remote sensing

MDPI

Article

Describing Land Cover Changes via Multi-Temporal Remote Sensing Image Captioning Using LLM, ViT, and LoRA

Javier Lamar León *, Vitor Nogueira, Pedro Salgueiro and Paulo Quaresma

Centro Algoritmi, LASI, University of Évora, 7005-854 Évora, Portugal; vbn@uevora.pt (V.N.); pds@uevora.pt (P.S.); pq@uevora.pt (P.Q.)
* Correspondence: jlamarleon@gmail.com or jlamarleon@uevora.pt

Highlights

What are the main findings?

- A novel Multimodal Vision Language Transformer (MVLT-LoRA-CC) is proposed, integrating a Vision Transformer, a Large Language Model, and Low-Rank Adaptation (LoRA) for efficient and interpretable remote sensing change captioning.
- A new Complementary Consistency Score (CCS) framework is introduced to jointly evaluate descriptive accuracy for change samples and stability recognition for no-change cases, offering a unified and semantically grounded assessment.

What are the implication of the main findings?

- The proposed MVLT-LoRA-CC achieves state-of-the-art performance on the LEVIR-CC dataset, improving generalization and semantic precision compared to vision-only methods.
- This framework establishes a scalable and context-aware approach for multimodal Earth observation, enhancing interpretability and robustness in environmental monitoring applications.

Abstract

Describing land cover changes from multi-temporal remote sensing imagery requires capturing both visual transformations and their semantic meaning in natural language. Existing methods often struggle to balance visual accuracy with descriptive coherence. We propose MVLT-LoRA-CC (Multi-modal Vision Language Transformer with Low-Rank Adaptation for Change Captioning), a framework that integrates a Vision Transformer (ViT), a Large Language Model (LLM), and Low-Rank Adaptation (LoRA) for efficient multi-modal learning. The model processes paired temporal images through patch embeddings and transformer blocks, aligning visual and textual representations via a multi-modal adapter. To improve efficiency and avoid unnecessary parameter growth, LoRA modules are selectively inserted only into the attention projection layers and cross-modal adapter blocks rather than being uniformly applied to all linear layers. This targeted design preserves general linguistic knowledge while enabling effective adaptation to remote sensing change description. To assess performance, we introduce the Complementary Consistency Score (CCS) framework, which evaluates both descriptive fidelity for change instances and classification accuracy for no change cases. Experiments on the LEVIR-CC test set demonstrate that MVLT-LoRA-CC generates semantically accurate captions, surpassing prior methods in both descriptive richness and temporal change recognition. The approach establishes a scalable solution for multi-modal land cover change description in remote sensing applications.

Academic Editors: Jiayi Pan and Xinghua Li

Received: 23 October 2025
Revised: 29 December 2025
Accepted: 31 December 2025
Published: 4 January 2026

Keywords: land cover change captioning; multi-temporal remote sensing; vision language transformers; Low-Rank Adaptation (LoRA); natural language generation

1. Introduction

The automatic generation of image captions represents a dynamic intersection of computer vision and natural language processing, aiming to produce coherent, human-like descriptions of visual content [1]. While traditionally applied in domains such as assistive technologies, social networks, and autonomous systems, this technology has increasing relevance in Earth observation and environmental analysis. In particular, image change captioning, the task of describing differences between pairs or sequences of multi-temporal remote sensing images, provides a powerful framework for characterizing land cover changes over time [2]. By translating complex spectral and spatial changes into descriptive narratives, automatic captioning facilitates the interpretation of raw satellite data while capturing the semantic essence of environmental transformations.

The primary objective in multi-temporal remote sensing captioning is the accurate description of land cover changes. These changes represent substantive shifts in land use and ecological processes, including deforestation, urban expansion, agricultural intensification, water body fluctuations, or vegetation regrowth. For an automated captioning system, effectively capturing and articulating these changes in natural language is critical for supporting downstream tasks, such as decision-making, environmental reporting, or change summarization.

A key challenge in this task lies in the semantic relationships between objects when describing land cover changes. Transformations often involve multiple interdependent entities; for example, the conversion of forest to agricultural land entails both the loss of vegetation and the emergence of cultivated areas. Human annotators emphasize different aspects of these changes: some highlight vegetation removal, others the emergence of built infrastructure, and others the broader ecological context. Automated captioning systems must therefore account for both object-level changes and higher-order relational patterns to produce semantically meaningful descriptions.

Evaluating the outputs of automated captioning systems also presents challenges. Generated captions must accurately depict complex spatial and temporal transformations, yet human descriptions exhibit significant variability in wording and emphasis. Conventional metrics, such as BLEU or ROUGE, often fail to capture semantically correct but lexically distinct captions, and may not adequately handle no change instances. Selecting robust evaluation strategies is therefore essential for assessing multi-temporal captioning systems.

Recent approaches have explored multi-temporal image captioning from several perspectives. Chg2Cap [3] employs a three-stage architecture with a Siamese CNN-based feature extractor, an attentive encoder, and a Transformer-based caption generator. RSICCformer [4] uses a fully Transformer-based architecture tailored for remote sensing change captioning. SAT-Cap [5] integrates spatial- and channel-level attention mechanisms with a difference-guided fusion module for efficient multi-temporal feature integration. While effective, these methods are constrained by dataset-specific vocabularies and pretraining objectives, which limit generalization and produce repetitive or template-like captions.

Moreover, existing models generally operate in a vision-only regime, lacking the ability to leverage external textual guidance. Recent advances in multi-modallearning [6] enable architectures that process both image and text inputs, allowing natural language prompts to guide the attention and descriptive focus of the model. This capability supports

more flexible and context-aware caption generation, moving beyond static, dataset-specific vocabularies toward semantically rich descriptions.

To address these limitations, we propose a multi-modal Vision Language Transformer (MVLT-LoRA-CC) that integrates a Vision Transformer (ViT) for spatial representation, a Large Language Model (LLM) for linguistic reasoning, and Low-Rank Adaptation (LoRA) for parameter-efficient fine tuning. By processing paired multi-temporal images and incorporating textual guidance, the model generates descriptive captions that accurately convey land cover changes. Leveraging universal pretrained vocabularies and cross-modal feature alignment, our approach achieves state-of-the-art performance on the LEVIR-CC dataset [4], providing a scalable and semantically robust framework for describing environmental transformations.

Finally, we propose the Complementary Consistency Score (CCS) framework to provide a unified evaluation strategy for both change and no change instances, addressing the limitations of conventional lexical metrics. This allows for a more semantically grounded assessment of multi-temporal captioning quality.

The remainder of this paper is structured as follows: Section 2 presents the proposed multi-modal Vision Language Transformer architecture, detailing the visual and linguistic components and the integration of LoRA. Section 2.5 describes the LEVIR-CC dataset and its relevance for multi-temporal remote sensing captioning. Section 2.6 introduces the CCS evaluation framework. Section 2.9 reports experimental results, including quantitative and qualitative assessments. Finally, Section 5 concludes with a summary of contributions and directions for future work.

2. Materials and Methods

To effectively address the task of monitoring land cover changes through multi-temporal remote sensing image captioning, we design a hybrid multimodal framework that integrates Vision Transformers (ViTs), Large Language Models (LLMs), and Low-Rank Adaptation (LoRA). The goal is to jointly process sequences of remote sensing images and textual inputs so that the model can generate captions describing both spatial patterns and their temporal evolution.

In contrast to standard single image encoders, our framework is tailored for multi-temporal inputs. Each image in the temporal sequence is decomposed into fixed-size patches and embedded using a lightweight convolutional projection [7]. The resulting patch tokens from all timestamps are concatenated into a unified token sequence, enabling the Vision Transformer to model temporal dynamics directly through its self-attention layers. Spatial and temporal consistency is maintained using positional encodings that include temporal indices, allowing the model to differentiate between acquisition times while preserving spatial alignment across the sequence.

To control information flow inside the visual module, we apply a BlockDiagonalMask so that attention is restricted to visual tokens. This maintains clean intra- and inter-temporal reasoning within the Vision Encoder. Rotary positional embeddings (RoPE) are employed to enrich spatial relationships without requiring explicit derivations of the underlying attention mechanism [8].

Bridging visual and textual modalities is achieved through a pretrained multimodal LLM following a cross-attention fusion paradigm similar to [9]. In this setup, visual tokens condition the language model during autoregressive caption generation, allowing textual queries to access multi-temporal visual information at each decoding step. A BlockDiagonalCausalMask enforces causal constraints for text tokens while keeping all visual tokens globally accessible, ensuring stable multimodal conditioning throughout the generation process.

Given that directly fine tuning large multimodal LLMs is computationally prohibitive, we integrate Low-Rank Adaptation (LoRA) [10–12] specifically into two components of our architecture: the multi-modal Transformer backbone and the visual adapter. In both modules, LoRA introduces compact trainable low-rank matrices into selected projection layers while keeping the pretrained parameters frozen. This setup significantly reduces memory consumption and enables efficient domain adaptation to multi-temporal remote sensing data, allowing the model to learn temporal change representations without full end-to-end fine tuning.

