INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

ORIGINAL ARTICLE

OPEN ACCESS

Received: 28/07/2025

Accepted: 14/08/2025

Published: 11/09/2025

Citation: Dhinesh A, Sumathy P, Vadivel A (2025) Performance of Generative AI used by ChatGPT and Attention Mechanism for Remote Sensing Image Captioning (RSIC): A Comparative and Comprehensive Analysis. Indian Journal of Science and Technology 18(33): 2740-2751. https://doi.org/10.17485/IJST/v18i33.1301

* **Corresponding author**.

ayy.dhinesh@gmail.com

Funding: None

Competing Interests: None

Published By Indian Society for Education and Environment (iSee)

ISSN
Print: 0974-6846
Electronic: 0974-5645

Performance of Generative AI used by ChatGPT and Attention Mechanism for Remote Sensing Image Captioning (RSIC): A Comparative and Comprehensive Analysis

A Dhinesh[1]*, P Sumathy[2], A Vadivel[3]

1 Research Scholar, Department of Computer Science, Bharathidasan University, Tiruchirappalli-620023, Tamilnadu, India
2 Assistant Professor, Department of Computer Science, Bharathidasan University, Tiruchirappalli-620023, Tamilnadu, India
3 Professor, Department of Artificial Intelligence and Data Science, GITAM School of Technology GITAM University, Bangaluru-561203, Karnataka, India

Abstract

Objective: Remote Sensing Image Captioning (RSIC) is a challenging research area that aims to automatically generate descriptive textual captions for Remote Sensing Images (RSIs) with recent advancements in deep learning, RSIC performance has significantly improved. **Methods:** This study presents a comparative analysis of two prominent approaches: Channel Attention Mechanism (CAM) and Generative AI (GenAI). The Channel Attention Mechanism emphasises on learning channel-wise with encoding-decoding, feature extraction and attention module. On the other hand, Generative AI, particularly models like GPT, leverages large-scale pretraining to diverse datasets for generates coherent and contextually relevant captions. It extracts vision-based feature and contains auto-tokenizer within the ViT Encoder-Decoder. **Findings:** This study evaluated on the RSICD dataset using metrics such as BLEU, METEOR, ROUGE-L, CIDEr, Precision, Recall, F1-Score, F2-Score, and Accuracy, GenAI outperforms CAM across all these metrics. Specifically, GenAI achieves a Precision of 0.90, Recall of 0.86, and Accuracy of 0.92. BLEU-4 scores for GenAI (0.7684) exceed CAM (0.4932), highlighting superior caption quality. However, both methods show limitations in accurately quantifying objects and specifying their coordinates within images. **Novelty:** The comparative analysis highlights the trade-off between traditional attention-based method and modern generative models.

Keywords: Channel Attention Mechanism; Encoder-Decoder; Feature Extraction; Generative AI; Pre-Trained Models; RSICD Dataset

1 Introduction

Remote Sensing Images (RSIs) are captured from the satellite, aircraft, drones, etc., and Remote Sensing Image Captioning (RSIC) is a process of generating textual description

of RSIs. It also describes the relationship between the scenes and objects present in the images[1]. It is an interdisciplinary field that uses the concept of computer vision and Natural Language Processing (NLP) and other contemporary fields[2],[3]. It focuses on to generate accurate and relevant textual description for visual content present in an image. Channel Attention Mechanism (CAM) is a method that uses Convolutional Neural Networks (CNNs) for feature extraction from Remote Sensing Images. It enhances the visual features by assigning weights to the attention pixel of different channels. It focusses on informative channels and detect relevant features by suppressing irrelevant information. As a result, quality of feature that is extracted is good for generating captions. This method is integrated with various architectures for generating image caption and shows promising results in terms of accuracy and relevance[4],[5]. On the other hand, Generative AI (GenAI) is a method that uses a wide range of techniques for generating human-like output and readable descriptions[6]. It uses transformer architecture based neural network to generate captions based on visual perception. It offers flexibility and creativity in caption generations and however, it may face challenges in detecting and captioning fine-grained visual details and maintaining coherence in longer descriptions.

Fig 1. Sample of Remote Sensing Image Captioning. (a) Many planes are parked near a large building on an airport. (b) Several buildings and green trees are around a piece of bare land. (c) Some planes are parked near an airport with parking lot

In Figure 1, we depict the sample of remote sensing images for captioning. These images are having pictures of planes that are parked near an airport. Figure 1(a) is captioned as "Many planes are parked near a large building on an airport", Figure 1(b) is captioned as "Several buildings and green trees are around a piece of bare land" and Figure 1(c) is captioned as "Some planes are parked near an airport with parking lot". It is observed from these captioning texts is that each one is captioned in different way and the commonality is the caption perform text. As a result, we present a comparative study of Channel Attention Mechanism and Generative AI is used by ChatGPT[7] in this paper. The reason for choosing these two methods is that ChatGPT has penetrated deep in the Remote Sensing Image Captioning research. Similarly, the Channel Attention Mechanism is considered as one of the popular approaches in Remote Sensing Image Captioning.

The rest of this paper is organized as follows, review methodology of the research, details on dataset and evaluation metrics are presented in Section 2. The results are compared and analysed in Section 3 and the paper is concluded in last Section of the paper.

