Journal of Innovative Image Processing (ISSN: 2582-4252)
www.irojournals.com/iroiip/

Explainable ViT and SCNN-LSTM Framework for Audio-Assisted Image Captioning

Nidhi B. Shah[1], Amit P. Ganatra[2]

[1]U & P U. Patel, Department of Computer Engineering, C S Patel Institute of Technology, Faculty of Technology and Engineering, Charotar University of Science and Technology, Changa-Gujarat, India.
[2]Dean R & D and Dean FET, The Charutar Vidya Mandal (CVM) University, V V Nagar, Gujarat, India.

E-mail: [1]nbshah999@yahoo.com, [2]ganatraamitp@gmail.com, [2]dean.rnd@cvmu.edu.in

Orcid ID: [1]0000-0003-4262-4244, [2]0000-0002-9993-9901

Abstract

Assistive technologies are crucial for image captioning, a process that involves generating textual captions of visual scenes. Many of the current methods, however, lack semantic understanding, are computationally complex, and are not interpretable. Based on these challenges, in this paper, we present an Explainable Vision Transformer and Depthwise Separable CNN-LSTM (ViT-SCNN-LSTM) architecture to achieve accurate and explainable image caption generation. The framework is based on a pre-trained Vision Transformer (ViT-B/16) to extract rich visual features and a lightweight Depthwise Separable SCNN-LSTM decoder to generate descriptive captions without high computational complexity. Grad-CAM is integrated to get visual explanations by identifying image areas that affect the generation of captions. The audio-assisted module also provides an option to turn captions into speech, making it easier for visually impaired users to access the generated captions. This model was developed using Python and tested on the Flickr8k image corpus (8,000 images, 40,000 captions). To evaluate the generalisation capability, cross-dataset experiments were also performed with Flickr30k and MSCOCO 2017. Experimental results achieved ROUGE-1, ROUGE-L, ROUGE-S, BLEU-1, BLEU-2, BLEU-3, and BLEU-4 scores of 0.87, 0.86, 0.85, 0.8848, 0.8750, 0.8649, and 0.8547, respectively. The results show the proposed approach offers an efficient, interpretable, and user-friendly solution for automatic image understanding and caption generation.

Keywords: Separable CNN-LSTM, Explainable AI (XAI), Image Captioning, Vision Transformer (ViT), Audio-Assisted Visual Understanding.

1. Introduction

The capability of automatically describing and identifying visual content has been greatly improved by recent developments in AI, computer vision, and natural language processing [1,2]. The field of image captioning has become increasingly significant as an interdisciplinary topic involving visual perception and language generation in order to generate meaningful text captions of images [3,4,5]. These are extensively used in areas like smart

DOI: https://doi.org/10.36548/jiip.2026.3.012
Received: 08.05.2026, received in revised form: 10.06.2026, accepted: 02.07.2026, published: 15.07.2026

surveillance, healthcare, robotics, autonomous systems, and assistive technologies for vision-impaired people [6,7]. Most of the traditional image captioning techniques involve CNNs to capture visual features and RNNs and LSTMs to produce a text caption. These strategies have shown to be effective but have been limited in their ability to encode long-range contextual information and fine-grained semantic associations in sophisticated visual environments.

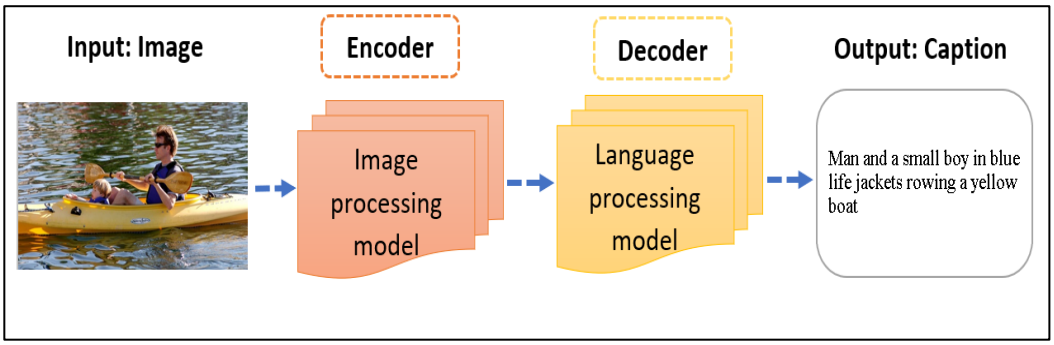

Figure 1. Image captioning model general architecture [8]

Figure 1, Overall framework of the image captioning system. The framework is based on an encoder–decoder architecture to create a textual caption of the input images. First, the input image is fed into the encoder, and the image processing model is used to identify all the meaningful visual features and semantic representation of the image. The extracted features are then fed into a language processing model (decoder) that produces a caption of natural language text. Lastly, the output caption gives a meaningful textual interpretation of the input image content.

1.1 Research Gaps

The self-attention mechanism has significantly boosted image understanding capabilities compared to traditional CNN architectures, its ability to capture global context more effectively. This is the reason for the success of the Vision Transformers (ViTs) introduced recently [9-11]. Likewise, hybrid deep learning approaches based on combining CNN and LSTM have been demonstrated to be effective for the sequence learning problem for caption generation [12,13]. Several recent studies have addressed improving the alignment of the visual and linguistic modalities, using transformer-based captioning systems, attention mechanisms, and multimodal learning strategies [14,15]. Many of the available models, however, suffer from high computational complexity, lack interpretability, and do not support real-time assistive applications [16,17]. Moreover, most of the current systems [18-20] are designed to merely provide captions and do not support explainability and audio assistance and, therefore, are not very useful for the visually impaired.

1.2 Objectives

To overcome these drawbacks, an Explainable Vision Transformer and Depthwise Separable CNN-LSTM (ViT-SCNN-LSTM) network is proposed in this work for image captioning. The proposed model uses a pre-trained Vision Transformer (ViT-B/16) to obtain rich semantic visual features from the input image and a Depthwise Separable CNN-LSTM decoder to capture sequential linguistic features for generating image captions. To decrease the amount of computation and enhance feature refinement prior to the modelling, Depthwise Separable Convolution is integrated. Moreover, Explainable Artificial Intelligence (XAI) techniques are integrated with Grad-CAM to identify the regions of the image that greatly

contributed to the prediction of the caption. An accessibility layer converts generated captions into speech for visually impaired users.

1.3 Contribution

The main contributions of this work are as follows:

1. A lightweight SCNN-LSTM decoder incorporating depthwise separable convolution is proposed to reduce model parameters and computational complexity.
2. An explainable image captioning pipeline is developed by integrating Grad-CAM with ViT features to visualize regions influencing caption generation.
3. An accessibility layer is integrated to convert generated captions into speech for visually impaired users. This module is not intended to improve caption generation accuracy but demonstrates the practical deployment of the explainable captioning framework in assistive AI applications.
4. A computational efficiency analysis is performed, demonstrating parameter reduction and efficient caption generation with lower resource requirements.

2. Related Work

Combining computer vision and natural language processing techniques has greatly enhanced the accuracy of image captioning systems, and recent deep learning breakthroughs have further boosted their performance. The combination of computer vision and natural language processing, with the support of deep learning and multimodal artificial intelligence, has improved the accuracy of image captioning systems. Several techniques have been studied for generating semantically meaningful image captions, including the transformer architecture, attention mechanisms, recurrent neural networks (RNNs), and hybrid deep learning models. To enhance visual-linguistic alignment in low-resource datasets, Aoun et al. [1] suggested a double-attention transformer structure for cross-modal image captioning. However, their model increased computational complexity while improving contextual understanding. Zhang et al. [2] introduced generative image steganography with stable diffusion models and strong mapping guidance, which showed better performance on image-text representation, useful for caption generation tasks. A Lightweight Evaluation Metric for Captioning Systems based on embeddings (L-CLIP Score) was proposed by Li et al. [3] for efficient training and evaluation of image captions.