The overall structure of the proposed framework is illustrated in Figure 1, highlighting the flow from multi-temporal visual inputs and text tokens through the shared multi-modal Transformer backbone with LoRA adapters.

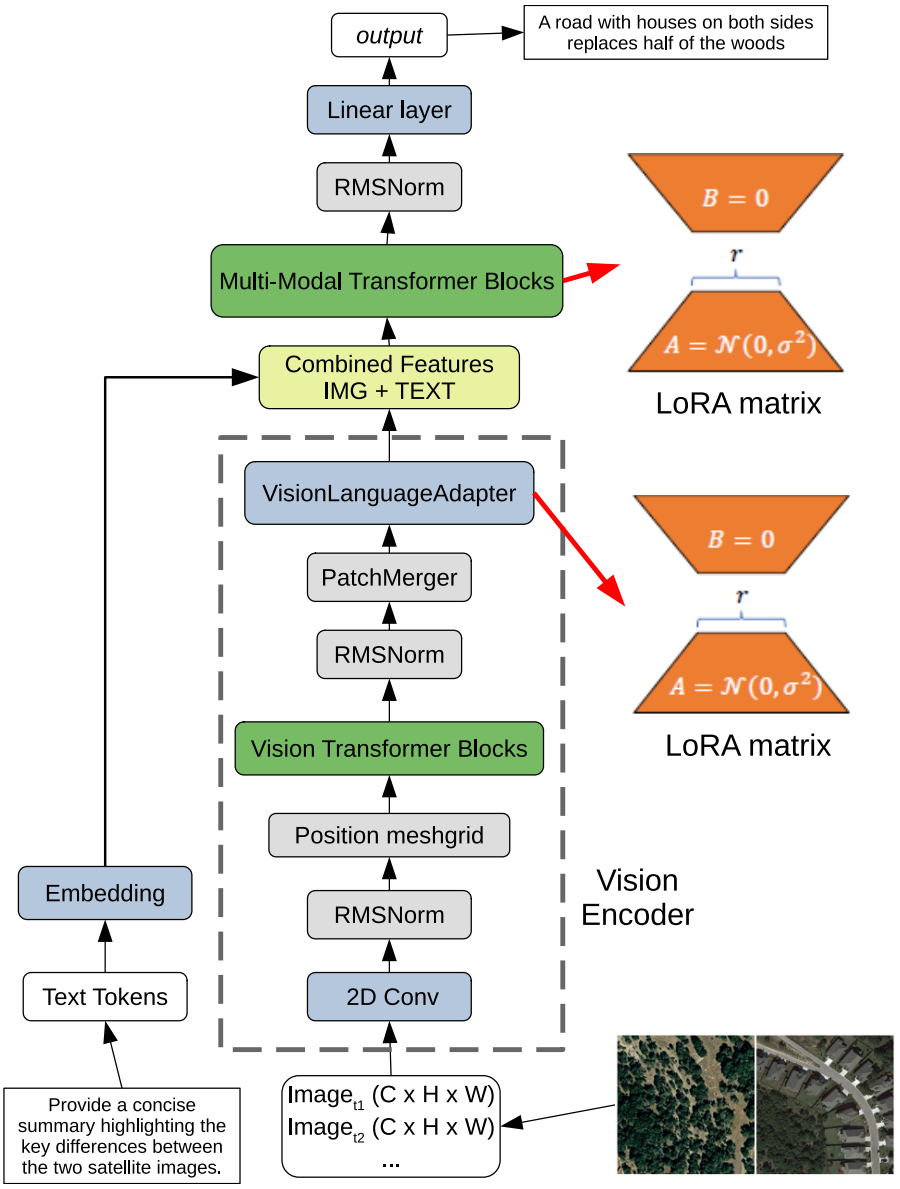

Figure 1. Overview of the proposed multi-temporal vision language architecture. Multi-temporal satellite images are encoded by a frozen Vision Transformer, producing visual token sequences that serve as inputs to the Vision Language Adapter. LoRA modules are selectively integrated only into the Vision Language Adapter and the multi-modal Transformer blocks, enabling efficient cross-modal alignment and multimodal reasoning without modifying the pretrained visual backbone.

2.1. Multi-Temporal Input Processing

Multi-temporal image sequences are represented as

$$\{I_t \in \mathbb{R}^{C\times H\times W}\}_{t=1}^{T}, \tag{1}$$

where T is the number of temporal observations, C denotes the number of channels, and H and W are the spatial dimensions. Each image is decomposed into patches via Conv2D projection and normalized with RMSNorm to produce patch embeddings $X_t \in \mathbb{R}^{N\times D}$, where N is the number of patches and D is the embedding dimension. These embeddings are processed by a stack of visual Transformer layers to capture spatial dependencies and preserve temporal structure. Optionally, a Patch Merger reduces spatial redundancy:

$$Z_t = \text{PatchMerger}(\tilde{X}_t^{(L_v)}). \tag{2}$$

Here, $\tilde{X}_t^{(L_v)} \in \mathbb{R}^{N\times D}$ denotes the output of the final visual Transformer layer, with L_v representing the number of Transformer encoder layers in the vision branch, and $Z_t \in \mathbb{R}^{N'\times D}$ is the merged representation ($N' < N$) that preserves high-level semantic information while reducing spatial redundancy.

2.2. Cross-Modal Fusion Mechanisms

Temporal visual embeddings Z_t are projected into a shared multimodal space via a lightweight Vision Language Adapter:

$$F_t = W_{\text{out}}\,\text{GELU}(Z_t W_{\text{in}}). \tag{3}$$

Here, W_{in} and W_{out} are the input and output projection matrices, respectively, $\text{GELU}(\cdot)$ is the activation function, and F_t is the resulting multimodal embedding.

Text tokens $\mathcal{T} = \{t_1, \ldots, t_L\}$ are embedded and concatenated with visual features to form a multi-modal sequence:

$$X_0 = [F_1; F_2; \ldots; F_T; \mathcal{T}]. \tag{4}$$

This sequence is processed by a multi-modal Transformer backbone that fuses visual and textual information, enabling rich cross-modal representations.

2.3. Autoregressive Multimodal Conditioning

A `BlockDiagonalCausalMask` enforces autoregressive dependencies on text tokens while allowing full access to all visual tokens:

$$M_{\text{causal}} = \begin{cases} 0, & \text{if token } i \text{ can attend to token } j; \\ -\infty, & \text{otherwise.} \end{cases}$$

Here, $j \le i$ for text tokens, or j corresponds to a visual token. This ensures that each generated word is conditioned on both prior words and the complete multi-temporal visual context. The final multi-modal representation is normalized and projected to the vocabulary space:

$$Y = \text{RMSNorm}(X^{(L_m)})W_o, \quad P(w_i \mid w_{<i}, I_{1:T}) = \text{softmax}(Y_i). \tag{5}$$

Here, W_o is the output projection matrix, $\text{RMSNorm}(\cdot)$ is the root mean square layer normalization, Y is the projected representation in the vocabulary space, and $P(w_i|w_{<i}, I_{1:T})$

is the probability of predicting token w_i given all previous predicted tokens $w_{<i}$ and the visual inputs $I_{1:T}$.

2.4. Specific LoRA Integration Rationale

To enable efficient fine tuning of the multi-modal architecture while avoiding unnecessary parameter overhead, LoRA is selectively applied only to modules that most strongly influence cross-modal alignment and language driven reasoning. Instead of distributing LoRA adapters uniformly across all Transformer layers, we adopt a task-oriented strategy that focuses on adapting the model components responsible for bridging visual and textual semantics. This selective integration minimizes computational cost while preserving the robust general-purpose representations learned during pretraining.

1. No LoRA in Vision Transformer Blocks.

Although Vision Transformer (ViT) blocks contain many parameters, we intentionally avoid inserting LoRA adapters into their Multi-Head Self-Attention (MHSA) or Feed-Forward Network (FFN) layers. This decision is motivated by two considerations:

(i) Stable low-level visual features. Pre-trained ViTs already provide strong spatial representations that generalize well across remote sensing domains. Introducing LoRA into the deep visual stack risks altering these stable features and may degrade the spatial consistency required for multi-temporal analysis.

(ii) Task-specific adaptation occurs after visual abstraction. Our method relies on aligning high-level visual semantics with language, which happens *after* the vision encoder. Since the ViT's role is to extract modality-specific spatiotemporal patterns, it is more effective to preserve it unchanged and apply LoRA only to the cross-modal components where semantic alignment is required.

Empirically, we find that retaining a frozen ViT backbone leads to more stable learning dynamics, reduces overfitting on domain-specific remote sensing datasets, and concentrates the learnable capacity of LoRA where it yields the most significant performance gains.

