\right]. \tag{16}$$

Here, we adopt the reverse KL divergence to prevent the policy model from deviating excessively from the reference model, thus ensuring stable training. Through our proposed DNPO, our model can leverage comprehensive preference information to achieve more stable preference training and generate higher-quality change captions.

4 Experiments

4.1 Metrics and Implementation Details

Benchmark metrics. We propose a new benchmark to evaluate model's ability of analyzing changes between bi-temporal images, and employ BLEU-1, BLEU-2, BLEU-3, BLEU-4 [33], ROUGE-1, ROUGE-2, and ROUGE-L [27] as evaluation metrics. BLEU-1 and ROUGE-1 measure the accuracy of key terms, BLEU-2 and ROUGE-2 reflect local semantic consistency, BLEU-3 and BLEU-4 assess sentence-level coherence, and ROUGE-L evaluates global semantic consistency.

Table 2: Performance comparison of different models on our ***out-of-domain*** RSICI test set for remote sensing image change captioning.

Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L	SBS
Qwen-VL-Max [1]	32.87	13.12	5.77	3.16	36.14	6.16	21.15	71.18
GLM-4.5V [14]	28.09	12.33	6.53	3.67	31.65	6.52	19.93	68.92
SenseChat-Vision [35]	27.88	12.48	6.35	3.69	31.89	6.67	19.65	69.02
Intern-S1 [21]	28.42	10.81	4.61	2.47	32.28	5.05	19.68	65.27
InternVL-3.5-241B [21]	28.67	11.08	4.66	2.47	32.58	5.22	19.70	65.45
Qwen3-VL-235B [41]	30.28	11.68	4.93	2.65	33.80	5.38	20.26	70.25
Qwen3-VL-32B [41]	29.33	10.95	4.63	2.49	33.88	5.19	20.35	64.69
Qwen2.5-VL-72B [2]	30.89	13.20	5.71	3.06	34.45	6.73	20.25	68.96
TEOChat-7B [22]	9.06	2.86	1.44	1.03	17.75	3.14	11.15	41.76
CCExpert-7B [48]	8.96	2.71	1.33	0.96	17.88	3.21	11.14	40.84
Qwen2.5-VL-7B [2]	11.19	3.39	1.50	0.84	20.33	2.09	12.64	56.48
RSICCLLM	**69.51**	**63.16**	**55.06**	**47.84**	**43.03**	**15.05**	**25.89**	**73.86**

Table 3: Comparison with existing remote sensing text-image datasets. Avg_L denotes the average number of words per caption. Temporal indicates whether bi-temporal images are included in one sample. Preference specifies whether the dataset contains preference information, and ChangeCap denotes whether this dataset is designed for the RSICC task.

Dataset	Year	#Image	Caption		Temporal	Preference	ChangeCap	Field
			#Captions (Avg_L)	Details				
SpaceNet 7 [42]	2021	2,389	N/A	✗	✓	✗	✗	Urban
S2Looking [37]	2021	5,000	N/A	✗	✓	✗	✗	Urban
QFabric [43]	2021	2,520	N/A	✗	✓	✗	✗	Urban
SpaceNet 8 [17]	2022	2,576	N/A	✗	✓	✗	✗	Disaster
LEVIR-CC [30]	2022	20,154	50,385 (40)	✗	✓	✗	✓	General
Dubai-CC [18]	2022	1,000	2,500 (35)	✗	✓	✗	✓	Urban
RSICap [19]	2023	2,585	2,585 (60)	✓	✗	✗	✗	General
RS5M [57]	2024	5M	5M (49)	✓	✗	✗	✗	General
VRSBench [25]	2024	29,614	29,614 (52)	✓	✗	✗	✗	General
WHU-CDC [38]	2024	14,868	37,170 (–)	✓	✓	✗	✓	Disaster
XLRS-Bench [44]	2025	934	934 (379)	✓	✗	✗	✗	General
RSCC [9]	2025	124,702	62,351 (72)	✓	✓	✗	✓	Disaster
RSICI (Ours)	2026	40,000	20,000 (77)	✓	✓	✗	✓	General
RSICP (Ours)	2026	20,000	10,000 (81)	✓	✓	✓	✓	General