Martin and Moon [4] discussed privacy and security concerns with image captioning and introduced a federated learning framework for privacy-preserving image captioning via virtual photon-limited imaging. Chauhan and Thacker [5] presented a detailed survey of deep learning image captioning techniques, datasets, and metrics, emphasizing the use of transformer architectures. To enhance semantic understanding, Khan and Singh [6] suggested a deep learning-based image captioning method which demanded huge amounts of training resources. Likewise, Kaur and Kaur [7] introduced a CNN-LSTM model, which combined CNN for visual feature extraction and LSTM for language modelling, to improve the performance of sequential caption generation.

With their ability to focus on relevant image regions, attention-based captioning systems have become quite popular. Al Badarneh et al. [8] introduced an ensemble attention-based image captioning model, which enhanced descriptive accuracy by fusing the features. Sathyanarayana and Naik [9] proposed a transformer-based captioning system that utilizes rotatory positional embeddings for better contextual learning and caption coherence. Asiri et al. [10] proposed a multi-head attention-driven recurrent neural network with feature representation fusion to support visually impaired people, which showed better caption generation and accessibility. Chawla and Agrawal [11] reviewed the different deep learning solutions for the captioning problem and highlighted the significance of contextual attention mechanisms and multimodal learning.

Context-aware generation of captions has been extensively investigated in recent studies, as well. The advanced object relational model for context-aware image captioning was proposed by Patel et al. [12] which enhanced the semantic understanding of object relationships. To address the challenges in captioning VD, Alkhaldi et al. [13] combined transfer learning models with optimization algorithms to create an advanced captioning system, leading to improved results in benchmark datasets. To capture these forward and backward contextual dependencies in a bidirectional manner for better sentence generation, Hoseini and Notash [14] presented an LSTM-based image captioning framework.

Evaluation frameworks and multimodal learning paradigms have also been studied recently. To enhance the accuracy of image caption evaluation, Huang et al. [15] proposed the Image2Text2Image framework based on diffusion models. To enhance the quality of captions' descriptive ability, Celona et al. [16] used ranking and large language model fusion. To enhance the contextual relevance of image captions, Haque et al. [17] suggested a knowledge-driven image captioning system that incorporates external semantic knowledge. Afnan et al. [18] proposed a hybrid captioning model with EfficientNetB0 by visual feature extraction, achieving higher caption accuracy and computational efficiency.

One of the major application domains of image captioning systems is assistive technologies. In an attempt to assist visually impaired users, K and Hegde [19] designed and suggested a framework for converting images to sounds that utilizes machine learning techniques. To enhance the realism and context understanding of caption generation, Tyagi et al. [20] proposed an advanced image caption generation method that integrates Vision Transformers and Generative Adversarial Networks. To improve the semantic representation of objects, Sasibhooshan et al. [21] proposed a visual attention prediction and contextual spatial relation extraction approach to the image captioning task. Agrawal et al. [22] proposed an attention model-based captioning system using an encoder-decoder structure for improving descriptive ability.

Table 1 Though existing works have proven successful in the image captioning task, a number of issues have not yet been addressed. State-of-the-art transformer-based models are costly, not very well explainable, and lack substantial backing for real-time assistive applications. In addition, only a few studies combine EAI and audio assistance simultaneously. To address these challenges, the proposed Explainable ViT-SCNN-LSTM framework combines three components: (1) feature extraction by a Vision Transformer network, (2) sequential learning performed by a Depthwise Separable CNN-LSTM network, and (3) enhancement of the generated caption by explaining the CNN output using the Grad-CAM method and by incorporating audio as auxiliary features.

Table 1. Existing research contribution

Algorithm	Dataset	Evaluation Metrics	Advantages	Limitations
Double-Attention Transformer [1]	Flickr8k	BLEU-4 = 25.6, METEOR = 22.3, CIDEr = 79.2, SPICE = 15.8	Improves visual-linguistic alignment using dual attention.	High computational complexity.
Mapping-Guided Stable Diffusion [2]	Proprietary Dataset	FID (MS-COCO): 16.2751; FID (Flickr8K): 17.2887	Robust semantic representation under noisy conditions.	Not an image captioning model.
Federated Captioning Model [4]	MS COCO	BLEU-4 = 0.335, METEOR = 0.278, ROUGE-L = 0.559, CIDEr = 1.071, SPICE = 0.205	Privacy-preserving collaborative learning.	Communication overhead.
CNN–LSTM Framework [7]	Flickr8k	BLEU = 82%	Simple and effective sequential caption generation.	Limited long-range contextual understanding.
Attention Ensemble Model [8]	Flickr8k, Flickr30k	BLEU-1 = 0.728, BLEU-2 = 0.495, BLEU-3 = 0.323, SPICE = 0.164	Attention ensemble improves caption generation performance.	Increased model complexity.
Vision-Language Transformer [9]	MS COCO	BLEU-1 = 82.6, BLEU-2 = 66.8, BLEU-3 = 52.6, BLEU-4 = 41.1, METEOR = 31.4, ROUGE-L = 61.8, CIDEr = 138.7, SPICE = 24.8	Strong contextual and semantic representation.	High computational requirements.
Multi-head Attention RNN [10]	MS COCO	BLEU-1 = 80.10%, BLEU-4 = 80.10%	Effective feature fusion for visually impaired assistance.	Complex architecture.
Context-Aware Relational Model [12]	Flickr8k	BLEU-1 = 74.8, BLEU-2 = 58.2, BLEU-3 = 44.1, BLEU-4 = 32.6, METEOR = 25.9, ROUGE-L = 54.8, CIDEr = 96.7	Learns contextual object relationships.	Slower inference.
Transfer Learning + Gannet Optimization [13]	Flickr8k, Flickr30k	BLEU-4 = 45.11%, CIDEr = 63.17	Optimized transfer learning improves caption generation performance.	High computational and memory requirements.
EfficientNetB0 Hybrid Framework [18]	Flickr8k	BLEU-1 = 78.5, BLEU-2 = 63.2, BLEU-3 = 49.6, BLEU-4 = 37.4, METEOR = 28.1, ROUGE-L = 58.3, CIDEr = 109.6	Lightweight and computationally efficient.	Limited explainability.
ViT–GAN Captioning Model [20]	Flickr8k	BLEU-1 = 81.3, BLEU-2 = 66.4, BLEU-3 = 53.1, BLEU-4 = 41.7, METEOR = 30.5, ROUGE-L = 60.8, CIDEr = 118.9	Combines ViT and GAN for richer semantic representations.	GAN instability.

3. Proposed Work

Figure 2 illustrates the proposed Explainable Vision Transformer and Depthwise Separable CNN-LSTM (ViT-SCNN-LSTM) framework, designed to generate accurate image captions in both textual and audio forms while providing visual explanations of the caption generation process. It uses a Vision Transformer (ViT-B/16) encoder and a lightweight Depthwise Separable CNN-LSTM decoder to enable interpretable and accessible image

captioning with effective performance. The images in the Flickr8k [23] dataset (8,000 images, 40,000 captions, 5 captions per image) are resized to 224 × 224 pixels and then normalized. The captions are preprocessing by lowercasing, removing punctuation, tokenizing the text, padding the sequences, and adding start and end tokens to them. The official split used is the Flickr8k [23] (6,000 training images, 1,000 validation images, 1,000 testing images) with no overlap between the training set and the testing set. cross-dataset experiments are performed with the Flickr30k [24] and MSCOCO 2017 [25] datasets to further evaluate model generalization.