2. Vision Language Adapter (Cross-Modal Alignment).

The Vision Language Adapter is the primary interface where visual embeddings are projected into the shared multimodal space. LoRA is applied to its input and output projection layers, enabling the model to adjust visual to textual mapping with minimal trainable parameters. These low-rank updates allow the adapter to refine semantic alignment between visual tokens and language features crucial for generating temporally grounded captions without modifying the underlying ViT or LLM backbones.

3. Multi-modal Transformer Blocks.

LoRA is also inserted into the Multi-Head Attention and FFN layers of the multi-modal side Transformer blocks. These layers perform the core multimodal reasoning: integrating temporal visual representations with textual context during autoregressive decoding. By applying LoRA here, the model can learn task-specific cross-modal dependencies (e.g., linking observed land cover change patterns to linguistic expressions) while keeping the large pretrained LLM frozen.

4. Exclusion of Convolutional Layers.

LoRA is not applied to the initial Conv2D patch embedding layer. Unlike dense linear transformations, convolutional filters encode strong spatial priors and have limited parameter redundancy, making low-rank decomposition suboptimal. Moreover, freezing the convolutional parameters preserves stable low-level spatial features across temporal observations, while subsequent Transformer and adapter layers learn the domain-specific

adaptation required for captioning land cover changes. This results in both computational efficiency and improved temporal consistency.

In summary, our LoRA strategy emphasizes adapting the parts of the model responsible for semantic fusion and reasoning, while preserving the pretrained modules responsible for low-level feature extraction and spatial representation. This targeted approach maximizes adaptation efficiency and performance in multi-temporal remote sensing captioning tasks.

The following pseudocode (see Algorithm 1) illustrates the high-level processing flow of the multi-modal Vision Language Transformer with LoRA integration. It demonstrates multi-temporal input processing, cross-modal fusion, and autoregressive decoding in a concise, Python-like format.

Algorithm 1 Multi-modal Vision Language Forward Pass with LoRA

Require: text_tokens, images, sequence lengths
Ensure: output logits over vocabulary
text_features $\leftarrow$ TokenEmbedding(text_tokens)
for each I_t in images **do**
    $X_t \leftarrow \text{Conv2D}(I_t)$
    $X_t \leftarrow \text{RMSNorm}(X_t)$
    $Z_t \leftarrow \text{VisionTransformer}(X_t)$
    $F_t \leftarrow \text{VisionLanguageAdapter}(Z_t)$ ▷ LoRA applied
end for
$X_0 \leftarrow \text{Concat}([F_1, \ldots, F_T, \text{text\_features}])$
$H \leftarrow \text{MultiModalTransformer}(X_0, \text{BlockDiagonalCausalMask})$ ▷ LoRA applied
$H \leftarrow \text{RMSNorm}(H)$
output_logits $\leftarrow HW_o$
return output_logits

2.5. Dataset Description: LEVIR-CC

To evaluate the proposed multi-modal vision language framework for land cover change monitoring, we adopt the LEVIR-CC dataset, a large-scale and richly annotated benchmark specifically designed for Remote Sensing Image Change Captioning (RSICC) tasks. This dataset provides a solid foundation for developing and validating models capable of generating natural language descriptions of land cover dynamics, effectively bridging visual change detection and semantic caption generation.

The LEVIR-CC dataset introduced in [4]. The dataset is constructed from multi-temporal, very high-resolution (VHR) remote sensing imagery collected primarily from Google Earth, with a ground sampling distance of approximately 0.5 m per pixel. It contains a total of 10,077 carefully aligned bitemporal image pairs, each of spatial size 256×256 pixels. Every image pair is accompanied by five human-written captions, resulting in more than 50,000 natural language descriptions that narrate the observed changes.

Each caption provides concise yet semantically rich descriptions of surface changes such as the construction or demolition of buildings, road expansions, vegetation increase or loss, and other anthropogenic modifications. The textual annotations are linguistically diverse, with an average caption length of approximately 11.6 words, covering a vocabulary of over 2800 unique tokens. The image pairs are temporally distributed across multiple years and include complex urban, suburban, and rural contexts, ensuring broad geographic and semantic diversity.

A key strength of LEVIR-CC lies in its scale, diversity, and annotation quality. The dataset's multi-domain coverage and fine-grained textual labels enable robust cross-modal learning between vision and language modalities. Furthermore, its balanced rep-

resentation of change and non-change samples allows for effective evaluation of both detection accuracy and descriptive reasoning.

Since its release, LEVIR-CC has evolved into a standard benchmark for multi-modal remote sensing research. Recent studies have used it to evaluate models for vision language pretraining, multi-modal reasoning, and semantic change detection [5,13,14]. Its structured design, large-scale human annotations, and well-aligned bitemporal image caption pairs make it particularly suitable for training and assessing models that aim to jointly capture spatial, temporal, and semantic correlations.

Given these attributes, LEVIR-CC provides an ideal experimental platform for assessing our proposed architecture. It enables rigorous evaluation of both visual encoding performance and cross-modal generative capabilities, offering a meaningful benchmark for understanding how Transformer-based vision language models can interpret and describe real-world land cover evolution.

To provide a clearer understanding of the dataset characteristics, Figure 2 (taken from [4]) presents representative examples from the LEVIR-CC benchmark, highlighting various types of land cover changes and their associated human-annotated captions.

Figure 2. Example samples from the LEVIR-CC dataset. Each bi-temporal image pair has a spatial size of 256 × 256 pixels at a spatial resolution of 0.5 m/pixel, accompanied by five human-annotated captions describing the detected land cover changes. Unlike LEVIR-CC, which focuses solely on building-related changes, LEVIR-CC encompasses multiple change types, including buildings, roads, and rivers while disregarding irrelevant variations such as illumination differences, shadows, or minor vegetation changes. (figure taken from [4]).

The LEVIR-CC dataset presents a balanced distribution between change and no change samples, as shown in Table 1. Each change image pair is annotated with five diverse captions describing specific land cover modifications, such as new constructions, vegetation loss, or infrastructure expansion. In contrast, no change samples share a uniform set of five identical captions indicating the absence of observable change. This design allows the dataset to support both descriptive and discriminative evaluation: Models can be assessed not only on their capacity to generate semantically rich captions for true changes but also on their ability to correctly identify and represent stable regions through consistent no change predictions.

Table 1. Number of bitemporal image pairs in the training, validation, and test sets of the LEVIR-CC dataset.

Change or Not	Training	Validation	Test	Total
change	3407	667	964	5038
no change	3408	666	965	5039
Total	6815	1333	1929	10,077

Dual Task Challenge in LEVIR-CC

The evaluation of models trained on LEVIR-CC requires special attention due to its inherent dual task nature. The dataset simultaneously supports two distinct but complementary objectives: identifying whether a bi-temporal image pair represents a change or no change scenario, and generating descriptive captions for samples containing actual changes. This duality complicates metric interpretation, as conventional captioning scores such as BLEU, CIDEr, METEOR, and SPICE are meaningful only for positive change cases, where textual descriptions convey semantic content. For no change samples, evaluation should instead focus on the model's ability to correctly recognize stability and consistently generate neutral no change expressions. Aggregating both types of samples into a single evaluation can obscure performance, since standard text similarity metrics are insensitive to the semantic polarity of negative statements. Therefore, a decoupled evaluation protocol separating change and no change cases provides a more faithful and interpretable assessment of multi-modal models designed for change captioning tasks.

2.6. Metrics Used in Image Captioning

In this section, we describe the most commonly used automatic evaluation metrics for image captioning. These include BLEU [15], METEOR [16], ROUGE [17], CIDEr [18], and SPICE [19]. Each metric captures different aspects of caption quality, from *n*-gram overlap to semantic content. The variability in human descriptions poses a fundamental challenge for automatic evaluation of captioning models. Metrics such as BLEU or ROUGE-L, which rely heavily on surface-level *n*-gram overlap, often penalize outputs that use different words or sentence structures despite accurately conveying the same semantic meaning. In the context of multi-temporal remote sensing, this issue becomes even more critical, since descriptions of land cover change may legitimately emphasize different aspects of the same transformation, such as the disappearance of vegetation versus the emergence of cropland. As a result, a model's prediction may be judged as poor under lexical overlap metrics even when it is semantically valid. More advanced measures like CIDEr and SPICE attempt to mitigate this limitation by incorporating consensus weighting or semantic parsing, but they too struggle to fully capture domain-specific nuances, such as the ecological and socioeconomic significance of certain changes. These challenges highlight the need for careful selection and interpretation of evaluation metrics when assessing captioning systems designed for environmental monitoring.

BLEU (*n*-gram Precision in Captioning): BLEU is based on the precision of *n*-grams between candidate and reference captions. While it is widely used due to its simplicity and comparability, it is less effective for short, variable captions where synonyms and paraphrasing are common.

