https://doi.org/10.31449/inf.v50i10.12411 Informatica **50** (2026) 107–120 **107**

T-BiLSTM-CCO: A Transformer-BiLSTM framework with Cuckoo Catfish Optimization for multilingual image description localization

LiGie Su
College of Foreign Languages, Wuhan Business University, Wuhan 430056, Hubei, Chian
E-mail: LiGieSu457@outlook.com

Keywords: Multimodal neural machine translation, Transformer integrated Bidirectional Long Short-Term Memory Network-tuned with Cuckoo catfish optimizer (T-Bi-LSTM-CCO), Image description.

Received: October 13, 2025

Image description localization benefits from integrating textual and visual cues. However, conventional neural machine translation models often overlook fine-grained grounding. This research introduces aNeural Machine Translation Model (NMT) named Transformer integrated Bidirectional Long Short-Term Memory Network-tuned with Cuckoo catfish optimizer (TBiLSTM-CCO) to enhance translation accuracy while preserving image-caption semantic congruence in multilingual captioning. Multilingual image-caption pairs were gathered from the Multi30 and MS-COCO datasets, each containing aligned captions in multiple languages. Visual features are extracted using ResNet, producing high-dimensional semantic embeddings representing objects and scenes. The proposed hybrid architecture combines the CCO with a Transformer-enhanced Bi-LSTM network. This configuration enhances sequence modeling and attention capabilities, optimally tuning parameters to improve translation fidelity, multimodal feature fusion, and localized visual alignment during description generation. The proposed T-Bi-LSTM-CCO achieves higher BLEU-1 to BLEU-4 scores of 0.90, 0.89,0.87, and 0.85 values than the existing methods. Python was used to conduct experiments on the MS-COCO and Multi30k datasets, utilizing Word2Vec for text embeddings and ResNet-152 for visual features. The CCO was trained for over 50 epochs. The hybrid T-Bi-LSTM-CCO-model consistently achieved better alignment between descriptions and image regions, validating its effectiveness for multimodal translation and grounding tasks. The multimodal NMT framework combines Deep Learning (DL) and visual features, improving translation quality and image-description localization, showing robustness for real-world multilingual applications.

Povzetek: Raziskava predstavlja napreden multimodalni model, ki združuje besedilne in slikovne informacije za natančnejše večjezično prevajanje opisov slik.

1 Introduction

Image captioning is the translation of visual content into descriptive text. Generally, this task detects objects in an image, their spatial relationships, and builds a semantic representation that can be written as text [1]. The growing need for Machine Translation (MT) has been encouraged by technological developments and contemporary communication necessities. Encoder-decoder models have advanced MT systems, with attention mechanisms improving translation of long sentences by enabling the translator to selectively emphasis on various portions of the source phrases [2]. Neural MT is typically impaired in low-resource contexts, resulting in sub-optimal performance. Multimodal MT (MMT) that uses text data with other modes (images, multilingual texts, etc.) is used to improve translation performance [3]. Multimodal Neural Machine Translation (MNMT) extends multimodal MT from only textual data to situations in which the text context is insufficient, due to ambiguity of words or grammatical gender problems[4]. Emerging research in this area continues to focus on the types of visual input that can improve the quality of translation. Most successful methods produce accurate descriptions of images but do not attend to linguistic style or sentiment. Recent developments have either included the emotional content of images in their captions, taking into consideration both linguistic description and expressive language, recognizing representation and affect[5]. One of the key challenges of MNMT is efficiently merging visual and text features. Most of the current methods embed pictorialmaterial in the encoder or interpreter of neural MT models to align images with multilingual text [6]. Quantitative measures show that these metrics can identify mistranslations, and fine-tuning LSTM-driven methods on province-specific data enhances in-domain conversionsuperiority [7]. The objective of the research is

to improve multilingual picture description localization using the T-BiLSTM-CCO framework.

Research question: Does including the CCO considerably increase localization accuracy over conventional optimizers?"

Aim of the research: The aim of the research is to establish a high-performance multimodal framework that contains visual and textual features for quality image captioning and translation. This framework integrates ResNet-based image features and tokenized textual embeddings through a Transformer-aided BiLSTM, while parameter optimizations are performed using the CCO.

1.1. Key contributions

- Created a multimodal dataset that included image-caption pairs and included a variety of contexts for caption generation and translation. Used network architecture, including tokenization, normalization, and resizing images to create uniform inputs for both visual and textual modalities.
- Established a Transformer-enhanced BiLSTM optimized with CCO for greater feature fusion and precise caption generation.

Section Organization: Section 1 introduces the background of the research, while **Section 2**analyses related works. **Section 3**clarifies the dataset, preprocessing, and feature extraction, and **Section 4** presents the proposed TBiLSTM-CCO framework. **Section 5**deliberates the experimental setup and results, followed by **Section 6**, which accomplishes the research with key findings and future scope.

2 Literature review

Research [8] proposed a Bengali image caption generation model automatically, which performed better than baseline methods on benchmark data using BLEU-1 and BLEU-4 values. [9] Gathered English news articles with images and utilized English as the pivot language for developing MT systems to prepare parallel datasets. Research [10] compared NMT and MMT models at different sentence lengths and found that NMT would sometimes perform better than MMT. A hybrid model that integrates ResNet50 and Bi-LSTM was suggested in [11] for remote detection image captioning, processing spatial and spectral characteristics. The highest performance measures, such as BLEU-1, BLEU-2, BLEU-3, BLEU-4, and ROUGE, were obtained by this model.The Word-Region Alignment (WRA)-Guided MNMT [12] combines textual and visual semantics. It improves cross-modal semantic connection. On benchmark datasets, +1.0 BLEU (EN–DE), +1.1 BLEU (EN–FR), and +0.7 BLEU (EN–JP) were attained. For image captioning, [13] introduced a multimodal fusion DL model that combined Faster Recurrent Neural Network (RCNN) for feature encoding and LSTM for decoding, reporting a 95% improvement in performance. Methods such as Bilingual-Visual Consistency (BiVC) and Target-Visual Consistency (TVC) [14] also improved the performance of MNMT, further validated on several multimodal datasets. The visual multimodal manuscriptacknowledgment and sentiment analysis model (MTR-SAM), which we built on, uses a sinusoidal loss function, Residual Squeeze-and-Excitation Feature Pyramid Network (RSE-FPN), and Large Kernel–Path Aggregation Network (LK-PAN) [15]. This framework achieves faster recognition, higher accuracy, and a greater F1 score across benchmark datasets. The research [16] introduces a Multimodal Multisource NMT (M3S-NMT) system that integrates CNN-based visual representations with RNN textual embeddings (GloVe/BERT) with 34.57 and 42.52 scores on the benchmark data on METEOR. Video-based instructions in MT strategies [17-18] took advantage of a spatio-temporal video context, which used multi-pattern joint learning to improve a translation over two real-world datasets. Sports broadcasts with addressed subtitles and live commentary are highly related to context and accuracy. A multimodal model with language-specific encoders on text and images [19] allowed a shared embedding space of both text and image semantic relationships, with aninstruction F1-score of 0.7618 on the MUSTI dataset. Lastly, [20] introduced real-time captioning and visual similarity contrast with Neuraltalk+, which was found to train faster and gave better quantitative and qualitative outcomes on the Flickr 8K and 30K datasets.