1.1 Background of the Study and Focus Area

Remote sensing image captioning is an important research area to generate captions and describe the objects in Remote Sensing Images. Natural Image Captioning (NIC)[8] refers to the process of captioning conventional image from natural images. There are three types of NIC and template-based is one of them that artificially generate sentences with fixed structure. Retrieval-based method generates images, its corresponding captions and, however, it is not suitable for generating flexible sentences. This issue is handled by the sequence generation-based method in which the image features are trained, caption models are combined and sequential relationship with adjacent words are maintained. However, this approach generates captions with syntactic errors.

The RSIC with channel attention mechanism has been proposed by Cheng et al., It uses the RSIC Datasets and theory of deep learning, computer vision and natural language processing approaches. CNNs extracts the features from the image and RNN generate caption. In Generative AI based models, realistic captions are generated using Deep Learning methods (CNNs and RNNs) and transformers[9]. It uses both vision and image transformer models. The image is directly processed, and captions

are generated autoregressively. Based on the above observation the focus area of methods is RSIC, and these methods should have improved usual understanding, which refers to better object recognition, scene understanding good fine-grained detail and effective representation of images. The Generative AI and Large Language Model (LLM) should focus on convergence rate, effective data utilization, capturing rich interplay and capabilities to quantify the objects along with their location coordinates. Exploring techniques for developing pre-trained models for a domain specific application with generalization capability and improved performance. The researchers should have a robust evaluation metrics similar to BLEU, METEOR, ROUGE and CIDEr for measuring the quality of captions that are generated. In the next section we comprehensively review these two methods and present the analysis.

2 Review Methodology

In this section, we present the comparative analysis and working principles of Channel Attention Mechanism and Generative AI.

2.1 Channel Attention Mechanism (CAM)

The functional diagram of this approach is depicted in Figure 2. The input image is processed by both CNN and LSTM and the structural features are extracted. Both of these features are presented to Channel Attention Mechanism [10] for generating the caption of the input image.

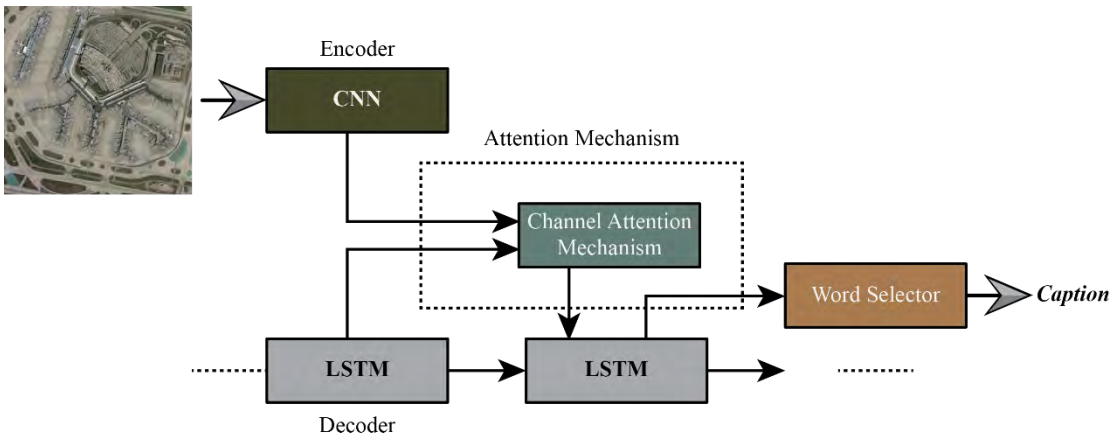

Fig 2. Functional diagram of Channel Attention Mechanism

2.1.1. Encoder - Decoder Architecture

The encoder-decoder architecture is a common framework used in image captioning tasks. The encoder is typically a Convolutional Neural Network (CNN) and extracts high-level features of input image. Pre-trained CNN architectures such as VGG, ResNet and Inception are used as encoders for their effectiveness in extraction feature from the images. The encoder transforms the input image into a fixed-length feature vector, which unfolds rich semantic information of the image. The decoder is typically a Recurrent Neural Network (RNN) or transformer-based architecture that generate captions based on the encoded image features. Long Short-Term Memory (LSTM) generates captions sequentially, i.e. one word at a time. At each time, the decoder uses the word that was generated in the previous iteration, encodes image features and predict the next word in the caption. The decoding process continues until the end-of-sequence token is generated or a maximum caption length is reached. The model is trained end-to-end using a dataset of image-caption pairs. The encoder and decoder are jointly trained to minimize the loss function that measures the difference between the predicted captions and the ground truth captions. The trained model processes the input image and generates a caption by the learned encoder-decoder architecture. The decoder generates captions using either greedy decoding or beam search. The greedy decoding chooses the word that has the highest probability and in contrast the beam search explores the multiple captions in parallel. Finally, the caption of the input image is generated by the model. The encoder-decoder architecture of image captioning leverages the strength of both CNN and RNN. While CNN is used for extracting the feature of the image, the RNN also called as transformer and it is used for generating sequential language. It allows the model to effectively capture the visual content of the input image and generate coherent and descriptive captions. In

addition, the modular nature of the architecture has the flexibility to experiment different encoder and decoder architectures [11] for improvs the performance.