METEOR (Recall Oriented Evaluation): METEOR improves upon BLEU by incorporating stemming, synonyms, and recall. It has been shown to correlate better with human judgments, especially at the sentence level, making it suitable for evaluating short captions.

ROUGE-L (Overlap Based Summarization Metric): Originally proposed for text summarization, ROUGE-L measures recall oriented overlap using *n*-grams and longest common

subsequences. Its focus on recall makes it less common in captioning but still useful for capturing content coverage.

CIDEr (Consensus-Based Evaluation): CIDEr was specifically designed for image captioning. It employs TF-IDF-weighted *n*-gram similarity to reward informative and distinctive words. It is the primary metric in many captioning benchmarks, as it correlates strongly with human judgments.

SPICE (Semantic Propositional Image Caption Evaluation): SPICE evaluates caption quality by comparing scene graphs of candidate and reference captions, focusing on semantic propositional content such as objects, attributes, and relationships. This semantic-based approach makes it more effective than lexical overlap metrics for tasks requiring semantic accuracy, and it correlates well with human judgments, particularly in contexts with paraphrasing or synonym use.

2.7. Importance of Metrics for Change Captioning

Given the specific challenges of multi-temporal remote sensing captioning, it is useful to consider the relative importance of these metrics for evaluating model performance. CIDEr and SPICE generally provide the most informative assessment, as they better capture semantic content and reward distinctive, contextually meaningful descriptions of land cover changes. METEOR also holds high relevance due to its consideration of synonyms and recall, which is critical when multiple valid expressions can describe the same transformation. BLEU and ROUGE-L, while still valuable for lexical consistency and content coverage, are less reliable in this context because they are highly sensitive to exact word matches and may penalize semantically correct but lexically diverse predictions. Therefore, for change captioning tasks, metrics emphasizing semantic fidelity and consensus (CIDEr, SPICE, METEOR) should be prioritized, while BLEU and ROUGE-L can serve as supplementary measures to complement the overall assessment (see Table 2).

Table 2. Summary of image captioning metrics and their relevance for change captioning tasks.

Metric	Focus	Relevance
CIDEr	TF–IDF weighted *n*-grams	Very High : rewards informative and semantically relevant words
SPICE	Semantic propositional content	Very High: focuses on meaning; aligns with human judgment
METEOR	Recall, synonyms, stemming	High: captures semantic similarity and paraphrases
BLEU	*n*-gram precision	Low Medium: sensitive to exact wording; penalizes valid paraphrases
ROUGE-L	Recall oriented *n*-grams and LCS	Medium: captures content coverage but sensitive to exact wording

According to Table 2, the following composite measures have been introduced in the state-of-the-art approach to address the limitations of individual metrics and provide a more robust evaluation framework:

$$S_{BMRC} = \frac{1}{4}\left(\text{BLEU-4} + \text{METEOR} + \text{ROUGE-L} + \text{CIDEr}\right), \tag{6}$$

$$S_{MC} = \frac{1}{2}\left(\text{METEOR} + \text{CIDEr}\right), \tag{7}$$

$$S_{MCS} = \frac{1}{3}\left(\text{METEOR} + \text{CIDEr} + \text{SPICE}\right). \tag{8}$$

The composite score S_{BMRC} offers a balanced integration of lexical and semantic evaluation criteria by combining BLEU-4, METEOR, ROUGE-L, and CIDEr. This formulation

jointly accounts for precision and recall, capturing both textual consistency and descriptive adequacy. The simplified variant S_{MC} focuses on meaning-oriented assessment through METEOR and CIDEr, providing a compact yet semantically robust indicator of performance. Meanwhile, S_{MCS} further strengthens the emphasis on semantic fidelity by incorporating SPICE alongside METEOR and CIDEr, making it especially relevant for tasks where the accurate conveyance of meaning and contextual relationships is critical.

Overall, these composite metrics are intended to counterbalance the weaknesses of individual measures, particularly their susceptibility to lexical variation, and deliver a more comprehensive and reliable evaluation framework for multi-temporal remote sensing image captioning models.

To illustrate the behavior and relevance of the previously described metrics in the context of multi-temporal change captioning, we conducted a set of experiments using the test set of the LEVIR-CC dataset. Each sample in the dataset is annotated with five reference captions describing the observed changes. In our evaluation setup, we follow a common protocol in captioning studies: One caption is treated as the predicted output from the model, while the remaining four captions serve as reference descriptions. This approach allows us to simulate the variability in human language and analyze how different metrics respond to lexical and semantic differences between the predicted caption and multiple valid references.

By applying BLEU, METEOR, ROUGE-L, CIDEr, and SPICE in this setting, we can observe the sensitivity of each metric to paraphrasing, synonym usage, and semantic coverage, providing insights into their suitability for evaluating change captioning tasks. The results for samples with only changes are presented in Table 3, while the results for samples with no changes are shown in Table 4.

Table 3. Evaluation results (%) on the LEVIR-CC test set for samples with only changes, using each caption as the predicted output and the remaining four as references. Values in bold indicate the best-performing results.

Prediction	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr	SPICE	S_{BMRC}	S_{MC}	S_{MCS}
Caption 1	62.20	42.90	28.90	19.20	23.70	43.40	42.00	23.00	32.10	32.85	29.60
Caption 2	62.50	41.80	26.80	17.30	23.20	43.20	38.30	22.50	30.50	30.75	28.00
Caption 3	62.00	43.30	29.90	20.10	25.00	44.20	46.70	22.70	**34.00**	**35.85**	**31.50**
Caption 4	66.10	46.40	31.60	22.00	23.40	43.70	41.60	20.40	32.70	32.30	2 8.50
Caption 5	60.70	40.90	26.30	17.30	22.70	42.30	37.40	21.60	29.90	30.05	27.20
Average	62.70	43.06	28.70	19.18	23.60	43.36	41.20	22.04	31.84	32.36	28.96

Table 4. Evaluation results (%) on the LEVIR-CC test set for samples with no changes, using each caption as the predicted output and the remaining four as references.

Prediction	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr	SPICE	S_{BMRC}	S_{MC}	S_{MCS}
Caption 1	50.00	0.00	0.00	0.00	5.20	25.00	-	0.00	-	-	-
Caption 2	20.00	0.00	0.00	0.00	11.10	16.20	-	20.00	-	-	-
Caption 3	28.60	0.00	0.00	0.00	10.90	19.10	-	20.00	-	-	-
Caption 4	28.60	0.00	0.00	0.00	10.90	19.10	-	20.00	-	-	-
Caption 5	25.00	0.00	0.00	0.00	15.70	25.00	-	0.00	-	-	-
Average	30.44	0.00	0.00	0.00	10.76	20.88	-	12.00	-	-	-

The evaluation results for no change samples, presented in Table 4, reveal a markedly different behavior compared to the change samples. In this case, the overall metric values are substantially lower, especially for BLEU and ROUGE-L, which depend heavily on

lexical overlap. This outcome is expected, as no change captions typically employ short and semantically equivalent expressions with considerable lexical variation (e.g., there is no difference vs. the two scenes seem identical), resulting in limited n-gram correspondence among captions.

It is important to note that the CIDEr metric was not computed for this subset, since many identical phrases occur repeatedly across samples, leading to undefined or non-informative TF–IDF weighting. Under such conditions, CIDEr cannot reliably capture distinctiveness or informativeness, which are central to its formulation. Conversely, METEOR and SPICE maintain comparatively higher and more-consistent values, reflecting their greater sensitivity to semantic similarity and propositional meaning.

Overall, the results emphasize the limitations of traditional overlap-based metrics in evaluating semantically equivalent no change captions. In fact, most state-of-the-art methods that report high evaluation scores on such samples do so by predicting a fixed no change phrase identical to the reference text. While this strategy ensures perfect metric alignment, it effectively transforms the problem from a caption generation task into a binary classification one simply detecting whether change occurs or not, thus overlooking the generative and descriptive objectives that define captioning-based approaches.

To further illustrate the evaluation challenges discussed, we conducted an additional experiment including both change and no change samples from the LEVIR-CC test set. For the latter, five equivalent expressions were employed to indicate the absence of change, each corresponding to one of the reference captions, as shown in Table 5. This setup enables us to analyze how lexical variability among semantically equivalent no change statements influences the behavior of automatic evaluation metrics. Specifically, we compare three configurations using caption 3, the best-performing case in Table 3, as the predicted text:

(i) Original configuration: The no change sentences differ in wording across captions.
(ii) Unified phrasing configuration: The predicted caption employs a unified phrasing identical to one of the reference sentences (“the two scenes seem identical”).
(iii) Minor lexical difference configuration: Introduces a subtle variation in the no change sentence by altering a single word (“is”) relative to caption 4, changing “no change has occurred” to “no change is occurred”.