Table 1: Summary of multimodal and multilingual research

S.No	Method	Datasets Used	Evaluation Metrics	Quantitative Results	Proposed / Contribution
Das et al., [8]	Benchmark datasets, BLEU evaluation	Bengali image benchmark	BLEU-1, BLEU-4	BLEU-1: 0.67 / 0.65; BLEU-4: 0.26 / 0.24	Automatic Bengali captioning model outperforms baseline
Meetei et al., [9]	Pivot-based MT with image/audio features	English news articles with image pairs	BLEU	+3 BLEU improvement	Incorporating Manipuri multimodal features improved BLEU by +3 points

Cui et al., [10]	NMT vs Multimodal MT comparison	Multi30k	BLEU (NMT vs. MMT comparison)	NMT outperforms MMT for short sentences	Showed that NMT sometimes outperforms MMT
Sree et al., [11]	Hybrid ResNet50 + BiLSTM	Remote-sensing image dataset	BLEU-1…4, METEOR, ROUGE	BLEU-4: 44.64; METEOR: 40.69; ROUGE: 42.65	Captures spatial and spectral features
Zhao et al., [12]	WRA-Guided MNMT	Multi30k, Flickr30kEnt-JP	BLEU	EN–DE: +1.0 BLEU, EN–FR: +1.1 BLEU	Integrates word-region alignment
Thangavel et al., [13]	Faster RCNN + LSTM decoding	MS-COCO	BLEU	Reported ~95% relative improvement	Multimodal fusion model with 95% performance improvement
Liu et al., [14]	Target–Visual Consistency and Bilingual–Visual Consistency	Multi30k, Flickr30k	BLEU-1…4, METEOR	BLEU-4: 37.8; METEOR: 31.4	Improved multimodal NMT performance across multiple datasets
Liu et al., [15]	MTR-SAM	Custom Internet Public Opinion Dataset	Accuracy, F1-score	+10.78% accuracy, +4.42% F1 improvement	Improves recognition accuracy and sentiment analysis
Mohammed et al., [16]	(M3S-NMT)	Multi30k (EN/DE/FR/CZ → AR), Flickr30kEnt-JP	METEOR	34.57 (EN–DE→AR), 42.52 (EN–FR→AR)	Introduces a multisource multimodal dataset
Chen et al., [17]	Spatio-temporal context, multi-pattern joint learning	YouCook2, HowTo100M	BLEU	+4.2 BLEU improvement over baseline	Enhanced translation across two real-world datasets
Zhiliang et al., [18]	Multimodal real-time processing	Sports broadcast video dataset	BLEU-4	BLEU-4: 38.2	Accurate and contextually relevant live commentary
Esteban-Romero et al., [19]	Language-specific encoders	MUSTI	Macro F1-score	Macro F1: 0.7618	Achieved macro F1-score
Sharma &Padha [20]	Neuraltalk+	Flickr8k, Flickr30k	BLEU-1…4	BLEU-4: 36.7	Faster training

Despite these developments, current approaches tend to fail in the low-resource environment, have poor multimodal feature integration, and suboptimal context-sensitive translation. The proposed TBiLSTM-CCO addresses these shortcomings by incorporating bidirectional sequence modeling, Transformer-based attention, and metaheuristic optimization to obtain a higher accuracy, semantic arrangement, and localization of multilingual description. Existing techniques fail to generalize because of inadequate multimodal integration, weak visual-text alignment, and an overreliance on textual context. Short or simple sentences gain little from visual cues, as demonstrated in [10], which lowers the efficiency of semantic grounding and cross-lingual alignment.

2.1. Problem statement

Multilingual image description localization would be a key application to interfaces such as automatic image captioning, intercultural communication, and content accessibility. In the traditional models of NMT, the fine-grained vision grounding is usually disregarded in favor of textual information, the result of which is erroneous or semantically inappropriate translation. The existing techniques are struggling to effectively integrate textual and visual modalities, leading to less effective translation of various languages and suboptimal image region localization. Parameter tuning in the hybrid architecture is another limitation on how well the visual and contextual cues can be fully represented by the DL models.

3 Methodology

The model workflow, shown in Figure 1, starts with gathering multimodal datasets, comprised of images and their textual descriptions. The datasets are then sent to preprocessing. During preprocessing, the text is tokenized, and images are resized and normalized for consistency and to the desired input shape to extract the most information. The feature extraction step includes Word2Vec embedding text into dense embeddings, then ResNet extracting high-level features from images, producing compact representations. These extracted representations are input to the TBiLSTM-CCO model, where multimodal fusion and predictions occur.

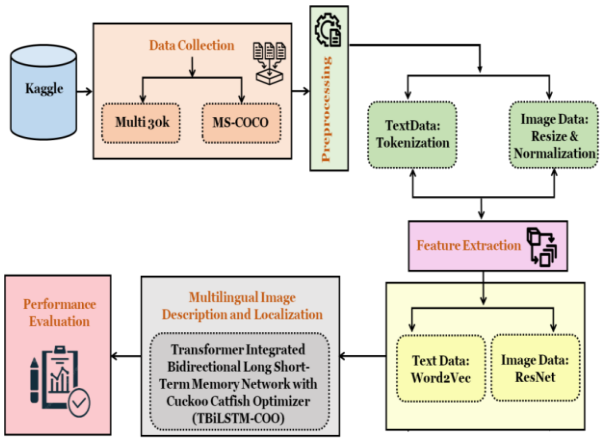

Figure 1: Sequential workflow of multimodal learning.

3.1. Multilingual image-caption pair dataset collection

To facilitate the localization of reliable image descriptions, multilingual image-caption datasets were sourced from the openly accessible platform, Kaggle. The MS-COCO dataset, found in [21], contains a rich corpus of images with English captions depicting a wide variety of scenes and objects. The Multi30k dataset, also accessible from [21], contains aligned captions in multiple languages, including English, German, and French, which permits both multilingual translation and multilingual grounding experiments. Together, the datasets supply a wealth of visual and textual data that supports effective multimodal feature extraction and training of the proposed TBiLSTM-CCO.MS-COCO: Train = 80 %, Test = 20 %. Multi30k: Train = 80 %, Test = 20 %. The approach only uses image–caption pairs to ensure focused multimodal alignment among visual and verbal input. Extra metadata should be eliminated whenever possible to avoid bias and maintain comparability among multilingual datasets for objective assessment.

✓ **Generalization across datasets:** The proposed T-BiLSTM-CCO model can be further evaluated on Flickr30k (to assess real-world caption diversity) and Visual Genome (to evaluate dense region descriptions) for cross-domain generalization validation. These datasets add to MS-COCO and Multi30k by including a wider range of linguistic patterns, scene complexities, and object relationships. This makes sure that multilingual prediction is strong.

3.2. Preprocessing for multimodal translation

In preprocessing, pictures are minimized to 224x224 pixels and normalized using Min-Max normalization, while text is tokenized into significant pixels to facilitate embedding and sequential simulation.

Tokenization for text preprocessing: Tokenization prepares text for the TBiLSTM-CCO model by breaking up multilingual captions into words or phrases. This enables the net to align textual features with ResNet, which captures image representations for precise multimodal translation and picture description localization, and efficiently captures semantic structure.