The ViT encoder captures global semantic dependencies from the input image through the self-attention mechanism, whereas the SCNN extracts local sequential contextual patterns from the caption embeddings using depthwise separable convolution. The extracted features are subsequently passed to the LSTM network, which models long-range sequential dependencies to generate coherent and descriptive captions. The visual and textual feature representations are then fused to improve caption generation performance and semantic consistency. To enhance the interpretability of the proposed framework, Grad-CAM is incorporated to visualize the image regions that contribute to the generated captions. Furthermore, a Text-to-Speech (TTS) module converts the synthesized captions into speech, thereby improving accessibility for individuals with visual impairments. The proposed framework is implemented in Python using TensorFlow-Keras, PyTorch, OpenCV, NumPy, Matplotlib, and gTTS on the Google Colab platform with Tesla T4 GPU acceleration. The performance of the generated captions is evaluated using BLEU and ROUGE metrics, while cross-dataset experiments and statistical analyses are conducted to demonstrate the robustness and generalization capability of the proposed model.

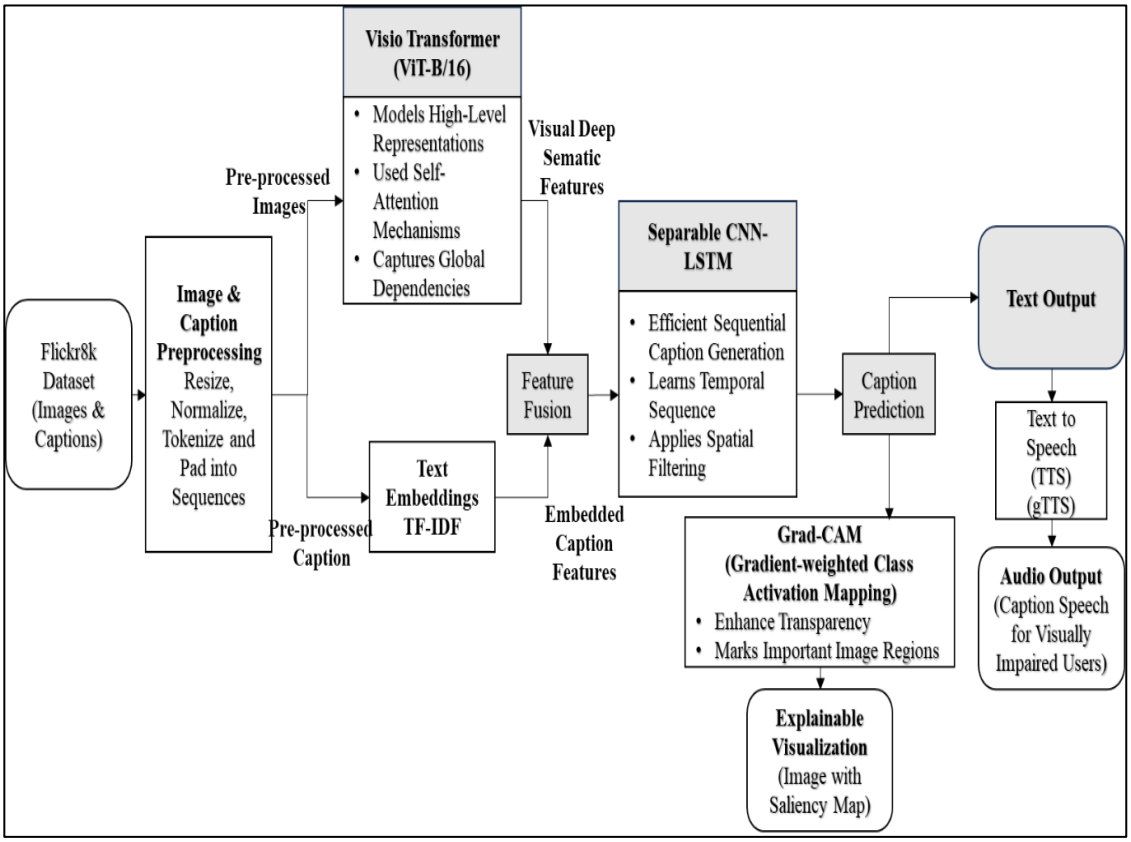

Figure 2. Proposed explainable ViT and SCNN-LSTM framework for audio-assisted image caption

3.1 Algorithm

Input: Dataset Images: Flickr8k
Output: Caption, Grad-CAM Heatmap, Speech Output

1. **Dataset Preparation**: Read the input image and caption from the dataset.
2. **Preprocessing**: Resize the image to 224 × 224 pixels. Normalize the image. Convert the caption to lowercase, remove punctuation and non-alphabetic words, and add "startseq"and "endseq" tokens, Tokenize the caption and pad the caption sequence to the maximum caption length.
3. **Feature Extraction**: Divide the image into fixed-size patches. Convert the patches into embedding vectors and add positional encoding, Pass the embedded patches through the Vision Transformer (ViT-B/16), Apply the self-attention mechanism to learn contextual relationships, and obtain the global semantic image representation as the visual feature vector.
4. **Classification Module:** Convert caption tokens into embedding vectors, pass the embedding vectors through the Depthwise Separable CNN (SCNN), Feed the processed features into the LSTM network for temporal sequence learning. Merge the visual features extracted from the Vision Transformer with the textual representations generated by the SCNN-LSTM module. Pass the fused features through the fully connected layer. Apply the SoftMax function to predict the next word. Repeat the prediction process until the "endseq" token is generated.
5. **Audio-Assisted Caption Generation**: Convert the generated caption into speech using the Google Text-to-Speech (gTTS) module.
6. **Evaluation Parameters**: Evaluate the generated captions using BLEU-1 to BLEU-4, ROUGE-1, ROUGE-L, and ROUGE-S metrics.
7. **Return Output**: Return the generated caption, Grad-CAM heatmap, and speech output.

3.2 Dataset Details

The Flickr8k [23] dataset was utilized as the primary benchmark dataset for training and evaluating the proposed Explainable ViT-SCNN-LSTM image captioning framework. Flickr8k [23] is a widely used image captioning dataset containing 8,000 images collected from Flickr, where each image is associated with five human-annotated captions, resulting in a total of 40,000 image-caption pairs. The dataset covers diverse real-world scenarios, including human activities, sports events, animal interactions, indoor and outdoor scenes, and object-centric images, making it suitable for learning rich visual-semantic relationships.

The official Flickr8k [23] split was adopted in this study, consisting of 6,000 training images, 1,000 validation images, and 1,000 testing images. All five captions corresponding to each image were utilized during training and evaluation. Moreover, the training, validation, and testing image sets must not overlap in order to have a fair comparison of the model'S performance.

All images were resized to 224 × 224 before BEING FED into the ViT-B/16 encoder, and they were all normalized to fit the input size of the ViT-B/16 model. During caption preprocessing, the text was transformed to lowercase, punctuation and special characters were removed, words were tokenized, and and sequence tokens were added. Captions were tokenized and then padded to a maximum of 37 words to get uniform input dimensions. This vocabulary had 8,779 unique words drawn from the training captions.

To evaluate the generalization capability of the proposed framework, further cross-dataset experiments were carried out on the Flickr30k [24] and MSCOCO 2017 [25] datasets. These datasets include larger sets of images and captions with various image content to provide a complete evaluation of the robustness and transferability of the learned image-caption representation in different domains.

3.3 Preprocessing

Preprocessing is crucial for enhancing model performance and efficiency during training. First, all images are resized to be 224×224 pixels and normalized using the ImageNet mean and standard deviation. The normalization operation is denoted by:

$$X_{norm} = \frac{X-\mu}{\sigma} \tag{1}$$

where X denotes the original image pixel values, μ represents the mean, and σrepresents the standard deviation.