In these cases, all no change samples are correctly predicted. In the first case, the samples are expressed through distinct yet semantically equivalent phrases. In the second, the model generates a no change sentence that exactly matches one of the reference texts, representing the most typical situation in practical model evaluation. The third configuration, despite involving only a minimal lexical modification, yields a considerable decrease in most metric scores. Together, these comparisons highlight the strong sensitivity of automatic metrics to surface-level textual similarity and their tendency to over penalize valid paraphrases, thus underscoring the inherent limitations of lexical overlap-based measures in dual task captioning scenarios.

Table 5. Reference sentences labeled as no change corresponding to each caption in the LEVIR-CC test set.

Caption	No Change Sentence
Caption 1	“there is no difference.”
Caption 2	“the two scenes seem identical.”
Caption 3	“the scene is the same as before.”
Caption 4	“no change has occurred.”
Caption 5	“almost nothing has changed.”

The results in Table 6 clearly demonstrate the sensitivity of traditional captioning metrics to lexical variation in no change expressions. Despite conveying the same semantic meaning, the original configuration with diverse no change phrases yields substantially lower scores across all metrics. When a single, consistent phrasing is used, performance values increase dramatically particularly for BLEU, ROUGE-L, and CIDEr, which rely on exact *n*-gram overlap. This outcome empirically supports the discussion in the previous section: Automatic metrics can undervalue semantically correct predictions in dual task captioning datasets like LEVIR-CC, where linguistic diversity naturally arises. Therefore, metric interpretation must account for this limitation to avoid misleading conclusions about model quality.

Table 6. Evaluation results (%) on the LEVIR-CC test set for both change and no change samples. Three configurations are compared: (i) Original variation: distinct paraphrases for no change captions; (ii) Unified phrasing: predicted captions use a phrasing identical to one of the reference sentences ("the two scenes seem identical"); (iii) Minor lexical difference: a single word change ("is") in the no change sentence. Results highlight the sensitivity of automatic metrics to lexical variability.

Configuration (No Change Phrasing)	B1	B2	B3	B4	M	R	C	S	S_{BMRC}	S_{MCS}
Original variation: distinct paraphrases across captions	49.40	30.80	20.70	13.80	19.80	31.70	28.60	21.30	23.50	23.30
Unified phrasing: identical text to one reference ("the two scenes seem identical")	73.50	60.30	50.20	41.80	36.90	72.10	153.40	38.60	76.10	76.30
Minor lexical difference: "no change has occurred" to "no change is occurred"	71.80	47.10	29.70	19.40	35.10	59.60	59.70	22.40	43.50	39.10

B1 = BLEU-1, B2 = BLEU-2, B3 = BLEU-3, B4 = BLEU-4, M = METEOR, R = ROUGE-L, C = CIDEr, S = SPICE.

2.8. Toward a More Suitable Evaluation Strategy for Dual Task Captioning

The experiments discussed above reveal a key limitation of traditional captioning metrics when applied to the LEVIR-CC dual task setting. Conventional measures such as BLEU, METEOR, ROUGE-L, CIDEr, and SPICE are well suited to evaluating textual descriptions of visual changes but become unreliable when assessing no change captions. In these cases, minor lexical variations between semantically equivalent sentences (e.g., "the scene is the same as before" vs. "the two scenes remain identical") can lead to substantial fluctuations in computed scores, even though both statements convey the same meaning. This inconsistency obscures the true performance of models that must simultaneously detect and describe changes.

To address this issue, we propose a two-branch evaluation protocol that separates the measurement of descriptive quality from the assessment of change detection accuracy:

- Change samples: For samples where a change is present, we continue to employ standard captioning metrics BLEU, METEOR, ROUGE-L, CIDEr, and SPICE to quantify the quality, fluency, and semantic correctness of the generated change descriptions.
- No change samples: For samples where no change is present, evaluation focuses on the model's ability to correctly recognize scene stability. Since this task represents a categorical rather than descriptive decision, traditional similarity-based captioning metrics are not appropriate. Instead, we evaluate such cases using **accuracy**, defined as the ratio of correctly identified no change samples to the total number of no change instances:

$$S_{\text{no-change}} = \frac{\text{TP}_{\text{no-change}}}{N_{\text{no-change}}}, \tag{9}$$

where $TP_{\text{no-change}}$ denotes the number of correctly predicted no change cases, and $N_{\text{no-change}}$ is the total number of samples labeled as no change. This formulation yields an interpretable measure of the model's capacity to detect temporal stability, independent of linguistic variability.

Finally, we define a unified evaluation index, denoted as the Change Captioning Score (CCS), which integrates both descriptive and detection performance using a balancing coefficient $\beta \in [0, 1]$. Each variant of the composite metric can be incorporated as the change-sensitive component:

$$CCS_{BMRC} = \beta \cdot S_{CCS\_BMRC} + (1 - \beta) \cdot S_{\text{no-change}} \quad (10)$$

$$CCS_{MC} = \beta \cdot S_{CCS\_MC} + (1 - \beta) \cdot S_{\text{no-change}} \quad (11)$$

$$CCS_{MCS} = \beta \cdot S_{MCS} + (1 - \beta) \cdot S_{\text{no-change}} \quad (12)$$

The parameter β allows flexible weighting according to dataset composition or application focus, for example, $\beta = 0.5$ for balanced test sets or $\beta = 0.7$ when precise change description is prioritized.

This hybrid evaluation framework mitigates the lexical sensitivity of traditional metrics and aligns the assessment process with the dual objectives of multi-modal change captioning: accurately identifying when a change occurs and effectively describing what has changed.

2.9. Experimental Setup

To evaluate the effectiveness of our proposed framework, we employ the Mistral-Small-3.1-24B-Instruct-2503 (https://huggingface.co/mistralai/Mistral-Small-3.1-24B-Instruct-2503, accessed on 17 October 2025) model as the foundational multi-modal Large Language Model (LLM) for fine-tuning with Low-Rank Adaptation (LoRA). This state-of-the-art, instruction-tuned model with 24 billion parameters is specifically designed to handle image–text inputs and produce textual outputs, making it well suited for tasks requiring joint reasoning across visual and linguistic modalities. By integrating LoRA, we efficiently adapt its pre-trained weights to the specialized domain of multi-temporal remote sensing image captioning, while preserving the model's general multi-modal knowledge and language generation capabilities. This approach ensures that the model retains its robust ability to process and reason about both visual and textual information, while being fine-tuned to accurately describe spatial and temporal land cover changes, all with significantly reduced computational overhead compared to full fine tuning.

2.10. LoRA Integration: Target Layer Selection

Following the integration strategy described in Section 2.4, LoRA is applied only to the Vision Language Adapter and the multi-modal Transformer blocks. Table 7 summarizes the parameter components of the multi-modal model, highlighting which modules remain frozen and which include LoRA parameters (rank = 128) during fine-tuning.

2.11. Model Architecture and Training Configuration

Key architectural parameters are summarized below; detailed layer shapes are provided in Table 7.

Model Parameters

- Model Dimension (d_{model}): 5120.
- Number TextTransformer Blocks: 40.
- Attention Heads (n_{heads}): 40.

- Head Dimension (d_{head}): 128.
- Hidden Dimension (d_{ff}): 32,768.
- Vocabulary Size: 131,072.
- Normalization Epsilon (ϵ_{norm}): 1×10^{-5}.
- Vision Transformer: patch size 14×14 and 24 transformer blocks.

Table 7. Compact parameter dimensions of the multi-modal base model. Frozen (✗) and trainable LoRA (✓) parameters are indicated. Repeated transformer blocks are summarized for brevity. The value 128 in bold refers to the rank of the LoRA matrices.

Component	Parameter	Shape	Trainable
Text Embedding	Token Embed	[131072, 5120]	✗
Vision embedding (patch-level)	Patch Embed (`patch_conv.w`)	[1024, 3, 14, 14]	✗
	Pre-LN (`ln_pre.w`)	[1024]	✗
VisionTransformerBlock 0	Attn LN (`attn_norm.w`)	[1024]	✗
	$W_Q/W_K/W_V/W_O$	[1024, 1024]	✗
	FFN Proj $(W_1/W_2/W_3)$	[4096, 1024] / [1024, 4096]	✗
	FFN LN	[1024]	✗
VisionTransformerBlocks 1–24 (same structure)	Repeated block parameters (same as Block 0)		✗
VisionLanguageAdapter	Norm (`norm.w`)	[1024]	✗
	Patch Merge (`merge_layer.w`)	[1024, 4096]	✗
	Linear 1 Base (`base.w`)	[5120, 1024]	✗
	Linear 1 LoRA (A/B)	[**128**, 1024] / [5120, **128**]	✓
	Linear 2 Base (`base.w`)	[5120, 5120]	✗
	Linear 2 LoRA (A/B)	[**128**, 5120] / [5120, **128**]	✓
multi-modal TransformerBlocks 0	Attn $(W_Q/W_K/W_V/W_O)$ Base	[4096, 5120], [1024, 5120]	✗
	Attn LoRA (A/B)	[**128**, 5120], [4096, **128**]	✓
	FFN $((W_1/W_2/W_3))$	[32768, 5120], [5120, 32768]	✗
	FFN LoRA (A/B)	[**128**, 5120], [32768, **128**]	✓
	FFN LN	[5120]	✗
multi-modal TransformerBlocks 1–40 (same structure)	Repeated block parameters (same as Block 0)		✓ (LoRA only)
	Final LN (`final_norm.w`)	[5120]	✗
	LM Head (`lm_head.w`)	[131072, 5120]	✗

On the other hand, Table 8 presents the parameter statistics of the multi-modal model under LoRA fine-tuning, highlighting the small proportion of trainable parameters. Table 9 outlines the training configuration used in this work.