Given that change captions are open-ended and allow multiple valid phrasings, we introduce *Sentence-BERT Similarity (SBS)* to evaluate the overall semantic similarity between the prediction and ground truth as follows:

$$\mathrm{SBS} = \frac{1}{N}\sum_{i=1}^{N}\cos\big(\phi(\hat{y}_i),\,\phi(y_i)\big), \tag{17}$$

where N is the dataset size, and $\phi(\cdot)$ denotes a Sentence-BERT [34] encoder. $\hat{y}_i$ is the predicted caption, and y_i denotes the reference. $\cos(\cdot,\cdot)$ represents the cosine similarity. This metric can quantify the semantic gap between the outputs and the ground-truth captions. A higher value indicates better generation performance on RSICC. We also design a metric to assess the hallucination robustness of our method against other models; details are provided in **Appendix B.1**

Implementation details. All models are trained on 2 NVIDIA A100 GPUs. More details are in **Appendix B.2**.

4.2 Comparative Evaluation

We conduct evaluations on the in-domain test set of RSICI, with comparisons to mainstream multimodal models shown in Table 1. The compared methods include both large-scale general-purpose models and specialized models designed

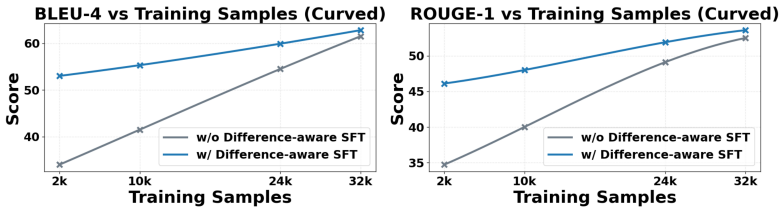

Fig. 4: Ablation on Difference-aware SFT with different training sizes. Each subfigure plots the metric score (vertical axis) against training size (horizontal axis).

Table 4: Ablation on Difference-aware SFT components. G denotes the geometric-texture statistics.

Method	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L
w/o G	62.54	53.13	28.20	38.11
w/o ΔS	61.90	52.89	28.14	37.85
w/o T	60.45	51.37	26.98	36.03
RSICCLLM	**62.78**	**53.54**	**28.94**	**38.30**

Table 5: Effectiveness of training strategies on various metrics.

Strategy	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L
Without SFT	3.21	17.27	2.79	10.94
SFT	61.68	52.31	27.55	37.51
SFT + DNPO	**62.78**	**53.54**	**28.94**	**38.30**

Table 6: Ablation on two negative sample construction strategies under DNPO. Here, S = SFT the model with RSICI; F = preference training with dataset only from info-filtered negative sample selection, R = preference training with dataset only from replacement-based negative sample construction.

Setting	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L
S	61.68	52.31	27.55	37.51
S+F	62.38	53.19	28.44	38.03
S+R	62.16	53.37	28.19	37.92
S+F+R	**62.78**	**53.54**	**28.94**	**38.30**

for change captioning. Clearly, our method, with only 7B parameters, achieves the best performance across all 8 metrics, with relative improvements of approximately 10%–60%. Overall, the results demonstrate that the proposed approach significantly surpasses both larger general-purpose and domain-specific models in terms of descriptive accuracy and semantic consistency.

To further evaluate the model's generalization ability, we conduct experiments on the out-of-domain test set of RSICI, with the results presented in Table 2. It can be observed that even when tested on out-of-domain data, RSICCLLM can still effectively analyze changes between images and surpasses other larger models across all evaluation metrics, demonstrating its strong generalization capability. Meanwhile, we use GPT-5.2 Thinking as the judge to evaluate four aspects—object, direction, action verb, and change magnitude—and report the average score across these attributes. Our model achieves the best performance. Detailed results are provided in **Appendix B.3**.

At this stage, we also compare our constructed datasets with other existing datasets, and the results are shown in Table 3. RSICI contains a total of 40,000 images paired with 20,000 captions, with an average caption length of 77 words. It is the largest instruction dataset for RSICC in the general domain, providing sufficient training data for models to learn how to understand and analyze changes in bi-temporal images. RSICP includes 20,000 images and represents the first released preference dataset in the RSICC field. This dataset enables the model to more precisely distinguish changes between remote sensing images and plays an important role in advancing research in this area.