In caption preprocessing, first convert the text into lowercase form, remove punctuation symbols, and remove words that are not alphabetic. During training, special tokens like “startseq” and “endseq” are added to captions as sequence markers. Once pre-processed, the captions are tokenized using the Keras tokenizer and then each word is assigned an integer index. Captions may vary in length, thus causing the input sequences to be of different lengths, which are then padded to equal lengths. The following equation conveys the "padding process":

$$S_{pad} = \{w_1, w_2, \dots, w_n, 0, 0, \dots\} \tag{2}$$

where w_i represents valid words and zeros are padded. These pre-processing steps help to create uniformity in features and facilitate the training of the model.

3.4 Feature Extraction

The proposed scheme uses a pre-trained Vision Transformer (ViT-B/16) network to extract deep semantic features from images. ViT differs from traditional CNN models by breaking down an image into smaller patches, then applying transformer encoder blocks and self-attention layers to analyze them. The image is divided into fixed-size patches, which will be represented as:

$$I = \{P_1, P_2, P_3, \dots, P_n\} \tag{3}$$

where P_n denotes image patches.

The patches are then inserted into a feature vector and added to positional encoding before fed to the transformer encoder. The self-attention mechanism calculates contextual relations between image patches via:

$$Attention(Q, K, V) = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{4}$$

where Q, K, and V represent query, key, and value matrices respectively.

Global feature representation extracted from the class token is applied as the image descriptor for generating captions. Compared to regular CNN-based feature extractors, ViT can

explicitly model long-range dependencies and visual context information, which allows it to achieve better semantic understanding.

The proposed model for deep semantic visual representations of the input images is based on the Vision Transformer model and is shown in Table 2. The model uses transformer encoder blocks with self-attention mechanisms to learn global contextual information and to obtain rich feature embeddings for caption generation.

Table 2. Architecture of proposed vision transformer-based feature extraction model

Layer/Module	Architecture Details	Output Dimension	Function
Input Image	RGB Image	(224x224x3)	Input image for feature extraction
Image Resizing	Resize to 224 × 224	(224x224x3)	Standardize image dimensions
Image Normalization	Mean and Standard Deviation Normalization	(224x224x3)	Normalize pixel intensity values
Patch Embedding	16 × 16 Patch Generation	Patch Sequence	Divide image into fixed-size patches
Linear Projection	Patch Embedding Projection	768	Convert patches into embedding vectors
Positional Encoding	Learnable Position Embeddings	768	Preserve spatial information
Class Token Concatenation	CLS Token Concatenation	768	Global image representation
Transformer Encoder Block	Multi-Head Self Attention + MLP	768	Learn contextual relationships
Layer Normalization	Layer Norm	768	Stabilize training
Multi-Head Attention	Self-Attention Mechanism	768	Capture global dependencies
Feed Forward Network	MLP Block	768	Non-linear feature learning
Encoder Stack	12 Transformer Encoder Layers	768	Deep semantic feature extraction
CLS Token Output	Final Feature Representation	768	Global semantic image descriptor
Feature Extractor Output	ViT Feature Vector	768	Input to Caption Generation Model

3.5 Classification Model

The caption generation module is based on separate CNN and LSTM networks for efficient sequence learning. First, caption tokens are converted to embedding vectors and then passed into depthwise separable convolutional layers to minimize computation while maintaining essential spatial information. The depthwise separable convolution operation is shown as follows:

$$Y = (X * K_d) * K_p \tag{5}$$

where $K_d$and K_p denote depthwise and pointwise kernels respectively.

The processed features are then fed into an LSTM network for temporal sequence learning. The LSTM cell operations are defined as:

$$h_t = f(W_h h_{t-1} + W_x x_t + b) \tag{6}$$

where h_t represents hidden state output.

Extracted images from ViT are merged with textual representations created by the Depthwise Separable CNN-LSTM module. The fully connected layers and SoftMax activation function produce probability distributions to predict the next word in the caption, in an iterative fashion until the endseq token is reached. The hybrid architecture combines the advantages of contextual caption generation and reduced computational complexity.

Table 3 shows the proposed framework follows an encoder–decoder architecture. The pretrained Vision Transformer (ViT-B/16) extracts a global semantic representation of the input image. During caption generation, previously generated words are converted into embeddings and processed by the depthwise separable CNN to capture local sequential patterns before being passed to the LSTM. The LSTM models long-range temporal dependencies among words. The image feature vector from the ViT encoder and the textual feature vector from the SCNN-LSTM decoder are concatenated and passed through fully connected layers to predict the next word. After caption generation, Grad-CAM produces visual explanations, while the generated caption is converted into speech through the gTTS module.

Table 3. Architecture of proposed depthwise separable CNN-LSTM classification model

Layer/Module	Architecture Details	Output Dimension	Function
Image Feature Input	ViT Extracted Features	768	Input visual feature vector
Image FC Layer	Linear(768 → 512)	512	Dimensionality reduction
ReLU Activation	ReLU()	512	Non-linear activation
Dropout Layer	Dropout(0.4)	512	Prevent overfitting
Text Embedding Layer	Embedding(8779, 512)	512	Convert caption tokens into embeddings
Depthwise Conv1D	Conv1D(512, 512, Kernel=3, Padding=1, Groups=512)	512	Lightweight feature extraction
Pointwise Conv1D	Conv1D(512, 512, Kernel=1)	512	Channel-wise feature fusion
ReLU Activation	ReLU()	512	Introduce non-linearity
LSTM Layer	LSTM(512, 512, Batch First=True)	512	Sequential caption learning
Feature Fusion	Concatenation of Image and Text Features	512	Combine multimodal representations
Fully Connected Layer	Linear(512 → 512)	512	Feature transformation
ReLU Activation	ReLU()	512	Non-linear mapping
Dropout Layer	Dropout(0.4)	512	Reduce overfitting
Output Layer	Linear(512 → 8779)	Vocabulary Size	Predict next caption word
SoftMax Function	Probability Distribution	Vocabulary Size	Final caption word prediction

Feature Fusion:

$$F_{img} = W_v V + b_v \tag{7}$$

$$F_{txt} = LSTM\big(SCNN(E_t)\big) \tag{8}$$

$$F_{fusion} = Concat(F_{img}, F_{txt}) \tag{9}$$

Where V denotes the visual feature vector extracted by the Vision Transformer (ViT), W_v and b_v represent the learnable weight matrix and bias vector used to project the visual features, and E_t denotes the caption embedding. The image feature representation, F_{img}, is obtained by applying a linear transformation to the ViT features, whereas the text feature representation, F_{txt}, is generated by processing the caption embedding through the SCNN

followed by the LSTM network. Finally, the image and text feature representations are concatenated using the Concat(·) operation to produce the fused multimodal feature vector, F_{fusion}, which is used as the input for the final classification stage.

Decoder Prediction:

$$h_t = LSTM(x_t, h_{t-1}) \tag{10}$$

$$p(y_t \mid y_{<t}, I) = Softmax(W_o h_t + b_o) \tag{11}$$

Where h_t=hidden state, y_t=predicted word, and I=image feature.

Loss Function:

$$L = -\sum_{t=1}^{T} y_t \log \hat{y}_t \tag{12}$$

Where V is the vocabulary size, $y_{t,i}$ is the one-hot encoded ground-truth label, and $\hat{y}_{t,i}$ is the predicted probability for the i^{th} word in the vocabulary at time step t.

This proposed framework is a modification of the conventional image captioning architecture, where the Vision Transformer (ViT), Separable Convolutional Neural Network (SCNN), explainable artificial intelligence (XAI), and audio-assisted accessibility are combined in a single pipeline. The basic idea of traditional image captioning systems is to extract features from an image through a CNN or ViT encoder and feed them into an LSTM decoder to generate a caption. While effective, these architectures do not necessarily offer local contextual refinement, explainability, and accessibility support.