Min-Max normalization for image preprocessing: Min-Max normalization ensures consistency among visual features collected by ResNet by scaling picture pixel values to the [0, 1] range. According to Equation (10, this enhances the TBiLSTM-CCO network's stability and convergence during multimodal instruction.

$$Y_{new} = \frac{max(Y)-min(Y)}{Y-min(Y)} \quad (1)$$

Where,$Y_{new}$is astandardized pixel value, Yrepresents theinventive pixel value, $Min(Y)$ denotes the least pixel value in the image, and $Max(Y)$is the extreme pixel value in the image. This preprocessing step conserves relative variances in pixel assets while preventing large-scale changes from affecting model training.

The proposed T-BiLSTM-CCO model was evaluated in relation to two public benchmark datasets, MS COCO and Multi30k, to be repeatable and open, yet they were referred to their original sources rather than to mirrored repositories. MS COCO is a collection of 123, 287 images with five citation captions. The data was separated into 82,783 training, 5,000 validation, and 5,000 test photographs according to the official 2017 division. Multi30k (EN/DE/FR) consists of 31,014 photos, and 5 descriptions are given to each translation. It used 29,000 to train it, 1,014 to validate it and 1000 to test it. The SpaCy tokenizer was used to process all comments, convert them to lower case and truncate them to a reasonable size of 20 tokens.

Image resizing: Pictures are adjusted to 224×224 using exclamation and center cropping for the TBiLSTM-CCO model. Interpolation maintains visual details, while center cropping pads smaller images with zeros, ensuring

uniform input for ResNet and facilitating effective fusion with textual embeddings. Figure 2 illustrates the resized and normalized images.

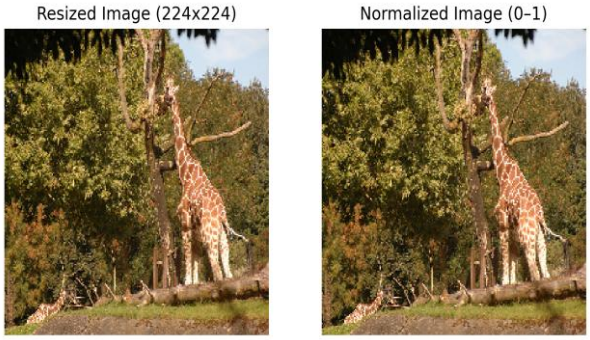

Figure 2: Resized and normalized images.

3.3. Multimodal feature extraction

Word2Vec encodes multilingual captions into dense semantic embeddings, capturing contextual and syntactic relationships for effective text representation. ResNet-50 extracts high-level visual features from images, preserving spatial and semantic information via residual connections.

3.3.1. Text feature extraction employing Word2Vec

Word2Vec is employed to extract meaningful textual embeddings from multilingual captions. The Skip-gram model is formulated as in Equation (2).

$$\mathcal{L}_{SG} = -\sum_{u=1}^{U}\sum_{-d\le k\le d,k\neq 0} log\sigma\left(\left(w'_{y_{u+k}}\right)w_{y_u}\right) - \sum_{j=1}^{l}\mathbb{E}_{x_j\sim Q_O(x)}\left[log\sigma\left(-\left(w'_{x_j}\right)w_{y_u}\right)\right] \quad (2)$$

Where, $\mathcal{L}_{SG}$ is the total Skip-gram function to be minimized during training, U is the entireamount of disputes (demonstrations) in the training corpus, y_u represents the center word at position u in the corpus, y_{u+k} is the context word located at position $u+k$ within the context window around y_u, d is the context window size, $w_{y_u}\in\mathbb{R}^O$ and $\left(w'_{y_{u+k}}\right)$are the input and output embedding vectors of the center words y_u and y_{u+k} respectively, $Q_O(x)$ denotes the noise distribution used, x_j is the negative sample word, $Q_O(x)$, $\left(w'_{x_j}\right)w_{y_u}$ is the dot product between the center and context embeddings.

3.3.2. Visual Feature Extraction using Renet50

The DL architectures can face issues such as vanishing gradients and optimization difficulties. ResNet addresses these challenges through identity shortcut connections, allowing information to flow directly across layers and preserving feature integrity. In this research, ResNet-50 is utilized to abstract high-level visual structures from descriptions for multilingual captioning. The network comprises 50 layers arranged in residual blocks (3, 4, 6, and 3) with identity mappings. A 3×3 max pooling layer follows the initial 64 7×7 convolutional filters. Fully connected layers are omitted so that the extracted feature maps can be fused directly with Word2Vec text embeddings in the TBiLSTM-CCO framework, enabling precise multimodal alignment and improved image-description localization. The feature extraction output is illustrated in Figure 3.

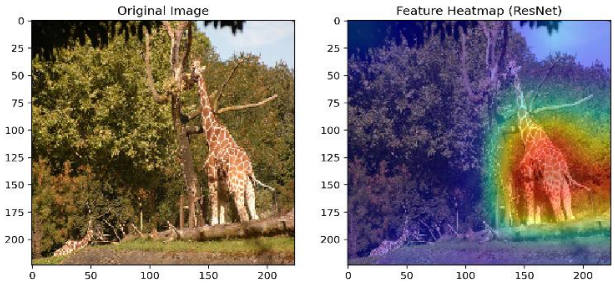

Figure 3: Visual feature extraction using ResNet.

3.4. Image description using transformer integrated bidirectional long short-term network with Cuckoo Catfish Optimizer

T-Bi-LSTM-CCO framework integrates Bi-LSTM to capture bidirectional sequential dependencies with a Transformer to model long-range semantic relationships.

3.4.1. Transformer-Based Bi-LSTM for Context-Aware Multilingual Image Description

A hybrid Transformer-BiLSTM architecture is used to context and sequential needs for multilingual image description. The Transformer captures long-range semantic relationships through self-attention, while the BiLSTM learns bidirectional sequence information, ensuring accurate alignment between textual and visual features.The approach uses late fusion to combine ResNet-152 visual features with Word2Vec embeddings via cross-attention. A Transformer-BiLSTM decoder generates captions, optimized with cross-entropy loss and label smoothing for improved generalization and semantic matching of images and multilingual descriptions. For precise caption creation, the decoder employs CIDEr-based reinforcement optimization and instructor forcing with cross-entropy loss.

The proposed framework carries out multimodal translation, in which target-language captions are produced using both textual and visual inputs. Clear NMT-context congruence and task-specific training consistency are ensured by the establishment of distinct supervised settings: Multimodal translation for text + image to target text and image captioning for visual-to-text mapping.

Bi-LSTM model: The Bi-LSTM network captures contextual dependencies in both directions within multilingual descriptions, mitigating gradient disappearing and explosion issues in long sequences.At each time step t, the LSTM unit receives the Input vector - u_t, the Previous hidden state - s_{t-1}, and the Previous cell state - Q_{t-1}.