4.3 Ablation Study

Ablation on Difference-aware SFT with different training size. Under the same implementation, we compare the model trained with the proposed Difference-aware SFT against the standard SFT on training sets of 2k, 10k, 24k, and 32k, evaluated by BLEU-4 and ROUGE-1. As shown in Fig. 4, both models benefit from more data, while the model with Difference-aware SFT consistently outperforms the baseline. Notably, the performance gain becomes more significant in low-data regimes, indicating Difference-aware SFT can effectively guide the model to understand changes in bi-temporal images, especially when training data are limited. More results are shown in **Appendix C.1**.

Ablation on different components in Difference-aware SFT. We analyze the contribution of the three modules in Difference-aware SFT separately. *w/o* G denotes replacing the geometric-texture statistics with simple averaging over convolutional outputs. The definitions of ΔS and T follow Sec. 3.3, and the results are reported in Table 4. Removing T leads to the largest performance drop, underscoring the importance of explicit change extraction. Removing ΔS also degrades performance, indicating that the convolutional operators can provide useful change cues. Eliminating G further reduces accuracy, showing the effectiveness of geometric-texture statistics. Overall, these results confirm the benefit of all three components for change perception and understanding.

Effectiveness of different training strategies. We evaluate the contribution of each training stage in Table 5. The base model performs poorly without fine-tuning, while Difference-aware Supervised Fine-tuning (SFT) brings substantial gains across all metrics. Adding the DNPO stage yields further improvements by exploiting constraints from positive–negative sample pairs to refine the model's outputs. We therefore construct a more challenging subset by manually selecting approximately 400 difficult samples, and find that DNPO brings a 7.4% improvement over the SFT stage. Detailed results are provided in **Appendix C.2**.

Ablation on two negative sample construction strategies under DNPO. To assess the impact of two negative sample construction strategies, we conduct an ablation as shown in Table 6. Clearly, both strategies individually yield consistent gains over the SFT-only baseline, while combining their preference datasets achieves the best performance, confirming the effectiveness of both strategies for constructing high-quality preference data.

Ablation on cross attention fusion. Cross-attention is used to fuse information from different views. T, obtained via the MLLM tokenizer, is better aligned with semantic information but lacks explicit spatial-geometric cues. In contrast, the convolution-derived ΔS focuses on spatial geometry but cannot capture semantics. Fusing T and ΔS therefore strengthens change perception and understanding. We modified the corresponding module in RSICCLLM, trained it with SFT only, and evaluated on the in-domain test set. The results are reported in Table 7, where T denotes $I_1 - I_2$.

Ablation Study on $\mathcal{L}_{\mathrm{S}}$. We conduct this ablation study to evaluate the impact of the loss term $\mathcal{L}_{\mathrm{S}}$. BLEU-4 and ROUGE-1 are adopted as evaluation metrics, and the results are shown in Fig. 5. As observed, the model trained with the $\mathcal{L}_{\mathrm{S}}$

Table 7: Ablation on cross attention fusion and ΔS injection strategies.

ΔS Injection	BLEU-1	ROUGE-1	SBS
Only ΔS	78.29	50.63	78.27
Concat ($[T, \Delta S]$)	78.63	50.81	78.67
Add ($T + \Delta S$)	79.14	51.07	79.38
Mul ($T \odot \Delta S$)	79.39	51.24	79.42
Cross-Attn	**81.28**	**53.54**	**80.38**

Table 8: Comparison of SBS, parameters and throughput. *samples/s* denotes the number of samples processed per second.

Method	SBS ↑	Param.↓	samples/s↑
Qwen3-VL-235B	71.70	235B	0.104
Qwen3-VL-72B	70.06	72B	0.213
InternVL-3.5-241B	72.98	241B	0.096
Ours	**82.72**	**7B**	**0.631**

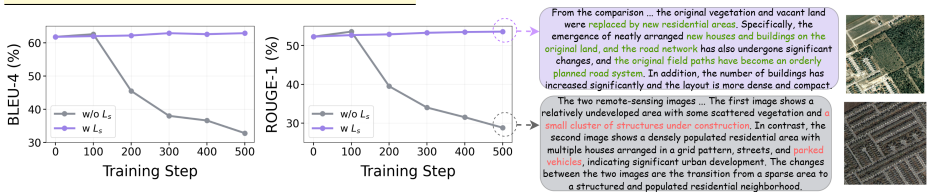

Fig. 5: Ablation study on whether to employ $\mathcal{L}_{\mathrm{S}}$.