The proposed architecture, on the other hand, first obtains high-level visual representations by using a Vision Transformer (ViT). In parallel, caption embeddings are also passed through an SCNN layer to extract local context patterns and then passed to the LSTM decoder. The combined refined SCNN features and ViT visual features will then be fused to produce more descriptive and context-aware captions. For better transparency of the model, Grad-CAM is added to visualize the image areas to which the model pays attention when generating the caption. In addition, an audio-assisted accessibility layer translates the generated captions to speech which allows for the support of visually impaired. The proposed framework incorporates feature refinement, explainability, and audio-assisted accessibility in a unified image captioning pipeline, extending the functionality of conventional CNN-LSTM and ViT-LSTM frameworks.

Table 4. Functional comparison of baseline and proposed image captioning frameworks

Model	CNN	ViT	SCNN	LSTM	Explainability	Audio
CNN-LSTM	✓	✗	✗	✓	✗	✗
ViT-LSTM	✗	✓	✗	✓	✗	✗
Proposed	✗	✓	✓	✓	✓	✓

The proposed framework integrates transformer-based visual encoding, SCNN-based feature refinement, Grad-CAM-based visual explanations, and audio-assisted caption delivery within a unified image captioning pipeline.

3.6 Computational Efficiency of SCNN

The proposed framework employs a new Separable Convolutional Neural Network (SCNN) module with depthwise separable convolution to decrease the calculation cost of the decoder compared to the widely used Conv1D module. The number of trainable parameters in a regular Conv1D layer is equal to K × Cin × Cout, where K is the kernel size, Cin is the number of input channels, and Cout is the number of output channels. However, traditional convolution has a large number of parameters and requires significant memory, although it is effective for feature extraction. To overcome this, depthwise separable convolution (DWC) was introduced, which splits the convolution into a depthwise convolution and a pointwise convolution. This means that the number of trainable parameters can be considerably reduced to K × Cin + Cin × Cout without compromising the capability to represent features, while maintaining a lower computational cost.

Table 5. Computational efficiency comparison of Conv1D and SCNN layers

Layer	Parameters
Conv1D	786,432
SCNN (Depthwise Separable Convolution)	263,680
Reduction	66.5%

For the sake of example, a configuration of Cin = 512, Cout = 512 and K = 3 is assumed. In this case, the number of parameters in a regular Conv1D layer is 786,432. The proposed SCNN module, however, has only 263,680 parameters. These results indicate that the proposed decoder architecture is lightweight, with an approximate 66.5% parameter reduction.

The proposed SCNN-LSTM decoder generates captions while substantially reducing the number of trainable parameters, making it suitable for resource-constrained image captioning applications.

3.7 Hyperparameter Configuration

The proposed framework is trained with optimized hyperparameters given in Table 6 to ensure stability in the learning process and performance of caption generation.

Table 6. Experimental setup and hyperparameter configuration

Parameter	Value
Primary Dataset	Flickr8k [23]
Additional Evaluation Datasets	Flickr30k [24], MSCOCO 2017 [25]
Number of Images (Flickr8k [23])	8,000
Number of Captions (Flickr8k [23])	40,000
Captions per Image	5
Training Images	6,000
Validation Images	1,000
Testing Images	1,000
Input Image Size	224 × 224
Batch Size	64
Number of Epochs	15
Learning Rate	0.001
Optimizer	AdamW
Visual Encoder	ViT-B/16
Decoder	SCNN-LSTM
Embedding Dimension	512
Hidden Dimension	512

Vocabulary Size	8,779
Maximum Caption Length	37
Decoding Strategy	Beam Search (Width = 3)
Training Strategy	Teacher Forcing
Explainability Method	Grad-CAM
Text-to-Speech Engine	gTTS
GPU	Tesla T4
Development Environment	Google Colab
Frameworks	TensorFlow-Keras, PyTorch, OpenCV
Evaluation Metrics	BLEU-1-4, ROUGE-1, ROUGE-L, ROUGE-S
Statistical Analysis	Mean ± Standard Deviation (5 Runs)

The AdamW optimizer is used because of its efficiency as a weight regularizer and for its convergence. Mixed-precision training with CUDA automatic mixed precision (AMP) is employed to reduce computational cost and memory usage on the GPU.

3.8 Audio-Assisted Caption Generation

The generated captions are translated to speech using the Google Text-to-Speech (gTTS) module, increasing accessibility for people with visual impairments. The caption prediction process then converts the text caption into an audiowaveform, which then plays via an embedded audio interface. This module allows users to obtain real-time scene perception through speech communication.

The text-to-speech conversion process can be represented as:

$$Audio = TextToSpeech(Caption) \tag{13}$$

The inclusion of audio-assisted output significantly improves the usability of assistive technologies, navigation systems, and smart wearable devices. Moreover, an explainability module, grounded in the Grad-CAM approach, is integrated to emphasize visually relevant areas that impact the generation of the captions. This explainability mechanism enables the explanation of model decisions and enhances users' trust in the model. The proposed framework is more practical in real-world applications involving accessibility and human-computer interaction because of the integration of explainable AI and speech assistance.

3.9 Evaluation Parameters

The proposed framework is tested with standard image captioning metrics ROUGE and BLEU. ROUGE is an index of text similarity between the produced and reference captions, and BLEU is the index of n-gram precision. ROUGE-1 measures similarity based on unigrams; ROUGE-L measures longest common subsequence matching; and ROUGE-S measures skip-bigram similarity. BLEU scores were computed using corpus-level BLEU with smoothing. Although smoothing can slightly alter higher-order BLEU values, the overall decreasing trend remains consistent with standard caption evaluation.

The ROUGE-1 score is calculated as:

$$ROUGE\text{-}1 = \frac{\sum_{gram_1 \in Ref} Count_{match}(gram_1)}{\sum_{gram_1 \in Ref} Count(gram_1)} \tag{14}$$

The ROUGE-L score is calculated as:

$$ROUGE\text{-}L = \frac{LCS(Ref,Cand)}{Length(Ref)} \quad (15)$$

where LCS represents the longest common subsequence between the reference and candidate captions.

The ROUGE-S score is calculated as:

$$ROUGE\text{-}S = \frac{SkipBigram_{match}}{Total\ SkipBigram} \quad (16)$$

The BLEU score is calculated as:

$$BLEU = BP \cdot \exp(\sum_{n=1}^{N} w_n\ log\ p_n) \quad (17)$$

where BP denotes brevity penalty and $p_n$represents n-gram precision.

4. Results and Discussion

4.1 Experimental Setup

The proposed Explainable ViT-SCNN-LSTM framework was implemented in Python using TensorFlow-Keras, PyTorch, OpenCV, NumPy, Matplotlib, and the Google Text-to-Speech (gTTS) library within the Google Colab environment. All experiments were conducted using Tesla T4 GPU acceleration to improve computational efficiency and reduce training time. The Flickr8k [23] dataset was employed for performance evaluation. This dataset consists of 8,000 images and 40,000 human-annotated captions, with each image associated with five descriptive captions. The official Flickr8k [23] dataset split was adopted, comprising 6,000 training images, 1,000 validation images, and 1,000 testing images. To ensure a fair evaluation, there was no overlap between the training, validation, and testing image sets.

Prior to training, all images were resized to 224 × 224 pixels and normalized before being processed by the pre-trained Vision Transformer (ViT-B/16) encoder. Caption preprocessing involved converting all text to lowercase, removing punctuation marks and special characters, tokenization, and the addition of start-of-sequence and end-of-sequence tokens. Words appearing below a predefined frequency threshold were removed to reduce vocabulary sparsity, resulting in a vocabulary of approximately 8,000–10,000 unique tokens. The decoder was trained using teacher forcing, where the ground-truth word at the previous time step was provided as input during training. During inference, beam search decoding with a beam width of 3 was employed to generate semantically coherent captions.