2.1.2. Image Feature Extraction from Remote Sensing Images

The Convolutional Neural Network (CNN) is predominantly used for extracting feature from Remote Sensing Images. The input images are pre-processed to fix their size, format, enhancing contrast and sharpness. Various well-known models such as Visual Geometry Group of (VGGNet), Residual Network (ResNet) or Inception Net are used for extracting feature from the input based on the complexity of feature. The simple features such as edges, texture and patterns are detected and extracted in the convolutional layers of CNN. In contrast, the complex and abstract features such as shape, structure, and context are detected and extracted by the deep layer of the CNN. The pre-trained CNN model on a large dataset like ImageNet is fine-tuned on the remote sensing image dataset and the feature is learned by employing transfer learning principle. The feature map generated by the CNN represents the spatial activations [12] across different channels to capture both local and global information. All the feature information that are extracted from various layers are aggregated to represent the RSI comprehensively. The LSTM network transformer architecture analyse the features and generates descriptive captions based on learned features.

2.1.3. Training and Validation Strategy

The RSICD dataset is divided into three subsets as training, testing and validation and the ratio is 70% ,15% and 15%, respectively. The ratio can be modified based on the size of dataset [13] and specific requirements. The ratio is choosing such that each image in the RSICD dataset is associated with one or more classes. While splitting the dataset, it is ensured that the images and the corresponding captions are logically linked in the relevant subset. The data are randomized before splitting to ensure that each subset is representative of the entire dataset. The model is trained on the training data, which is monitored using epochs and loss as shown in Figure 3. It is observed from Figure 3 that the value of loss is minimized while the number of epochs increases and after 17 epochs both the training and validation loss are saturating.

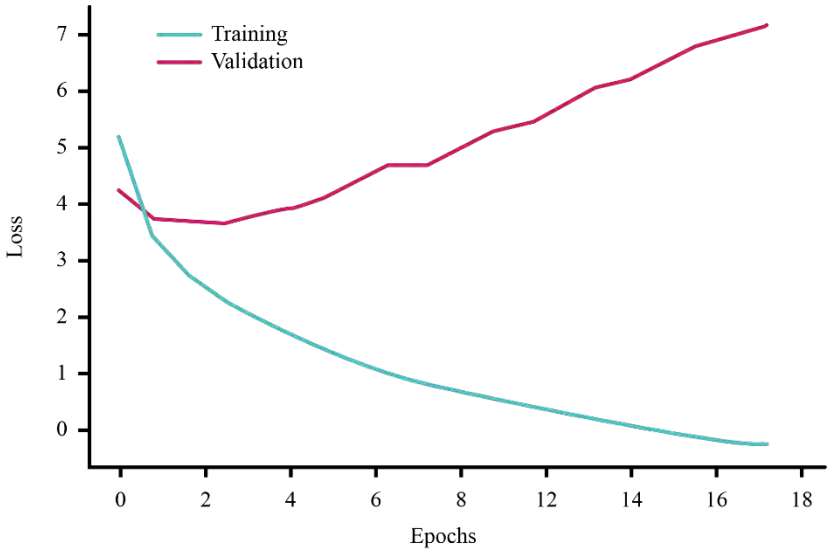

Fig 3. Training and Validation loss on RSICD

Categorical cross-entropy loss [14] is a measure that calculates the difference between the predicted and ground truth captions. The model is optimized to minimize the difference between generated and ground truth captions during the training phase. The learning is supervised using validation data after every epoch in the validation phase. Hyperparameters are chosen based on the model's performance. The captions for the test images are generated based on the trained data. Evaluation metrics such as BLUE, METEOR, ROUGE and CIDEr are used for measuring the quality of generated captions and validation loss [15].

2.2 Generative AI (GenAI)

The concept of Generative AI in Remote Sensing Image Captioning generate description of RSIC using pre-trained models by understanding the objects present in RSI. In this section, we present the feature extraction using Vision Transformer, decoding strategy for generating caption using Generative AI models.

2.2.1. Vision Transformer for Image Feature Extraction

Vision Transformer (ViT) architecture[16] extract features from input RSI using CNN. The Vision Transformer is pre-trained on a large-scale image dataset to learn meaningful representations of image content. The model is pre-trained on a specific dataset for adapting to the data distribution. The Recurrent Neural Networks (RNNs), Transformer-Based Large Language Models (ChatGPT) are trained on RSICD dataset and the features are extracted. The procedure to generate caption using Generative AI is depicted in Figure 4. The input image is fed to CNN based ViT feature extractor. The CNN extracts the structural information is fed into Auto-Tokenizer (LLM) and the captions are generated.