constraint performs slightly worse at the early stage compared to the one trained only with $\mathcal{L}_{\mathrm{DPO}}$. However, the $\mathcal{L}_{\mathrm{DPO}}$-only model quickly reaches its peak and then suffers a sharp performance drop as training continues, eventually leading to collapse. In contrast, the model incorporating $\mathcal{L}_{\mathrm{S}}$ exhibits a more stable training process and ultimately achieves better and more consistent performance through the joint effect of $\mathcal{L}_{\mathrm{S}}$ and $\mathcal{L}_{\mathrm{DPO}}$. These results suggest that $\mathcal{L}_{\mathrm{S}}$ helps prevent π_θ from deviating excessively from π_{ref}, mitigating the influence of data noise and improving the stability and effectiveness of preference training.

About inference cost and model comparison. As shown in Table 8, We compare our model with Qwen3-VL-235B, Qwen3-VL-72B, and InternVL-3.5-241B in terms of parameters, throughput, and performance. Despite using only a 7B model, we outperform the second-best model, Qwen3-VL-235B, by about 11 SBS points, while achieving a $\sim$6$\times$ higher prediction throughput in samples per second. The experimental results compared with existing deep learning methods in the RSICC literature are provided in **Appendix C.3**.

4.4 Visual Analysis

To further analyze the model's behavior in the change captioning process, we visualize the outputs as shown in Fig. 6. The upper example compares RSICCLLM with the domain-specific model TEOChat. Compared with TEOChat, our method better captures the changes between bi-temporal images. The lower example compares RSICCLLM with the baseline Qwen2.5-VL-7B, where RSICCLLM shows substantially improved change captioning ability, perceiving fine-grained changes, correctly localizing them, and producing more detailed descriptions. Additional visualizations are provided in **Appendix D**.

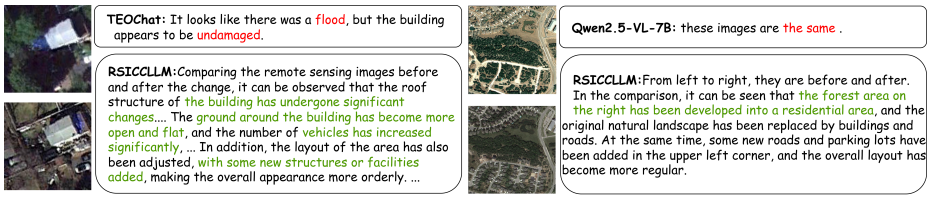

Fig. 6: Visual analysis of change captioning from different models for qualitative comparison.

5 Conclusion

In this paper, we propose RSICCLLM, a domain-adapted post-training framework for vision-language models. By integrating Difference-aware Supervised Fine-tuning and Dual-Negative Preference Optimization, our method effectively aligns visual changes with high-quality descriptions. In addition, we design a new data generation strategy and release a large-scale instruction dataset RSICI. Meanwhile, we also establish the first preference dataset RSICP for this task. Extensive experiments demonstrate that RSICCLLM achieves superior performance across multiple metrics, validating the effectiveness of our approach.

Acknowledgement. This work was supported in part by the National Natural Science Foundation of China under Grant Nos. 62576076, 62476264, and 62406312; the Guangdong Basic and Applied Basic Research Foundation under Grant No. 2023A1515140037; the CCF-Tencent Rhino-Bird Open Research Fund; the Guangdong Research Team for Communication and Sensing Integrated with Intelligent Computing under Project No. 2024KCXTD047; and the Youth Key Project of the Chinese Academy of Sciences under Grant No. GFQN-2026-34. The authors also acknowledge the SongShan Lake HPC Center (SSL-HPC) at Great Bay University for providing computational resources.