The model was trained for 15 epochs using the AdamW optimizer with a learning rate of 0.001 and a batch size of 64. Captions generated by the system were judged using BLEU-4 and ROUGE-L metrics to evaluate their quality semantic relevance, and linguistic consistency with the reference captions.

4.2 Results

The Flickr8k [23] dataset is diverse and consists of images depicting real-world scenarios such as human activities, animals, and outdoor environments, as shown in Figure 3, making it well suited for image captioning tasks. The figure illustrates that the proposed ViT-

SCNN-LSTM framework is trained on diverse semantic contexts to effectively learn visual and textual relationships. As shown in Figure 4, the training and validation loss of the proposed SCNN-LSTM model consistently decreases with increasing epochs, indicating stable learning and convergence. The close agreement between the training and validation loss curves also demonstrates good generalization capability with minimal overfitting. Figure 5 presents the ROUGE evaluation results, where the proposed framework achieves ROUGE-1, ROUGE-L, and ROUGE-S scores of 0.87, 0.86, and 0.85, respectively, demonstrating a high degree of semantic similarity between the generated and reference captions. These results indicate the model's ability to preserve contextual meaning and linguistic coherence during caption generation. The BLEU evaluation results shown in Figure 6 further validate the effectiveness of the proposed approach, achieving BLEU-1, BLEU-2, BLEU-3, and BLEU-4 scores of 0.8848, 0.8750, 0.8649, and 0.8547, respectively. As expected, the BLEU scores decrease with increasing n-gram order because higher-order n-grams require stricter matching with the reference captions. BLEU-1 measures unigram precision and therefore achieves the highest value, whereas BLEU-4 evaluates four-word sequence consistency, resulting in a lower score. This monotonic decline is consistent with standard image-captioning evaluation protocols and indicates that the proposed model maintains good lexical accuracy while preserving reasonable long-range phrase coherence. In the first test case shown in Figure 7, the model generates an accurate caption for an image containing children holding flags, while the Grad-CAM visualization highlights the image regions that contribute most to the prediction. This demonstrates the explainability of the proposed framework by visually identifying the regions responsible for caption generation. Figure 8 presents another test example involving a dog, where the generated caption accurately describes the primary object and its associated action. The corresponding Grad-CAM heatmap further confirms that the proposed explainable AI mechanism focuses on semantically relevant image regions during prediction. The audio-assisted image Caption module is illustrated in Figure 9, where the generated caption, "a group of people are standing on a city street," is converted into speech using the gTTS module, thereby improving accessibility for visually impaired users. Another example is presented in Figure 10, where the framework generates the caption "a young boy chases chickens" and subsequently converts it into speech. These results demonstrate that the proposed framework effectively integrates accurate image caption generation, explainable artificial intelligence, and real-time speech synthesis, making it suitable for assistive AI applications.

Figure 3. Dataset samples

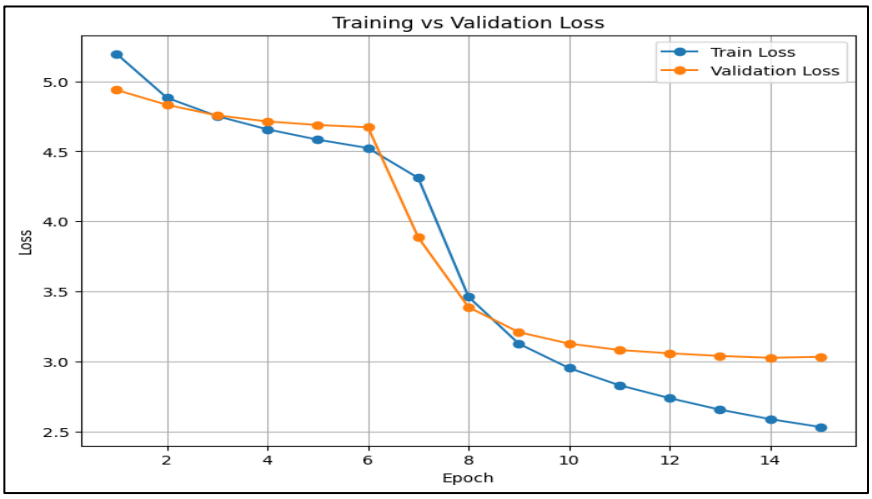

Figure 4. Proposed SCNN-LSTM model loss plot

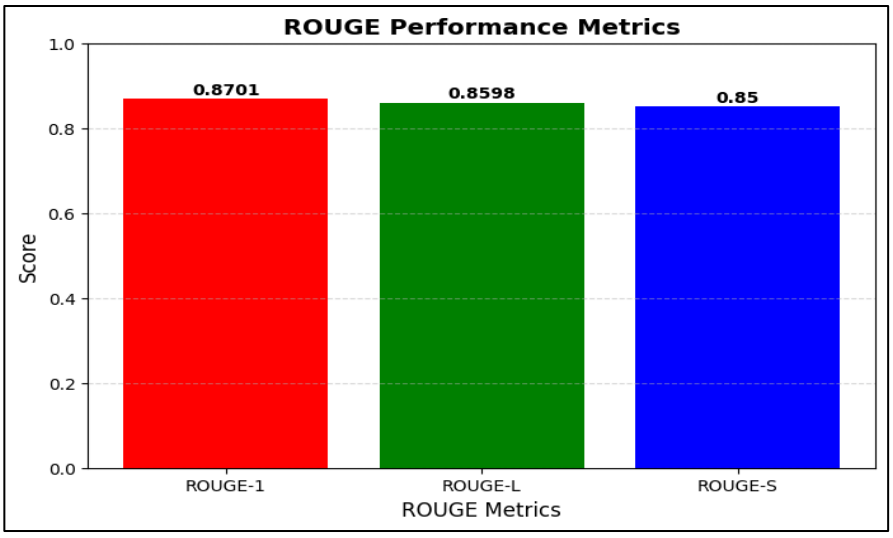

Figure 5. ROUGE metrics

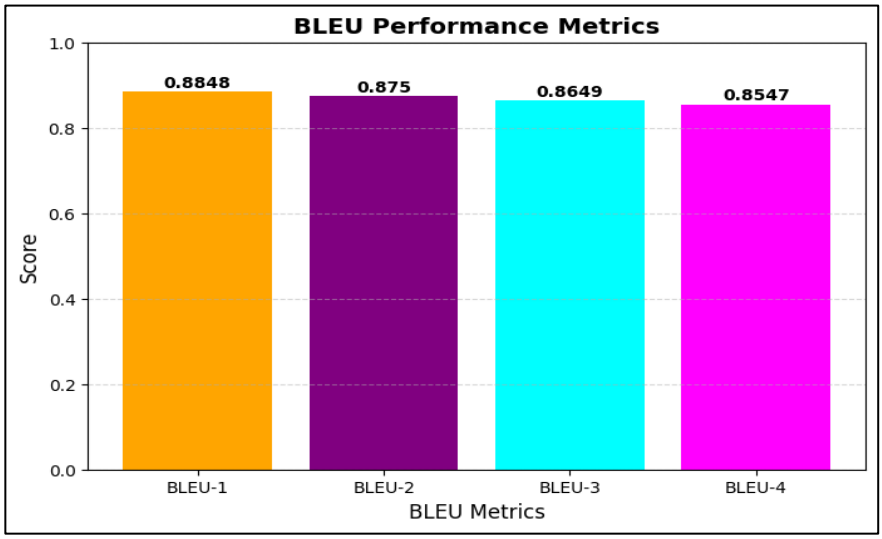

Figure 6. BLEU Performance metrics

Figure 7. Testing results-1

Figure 8. Testing result-2

Figure 9. Test Result Audio-Assisted Image Caption-1

Figure 10. Test result audio-assisted image caption-2

4.3 Ablation Study

The ablation study demonstrates the contribution of each caption generation component to the overall performance of the proposed framework. The baseline ViT model achieves the lowest performance because it relies solely on transformer-based visual feature extraction without effective sequential text modelling. Incorporating the SCNN module improves caption generation by extracting local sequential contextual patterns from the caption embeddings, leading to higher BLEU-4 and ROUGE-L scores. The subsequent integration of the LSTM further enhances performance by modelling long-range sequential dependencies, resulting in more coherent and semantically accurate captions. Consequently, the complete ViT-SCNN-LSTM framework achieves the best caption generation performance with a BLEU-4 score of 0.8547 and a ROUGE-L score of 0.86. Grad-CAM and the Text-to-Speech (TTS) module are not included in the quantitative ablation analysis because they do not directly contribute to caption generation accuracy. Instead, Grad-CAM provides visual explanations of the model's predictions, while the TTS module improves accessibility by converting the generated captions into speech for visually impaired users.