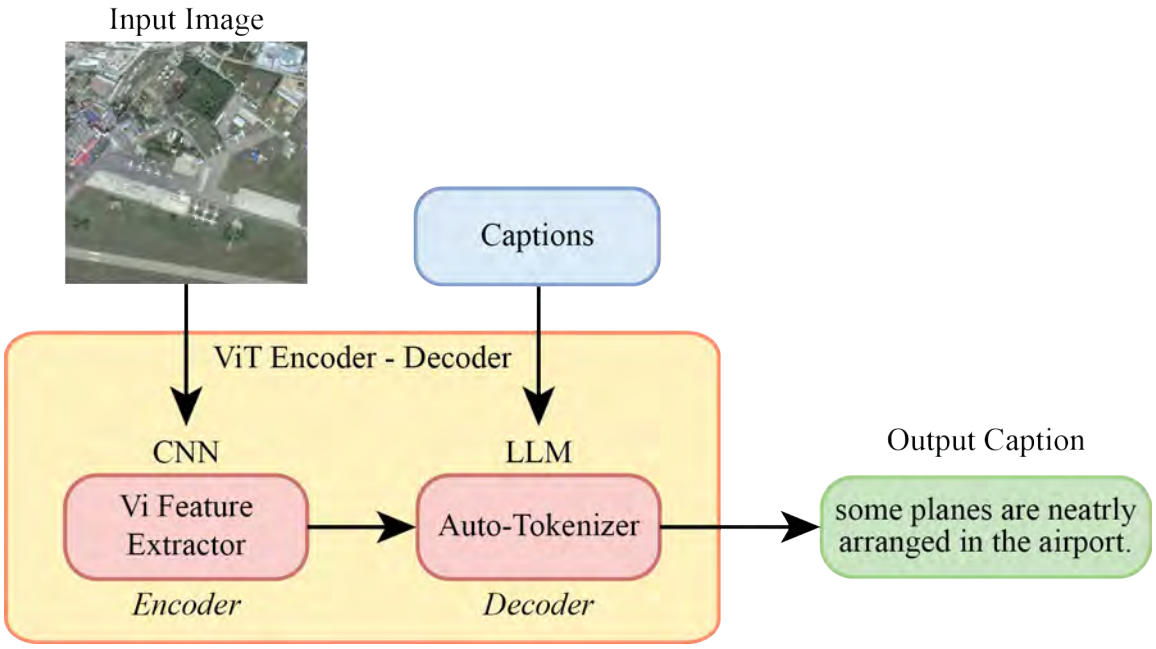

Fig 4. Process to generate caption using GenAI

2.2.2. Decoding Strategy to Generate Caption

The model is trained and optimized to generate captions that are semantically relevant to the input. Both combined Vision Transformer and Generative AI models are trained end-to-end. Cross-entropy loss or reinforcement learning-based objectives are used to optimize the model parameters. The training process is monitored and hyperparameters such as learning rate, batch size and regularization techniques are adjusted. Its performance is enhanced with suitable techniques for generating more diverse and contextually relevant captions using beam search or diverse decoding strategies. Finally, both Vision Transformer[17] and Generative AI techniques[18] are combined for generating captions, found that it is effective for wide range of images and leverages the strengths of both the methods.

2.3 Dataset

The Remote Sensing Image Captioning Dataset (RSICD) is specifically designed and developed for remote sensing image captioning. Unlike traditional image captioning datasets that focus only on everyday scenes, RSICD contains images typically has landscapes, urban areas or natural environments. In Figure 5, we depict the varieties of Remote Sensing Images for clarity.

RSICD consists of high-resolution remote sensing images that are captured from satellites or aerial platforms. Different scenes such as urban areas, landscapes, agricultural fields, water bodies, etc, are presented in Figures 5 (a)-(h). The diversity of

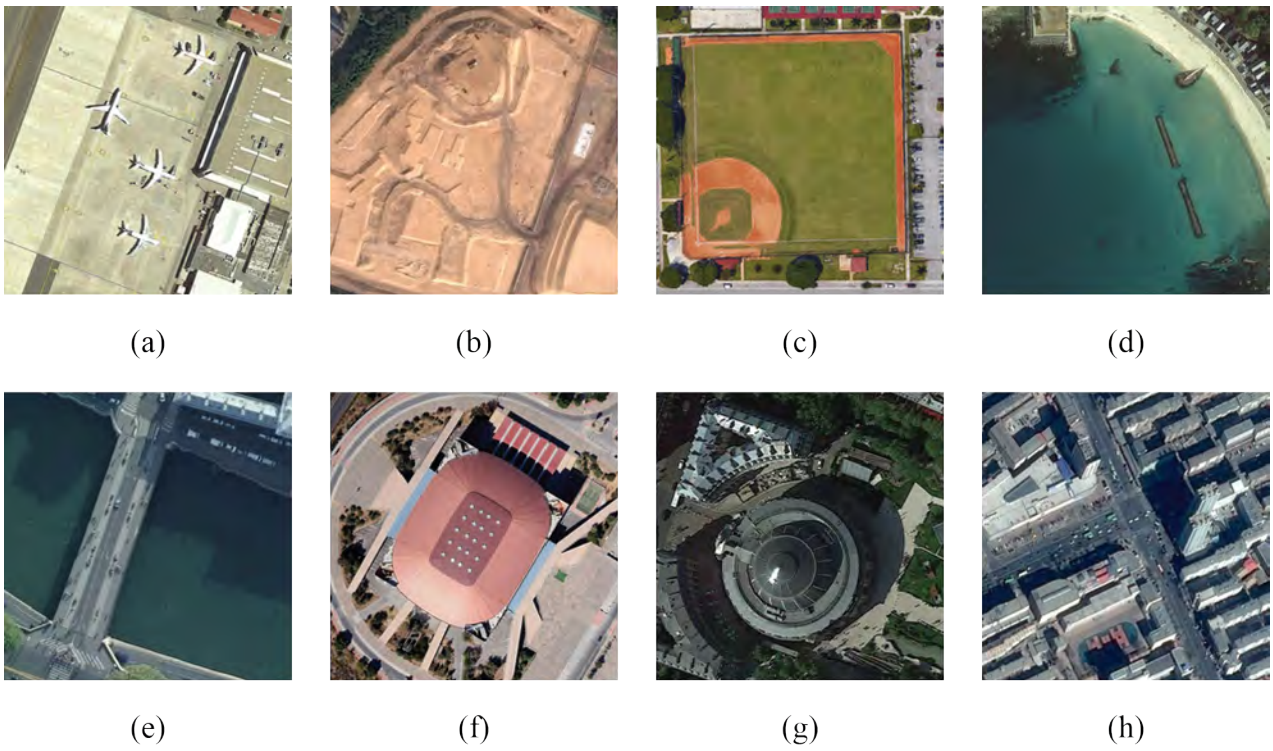

Fig 5. Remote Sensing Images from RSICD (a) Airport (b) Bare Land (c) Baseball Field (d) Beach (e) Bridge (f) Center (g) Church (h) Commercial

scenes and viewpoints in RSICD brings unique challenges for image captioning algorithms [19] compared to conventional image dataset.