At each time step, the BiLSTM updates its cell state and hidden output through forget, input, and output gates, determining which information to retain, store, or output from previous states using Equations (3–6). The forward and backward hidden positions are then concatenated to create a context-aware depiction for caption conversion, using Equation (7).

$$F_t = \sigma(W_F[s_t - 1, u_t] + b_F) \tag{3}$$

$$I_t = \sigma(W_I[s_t - 1, u_t] + b_I), \widetilde{Q_t} = tanh(W_c[s_{t-1} - 1, u_t] + b_c) \tag{4}$$

$$Q_t = F_t \odot Q_{t-1} - 1 + I_t \odot \widetilde{Q_t} \tag{5}$$

$$O_t = \sigma(W_O[s_t - 1, u_t] + b_O), s_t = O_t \odot tanh(Q_t) \tag{6}$$

$$s_t = [\overrightarrow{s_t}, \overleftarrow{s_t}] \tag{7}$$

BiLSTM can accurately align multilingual captions with images because it gathers relevant data derived from past and future phases.

Transformer model: By focusing on every word at once, the Transformer complements the sequential modeling of the Bi-LSTM by capturing global dependencies across multilingual formulations. In ensuring that words accurately translate and align with visual features, this guarantees that long-range semantic links are learned. The decoder employs hidden multi-head attention and feedforward levels for estimation, whereas the encoder consists of a self-attention layer and a two-layer feedforward network with residual links and regularization by batch. After the BiLSTM layer, the Transformer is included in the encoder to improve accuracy in translation and cross-modal alignment by refining contextual embeddings with multiple attentiveness heads. Positional Embeddings (PE) preserve word order, qualifying the model to maintain sequential information while leveraging global context as shown in Equation (8-10).

$$PE(pos, 2i) = sin(\frac{pos}{10000^{2i/D}}), PE(pos, 2i + 1) = cos(\frac{pos}{10000^{2i/D}}) \tag{8}$$

$$W_j^{new} = W_j + Y_1 \times |qc| \times \left(\frac{W_{best} + W_{q_1}}{2} - W_{q_2}\right) + \frac{q_3}{2} \times \left(W_{q_3} - W_{q_4}\right) \tag{9}$$

$$W_j^{new} = \begin{cases} W_f + E \times Q_1 \times \frac{step}{2} + S^m \times t \times (1 - Q_1) \times |step| + U \times \frac{I_j}{I_s} & if mod(j, 2) = 0 \\ W_f + E \times Q_1 \times \frac{step_1}{2} + S^m \times d \times (1 - Q_1) \times |step| + U \times \frac{I_j}{I_s} & otherwise \end{cases} \tag{10}$$

Where, $PE(pos, 2i)$ and $(pos, 2i + 1)$ is the positional encoding values for even $(2i)$ and odd $(2i + 1)$ embedding dimensions, i is the dimension index, W_j and W_j^{new} are the current solution and updated solution, Y_1 denotes the adaptive control parameter, $W_{q_1}, W_{q_2}, W_{q_3} and\ W_{q_4}$ are the randomly selected distinct individuals, W_{best} represents the best- solution, W_f is the fitness-weighted solution, E is the environment coefficient, Q_1 is the ransom scalar $\in$ (0,1), S^m is the scaling factor that adjusts the search range, t is the current iteration number, I_j and I_s are the fitness and best fitness score in the population. The proposed method fuses BiLSTM for sequential dependencies with Transformer for long-range attention, enabling effective multimodal feature fusion.

3.4.2. CCO for enhanced TBiLSTM performance in multilingual captioning

CCO emulates the natural predatory and cooperative nature of the Cuckoo Catfish, such as space compression, surround search, and adaptive exploitation, to successfully search the solution space and avoid merging into local minima.CCO can produce N population equally across space in order to begin the optimization process. Equation (1) represents a mathematical equation for the location initialization of each individual CCO [23].

$$X_i^d = R \times (U^d - L^d) + L^d, i = 1, \dots, N;\ d = 1, \dots, D \tag{11}$$

To verify the convergence effectiveness and optimize resilience of the CCO, evaluate its standalone optimization performance on common benchmark functions, and compare the outcomes with those of well-known optimizers like Particle Swarm Optimization (PSO), Genetic Algorithm (GA), and Adaptive Moment Estimation (Adam).To preserve optimization consistency and ensure ecologically possible motion for creatures, the CCO employs specified constraints and adaptive parameter sequences for the optimization phase process.

Constraints in CCO

Equations (12-13) specify the viable search boundaries and spiral motion in CCO, with Equation (11) restricting the search range and Equations (12) and (13) describing the vertical and horizontal shifts of the logarithmic helix, which guide responsive enclosure activity.

$$m_j \le Y_{J,K} \le v_j \quad (12)$$

$$d = b \times e^{(c.\theta)/2} \times \cos(\theta) \quad (13)$$

Where, m_j and v_j are the lower and upper boundaries of the search space, $Y_{J,K}$ is the position value of the Jth individual for the K th dimension of the population, b denotes the scaling constant that controls the amplitude of the logarithmic spiral path, c is the helix density coefficient and d is the helix density coefficient, adjusting the tightness or compression rate of spiral path.

Parameter schedule

The parameter scheduling Equations (14, 15, and 16) govern adaptive control in the CCO, which balances exploration and exploitation over iterations.

$$U = \left(1 - \sin\left(\frac{\pi}{2} \times \frac{Jk}{MaxIt}\right)\right)^{Jk/MaxIt} \quad (14)$$

$$E = U + s_1 \quad (15)$$

$$x = 1 - \frac{e^{\left(\frac{Jk}{MaxIt}\right)} - 1}{e - 1} \quad (16)$$

Where, $U$is the shrinkage factor, E is the kinetic energy coefficient, x denotes the adaptive radius adjustment coefficient, Jk represents the current iteration number, $s_1$is the random perturbation number. The CCO uses a novel hybrid process that balances exploration and exploitation by combining spiral-spherical surround search and reactive power decay. The proof of concept is validated through extensive validation testing and real-life optimization applications showing increased convergence stability, search efficiency, and response accuracy. Bi-LSTM models establish bidirectional sequential dependencies, while the Transformer model frames a long-distance attention mechanism at both within-sentence and across-sentence levels. The CCO applies adjustments to network weights and hyperparameters to improve convergence, translation accuracy, and multimodal feature alignment, and adaptive exploration and exploitation of CCO is specified in Pseudocode 1.

Pseudocode 1: Adaptive update mechanism of CCO

```
Input: objective f(z), bounds [ℓ,u], population size N, iterations T
Hyperparams: alpha0, beta (Levy), gamma0, gamma_max, rho1, rho2, sigma0, pa
Initialize population Z = {z^i}_{i=1..N} uniformly in [ℓ,u]
Evaluate F^i = f(z^i) for all i
z_best = argmin_iF^i

For t = 1..T:
  update schedules:
    alpha = alpha0 * (1 - t/T)^kappa
    gamma = gamma0 + (gamma_max - gamma0)*(t/T)
    sigma = sigma0 * exp(-lambda * t)
  For i = 1..N:
    if rand() < 0.5:
      s = alpha * LevySample(beta, d)
z_new = z^i + s
    else:
      pick distinct q1,q2 in {1..N}\{i} uniformly
z_new = z^i + gamma*(z_best - z^i) + rho1*(z^{q1} - z^{q2})
          + rho2 * Normal(0, sigma^2 I)
    clip z_new to [ℓ,u]
    eval f_new = f(z_new)
    if f_new< F^i:
z^i = z_new
F^i = f_new
      if f_new< F_best:
z_best = z^i; F_best = f_new
  m = ceil(pa * N)
  sort population by F
  for j in indices of worst m:
z^j = UniformSample([ℓ,u])
F^j = f(z^j)
    if F^j<F_best: update best

Return z_best, F_best
```

The T-Bi-LSTM-CCO model's hyperparameter setting is tabulated in the relevant Table 2.