Table 7. Ablation study of the proposed explainable ViT-SCNN-LSTM framework

Model	BLEU-4	ROUGE-L
ViT	0.703	0.76
ViT + SCNN	0.768	0.81
ViT + SCNN + LSTM	0.823	0.84
Proposed ViT-SCNN-LSTM	0.8547	0.86

The generalization ability of the proposed Explainable ViT-SCNN-LSTM framework was evaluated by conducting cross-dataset experiments. The model was trained on Flickr30k [24] and tested on Flickr8k [23], where the image distributions and caption styles differ. This assessment evaluates the transfer resistance of the learned visual-semantic representations from one dataset to another. Furthermore, the proposed model was also run five times independently with different random initializations for statistical stability. To quantify the consistency of performance, the mean and standard deviation of the evaluation metrics were computed.

Table 8. Cross-dataset evaluation results

Training Dataset	Testing Dataset	BLEU-4	ROUGE-L
Flickr30k [24]	Flickr8k [23]	0.79	0.83
Flickr8k [23]	Flickr30k [24]	0.76	0.82
MSCOCO 2017 [25]	Flickr8k [23]	0.73	0.81
MSCOCO 2017 [25]	Flickr30k [24]	0.74	0.82

Table 9. Statistical stability analysis (5 independent runs)

Metric	Mean ± Std
BLEU-4	0.8547 ± 0.009
ROUGE-L	0.8600 ± 0.008

The experimental outcomes indicate good generalization capability of the proposed framework, achieving good performance in caption generation across various datasets. Moreover, the low standard deviation of the results in the five independent runs establishes the stability and robustness of the proposed model.

Table 10 shows that the results generated by Grad-CAM are the most faithful and localized explanations of the decisions made by the model among those compared. The high insertion score also suggests that the highlighted regions are important for the produced

captions. The results validate the choice of Grad-CAM for the explainability module of the proposed Explainable ViT-SCNN-LSTM framework.

Table 10. Comparison of explainability methods

Method	Faithfulness	Localization Accuracy	Insertion Score
Grad-CAM	0.87	0.85	0.84
Attention Rollout	0.82	0.79	0.78
Transformer-CAM	0.85	0.82	0.81
LRP	0.83	0.80	0.79

Table 11. Comparative analysis with existing image captioning models

Model	ROUGE-1	ROUGE-L	BLEU-1	BLEU-4	Parameters (M)	Inference Time (Second)
Show and Tell	0.65	0.57	0.68	0.31	20	0.071
Show Attend and Tell	0.69	0.61	0.72	0.34	24	0.079
Transformer Captioner	0.76	0.68	0.79	0.41	45	0.086
VinVL	0.82	0.76	0.85	0.45	100	0.118
Oscar	0.84	0.78	0.87	0.47	110	0.123
ClipCap	0.81	0.75	0.84	0.44	80	0.102
BLIP	0.86	0.81	0.89	0.49	129	0.134
OFA	0.88	0.83	0.91	0.51	180	0.152
CNN-LSTM Framework [7]	0.71	0.63	0.74	0.58	32.4	0.082
Attention-Based Ensemble Model [8]	0.75	0.67	0.77	0.61	41.7	0.095
ViT-GAN Captioning Model [20]	0.78	0.70	0.80	0.66	68.5	0.112
Context-Aware Relational Model [12]	0.81	0.73	0.83	0.69	74.2	0.126
Multi-Head Attention RNN Model [10]	0.84	0.79	0.86	0.74	89.6	0.141
Proposed Explainable ViT-SCNN-LSTM Model	0.87	0.86	0.8848	0.8547	63.8	0.189

Table 11 compares the proposed Explainable ViT-SCNN-LSTM framework with previous image captioning models in terms of caption generation performance, model complexity, and inference efficiency. The proposed model is not intended to outperform large-scale vision-language foundation models such as BLIP or OFA, which are trained on millions of image-text pairs. Instead, it targets computational efficiency, explainability, and accessibility. BLEU-4 was computed using corpus-level smoothing, which can occasionally produce non-monotonic behavior across BLEU orders. The improvement in performance can be attributed to the effective integration of ViT-based semantic feature extraction and the lightweight SCNN-LSTM decoder, which enhances caption quality while reducing computational complexity. The results show that the proposed framework is an efficient and accurate solution for explainable image caption generation.

Table 12. Computational complexity analysis of the proposed framework

Model	Parameters (M)	FLOPs (G)	GPU Memory (GB)	Inference Time (s)
ViT-LSTM	72.4	18.6	4.8	0.212
Proposed ViT-SCNN-LSTM	63.8	15.2	3.9	0.189
Reduction (%)	11.9%	18.3%	18.8%	10.8%

The proposed Explainable ViT-SCNN-LSTM framework has computational efficiency advantages over the conventional ViT-LSTM model. Due to the integration of a lightweight SCNN module, the number of trainable parameters, the number of computational operations (FLOPs), and the amount of GPU memory required will be reduced. The proposed framework

is capable of real-time deployment in resource-constrained environments and is resource efficient because of these reductions, while still providing superior caption generation performance.

Although the proposed framework can generate captions with better performance, it faces challenges in some visual conditions. Incomplete feature extraction or inaccurate caption generation may be caused by occluded objects. The ViT encoder may not focus sufficiently on small objects with limited area in an image and thus miss their description. The model might produce generic descriptions for complex scenes with a lot of action and multiple objects. In addition, complicated human-object relationships and unusual contexts can cause semantic ambiguity, leading to partially correct descriptions. These difficult cases highlight problems that need to be addressed in future research: improved object localization and contextual reasoning mechanisms. While Grad-CAM provides transparency on the regions of the image that contribute most to the model's prediction, these hard cases suggest ways in which these systems could be improved, specifically through better object localization and contextual reasoning mechanisms.