References

1. Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., Zhou, J.: Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. arXiv preprint arXiv:2308.12966 (2023)
2. Bai, S., et al.: Qwen2.5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
3. Bai, S., Li, M., Liu, Y., Tang, J., Zhang, H., Sun, L., Chu, X., Tang, Y.: Univg-r1: Reasoning guided universal visual grounding with reinforcement learning (2025). https://doi.org/10.48550/arXiv.2505.14231, https://arxiv.org/abs/2505.14231
4. Ballard, D.H.: Generalizing the hough transform to detect arbitrary shapes. Pattern recognition **13**(2), 111–122 (1981)
5. Chen, H., Shi, Z.: A spatial-temporal attention-based method and a new dataset for remote sensing image change detection. Remote sensing **12**(10), 1662 (2020)
6. Chen, K., Chen, B., Liu, C., Li, W., Zou, Z., Shi, Z.: Rsmamba: Remote sensing image classification with state space model. IEEE Geoscience and Remote Sensing Letters (2024)
7. Chen, L., Gao, H., Liu, T., Huang, Z., Sung, F., Zhou, X., Wu, Y., Chang, B.: G1: Bootstrapping perception and reasoning abilities of vision-language model via reinforcement learning (2025). https://doi.org/10.48550/arXiv.2505.13426, https://arxiv.org/abs/2505.13426
8. Chen, X., Zhu, M., Liu, S., Wu, X., Xu, X., Liu, Y., Bai, X., Zhao, H.: Mico: Multi-image contrast for reinforcement visual reasoning (2025). https://doi.org/10.48550/arXiv.2506.22434, https://arxiv.org/abs/2506.22434
9. Chen, Z., Wang, C., Zhang, N., Zhang, F.: Rscc: A large-scale remote sensing change caption dataset for disaster events. arXiv preprint arXiv:2509.01907 (2025)
10. Cheng, G., Huang, Y., Li, X., Lyu, S., Xu, Z., Zhao, H., Zhao, Q., Xiang, S.: Change detection methods for remote sensing in the last decade: A comprehensive review. Remote Sensing **16**(13), 2355 (2024)

11. Chouaf, S., Hoxha, G., Smara, Y., Melgani, F.: Captioning changes in bi-temporal remote sensing images. In: 2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS. pp. 2891–2894. IEEE (2021)
12. Dong, S., Wang, L., Du, B., Meng, X.: Changeclip: Remote sensing change detection with multimodal vision-language representation learning. ISPRS Journal of Photogrammetry and Remote Sensing **208**, 53–69 (2024)
13. Geng, Z., Wang, Y., Ma, Y., Li, C., Rao, Y., Gu, S., Zhong, Z., Lu, Q., Hu, H., Zhang, X., et al.: X-omni: Reinforcement learning makes discrete autoregressive image generative models great again. arXiv preprint arXiv:2507.22058 (2025)
14. GLM-V Team: Glm-4.5v and glm-4.1v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. arXiv preprint arXiv:2507.01006 (2025)
15. Gu, Z., Zeng, M.: The use of artificial intelligence and satellite remote sensing in land cover change detection: Review and perspectives. Sustainability **16**(1), 274 (2024)
16. Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., et al.: Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv preprint arXiv:2501.12948 (2025)
17. Hänsch, R., Arndt, J., Lunga, D., Gibb, M., Pedelose, T., Boedihardjo, A., Petrie, D., Bacastow, T.M.: Spacenet 8-the detection of flooded roads and buildings. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1472–1480 (2022)
18. Hoxha, G., Chouaf, S., Melgani, F., Smara, Y.: Change captioning: A new paradigm for multitemporal remote sensing image analysis. IEEE Transactions on Geoscience and Remote Sensing **60** (2022)
19. Hu, Y., Yuan, J., Wen, C., Lu, X., Li, X.: Rsgpt: A remote sensing vision language model and benchmark. arXiv preprint arXiv:2307.15266 (2023)
20. Hu, Y., Yuan, J., Wen, C., Lu, X., Liu, Y., Li, X.: Rsgpt: A remote sensing vision language model and benchmark. ISPRS Journal of Photogrammetry and Remote Sensing **224**, 272–286 (2025)
21. Intern-S1 Team: Intern-s1: A scientific multimodal foundation model. arXiv preprint arXiv:2508.15763 (2025)
22. Irvin, J.A., Liu, E.R., Chen, J.C., Dormoy, I., Kim, J., Khanna, S., Zheng, Z., Ermon, S.: Teochat: A large vision-language assistant for temporal earth observation data. arXiv preprint arXiv:2410.06234 (2024)
23. Kuckreja, K., Danish, M.S., Naseer, M., Das, A., Khan, S., Khan, F.S.: Geochat:grounded large vision-language model for remote sensing. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 27831–27840 (2024). https://doi.org/10.1109/CVPR52733.2024.02629
24. Lebedev, M.A., Vizilter, Y.V., Vygolov, O.V., Knyaz, V.A., Rubis, A.Y.: Change detection in remote sensing images using conditional adversarial networks. In: The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences. vol. XLII-2, pp. 565–571 (2018). https://doi.org/10.5194/isprs-archives-XLII-2-565-2018
25. Li, X., Ding, J., Elhoseiny, M.: Vrsbench: A versatile vision-language benchmark dataset for remote sensing image understanding. Advances in Neural Information Processing Systems **37**, 3229–3242 (2024)
26. Li, Y., Tian, W., Jiao, Y., Chen, J., Qian, T., Zhu, B., Zhao, N., Jiang, Y.G.: Look before you decide: Prompting active deduction of mllms for assumptive reasoning (2024). https://doi.org/10.48550/arXiv.2404.12966, https://arxiv.org/abs/2404.12966

27. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: Text Summarization Branches Out: Proceedings of the ACL-04 Workshop (2004)
28. Liu, C., Chen, K., Qi, Z., Liu, Z., Zhang, H., Zou, Z., Shi, Z.: Pixel-level change detection pseudo-label learning for remote sensing change captioning. In: IGARSS 2024-2024 IEEE International Geoscience and Remote Sensing Symposium. pp. 8405–8408. IEEE (2024)
29. Liu, C., Yang, J., Qi, Z., Zou, Z., Shi, Z.: Progressive scale-aware network for remote sensing image change captioning. In: IGARSS 2023-2023 IEEE International Geoscience and Remote Sensing Symposium. pp. 6668–6671. IEEE (2023)
30. Liu, C., Zhao, R., Chen, H., Zou, Z., Shi, Z.: Remote sensing image change captioning with dual-branch transformers: A new method and a large scale dataset. IEEE Transactions on Geoscience and Remote Sensing **60**, 1–20 (2022)
31. Liu, C., Zhao, R., Chen, J., Qi, Z., Zou, Z., Shi, Z.: A decoupling paradigm with prompt learning for remote sensing image change captioning. IEEE Transactions on Geoscience and Remote Sensing **61**, 1–18 (2023)
32. Liu, N., Xu, X., Su, Y., Zhang, H., Li, H.C.: Pointsam: Pointly-supervised segment anything model for remote sensing images. IEEE Transactions on Geoscience and Remote Sensing (2025)
33. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL) (2002)
34. Reimers, N., Gurevych, I.: Sentence-bert: Sentence embeddings using siamese bert-networks. arXiv preprint arXiv:1908.10084 (2019)
35. SenseTime: Sensechat-vision. `https://www.sensetime.com/en/news-detail/51167493` (2024), official announcement / product page
36. Shen, J., Zhao, H., Gu, Y., Gao, S., Liu, K., Huang, H., Gao, J., Lin, D., Zhang, W., Chen, K.: Semi-off-policy reinforcement learning for vision-language slow-thinking reasoning. arXiv preprint arXiv:2507.16814 (2025)
37. Shen, L., Lu, Y., Chen, H., Wei, H., Xie, D., Yue, J., Chen, R., Lv, S., Jiang, B.: S2looking: A satellite side-looking dataset for building change detection. Remote Sensing **13**(24), 5094 (2021)
38. Shi, J., Zhang, M., Hou, Y., Zhi, R., Liu, J.: A multi-task network and two large scale datasets for change detection and captioning in remote sensing images. IEEE Transactions on Geoscience and Remote Sensing (2024)
39. Shi, Q., Liu, M., Li, S., Liu, X., Wang, F., Zhang, L.: A deeply supervised attention metric-based network and an open aerial image dataset for remote sensing change detection. IEEE Transactions on Geoscience and Remote Sensing **60**, 1–16 (2022). `https://doi.org/10.1109/TGRS.2021.3085870`
40. Team, K.: Kimi-vl technical report (2025). `https://doi.org/10.48550/arXiv.2504.07491`, `https://arxiv.org/abs/2504.07491`
41. Team, Q.: Qwen3-vl. `https://github.com/QwenLM/Qwen3-VL` (2025), code repository
42. Van Etten, A., Hogan, D., Manso, J.M., Shermeyer, J., Weir, N., Lewis, R.: The multi-temporal urban development spacenet dataset. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6398–6407 (2021)
43. Verma, S., Panigrahi, A., Gupta, S.: Qfabric: Multi-task change detection dataset. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1052–1061 (2021)