Table 8. Parameter statistics of the multi-modal base model. The table reports the total parameter count, trainable (LoRA) subset, and frozen parameters. LoRA adapters (rank = 128) enable fine-tuning only 3.0% of the total parameters.

Parameter Type	Count	% of Total
Total Parameters	24.75B	100%
• Trainable Parameters (LoRA)	741.3M	3.0%
• Frozen Parameters	24.01B	97.0%

Table 9. Training configuration used for LoRA fine-tuning of the multi-modal Mistral-24B model. All settings are derived directly from the training script. Lazy data loading and random sampling were applied to improve efficiency.

Parameter	Value/Description
LoRA Configuration	Rank = 128, α = 32, dropout = 0.1
Sequence Length	2048 tokens

Table 9. *Cont.*

Parameter	Value/Description
Batch Size	10
Training Epochs	5
Learning Rate	3×10^{-6}, cosine with restarts scheduler
Optimizer	AdamW (weight decay = 0.01, max grad norm = 1.0)
Loss Function	Cross entropy with masking applied.

2.12. Data Preparation and Input Formatting

For reproducibility, we detail the complete preprocessing and prompting pipeline used for multi-temporal caption training. All textual fields were normalized using Unicode NFC, lowercased, and cleaned to remove repeated whitespace and non-printable characters. Tokenization followed the native tokenizer of the Mistral-Small-3.1-24B-Instruct model (vocabulary size 131,072), ensuring consistency with the pretraining distribution.

The LEVIR-CC dataset is used not only for evaluation but also as the primary supervised dataset for model training, validation, and testing. We strictly follow the official data partitioning provided by the LEVIR-CC dataset website, which includes fixed and non-overlapping training, validation, and test splits. The training split is used to optimize the LoRA parameters, the validation split is employed for model selection and early stopping, and the test split is reserved exclusively for final performance evaluation, ensuring fair comparison and reproducibility across studies.

Each training sample contains a pair of satellite images corresponding to two timestamps. Following the conversational instruction format of the underlying model, we construct a message-based prompt in JSON form with explicit `user` and `assistant` roles. The user message contains a fixed instruction and both temporal images:

`text-prompt:` "Provide a concise summary highlighting the key differences between the two satellite images."
`Image_T1: <imageA>`
`Image_T2: <imageB>`

The assistant message contains the target caption. The dataset provides multiple reference captions per image pair; therefore, each caption is treated as an independent training instance. The mapping follows the structure

```
''messages'': [
{''role'': ''user'', ''content'': [text-prompt, imageA, imageB]},
{''role'': ''assistant'', ''content'': [reference caption]} ]
```

This structure directly mirrors the training implementation, where, for each image pair, we iterate over all available human written captions and create one training example per caption. This increases data diversity and stabilizes instruction-following behavior during fine tuning.

At inference time, the same instruction is used, but the assistant content is omitted, enabling autoregressive caption generation conditioned solely on the two temporal images. No additional prompt engineering or explicit change/no-change heuristics are applied; the model learns to infer change patterns implicitly from the instruction and training corpus. This unified prompting strategy ensures consistent alignment between preprocessing, training, and inference.

2.13. Computational Environment

The training was performed using 3 NVIDIA A100 GPUs (40 GB each) in a distributed setup. Mixed precision computation (`bfloat16`) and `torch.distributed` data parallelism

were employed to balance memory usage and throughput during LoRA fine-tuning of the Mistral-Small-3.1-24B-Instruct model. This configuration provided sufficient computational capacity to handle the multi-modal inputs and long sequence lengths.

The fine-tuning experiments were conducted using Python 3.10 and the PyTorch framework, built upon the Hugging Face Transformers ecosystem. The workflow integrated key libraries including `transformers`, `datasets`, and `peft` for LoRA-based parameter-efficient adaptation. Image preprocessing and feature handling were managed through `torchvision` and `Pillow (PIL)`.

For quantitative assessment, we employed the Microsoft COCO Caption Evaluation toolkit, available at (https://github.com/jiasenlu/coco-caption/tree/master, accessed on 17 October 2025). This toolkit is widely adopted for benchmarking captioning and multi-modal generation tasks, providing standardized implementations of key metrics such as BLEU, METEOR, ROUGE-L, CIDEr, and SPICE.

3. Results

This section presents an empirical evaluation of the proposed multimodal Vision Language Transformer framework for multi-temporal remote sensing image change captioning. Leveraging the LEVIR-CC dataset (https://github.com/Chen-Yang-Liu/LEVIR-CC-Dataset, accessed on 17 October 2025), we systematically assess the model's capacity to generate semantically rich and spatially accurate captions that reflect land cover changes over time.

Tables 10 and 11 summarize the quantitative comparison between the proposed method and representative state of the art (SoTA) approaches on the LEVIR-CC dataset. Table 10 reports the results for change samples, evaluated using standard lexical and semantic metrics, as well as the aggregated indicators S_{BMRC}, S_{MC}, and S_{MCS}.

Table 10. Quantitative comparison (%) of the proposed method against state of the art (SoTA) approaches for samples with only changes on the LEVIR-CC test set. S_{BMRC} is the average of BLEU-4, METEOR, ROUGE-L, and CIDEr; S_{MC} represents the mean of METEOR and CIDEr; and S_{MCS} denotes the average of METEOR, CIDEr, and SPICE. Values shown in bold indicate the best scores.

Method	B1	B2	B3	B4	M	R	C	S	S_{BMRC}	S_{MC}	S_{MCS}
Caption 3	62.00	43.30	29.90	20.10	25.00	44.20	46.70	22.70	34.00	35.85	31.50
Chg2Cap [3]	**77.32**	**63.26**	50.06	39.09	25.73	52.02	58.30		43.79	42.02	-
RSICCformer [4]	75.94	61.25	47.85	37.08	25.88	52.75	60.59		44.24	43.14	-
SAT-Cap [5]	77.12	63.20	**51.20**	**41.46**	26.28	53.23	68.91		47.47	47.60	-
Ours	75.00	60.10	47.10	36.70	**30.00**	**54.50**	**72.30**	25.70	**48.40**	**51.10**	42.66

Table 11. Quantitative comparison (%) between the proposed method and state of the art (SoTA) approaches for both change and no change samples on the LEVIR-CC test set, evaluated using the composite metrics introduced in Section 2.8. The complementary consistency scores (CCS) combine both branches using $\beta = 0.5$, offering a unified indicator of overall captioning consistency and multi-modal robustness. Values shown in bold indicate the best scores.

Method	S_{BMRC}	S_{MC}	S_{MCS}	$S_{\text{no-change}}$	CCS_{BMRC}	CCS_{MC}	CCS_{MCS}
Caption 3	34.00	35.85	31.50	100.0	67.00	67.93	65.75
RSICCformer [4]	44.24	43.14	-	94.48	69.36	68.81	-
Chg2Cap [3]	43.79	42.02	-	**98.22**	71.00	70.12	-
SAT-Cap [5]	47.47	47.60	-	97.80	72.64	72.70	-
Ours	**48.40**	**51.10**	42.66	96.89	**72.65**	**74.00**	69.78

The proposed model achieves superior performance across the majority of semantically oriented metrics, particularly METEOR, ROUGE-L, and CIDEr, indicating improved descriptive precision and contextual understanding. While SAT-Cap attains slightly higher BLEU scores due to greater lexical overlap with reference captions, our model exhibits a more balanced performance overall. The consistent gains in both S_{BMRC} and S_{MC} highlight stronger agreement between lexical accuracy and semantic coherence, underscoring the effectiveness of the multi-modal fusion- and LoRA-based adaptation strategy.

Table 11 extends the evaluation to include no change samples via the $S_{\text{no-change}}$ score, and integrates both sample ranges using the Complementary Consistency Scores (CCS) defined in Section 2.8. For this analysis, the balancing coefficient is set to $\beta = 0.5$, providing equal emphasis on descriptive fidelity and categorical accuracy.