The RSICD data contains a total of 10921 images. Among these, it is observed that 724 images are having 5 captions, 1495 images are captioned 4 times, 3 captions are extracted from 2182 images, 1667 images are captioned 2 times, and 4853 images are captioned only once. Generating accurate and descriptive captions from remote sensing images have challenges including handling different types of terrain and land cover, recognizing specific objects or landmarks and understanding complex spatial relationships.

The RSICD dataset is a resource for studying and advancing the state-of-the-art in remote sensing image captioning, with implications for a wide range of real-world applications such as military, agriculture, land covering, city planning etc., In Table 1, the classes detail of the RSICD is presented. It is observed that all classes are labelled on its name and containing images ranging from 240 to 1031 images. Below, sample of annotations for images are presented for clarity and understanding.

In Figure 6, Images belong to RSICD along with captions are presented. In Figure 6(b), the captions for Image ID:31 is presented. It is observed that there are five captions and each of them are different, in terms of their semantics. Similarly, Figure 6(d) presents the caption for Image ID:0 and found that all the captions are same. This is due to the fact that the content of these two images is different in terms of color, texture and objects.

2.4 Evaluation Metrices

We have used BLEU, METEOR, ROUGE and CIDEr as evaluation metrices for evaluating the performance of the methods under consideration.

2.4.1. Bilingual Evaluation Understudy (BLEU)

BLEU measures the similarity between the generated and reference captions using n-gram overlap and computes the precision of n-grams (typically up to 4-grams). Multiple reference captions are often available for each image. The BLEU score is ranging from 0 to 1, where higher score indicates the higher degree of matching between the reference and ground truth captions. The

Table 1. Sample image classification of RSICD

Scene	Number	Scene	Number
Airport	420	Mountain	340
Bare Land	310	Park	350
Baseball Field	276	School	300
Beach	400	Square	330
Bridge	459	Parking	390
Center	260	Playground	1031
Church	240	Pond	420
Commercial	350	Viaduct	420
Dense Residential	410	Port	389
Desert	300	Railway Station	260
Farm Land	370	Resort	290
Forest	250	River	410
Industrial	390	Sparse Residential	300
Meadow	280	Storage Tanks	390
Medium Residential	290	Stadium	290

(a)

1. a airport in the middle while with some sparse plants in it.
2. some planes neat arranged in the airport.
3. some light meadow besides the airport in it.
4. some light red and white buildings besides the airport in it.
5. many planes are parked in an airport near many buildpiece of meadow.

(b)

(c)

1. many planes are parked next to a long building in an airport.
2. many planes are parked next to a long building in an airport.
3. many planes are parked next to a long building in an airport.
4. many planes are parked next to a long building in an airport.
5. many planes are parked next to a long building in an airport.

(d)

Fig 6. Example of images in the RSICD dataset along with annotation. (a) Image ID:31 (b) Many Captions (c) Image ID:0 (d) One Captions

BLEU is calculated as given below in **Eq.1**.

$$BLEU = BP.\exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \tag{1}$$

In **Eq.1**, *BP* is Brevity Penalty, *N* is maximum number of *n*-gram, w_n is weight for each modified precision and p_n is modified precision for each gram.

2.4.2. Metric for Evaluation of Translation with Explicit Ordering (METEOR)

METEOR assesses the quality of generated captions by considering n-gram overlap, synonyms and paraphrase. It incorporates various linguistic resources and aligns them with the generated and reference captions. METEOR scores range from 0 to 1 with higher scores indicating better semantic similarity between the generated and reference captions. It is calculated as given below in **Eq.2**.

$$METEOR = \frac{(1-\gamma).Precision.Recall}{\gamma.Precision + (1-\gamma).Recall}(1-\beta.Penalty) \tag{2}$$

In **Eq.2** Precision is the accuracy of words in the candidate caption with respect to the reference caption, Recall measures the number of reference words are captured by the candidate caption, penalty is the penalty for unaligned words and β is the parameter that controls the balance between precision and recall.

2.4.3. Recall-Oriented Understudy for Gisting Evaluation (ROUGE)

ROUGE evaluates the quality of generated captions by measuring the overlap of n-grams and word sequences with the reference captions. It utilizes ROUGE-Longest Common Subsequence (ROUGE-L) metrics, and its scores is ranges from 0 to 1, where higher score indicates higher similarity between the generated and reference captions. It is calculated using **Eq.3**, where $ROUGH_L(i,j)$ is alignment score between C_i and R_i, where $C_i$is candidate word and R_i is reference word.

$$ROUGE_L(i,j) = \frac{min(C_i, R_j)}{R_j} \tag{3}$$

2.4.4. Consensus-based Image Description Evaluation (CIDEr)

CIDEr captures the consensus among human annotators by considering multiple reference captions for each image. It computes a similarity score based on the cosine similarity between generated n-gram vectors and reference captions. Its scores are not bounded and can range from 0 to a large value, with higher scores indicating higher similarity based on human judgments. The CIDEr is calculated using **Eq. 4**, where *r* and *c* represents reference and candidate captions.

$$CIDEr(r,c) = \frac{\sum_{i=1}^{N} cosine\_similarity(r_i, c)}{N} \tag{4}$$

These evaluation metrics provide quantitative measures of the performance of caption generation models on the RSICD dataset. By using multiple metrics, researchers can gain more comprehensive understanding of the strengths and weaknesses of different models and approaches in remote sensing image captioning.