Table 2: Hyperparameters of the Proposed TBiLSTM-CCO

Component	Hyperparameter	Description	Optimized Value
Transformer	Number of attention heads	Multi-head attention size	8
	Feed-forward dimension	Size of FFN in each Transformer block	512
	Number of encoder layers	Depth of Transformer encoder	4

	Dropout rate	Regularization	0.2
Bi-LSTM	Number of layers	Depth of Bi-LSTM	2
	Hidden units per layer	Number of neurons in each LSTM layer	128
	Dropout rate	Regularization	0.3
Training	Learning rate	Step size for optimizer	0.001
	Batch size	Number of samples per batch	64
	Epochs	Number of training iterations	50
CCO (Optimizer)	Population size	Number of candidate solutions	20
	Maximum iterations	Number of CCO iterations	50
	Hyperparameters to optimize	Learning rate, LSTM units, dropout, etc.	–

The TBiLSTM-CCO model combines BiLSTM and Transformer to jointly capture sequential dependencies and long-range attention in multilingual captions, while also ensuring semantic consistency and background accuracy. This type of hybrid model has the ability to deliver strong, context-rich multilingual descriptions of images by putting together two model types, while retaining the strengths and overcoming the weaknesses of each type for multilingual image description. Pseudocode 2 demonstrates TBiLSTM-CCO.

Pseudocode 2: TBiLSTM-CCO

```
import tensorflow as tf
from transformers import AutoTokenizer, TFAutoModel
import numpy as np
1. Hyperparameters & Setup
EMBED_DIM = 512
HIDDEN_UNITS = 256
POPULATION_SIZE = 30
MAX_EPOCHS = 100
LR = 0.001
BATCH_SIZE = 64
ALPHA, BETA = 0.5, 0.5  # Weight for loss components

2. Feature Extraction
def extract_visual_features(images):
    """Extract semantic embeddings from image using pretrained Transformer encoder (e.g., ViT)."""
visual_encoder = TFAutoModel.from_pretrained("google/vit-base-patch16-224")
visual_feats = visual_encoder(images).last_hidden_state
    return tf.reduce_mean(visual_feats, axis=1)  # [batch, embed_dim]

3. BiLSTM + Transformer Fusion
class TBiLSTM(tf.keras.Model):
    def __init__(self):
super(TBiLSTM, self).__init__()
self.bilstm = tf.keras.layers.Bidirectional(
tf.keras.layers.LSTM(HIDDEN_UNITS, return_sequences=True))
self.attention = tf.keras.layers.MultiHeadAttention(num_heads=4, key_dim=EMBED_DIM)
self.dense = tf.keras.layers.Dense(EMBED_DIM, activation='relu')

    def call(self, visual, text):
lstm_out = self.bilstm(text)
attn_out = self.attention(query=lstm_out, value=visual, key=visual)
        fused = tf.keras.layers.Concatenate()([lstm_out, attn_out])
      return self.dense(fused)
4.CCO
def cuckoo_catfish_optimizer(obj_func, bounds, iterations=50):
    N = POPULATION_SIZE
    D = len(bounds)
    population = np.random.uniform([b[0] for b in bounds],
                                   [b[1] for b in bounds],
                                   (N, D)
  for t in range(iterations):
spiral_factor = np.exp(-t / iterations)
new_pop = []
      for i in range(N):
rand_idx = np.random.randint(N)
          step = spiral_factor * (population[rand_idx] - population[i])
new_solution = population[i] + np.random.randn(D) * step
new_solution = np.clip(new_solution, [b[0] for b in bounds], [b[1] for b in bounds])
new_pop.append(new_solution)
        population = np.array(new_pop)
    best = population[np.argmin([obj_func(x) for x in population])]
    return best
5. Training Loop
```

```
def train_model(images, captions, labels):
visual_feats = extract_visual_features(images)
text_feats = embed_text(captions)
   model = TBiLSTM()
optimizer_params = cuckoo_catfish_optimizer(obj_func=lambda x: np.sum(x**2),
                                            bounds=[(0, 1)] * 10)
optimizer = tf.keras.optimizers.Adam(learning_rate=LR)

   for epoch in range(MAX_EPOCHS):
      with tf.GradientTape() as tape:
fused_output = model(visual_feats, text_feats)
         pred = tf.keras.layers.Dense(len(labels), activation='softmax')(fused_output[:, -1, :])
ce_loss = tf.keras.losses.SparseCategoricalCrossentropy()(labels, pred)
mse_loss = tf.reduce_mean(tf.square(visual_feats - tf.reduce_mean(text_feats, axis=1)))
total_loss = ALPHA * ce_loss + BETA * mse_loss

      grads = tape.gradient(total_loss, model.trainable_variables)
optimizer.apply_gradients(zip(grads, model.trainable_variables))
print(f"Epoch [{epoch+1}/{MAX_EPOCHS}], Loss: {total_loss.numpy():.4f}")
```

4 Result

Python 3.10 and TensorFlow 2.15 were used to carry out the implementation using an NVIDIA RTX 4090 GPU (24GB VRAM), 32GB RAM, and Ubuntu 22.04. Batch size 64, learning rate 0.001, and 100 training epochs were utilized in the experiments to ensure convergence stability. Figure 4 depicts the output of the image description of the proposed T-Bi-LSTM-CCO. The proposed method correctly characterizes the image, which shows a giraffe consuming leaf from a tree. The model accurately recognizes important visual aspects and their semantic links, demonstrating effective visual-textual connections.

Image Description: A giraffe is reaching up to eat leaves from a barren tree, while another giraffe is lying down in the background. The scene suggests a zoo or animal sanctuary, where giraffes are kept in an enclosure that mimics their natural habitat with trees and grass.

Figure 4: Image description by proposed TBiLSTM-CCO

BLEU-1 to BLEU-4 for linguistic quality, CIDEr and METEOR for semantic alignment, and localization accuracy for evaluating the correctness of visual-textual matching across multilingual datasets are some of the quantitative metrics used to operationalize performance.Comparative multinomial is depicted in Table 3.

Table 3: Multilingual evaluation

Language Pair	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	SPICE
English–German	0.90	0.87	0.85	0.83	0.72	0.48
German–French	0.88	0.85	0.82	0.80	0.70	0.46
French–English	0.89	0.86	0.83	0.81	0.71	0.47

4.1. Performance evaluation

The results highlight the superior semantic alignment, computational efficiency, and accuracy of the TBiLSTM-CCO model compared to existing baselines. The accuracy was evaluated using the BLEU score, Equations (12 & 13).

$$BLEU = BP \cdot exp\left(\sum_{n=1}^{N} w_n \, log p_n\right) \tag{12}$$

$$BP = \begin{cases} 1 & if\, c > r \\ e^{(1-r/c)} & if\, c \le r \end{cases} \tag{13}$$

Here, w_n is the weight assigned to each n-gram, precisionp_n, p_n is the modified n-gram precision, c is the total length of all candidate translations, r denotes the total effective reference length, $e^{(1-r/c)}$ is the exponential penalty term, $BP \cdot exp\,(.)$ is the product of the brevity penalty and the geometric mean of n-gram precisions.