In this complementary evaluation, the proposed framework achieves the highest consistency across both change and no change scenarios, reaching $\text{CCS}_{BMRC} = 72.65$ and $\text{CCS}_{MC} = 74.00$. These results confirm the model's capacity to generalize effectively, maintaining coherent and contextually appropriate descriptions even in temporally stable scenes. By contrast, previous approaches often reach high $S_{\text{no-change}}$ values only when the output captions exactly replicate reference expressions, thereby reducing the generative nature of the task to simple classification.

4. Discussion

A key advantage of the proposed approach lies in the integration of newly learned knowledge into the pre-trained language model without disrupting its universal linguistic space. The large language model employed here retains its general purpose vocabulary and semantic associations, allowing it to describe remote sensing phenomena using natural and contextually meaningful language rather than dataset-specific terms. This contrasts sharply with many SoTA captioning models, which rely on restricted vocabularies or dataset-dependent token dictionaries. The proposed hybrid ViT–LLM architecture, enhanced through LoRA fine-tuning, therefore bridges specialized remote sensing semantics with a general linguistic foundation, promoting better cross-domain generalization and more human-aligned textual descriptions of land cover dynamics.

Table 10 highlights an important observation regarding the evaluation of methods on samples with only changes. While some approaches, such as *Chg2Cap*, may achieve competitive overall performance when considering combined tasks, a closer look at metrics focused purely on text generation reveals a different trend. Under this isolated evaluation, *RSICCformer* outperforms *Chg2Cap* in both S_{BMRC} (44.24 vs. 43.79) and S_{MC} (43.14 vs. 42.02), demonstrating superior descriptive fidelity and semantic coherence. This underscores the importance of disentangling classification and generation tasks in order to fairly assess methods' capabilities of producing accurate and meaningful captions. Notably, our proposed approach further surpasses all baselines across most generation-focused metrics, reinforcing its effectiveness in capturing the nuanced changes in the scene.

In this evaluation, Caption 3 is treated as the prediction while the other four captions (see LEVIR-CC Section 2.5) serve as references. Caption 3 reflects the accurate description provided by a human expert, with all captions being correct and valid. However, it often employs different words to describe the changes, so the differences mainly affect semantics rather than lexical choice. Despite its expert quality, Caption 3 does not outperform any of the automatic generation methods across the considered metrics. Notably, METEOR demonstrates the closest alignment with Caption 3, indicating its particular sensitivity to semantic adequacy and meaning preservation, which are central to the human expert's description.

4.1. Ablation Study: Zero-Shot Baselines vs. LoRA Fine Tuning

To further quantify the contribution of the proposed LoRA-based adaptation, we conducted an ablation study comparing our model to two advanced, open-source, vision language models evaluated in a zero-shot setting: Mistral-Small-3.1-24B-Instruct-2503 and Qwen2-VL-72B-Instruct [20]. Although both models support general-purpose image captioning and visual reasoning, they are trained for broad multimodal tasks rather than the specialized objective of multi-temporal remote sensing change captioning.

To ensure a fair comparison, we designed a unified instruction prompt tailored to the expected answering behavior of general models:

"Analyze the two satellite images. Describe only significant object changes (appearances/disappearances) in up to 20 words. Ignore irrelevant differences. If no change, respond with: 'the two scenes seem identical', 'the scene is the same as before', or 'there is no difference'."

This prompt was carefully crafted to (i) constrain the output length, (ii) emphasize object-level transformations, and (iii) standardize the no change responses so that zero-shot models could produce captions compatible with the LEVIR-CC evaluation protocol. For comparability, metrics were computed only on samples with changes, ensuring that evaluations reflect true land cover transformation detection.

Despite using a prompt designed to favor them, the zero-shot baselines performed substantially worse than our LoRA-enhanced model, as reported in Table 12. The baseline models struggle to localize or describe land cover transformations, often generating generic or spatially inconsistent captions. In contrast, our model, which incorporates task-specific LoRA modules selectively inserted into the cross-modal alignment layers, achieves significantly higher scores across all captioning metrics.

These results confirm that the performance improvements stem directly from the proposed architectural contributions and LoRA-based adaptation, rather than from the intrinsic capabilities of the underlying language model. The ablation demonstrates that domain-aligned multimodal fine tuning is essential for generating accurate and semantically faithful descriptions of land cover changes.

Table 12 summarizes the results on the LEVIR-CC test set. As expected, both zero-shot models show limited accuracy due to the domain-specific nature of land cover change description. In contrast, our LoRA-enhanced model substantially improves performance across all captioning metrics, demonstrating that the gains stem directly from the proposed fine tuning mechanism. These findings confirm the importance of task-adapted parameter updates for capturing subtle and diverse land cover transitions in multi-temporal remote sensing imagery.

Table 12. Ablation study comparing zero-shot models with the proposed LoRA-enhanced framework on the change only subset of the LEVIR-CC test set. Scores are computed using standard captioning metrics and quantify descriptive accuracy exclusively on samples that contain observable land cover changes. Values shown in bold indicate the best scores.

Method	B1	B2	B3	B4	M	R	C	S_{BMRC}	S_{MC}
Mistral 24B (zero-shot)	26.1	13.0	6.8	4.2	13.2	22.7	7.0	11.8	8.7
Qwen2-VL-72B (zero-shot)	27.8	11.7	5.1	2.3	9.5	18.0	2.5	8.1	6.0
Ours	75.0	60.1	47.1	36.7	**30.0**	**54.5**	**72.3**	**48.4**	**51.1**

4.2. Ablation Study on LoRA Placement

To further evaluate the influence of LoRA placement on multi-temporal change captioning performance, we conduct an ablation study comparing three configurations: (A) the proposed selective LoRA integration, where adapters are inserted only into the multi-modal

Transformer Blocks and the Vision Language Adapter; (B) LoRA applied to all linear layers in the multimodal architecture, excluding Conv2D projections; and (C) LoRA applied to all linear layers and Conv2D blocks in both the vision tower and multimodal projector.

Following the evaluation protocol used throughout the paper, all metrics are computed exclusively on samples containing actual changes, ensuring that the results reflect the models' ability to describe meaningful land cover transformations.

The ablation results (see Table 13) show that the proposed selective LoRA placement (Config. A) yields the strongest overall performance. It achieves the highest scores across standard captioning metrics (BLEU, METEOR, ROUGE-L, and CIDEr) as well as the change-specific measures S_{BMRC}, S_{MC}, and S_{MCS}, indicating that adapting only the Vision Language Adapter and the multi-modal Transformer Blocks most effectively enhances temporal reasoning and semantic alignment.

Table 13. Ablation results for different LoRA placement strategies. Metrics computed using only samples with actual changes. Values shown in bold indicate the best scores.

LoRA Config	B1	B2	B3	B4	M	R	C	S	S_{BMRC}	S_{MC}	S_{MCS}
A_{Ours}	**75.0**	**60.1**	**47.1**	**36.7**	**30.0**	**54.5**	**72.3**	25.7	**48.4**	**51.1**	**42.7**
$B_{AllLinearLayers}$	71.7	56.4	43.5	33.5	29.7	51.5	65.7	24.8	45.1	47.7	40.1
$C_{Linear+Conv2D}$	73.5	57.7	44.2	33.4	29.6	52.1	60.8	**26.6**	44.0	45.2	39.0

Applying LoRA to all linear layers (Config. B) consistently reduces performance, suggesting that many of these projections do not benefit from adaptation and may introduce noise that weakens temporal fusion and description accuracy. Extending LoRA to both linear layers and Conv2D modules (Config. C) does not improve results either. While SPICE increases slightly, CIDEr and the CCS metrics drop, indicating that modifying early convolutional filters disrupts stable visual features needed for reliable cross-temporal alignment.

Overall, the results confirm that LoRA must be applied selectively: Preserving early visual representations while adapting high-level vision language components is essential for accurate and coherent multi-temporal change descriptions.

4.3. Qualitative Evaluation

To provide an intuitive understanding of the behavior of the proposed MVLT-LoRA-CC model, we conduct a qualitative evaluation using visual examples from the test split of the LEVIR-CC dataset. Eight representative samples are selected to cover a range of land cover types, including vegetation, roads, parking areas, and residential structures, as well as varying degrees of change intensity.

Good Predictions

Figures 3–6 illustrate four examples in which the model successfully captures the underlying multi-temporal changes.

Figure 3 (test_000280.png): Although the prediction differs lexically from the ground truth, it conveys the same underlying change vegetation removal and the addition of built structures demonstrating the model's ability to express similar content through alternative phrasings.

Figure 4 (test_000047.png): This example shows that the model identifies a limited set of changes, primarily the new structure and its road connection.

Figure 5 (test_000104.png): The prediction correctly identifies the replacement of trees by a road and houses, using slightly different wording and a more approximate measure of the affected area.

Figure 6 (test_000540.png): The model detects the main changes removal of vegetation and construction of buildings but uses more general terms (“many houses around” vs. “two lines of buildings and a house”), illustrating semantic generalization.