3 Overview

Result and Findings:

The comparative analysis demonstrates that the Channel Attention Mechanism and Generative AI can generate captions from RSI. We have observed the differences in their performance in terms of qualitative aspects. The channel attention mechanism tends to produce captions that are more relevant and semantically accurate. In contrast, the Generative AI models generate captions that are more diverse and imaginative. However, both of these models fail to quantify the number of objects present in RSI and linking the coordinates of the objects based on the analysis.

The caption is generated by both Channel Attention Mechanism and Generative AI model is presented in Table 2. However, the quantification objects is missing from captions generated by both of the methods. The advantages and disadvantages of these schemes are summarized in Table 3.

The Table 4 presents the performance of channel attention mechanism and Generative AI is generating captions of RSI. We have used various well-known performance metrics discussed in Section 2.4. The performance of both of the methods are good

Table 2. Captions generated by Channel Attention Mechanism and Generative AI

Image Name	Input Image	Channel Attention Mechanism *(Output Caption)*	Generative AI *(Output Caption)*
airport_12.jpg Img ID: 23		Some planes are parked near building in an air-port	A large jetliner sitting on top of an airport tarmac
airport_27.jpg Img ID: 189		Some planes are in an airport near some green meadows	An aerial view of a large building with a clock on it
airport_60.jpg Img ID: 319		The airport was built on the tarmac has some buildings and runways	An aerial view of a city at night

Table 3. Advantages and Drawbacks present in the Channel Attention Mechanism and Generative AI

Method	Channel Attention Mechanism	Generative AI
Advantages	• High potential for enhancement with new deep learning methods. • No asyntactic and sematic errors. • Provide the beast performance to align the image elements with words. • Each and every word generation is interpretable.	• Generated well defined sentences that doesn't in the corpus. • Automatically fine-tune the model. • May generate the sentence without grammatical errors. • It is directly optimizing the evolution metrices at test time.
Disadvantages	• False matches between RSI features and text in small size database. • Need very high quality for word annotation. • Poor performance in object quantification and identifying coordinates is not good.	• Unable to generate flexible and the versatile sentence (overfitting). • Highest space and time complexity. • Unable to detect and identify objects effectively. Its quantification of objects along with their coordinates is not good.

Table 4. Performance of Evaluation metrics

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE_L	CIDEr
Channel Attention Mechanism	0.6768	0.6537	0.5610	0.4932	0.4009	0.6541	1.6679
Generative AI	0.7701	0.7366	0.7395	0.7684	0.7028	0.7416	1.7340

with CIDEr and Generative AI is having edge over Channel Attention Mechanism. Similarly, as per METEOR and ROUGE_L also the performance of GenAI is encouraging. The BLEU variants also show that the GenAI is doing well in generating captions of RSI.

The experiment is further consolidated by measuring various metrices, like Precision, Recall, F1-Score, F2-Score and Accuracy. The Precision measures how many generated words are relevant, while Recall measures how much of the reference caption's content is captured. The F1 score balances precision and recall, providing a harmonic mean, whereas the F2 Score emphasizes recall more than precision. Accuracy measures the overall correctness of the generated captions. These metrics prove to be especially valuable when evaluating the outputs generated by Generative AI and Channel Attention Mechanism. A systematic evaluation of effectiveness each method is conducted, focusing on aspects such as word accuracy, relevance, and the overall quality of the generated image captions. Table 5 presents the result in terms of Precision, Recall, F1-Scores, F2-Scores

and Accuracy: for few images from each classification and its captions generated when Channel Attention Mechanism and the Generative AI are used separately. As presented in Table 1, the RSICD contains 30 classes having different count of images we have used 10% of images from each category of the dataset and these images are randomly selected from the category. The captions are generated for each image by Channel Attention Mechanism and Generative AI (Table 5 and 6).

Table 5. Performance of Channel Attention Mechanism on different categories of RSICD