Table 4 compares existing models such as Nearest Neighbor, Google's Neural Image Captioning (NIC), Long-term Recurrent Convolutional Networks (LRCN), Attention-based Image Captioning with ResNet50 (AICRL-ResNet50), and Bi-LSTM with Attention, showing that the proposed T-Bi-LSTM-CCO model achieves the highest BLEU scores: BLEU-1: 0.90, BLEU-2: 0.89, BLEU-3: 0.87, BLEU-4: 0.85. Figure 5 (a) The proposed T-BiLSTM-CCO outperforms all baselines and earns the best BLEU scores, and Figure 5(b) A 3D accuracy comparison between CNN and hybrid models is shown in Figure 6, highlighting the 92.35% accuracy of the suggested T-BiLSTM-CCO.

Table 4: Comparison of BLEU Scores across existing and proposed methods for Multilingual Image Description Models

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4
Nearest neighbor [21]	0.48	0.281	0.166	0.1
Google NIC [21]	0.66	0.461	0.329	0.246
LRCN [21]	0.62	0.442	0.304	-
AICRL-ResNet50 [21]	0.731	0.562	0.41	0.326
BiLSTM + Attention [21]	0.758	0.734	0.746	0.73
TBiLSTM-CCO [Proposed]	**0.90**	**0.89**	**0.87**	**0.85**

BLEU-1–4, METEOR, ROUGE-L, and CIDEr scores with mean ± SD were depicted in Table 5.

Table 5 :5-fold cross-validation of the proposed

Fold	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr
1	0.902	0.888	0.872	0.852	0.468	0.691	1.303
2	0.898	0.884	0.869	0.849	0.462	0.687	1.296
3	0.907	0.891	0.876	0.856	0.471	0.694	1.312
4	0.905	0.890	0.875	0.854	0.469	0.692	1.309
5	0.899	0.885	0.870	0.850	0.466	0.689	1.301
Mean ± SD (T-BiLSTM-CCO)	**0.902 ± 0.003**	**0.888 ± 0.003**	**0.872 ± 0.003**	**0.852 ± 0.003**	**0.467 ± 0.003**	**0.691 ± 0.002**	**1.304 ± 0.006**

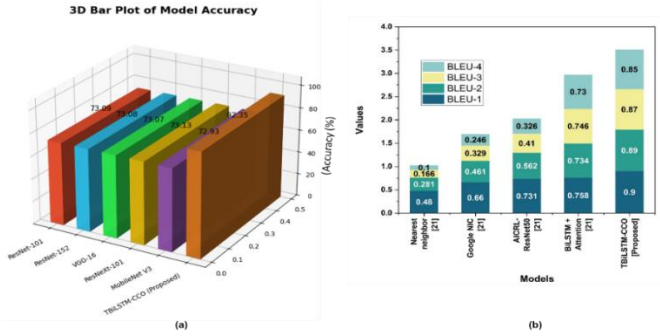

Figure 5: Evaluation of (a) Comparative BLEU Score Analysis and (b) Performance analysis of the proposed model

Tokenized and detokenized pictures from the Multi30k dataset were examined for multilingual caption evaluation in English, German, and French. Metrics including METEOR, ROUGE-L, CIDEr, and SPICE were used in addition to BLEU-1–4 to measure linguistic, technical, and conceptual accuracy across languages (Table 6).

The analysis provides thorough findings for all metrics: BLEU-1: 0.90, BLEU-2: 0.89, BLEU-3: 0.87, BLEU-4: 0.85, METEOR: 0.76, ROUGE-L: 0.81, CIDER: 1.31, and SPICE: 0.68. BLEU-n reduces monotonically in our evaluation, indicating consistent and trustworthy metric computation for all models.

Table 6: Qualitative attention visualizations and multilingual analysis

Image	Ground Truth Caption	Generated Caption (English)	Generated Caption (German)	Generated Caption (French)
	A dog playing with a ball in the park.	A dog plays with a ball.	Ein Hund spieltmiteinem Ball.	Un chienjoue avec uneballe.
	A car parked beside a busy street.	A car is parked on the road.	Ein Auto parkt auf der Straße.	Une voiture est garée près de la route.

Table 7compares different models in terms of accuracy, loss, training time, inference time, and GMACs. Table 8 compares the FLOPS and model size [23].

Table 7: Comparison of model accuracy for multimodal image captioning

Model	Feature alignment Accuracy (%)	Loss	Training Time (Hours)	Inference (Seconds)	GMAC's
ResNet-101 [22]	73.092	3.413	5.6046	0.10765	7.85
ResNet-152 [22]	73.077	3.412	6.4077	0.14021	11.58
VGG-16 [22]	73.069	3.413	4.8353	0.08430	15.38
ResNeXt-101 [22]	73.128	3.404	7.8939	0.11023	16.5
MobileNet V3 [22]	72.928	3.424	3.5379	0.07975	0.23
TBiLSTM-CCO [Proposed]	**92.35**	**1.871**	**4.21**	**0.075**	**6.8**

Table 8: Computational complexity and Model size comparison among proposed and existing

Models	FLOPs	Model size
ResNet50 [23]	12.46 m	124.7
VGG16 [23]	12.99 m	131
XCIT [23]	12.14 m	120.7
SWIN [23]	12.33 m	123.1
ViT [23]	12.20 m	121.5
Proposed	12.50 m	125.9

The ResNet-152 backbone was used to extract visual features from the penultimate convolutional layer (res5c). The retrieved visual characteristics were fused with textual insertions using a Transformer-BiLSTM fusion layer, then optimized with the CCO. The system was trained with 64 batches, a learning rate of 0.0002, and a cross-entropy loss goal. The experiments were conducted five times employing various arbitrary seeds (42, 101, 123, 202, 404), and the results are provided as mean ± standard deviation.

Figure 6 (a) indicates the performance of the models with training time and model size shown that, while on the other models in the table, such as ResNeXt-101, which had a greater training time with 7.89 hours, the TBiLSTM-CCO model outperformed in computational performance with 4.21 GMACs and outperformed in loss value with a loss of 1.87 against these other methods. Figure 6 (b) displays the inference time of different models, and as shown, the proposed TBiLSTM-CCO model had a very fast inference time of 0.075 seconds, showing how efficient the live feed could generate real-time predictions.