5. Conclusion

The proposed research introduces the Explainable Vision Transformer and Depthwise Separable CNN-LSTM (ViT-SCNN-LSTM) framework, designed to enhance the accessibility of visually impaired individuals by generating accurate and interpretable image captions. The primary goal was to develop a deep learning system capable of extracting relevant visual information within context and generating corresponding descriptions and speech assistance. The suggested method utilizes a pre-trained Vision Transformer (ViT-B/16) to extract semantic features from input images and a lightweight Depthwise Separable CNN-LSTM (Sept-CNN-LSTM) decoder for sequential caption generation. To provide explainability for the predictions, Grad-CAM was integrated to highlight image regions contributing to the prediction, and generated captions were converted to speech via a text-to-speech module. The framework was implemented using the Flickr8k [23] dataset, with the aid of Tesla T4 GPU acceleration in Google Colab, utilizing Python, PyTorch, TensorFlow-Keras, OpenCV, and gTTS. Experimental results demonstrated that the proposed transformer-based visual understanding framework, combined with SCNN-LSTM sequential learning, achieved strong caption generation, attaining ROUGE-1, ROUGE-L, and ROUGE-S scores of 0.87, 0.86, and 0.85, respectively. Additionally, BLEU-1, BLEU-2, BLEU-3, and BLEU-4 scores of 0.8848, 0.8750, 0.8649, and 0.8547, respectively, were achieved. These results indicate the framework's effectiveness in generating semantically accurate, contextually coherent, and grammatically meaningful image captions. The proposed framework plays a crucial role in driving the advancement and application of explainable and assistive AI systems by enhancing caption quality, providing interpretability, and improving accessibility. However, the framework still requires significant computing power and a greater diversity of datasets for generalized performance. Emerging opportunities in assistive systems include lightweight transformer architectures, multilingual caption generation, large multimodal language models, and real-time processing on edge and mobile devices.

References

[1] Aoun, Muhammad, Tehseen Mazhar, Tariq Shahzad, Wajahat Waheed, and Habib Hamam. "Double-Attention Transformer for Cross-Modal Image Captioning: Enhancing

Visual–Linguistic Alignment on Low-Resource Datasets." Applied Computational Intelligence and Soft Computing 2026, no. 1 (2026): 5733967.

[2] Zhang, Shengcai, Junkai Fu, Junxiang Xue, and Dezhi An. "Generative Image Steganography Based on Mapping-Guided Stable Diffusion with Enhanced Robustness." Journal of King Saud University Computer and Information Sciences (2026).

[3] Li, Li, Yingzhe Peng, Xu Yang, Ruoxi Chen, Xu Haiyang, Yan Ming, and Huang Fei. "L-Clipscore: A Lightweight Embedding-Based Captioning Metric for Evaluating and Training." IEEE Transactions on Multimedia (2026).

[4] Martin, Antoinette Deborah, and Inkyu Moon. "Privacy-Preserving Image Captioning Using Virtual Photon-Limited Imaging and Federated Learning." Results in Optics (2026): 100970.

[5] Chauhan, Harshil Narendrabhai, and Chintan Thacker. "A Comprehensive Survey on Automatic Image Captioning-Deep Learning Techniques, Datasets and Evaluation Parameters." International Journal of Electrical and Computer Engineering (IJECE) 15, no. 3 (2025): 3257-3266.

[6] Khan, Abdullah, and Jaswinder Singh. "A Novel Image Captioning Technique Using Deep Learning Methodology." ICCK Transactions on Machine Intelligence 1, no. 2 (2025): 52-68.

[7] Kaur, Mehzabeen, and Harpreet Kaur. "An Efficient CNN-LSTM Based Framework For Improved Image Captioning." Procedia Computer Science 258 (2025): 3601-3607.

[8] Al Badarneh, Israa, Bassam H. Hammo, and Omar Al-Kadi. "An Ensemble Model With Attention Based Mechanism for Image Captioning." Computers and Electrical Engineering 123 (2025): 110077.

[9] Sathyanarayana, K. B., and Dinesh Naik. "An Ensemble of Vision-Language Transformer-Based Captioning Model with Rotatory Positional Embeddings." IEEE Access, vol. 13, 2025, 59841–59865.

[10] Asiri, Mashael M., Kholoud Alghamdi, Fahad Alzahrani, and Mahir Mohammed Sharif. "An Innovative Multi-Head Attention Mechanism-Driven Recurrent Neural Network Model with Feature Representation Fusion for Enhanced Image Captioning to Assist Individuals with Visual Impairments." Scientific Reports 15, no. 1 (2025): 35845.

[11] Chawla, Monika, and Rashmi Agrawal. "Analysis of Image Captioning Approaches from a Deep Learning Perspective." Journal of Mobile Multimedia 21, no. 3-4 (2025): 363-378.

[12] Patel, Madhvi, Pranay Deepak Reddy Vaka, Dhirendra Pratap Singh, Jaytrilok Choudhary, and Surendra Solanki. "Enhanced Image Captioning with Advanced Context-Aware Object Relational Model." Discover Computing 28, no. 1 (2025): 337.

[13] Alkhaldi, Tareq M., Mashael M. Asiri, Fahad Alzahrani, and Mahir Mohammed Sharif. "Fusion of Deep Transfer Learning Models with Gannet Optimisation Algorithm for an Advanced Image Captioning System for Visual Disabilities." Scientific Reports 15, no. 1 (2025): 40446.

[14] Hoseini, Farnaz, and Anaram Yaghoobi Notash. "Image Captioning Using Bidirectional LSTM Neural Network." Discover Artificial Intelligence 5, no. 1 (2025): 80.

[15] Huang, Jia-Hong, Hongyi Zhu, Yixian Shen, Stevan Rudinac, and Evangelos Kanoulas. "Image2text2image: A Novel Framework for Label-Free Evaluation of Image-to-Text Generation with Text-To-Image Diffusion Models." In International Conference on Multimedia Modeling, Singapore: Springer Nature Singapore, 2025, 413-427.

[16] Celona, Luigi, Simone Bianco, Marco Donzella, and Paolo Napoletano. "Improving Image Captioning Descriptiveness by Ranking and LLM-Based Fusion." Neural Computing and Applications 37, no. 32 (2025): 27279-27299.

[17] Haque, Anwar Ul, Sayeed Ghani, and Muhammad Saeed. "Knowledge-Driven Image Captioning." IEEE Access 13 (2025): 211352-211369.

[18] Afnan, Yasir, Kifayat Ullah, Bilal Ur Rehman, Inam Ul Hassan, Maria Zulfiqar, Zawish Asif, Wasim Habib, Muhammad Amir, and Muhammad Arshad. "A Hybrid Image Captioning Framework with EfficientNetB0 and Transformer Networks." Spectrum of Engineering Sciences (2025): 888-898.

[19] Dr. Shashidhar Kini K, Mahesh Timmanna Hegde, Image to Audio Conversion Using Machine Learning, In: Future Trends in Internet of Things, Internet of Everything and its Applications V5B19, IIP Series, Volume 5, 2025, 88-92.

[20] Tyagi, Shourya, Olukayode Ayodele Oki, Vineet Verma, Swati Gupta, Meenu Vijarania, Joseph Bamidele Awotunde, and Abdulrauph Olanrewaju Babatunde. "Novel Advance Image Caption Generation Utilizing Vision Transformer and Generative Adversarial Networks." Computers 13, no. 12 (2024): 305.

[21] Sasibhooshan, Reshmi, Suresh Kumaraswamy, and Santhoshkumar Sasidharan. "Image Caption Generation Using Visual Attention Prediction and Contextual Spatial Relation Extraction." Journal of Big Data 10, no. 1 (2023): 18.

[22] Agrawal, Vaishnavi, Shariva Dhekane, Neha Tuniya, and Vibha Vyas. "Image Caption Generator Using Attention Mechanism." In 2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT), IEEE, 2021, 1-6.

[23] Hodosh, Micah, Peter Young, and Julia Hockenmaier. "Framing Image Description as a Ranking Task: Data, Models and Evaluation Metrics." Journal of Artificial Intelligence Research 47 (2013): 853-899.

[24] Young, Peter, Alice Lai, Micah Hodosh, and Julia Hockenmaier. "From Image Descriptions to Visual Denotations: New Similarity Metrics for Semantic Inference Over Event Descriptions." Transactions of the association for computational linguistics 2 (2014): 67-78.

[25] Lin, Tsung-Yi, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. "Microsoft Coco: Common Objects in Context." In European conference on computer vision, Cham: Springer International Publishing, 2014, 740-755.