Figure 3. Example of a correct prediction on the test set (image: `test_000280.png`). Image A earlier acquisition and Image B later acquisition. GT: A mall and a parking lot are built opposite a residential area through a road replacing the plants. Pred: The vegetation has been removed, and a road with a building and a parking lot has been built at the bottom along with a row of houses at the top.

Figure 4. Example of a correct prediction on the test set (image: `test_000047.png`). Image A earlier acquisition and Image B later acquisition. GT: A building with a path connected to the road appears. Pred: A house is built at the end of the road.

Figure 5. Example of a correct prediction on the test set (image: `test_000104.png`). Image A earlier acquisition and Image B later acquisition. GT: Many trees are replaced by a road and two rows of houses. Pred: A road with houses on both sides replaces half of the woods.

Figure 6. Example of a correct prediction on the test set (image: `test_000540.png`). Image A earlier acquisition and Image B later acquisition. GT: Many trees are replaced by two lines of buildings on the sides and a house at the bottom. Pred: The vegetation has been removed, and a road with many houses around it has been built.

Bad Predictions

Figures 7–10 present four examples highlighting typical failure cases.

Figures 7–9 (test_000096.png, test_000388.png, test_000968.png): The model struggles with sparse or subtle changes, as illustrated by missed detections of a villa appearing by the road, several bungalows scattered in a forest, or the disappearance of two small houses, where it predicted that "the scene is the same as before." These failures highlight the difficulty of identifying minor or distributed changes within largely unchanged areas. One contributing factor may be the Vision Transformer's patch size (e.g., 14×14 pixels), which can dilute small-scale features during embedding, making subtle changes less salient to the model.

Figure 10 (test_000373.png): Although this prediction might be considered "bad" by automated metrics due to low overlap with the ground truth, it still captures meaningful changes in the scene. The model correctly identifies the emergence of a new building and even adds a plausible detail: a swimming pool that is not in the GT. This illustrates that natural language predictions can convey semantically valid information about changes, even when they differ from the exact GT phrasing or object placement.

Figure 7. Example of an incorrect prediction on the test set (image: `test_000096.png`). Image A earlier acquisition and Image B later acquisition. GT: A villa appears on the side of the road on the right. Pred: The scene is the same as before.

Figure 8. Example of an incorrect prediction on the test set (image: `test_000388.png`). Image A earlier acquisition and Image B later acquisition. GT: Several bungalows are scattered in the forest. Pred: The scene is the same as before.

Figure 9. Example of an incorrect prediction on the test set (image: `test_000968.png`). Image A earlier acquisition and Image B later acquisition. GT: Two small houses disappear. Pred: The scene is the same as before.

Figure 10. Example of an incorrect prediction on the test set (image: `test_000373.png`). Image A earlier acquisition and Image B later acquisition. GT: More trees are built, and a building appears at the end of the road. Pred: A house with a swimming pool appears on the right side of the scene.

5. Conclusions

This work presents MVLT-LoRA-CC, a multi-modal Vision Language Transformer framework enhanced with Low-Rank Adaptation (LoRA) for multi-temporal remote sensing image captioning focused on land cover change description. By integrating paired temporal imagery with large pre-trained language models, the proposed approach effectively bridges spatiotemporal change detection and natural language generation. LoRA is selectively integrated into the Vision Language Adapter and multi-modal Transformer Blocks, enabling efficient fine tuning that preserves general linguistic knowledge while adapting the model to the specialized task of describing land cover changes. Compared with prior methods that rely on limited vocabularies or rigid architectures, our model demonstrates stronger semantic generalization and more expressive, contextually accurate change descriptions.

Extensive experiments, including quantitative evaluation and detailed qualitative examples, show that MVLT-LoRA-CC captures a wide range of changes and produces coherent, semantically meaningful captions. The Complementary Consistency Score (CCS) metrics CCS_{BMRC}, CCS_{MC}, and CCS_{MCS} provide a robust evaluation framework for both change and no change scenarios, highlighting the model's ability to distinguish subtle modifications.

Overall, the results underscore the value of integrating universal language models into remote sensing pipelines, enhancing interpretability, robustness, and generalization while reducing reliance on dataset-specific linguistic patterns. The ability to generate natural language descriptions of multi-temporal changes positions the model as a promising tool for applications such as disaster assessment, deforestation monitoring, and urban or infrastructure evolution tracking.

Future work will explore multi-sensor and multi-temporal fusion (e.g., SAR, hyperspectral), instruction-tuned or prompt-based adaptation for controllable caption generation, and evaluation on cross-dataset and multilingual benchmarks to further assess the robustness and universality of the proposed approach.

Author Contributions: Conceptualization, methodology, writing—review and editing, and code, J.L.L.; review and editing, P.Q., P.S. and V.N. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: No new proprietary data were created for this study. All experiments were conducted using publicly available remote sensing datasets, which were accessed in their original form without modification. The data supporting the findings of this work are fully accessible from their respective public sources. No restrictions apply.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

ViT	Vision Transformer
LLM	Large Language Model
LoRA	Low-Rank Adaptation
BLEU	Bilingual Evaluation Understudy
METEOR	Metric for Evaluation of Translation with Explicit ORdering
ROUGE-L	Recall-Oriented Understudy for Gisting Evaluation
CIDEr	Consensus-based Image Description Evaluation
SPICE	Semantic Propositional Image Caption Evaluation
LEVIR-CC	LEVIR Change Captioning Dataset

References

1. Vinyals, O.; Toshev, A.; Bengio, S.; Erhan, D. Show and Tell: A Neural Image Caption Generator. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015 ; pp. 3156–3164.
2. Daudt, R.C.; Le Saux, B.; Boulch, A. Fully Convolutional Siamese Networks for Change Detection. In Proceedings of the 2018 25th IEEE International Conference on Image Processing (ICIP), Athens, Greece, 7–10 October 2018; pp. 4063–4067.
3. Chang, S.; Ghamisi, P. Changes to captions: An attentive network for remote sensing change captioning. *IEEE Trans. Image Process.* **2023**, *32*, 6047–6060. [CrossRef] [PubMed]
4. 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 Trans. Geosci. Remote Sens.* **2022**, *60*, 1–20. [CrossRef]
5. Wang, Y.; Yu, W.; Ghamisi, P. Change Captioning in Remote Sensing: Evolution to SAT-Cap–A Single-Stage Transformer Approach. *arXiv* **2025**, arXiv:2501.08114.
6. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning Transferable Visual Models from Natural Language Supervision. In Proceedings of the Proceedings of the International Conference on Machine Learning (ICML), Virtual, 18–24 July 2021; pp. 8748–8763.
7. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; others. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
8. Su, J.; Ahmed, M.; Lu, Y.; Pan, S.; Bo, W.; Liu, Y. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing* **2024**, *568*, 127063. [CrossRef]
9. Alayrac, J.B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Hasson, Y.; Lenc, K.; Mensch, A.; Millican, K.; Reynolds, M.; et al. Flamingo: A visual language model for few-shot learning. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 23716–23736. [CrossRef]
10. Aghajanyan, A.; Zettlemoyer, L.; Gupta, S. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. *arXiv* **2020**, arXiv:2012.13255. [CrossRef]
11. Hu, E.J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Cao, Y.; Wang, S.; Wang, L. LoRA: Low-Rank Adaptation of Large Language Models. *arXiv* **2021**, arXiv:2106.09685.
12. Wang, S.; Yu, L.; Li, J. LoRA-GA: Low-Rank Adaptation with Gradient Approximation. *arXiv* **2024**, arXiv:2407.05000.
13. Yang, Y.; Liu, T.; Pu, Y.; Liu, L.; Zhao, Q.; Wan, Q. Remote sensing image change captioning using multi-attentive network with diffusion model. *Remote Sens.* **2024**, *16*, 4083. [CrossRef]
14. 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. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5648916. [CrossRef]
15. 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, Philadelphia, PA, USA, 6–12 July 2002; pp. 311–318.
16. Banerjee, S.; Lavie, A. 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, Ann Arbor, MI, USA, 29 June 2005; pp. 65–72.
17. Lin, C.-Y. ROUGE: A package for automatic evaluation of summaries. In Proceedings of the Workshop on Text Summarization Branches Out (WAS 2004), Barcelona, Spain, 25–26 July 2004; pp. 74–81.
18. Vedantam, R.; Zitnick, C.L.; Parikh, D. CIDEr: Consensus-based image description evaluation. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 4566–4575.
19. Anderson, P.; Fernando, B.; Johnson, M.; Gould, S. SPICE: Semantic propositional image caption evaluation. In Proceedings of the European Conference on Computer Vision (ECCV), Amsterdam, The Netherlands, 8–16 October 2016; pp. 382–398.
20. Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. Qwen2.5-vl technical report. *arXiv* **2025**, arXiv:2502.13923. [CrossRef]