44. Wang, F., Wang, H., Guo, Z., Wang, D., Wang, Y., Chen, M., Ma, Q., Lan, L., Yang, W., Zhang, J., et al.: Xlrs-bench: Could your multimodal llms understand extremely large ultra-high-resolution remote sensing imagery? In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 14325–14336 (2025)
45. Wang, J., Lin, K.Q., Cheng, J., Shou, M.Z.: Think or not? selective reasoning via reinforcement learning for vision-language models (2025). https://doi.org/10.48550/arXiv.2505.16854, https://arxiv.org/abs/2505.16854
46. Wang, L., Zhang, M., Gao, X., Shi, W.: Advances and challenges in deep learning-based change detection for remote sensing images: A review through various learning paradigms. Remote Sensing **16**(5), 804 (2024)
47. Wang, Z., Guo, X., Stoica, S., Xu, H., Wang, H., Ha, H., Chen, X., Chen, Y., Yan, M., Huang, F., et al.: Perception-aware policy optimization for multimodal reasoning. arXiv preprint arXiv:2507.06448 (2025)
48. Wang, Z., Wang, M., Xu, S., Li, Y., Zhang, B.: Ccexpert: Advancing mllm capability in remote sensing change captioning with difference-aware integration and a foundational dataset. arXiv preprint arXiv:2411.11360 (2024)
49. Xiao, W., Gan, L., Dai, W., He, W., Huang, Z., Li, H., Shu, F., Yu, Z., Zhang, P., Jiang, H., Wu, F.: Fast-slow thinking for large vision-language model reasoning (2025). https://doi.org/10.48550/arXiv.2504.18458, https://arxiv.org/abs/2504.18458
50. Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al.: Qwen3 technical report. arXiv preprint arXiv:2505.09388 (2025)
51. Yang, C., An, Z., Huang, L., Bi, J., Yu, X., Yang, H., Diao, B., Xu, Y.: Clip-kd: An empirical study of clip model distillation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15952–15962 (2024)
52. Yang, S., Li, J., Lai, X., Yu, B., Zhao, H., Jia, J.: Visionthink: Smart and efficient vision language model via reinforcement learning. arXiv preprint arXiv:2507.13348 (2025)
53. Yao, K., Xu, N., Yang, R., Xu, Y., Gao, Z., Kitrungrotsakul, T., Ren, Y., Zhang, P., Wang, J., Wei, N., et al.: Falcon: A remote sensing vision-language foundation model. arXiv preprint arXiv:2503.11070 (2025)
54. Yu, Z., Zhao, C., Wang, Z., Qin, Y., Su, Z., Li, X., Zhou, F., Zhao, G.: Searching central difference convolutional networks for face anti-spoofing. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5295–5305 (2020)
55. Zhan, Y., Xiong, Z., Yuan, Y.: Skyeyegpt: Unifying remote sensing vision-language tasks via instruction tuning with large language model. ISPRS Journal of Photogrammetry and Remote Sensing **221**, 64–77 (2025)
56. Zhang, X., Gao, Z., Zhang, B., Li, P., Zhang, X., Liu, Y., Yuan, T., Wu, Y., Jia, Y., Zhu, S.C., Li, Q.: Chain-of-focus: Adaptive visual search and zooming for multimodal reasoning via rl (2025). https://doi.org/10.48550/arXiv.2505.15436, https://arxiv.org/abs/2505.15436
57. Zhang, Z., Zhao, T., Guo, Y., Yin, J.: Rs5m and georsclip: A large scale vision-language dataset and a large vision-language model for remote sensing. IEEE Transactions on Geoscience and Remote Sensing (2024)
58. Zhu, Y., Li, L., Chen, K., Liu, C., Zhou, F., Shi, Z.: Semantic-cc: Boosting remote sensing image change captioning via foundational knowledge and semantic guidance. arXiv preprint arXiv:2407.14032 (2024). https://doi.org/10.48550/arXiv.2407.14032

59. Zou, S., Wei, Y., Xie, Y., Lao, M., Luan, X.: Remote sensing image change captioning: A comprehensive review: S. zou et al. International Journal of Multimedia Information Retrieval **14**(3), 26 (2025)