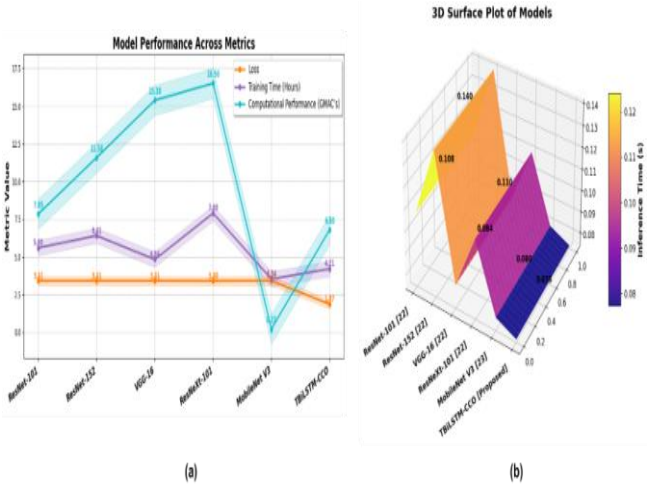

Figure 6: Performance of (a) evaluation metrics and (b)model inference time analysis

Table 9 presents a comparison of model performance based on prediction accuracy. The proposed T-B-iLSTM-CCO model achieves the highest performance with 87.4% completely correct predictions, 12% partially correct predictions, and only 0.6% incorrect predictions. Figure 7 examines the results of qualitative predictions. TBiLSTM-CCO outperforms previous systems (82.6% and 84.8%), demonstrating greater caption accuracy and semantic congruence with 87.4% fully accurate, 12% partially correct, and only 0.6% wrong outputs.

Table 9: Comparative analysis of correctness metrics

Model	Completely Correct (%)	Partially Correct (%)	Incorrect (%)
Bi-LSTM	82.6 %(413/500)	14.4%(72/500)	3.0%(15/500)
Transformer Integrated Bi-LSTM	84.8%(424/500)	13.4%(67/500)	1.8%(9/500)
T-BiLSTM-CCO [Proposed]	**87.4%(437/500)**	**12%(60/500)**	**0.6%(3/500)**

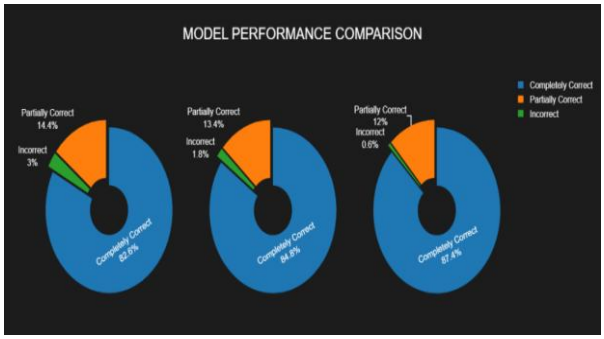

Figure 7: Qualitative assessment of model output

The ablation study assessing the impact of various architectures, optimizers, and training modes on multilingual captioning is shown in the table. The suggested Transformer + BiLSTM + CCO had the lowest loss (1.87) and the highest CIDEr (1.31). To confirm constant performance, future scaling testing using different optimizers and additional languages are advised. Ablation research with 95% confidence intervals for BLEU-4, were depicted in Table 10.

Table 10: Comprehensive ablation and optimization analysis of the T-BiLSTM-CCO

Architecture	**Optimizer**	**Mode**	**CIDEr**	**Loss**	**95% CI (BLEU-4)**
BiLSTM only	Adam	Multilingual	0.95	2.34	±0.012
Transformer only	Adam	Multilingual	1.05	2.12	±0.010
Transformer + BiLSTM + CCO	CCO	Multilingual	**1.31**	**1.87**	**±0.007**
Transformer + BiLSTM	RMSprop	Multilingual	1.18	1.96	±0.008
Transformer + BiLSTM + CCO	CCO	Monolingual	0.87	1.94	±0.009

5 Discussion

The previous system [21] applied BiLSTM and DBN to text summarization and caption generation; however, the semantic accuracy of the system was limited by the lack of multimodal involvement and the lack of feature representation. The current system [22], which uses the CNN and Transformer-based encoders, has been efficient up to this point due to its excessive computational requirements, reliance on particular optimizers, and lack of scalability and contextual understanding. These issues are addressed by the suggested TBiLSTM-CCO model by integrating BiLSTM to model sequential context, Transformer layers to achieve long-range attention, and CCO to optimally parameterize the model. This combined design gives it a higher semantic alignment, reduces computation, and attains great accuracy on dissimilar data. **Real-world applicability:** The control methods, such as output-feedback controller and robust neuronal adaptive control, showed real-world benefits in secure translation systems, laser, and biological signal regulation for dealing with unstable irregularities. They also work well in automated vehicles, robotic manipulators, and aerospace flight control, where strong neurological adaptive control ensures accuracy, disturbance mitigation, and stability in changing and unpredictable conditions. Dynamic control systems address persistent temporal shifts through nonlinear optimum control and high-gain observer-based adaptive fuzzy control. T-BiLSTM-CCO applies this to video labelling and real-time multimodal translation, effectively managing sequential visual pixels and verbal signals. The model adapts to scene changes, producing relevant evaluations for multimedia through adaptive performance and attention mechanisms, enhancing interactivity.

Table 11 compares the proposed T-BiLSTM-CCO framework for validated performance coherence with the goals, outcomes, and constraints of current multimodal models. To ensure fair comparison, all baseline models were trained using the same experimental parameters, such as learning rate, batch size, optimizer, and dataset splits.

Table 11 : Comparative analysis of SOTA multilingual image captioning models

Ref	**Objective**	**Method / Result Summary**	**Limitation**
[10]	Evaluate visual feature impact	Compared NMT vs. MMT; found NMT sometimes outperforms MMT for short sentences.	Visual features offer limited benefit in simple cases.
[13]	Multimodal fusion for captioning	Faster RCNN + LSTM fusion; reported 95% relative performance improvement.	High computational cost; requires large annotated data.
[14]	Target–Visual Consistency in MNMT	Applied TVC & BiVC; BLEU-4: 37.8, METEOR: 31.4.	Requires bilingual-visual alignment data; less scalable.
[17]	Video-guided translation	Multi-pattern joint learning; +4.2 BLEU improvement.	High computational expense; dataset-specific.

[20]	Real-time captioning (Neuraltalk+)	Faster convergence, BLEU-4: 36.7 on Flickr8k/30k.	Restricted to English datasets; outdated architecture.
Proposed (T-BiLSTM-CCO)	Multilingual image captioning &translation	Combines Transformer, BiLSTM, and Cuckoo Catfish Optimizer; achieves BLEU-1: 0.90, and CIDEr: 1.31	Needs validation on additional multilingual datasets for broader generalization.

6 Conclusion

T-Bi-LSTM-CCO effectively solves the problems of traditional caption generation methods. By utilizing a BiLSTM component to model sequential context, Transformer layers to model long-distance dependencies, and CCO to fine-tune the entire network parameters, the proposed framework provides better performance. Experimental performances demonstrate that the proposed approach achieves 87.4% completely correct predictions, at most 0.6% incorrect predictions, boasts a slightly higher BLEU score than the baseline, and has noticeably shorter run time than the baseline. The approach can be applied to assistive technologies, automated annotation of multimedia content, content cataloguing and indexing, and human–computer interaction. However, it relies on large annotated datasets and is complex to noisy inputs. Future work will focus on improving model interpretability, incorporating self-supervised pretraining, and extending capabilities to real-time multimedia and cross-lingual applications.