Classification	No. of Images	Precision	Recall	F1	F2	Accuracy
Airport	42	0.78	0.70	0.74	0.72	0.82
Bare land	31	0.76	0.68	0.72	0.70	0.81
Baseball Field	27	0.73	0.66	0.69	0.68	0.79
Beach	40	0.74	0.67	0.70	0.69	0.80
Bridge	45	0.75	0.68	0.71	0.70	0.82
Center	26	0.80	0.72	0.76	0.74	0.85
Church	24	0.77	0.70	0.73	0.72	0.83
Commercial	35	0.72	0.64	0.68	0.66	0.79
Dense Residential	41	0.74	0.67	0.70	0.69	0.80
Desert	30	0.78	0.70	0.74	0.73	0.84
Farmland	37	0.76	0.68	0.72	0.71	0.83
Forest	25	0.73	0.66	0.69	0.68	0.80
Industrial	39	0.79	0.72	0.75	0.74	0.85
Meadow	28	0.75	0.69	0.72	0.71	0.82
Medium Residential	29	0.74	0.67	0.70	0.69	0.80
Mountain	34	0.76	0.70	0.73	0.72	0.83
Park	35	0.75	0.68	0.71	0.70	0.82
School	30	0.77	0.70	0.73	0.72	0.84
Square	33	0.73	0.66	0.69	0.68	0.80
Parking	39	0.74	0.67	0.70	0.69	0.81
Playground	103	0.75	0.68	0.71	0.70	0.82
Pond	42	0.76	0.69	0.72	0.71	0.83
Viaduct	42	0.74	0.66	0.70	0.69	0.81
Port	38	0.79	0.72	0.75	0.74	0.85
Railway Station	26	0.77	0.70	0.73	0.72	0.84
Resort	29	0.75	0.68	0.71	0.70	0.81
River	41	0.74	0.66	0.70	0.69	0.80
Sparse Residential	30	0.68	0.60	0.64	0.62	0.78
Storage Tanks	39	0.71	0.66	0.68	0.67	0.80
Stadium	29	0.72	0.65	0.68	0.67	0.80
	1089	**0.75**	**0.67**	**0.71**	**0.69**	**0.81**

As per Table 5, the evaluation is presented for each category of the dataset. The total number of images is 1089 and average of Precision, Recall, F1-Score, F2-Score and Accuracy are 0.75, 0.67, 0.71, 0.69, 0.81 respectively.

Similar to above, the performance of Generative AI concept is evaluated and is presented in Table 6. The total number of images is 1089 and average of Precision, Recall, F1-Score, F2-Score and Accuracy are 0.90, 0.86, 0.88, 0.87, 0.92 respectively. Based on the result, it is observed that the performance of Generative AI is encouraging, and this is due to its capability to work with millions of weights to generate the captions. However, it has failed in generating information of object counts and its coordinate information.

4 Conclusion

This study compares two main methods for Remote Sensing Image Captioning (RSIC): Channel Attention Mechanism (CAM) and Generative AI (GenAI) models based on ChatGPT, using the RSICD dataset. Results show Generative AI outperforms CAM

Table 6. Performance of Generative AI on different categories of RSICD

Classification	No. of Images	Precision	Recall	F1	F2	Accuracy
Airport	42	0.91	0.88	0.89	0.88	0.94
Bare land	31	0.89	0.85	0.87	0.86	0.92
Baseball Field	27	0.90	0.86	0.88	0.87	0.91
Beach	40	0.92	0.89	0.91	0.90	0.95
Bridge	45	0.88	0.84	0.86	0.85	0.91
Center	26	0.93	0.90	0.92	0.91	0.96
Church	24	0.90	0.87	0.88	0.87	0.92
Commercial	35	0.89	0.85	0.87	0.86	0.91
Dense Residential	41	0.91	0.88	0.89	0.88	0.93
Desert	30	0.92	0.88	0.90	0.89	0.95
Farmland	37	0.90	0.87	0.88	0.88	0.92
Forest	25	0.88	0.85	0.86	0.85	0.91
Industrial	39	0.92	0.89	0.90	0.90	0.94
Meadow	28	0.91	0.87	0.89	0.88	0.93
Medium Residential	29	0.89	0.86	0.87	0.86	0.92
Mountain	34	0.93	0.89	0.91	0.90	0.96
Park	35	0.90	0.86	0.88	0.87	0.93
School	30	0.92	0.88	0.90	0.89	0.94
Square	33	0.90	0.86	0.88	0.87	0.92
Parking	39	0.89	0.86	0.87	0.86	0.91
Playground	103	0.91	0.87	0.89	0.88	0.93
Pond	42	0.92	0.88	0.90	0.89	0.94
Viaduct	42	0.90	0.86	0.88	0.87	0.92
Port	38	0.92	0.89	0.91	0.90	0.94
Railway Station	26	0.91	0.88	0.89	0.89	0.93
Resort	29	0.88	0.84	0.86	0.85	0.90
River	41	0.89	0.86	0.87	0.86	0.92
Sparse Residential	30	0.87	0.82	0.84	0.83	0.89
Storage Tanks	39	0.89	0.86	0.87	0.87	0.92
Stadium	29	0.89	0.86	0.87	0.86	0.91
	1089	**0.90**	**0.86**	**0.88**	**0.87**	**0.92**

across key metrics, including Precision (0.90 vs. 0.75), Recall (0.86 vs. 0.67), F1-Score (0.88 vs. 0.71), F2-Score (0.87 vs. 0.69), and Accuracy (0.92 vs. 0.81). BLEU-4 scores further demonstrate GenAI's superiority in producing semantically rich captions (0.7684 vs. 0.4932), alongside substantially better METEOR (0.7028 vs. 0.4009), ROUGE-L (0.7416 vs. 0.6541), and CIDEr (1.7340 vs. 1.6679) scores despite CAM has better interpretability and aligns image features well with words. However, both methods struggle to quantify the objects accurately and specify their coordinates within images, which limits their usefulness in precise applications. CAM's performance can be affected by dataset size and annotation quality, while GenAI requires high computational resources. Future work should focus on integrating object detection and spatial reasoning into captioning models to address these issues. Building larger, well-annotated datasets and developing remote sensing-specific transformer models could boost performance. Combining CAM's strength in feature extraction with GenAI's powerful language capabilities may lead to better results. Enhancing interpretability and reducing the complexity of GenAI models also remain important goals. GenAI excels in RSIC captioning due to large-scale pretraining on diverse datasets, enriching semantic understanding and sentence variety. Its transformer-based architecture (GPT + Vision Transformer) captures global context better than CNN-RNN, ensuring coherence. Auto-tokenization and multimodal fusion enable smooth integration of visual and textual data. Overall, this research highlights the potential and challenges of applying advanced GenAI models to RSIC, encouraging further

studies in this interdisciplinary field to improve automated image understanding for practical uses in monitoring, planning and disaster management.

References

1) Wei T, Yuan W, Luo J, Zhang W, Lu L. VLCA: vision-language aligning model with cross-modal attention for bilingual remote sensing image captioning. *Journal of Systems Engineering and Electronics*. 2023;34(1):9–1. http://doi.org/10.23919/JSEE.2023.000035.
2) Sebaq A, ElHelw M. RSDiff: Remote Sensing Image Generation from Text Using Diffusion Model. *Neural Computing & Applications*. 2024;36:23103–23111. https://doi.org/10.1007/s00521-024-10363-3.
3) Ye X, Wang S, Gu Y, Wang J, Wang R, Hou B. A Joint-Training Two-Stage Method for Remote Sensing Image Captioning. *IEEE Transactions on Geoscience and Remote Sensing*. 2022;60:1–16. https://doi.org/10.1109/TGRS.2022.3224244.
4) Li Y, Tao C, Liu M, Zhang X, Wang G, Zhang T, et al. Feature refinement and rethinking attention for remote sensing image captioning. *Scientific Reports*. 2025;15:8742. https://doi.org/10.1038/s41598-025-93125-y.
5) Cheng Q, Huang H, Xu Y, Zhou Y, Li H, Wang Z. NWPU-Captions Dataset and MLCA-Net for Remote Sensing Image Captioning. *IEEE Transactions on Geoscience and Remote Sensing*. 2022;60:1–19. https://doi.org/10.1109/TGRS.2022.3201474.
6) Feng J, Wang H, Dong S. A question-type guided and progressive self-attention network for remote sensing visual question answering. *Earth Science Information*. 2025;18:409. https://doi.org/10.1007/s12145-025-01917-7.
7) 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*. 2022;60:1–20. https://doi.org/10.1109/TGRS.2022.3218921.
8) Guo H, Zhou X, Yang P. Feature Enhancement Based Oriented Object Detection in Remote Sensing Images. *Neural Processing Letters*. 2024;56:244. https://doi.org/10.1007/s11063-024-11699-6.
9) Chen Z, Wang J, Ma A, Zhong Y. TypeFormer: Multiscale Transformer with Type Controller for Remote Sensing Image Caption. *IEEE Geoscience and Remote Sensing Letters*. 2022;19:1–5. https://doi.org/10.1109/LGRS.2022.3192062.
10) Sumbul G, Nayak S, Demir B. SD-RSIC: Summarization-Driven Deep Remote Sensing Image Captioning. *IEEE Transactions on Geoscience and Remote Sensing*. 2021;59(8):6922–6934. https://doi.org/10.1109/TGRS.2020.3031111.
11) Wang B, Zheng X, Qu B, Lu X. Retrieval Topic Recurrent Memory Network for Remote Sensing Image Captioning. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*. 2020;13:256–270. https://doi.org/10.1109/JSTARS.2019.2959208.
12) Zhao R, Shi Z, Zou Z. High-Resolution Remote Sensing Image Captioning Based on Structured Attention. *IEEE Transactions on Geoscience and Remote Sensing*. 2022;60:1–14. https://doi.org/10.1109/TGRS.2021.3070383.
13) Wang Q, Huang W, Zhang X, Li X. Word–Sentence Framework for Remote Sensing Image Captioning. *IEEE Transactions on Geoscience and Remote Sensing*. 2021;59(12):10532–10543. https://doi.org/10.1109/TGRS.2020.3044054.
14) Hoxha G, Melgani F, Demir B. Toward Remote Sensing Image Retrieval Under a Deep Image Captioning Perspective. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*. 2020;13:4462–4475. https://doi.org/10.1109/JSTARS.2020.3013818.
15) Huang W, Wang Q, Li X. Denoising-Based Multiscale Feature Fusion for Remote Sensing Image Captioning. *IEEE Geoscience and Remote Sensing Letters*. 2021;18(3):436–440. https://doi.org/10.1109/LGRS.2020.2980933.
16) Ueda A, Yang W, Sugiura K. Switching Text-Based Image Encoders for Captioning Images with Text. *IEEE Access*. 2023;11:55706–55715. https://doi.org/10.1109/ACCESS.2023.3282444.
17) Yuan Z, Li X, Wang Q. Exploring Multi-Level Attention and Semantic Relationship for Remote Sensing Image Captioning. *IEEE Access*. 2020;8:2608–2620. https://doi.org/10.1109/ACCESS.2019.2962195.
18) Zhang K, Li P, Wang J. A Review of Deep Learning-Based Remote Sensing Image Caption: Methods, Models, Comparisons and Future Directions. *Remote Sensing*. 2024;16(21):4113. https://doi.org/10.3390/rs16214113.
19) Ricci R, Bazi Y, Melgani F. Machine-to-Machine Visual Dialoguing with ChatGPT for Enriched Textual Image Description. *Remote Sensing*. 2024;16(3):441. https://doi.org/10.3390/rs16030441.