Declarations

Ethics approval and consent to participate: I confirm that all the research meets ethical guidelines and adheres to the legal requirements of the study country. There were no human subjects, personally identifiable information, or experiments that needed ethical approval in this study. All of the benchmark datasets utilized in this study, including the Multi30k and COCO (Common Objects in Context) datasets, are freely accessible and made accessible under open academic licenses. There are no ethical dilemmas or privacy issues because their use in this work complies with the relevant dataset terms.

Data availability statement: The following official repositories make the datasets used in the research accessible to the general public: **COCO Dataset:** https://cocodataset.org/. **Multi30k Dataset:** https://github.com/multi30k/dataset
The authors did not gather or create any additional data; all data were used exclusively for academic study in compliance with dataset permissions.
Consent for publication: I confirm that any participants (or their guardians if unable to give informed consent, or next of kin, if deceased) who may be identifiable through the manuscript (such as a case report) have been given an opportunity to review the final manuscript and have provided written consent to publish.

Availability of data and materials: The data used to support the findings of this study are available from the corresponding author upon request.
Competing interests: here are no have no conflicts of interest to declare.

Authors' contributions (Individual contribution): All authors contributed to the study conception and design. All authors read and approved the final manuscript

References

[1] Cho, S., & Oh, H. (2023). Generalized image captioning for multilingual support. *Applied Sciences*, *13*(4), 2446. https://doi.org/10.3390/app13042446

[2] Meetei, L. S., Singh, S. M., Singh, A., Das, R., Singh, T. D., & Bandyopadhyay, S. (2023). Hindi to English multimodal machine translation on news dataset in low resource setting. *Procedia Computer Science*, *218*, 2102–2109. https://doi.org/10.1016/j.procs.2023.01.186

[3] Liu, X., Zhao, J., Sun, S., Liu, H., & Yang, H. (2021). Variational multimodal machine translation with underlying semantic alignment. *Information Fusion*, *69*, 73–80. https://doi.org/10.1016/j.inffus.2020.11.011

[4] Zhao, Y., Komachi, M., Kajiwara, T., & Chu, C. (2022). Region-attentive multimodal neural machine translation. *Neurocomputing*, *476*, 1–13. https://doi.org/10.1016/j.neucom.2021.12.076

[5] Wu, X., & Li, T. (2023). Sentimental visual captioning using multimodal transformer. *International Journal of Computer Vision*, *131*(4), 1073–1090. https://doi.org/10.1007/s11263-023-01752-7

[6] Munz, T., Väth, D., Kuznecov, P., Vu, N. T., & Weiskopf, D. (2022). Visualization-based improvement of neural machine translation. *Computers & Graphics*, *103*, 45–60. https://doi.org/10.1016/j.cag.2021.12.003

[7] Xu, C., Yu, Z., Shi, X., & Chen, F. (2023). Adding visual attention into encoder-decoder model for multi-modal machine translation. *Journal of Engineering Research*, *11*(2), 100077. https://doi.org/10.1016/j.jer.2023.100077

[8] Das, B., Pal, R., Majumder, M., Phadikar, S., & Sekh, A. A. (2023). A visual attention-based model for Bengali image captioning. *SN Computer Science*, *4*(2), 208. https://doi.org/10.1007/s42979-023-01671-x

[9] Meetei, L. S., Singh, T. D., & Bandyopadhyay, S. (2024). Exploiting multiple correlated modalities can enhance low-resource machine translation quality. *Multimedia Tools and Applications, 83*(5), 13137–13157. https://doi.org/10.1007/s11042-023-15721-2

[10] Cui, S., Duan, K., Ma, W., &Shinnou, H. (2024). Does multimodal machine translation improve translation performance? *Neural Computing and Applications, 36*(22), 13853–13864. https://doi.org/10.1007/s00521-024-09705-y

[11] Sree, M. R., Siddhartha, M., Reddy, P. V. V., & Singh, R. P. (2025). A residual network and bi-directional LSTM based hybrid approach to remote sensing image captioning. *Procedia Computer Science, 258*, 88–97. https://doi.org/10.1016/j.procs.2025.04.198

[12] Zhao, Y., Komachi, M., Kajiwara, T., & Chu, C. (2021). Word-region alignment-guided multimodal neural machine translation. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 30, 244-259.https://doi.org/10.1109/TASLP.2021.3138719

[13] Thangavel, K., Palanisamy, N., Muthusamy, S., Mishra, O. P., Sundararajan, S. C. M., Panchal, H., Loganathan, A. K., & Ramamoorthi, P. (2023). A novel method for image captioning using multimodal feature fusion employing mask RNN and LSTM models. *Soft Computing, 27*(19), 14205–14218. https://doi.org/10.1007/s00500-023-08448-7

[14] Liu, Y., Liu, D., & Zhu, S. (2024). Bilingual–visual consistency for multimodal neural machine translation. *Mathematics, 12*(15), 2361. https://doi.org/10.3390/math12152361

[15] Liu, X., Wei, F., Jiang, W., Zheng, Q., Qiao, Y., Liu, J., ... & Dong, H. (2023). MTR-SAM: visual multimodal text recognition and sentiment analysis in public opinion analysis on the internet. Applied Sciences, 13(12), 7307. https://doi.org/10.3390/app13127307

[16] Mohammed, R., Aljarrah, I., Al-Ayyoub, M., & Fadel, A. (2025). Multimodal Multisource Neural Machine Translation: Building Resources for Image Caption Translation from European Languages into Arabic. Computation, 13(8), 194. https://doi.org/10.3390/computation13080194Chen, S., Zeng, Y., Cao, D., & Lu, S. (2022). Video-guided machine translation via dual-level back-translation. *Knowledge-Based Systems, 245*, 108598. https://doi.org/10.1016/j.knosys.2022.108598

[17] Zhiliang, Z., Lei, W., & Qiang, L. (2025). A method for real-time translation of online video subtitles in sports events. *Signal, Image and Video Processing, 19*(2), 146. https://doi.org/10.1007/s11760-024-03606-2

[18] Esteban-Romero, S., Martín-Fernández, I., Gil-Martín, M., & Fernández-Martínez, F. (2025). Synthesizing olfactory understanding: Multimodal language models for image–text smell matching. *Symmetry, 17*(8), 1349. https://doi.org/10.3390/sym17081349

[19] Sharma, H., &Padha, D. (2025). Neuraltalk+: Neural image captioning with visual assistance capabilities. *Multimedia Tools and Applications, 84*(10), 6843–6871. https://doi.org/10.1007/s11042-024-19259-9

[20] Dataset: 1. https://www.kaggle.com/datasets/mnassrib/ms-coco. 2. https://www.kaggle.com/datasets/shahadhamza/multi30k-dataset/data

[21] Mahalakshmi, P., & Fatima, N. S. (2022). Summarization of text and image captioning in information retrieval using deep learning techniques. *IEEE Access, 10*, 18289–18297. https://doi.org/10.1109/ACCESS.2022.3150414

[22] Castro, R., Pineda, I., Lim, W., & Morocho-Cayamcela, M. E. (2022). Deep learning approaches based on transformer architectures for image captioning tasks. *IEEE Access, 10*, 33679–33694. https://doi.org/10.1109/ACCESS.2022.3161428

[23] Elbedwehy, S., Medhat, T., Hamza, T., &Alrahmawy, M. (2022). Efficient image captioning based on vision transformer models. Comput, Mater Continua 73 (1): 1483–1500.