International Journal of Engineering Trends and Technology
ISSN: 2231–5381 / https://doi.org/10.14445/22315381/IJETT-V74I1P126

Volume 74 Issue 1, 333-354, January 2026
© 2026 Seventh Sense Research Group®

Original Article

AttenTAVO-Cap: A Hybrid Deep Learning and Metaheuristic Approach for Image Captioning

Chengamma Chitteti[1], K. Reddy Madhavi[2]

[1,2]School of Computing, Department of CSE, Mohan Babu University, Andhra Pradesh, Tirupati, India.

[1]Corresponding Author : sailusrav@gmail.com

Received: 04 August 2025 Revised: 17 November 2025 Accepted: 25 November 2025 Published: 14 January 2026

Abstract *- Image captioning, a problem at the intersection of natural language processing and computer vision, remains a difficult problem due to the inherent challenge in converting visual semantics to semantically rich text descriptions. Metaheuristic optimization in combination with neural network architectures has recently been shown to have excellent potential in bridging this gap. In this work, we present AttenTAVO-Cap, a novel hybrid image captioning model integrating an Attention-based Convolutional Neural Network (CNN) and Bi-directional Gated Recurrent Unit (Bi-GRU) architecture with the recently proposed Taylor African Vulture Optimization (TAVO) algorithm. The TAVO algorithm, inspired by African vultures' cooperative hunting behavior and augmented by Taylor series convergence properties, is utilized to optimize model hyperparameters very effectively. To completely assess the performance, experiments were conducted on two benchmark standards, Flickr8k and Flickr30k, with three versions of optimizers: TAVO, Genetic Algorithm (GA), and Particle Swarm Optimization (PSO). The outcome validated that AttenTAVO-Cap (TAVO) performed better than all the other models on a suite of evaluation metrics overall, with a BLEU-4 score of 0.29, METEOR of 38, and CIDEr of 194 and ROUGE-L of 67 on the Flickr8k corpus, and 0.29, 35, 191, and 63, respectively, on Flickr30k. Compared to baseline approaches, such as HABGRU + AVOA, the approach outlined here made considerable improvements, especially in semantic alignment and human-consensus-based measures. Results exhibit that hybrid Deep Learning (DL) and nature-inspired optimization can produce captions that are more accurate and human-like. Additionally, the present study provides possibilities to explore the explainability and generalizability of captioning models.*

Keywords *- Deep Learning, Flickr8k, Flickr30k, Genetic Algorithm (GA), Image Captioning, Metaheuristic Optimization, RoBERTa Embeddings, Taylor-African Vulture Optimization Algorithm (TAVO), Particle Swarm Optimization (PSO), Neural Architecture Optimization, Visual Attention, Bidirectional LSTM (BiLSTM).*

1. Introduction

In the fields of computer vision and natural language processing, image captioning, the act of describing an input image in text, is a well-studied research topic. It aims to convert more expressive visual information into natural language so that images can be processed by machines like humans. This capability is crucial in assistive technology for the blind, intelligent image search, self-driving cars, and human-robot communication. Early work in this domain utilized encoder-decoder models, where CNNs were employed to learn hierarchical spatial and semantic features from images and Recurrent Neural Networks (RNNs), specifically Long Short-Term Memory (LSTM) units, to generate coherent sequences of text [11, 15]. Though they performed well at first, these approaches were thoroughly challenged in model parameter tuning for good models, semantic faithfulness of the produced captions, and scalability to large and varied datasets. Current advancements in image captioning have followed two broad trends: metaheuristic model tuning for deep learning configurations and vision-language models based on transformers. Metaheuristic techniques based on natural inspiration, such as the African Vulture Optimization Algorithm (AVOA) or its Taylor-series improved version, Taylor-series enhanced VOA (TAVOA), have been rendering good results when combined with highly capable CNN architectures such as InceptionResNetv2 or sequence models such as hybrid attention bidirectional GRUs. There are many advantages to metaheuristic model tuning over gradient-based model tuning in gradient descent, which include combating overfitting, escaping the local minima, or facilitating smooth objectives even for complex, high-dimensional, and non-convex objective functions as exist in training tasks. Metaheuristic-based image captioning models remain relatively uninvestigated, especially when combined with reliable semantic representations of language. At the same time, transformer models have radically changed the field of image captioning with the introduction of self-attention mechanics that are well-suited for modeling complex interactions between visual and text data. Models incorporating multi-view vision features, attention layers, and vision-and-language scale AI pretraining models achieved new state-of-the-art performance on scaled-up vision-and-language corpora such as the MS COCO or Flickr30k data sets [1-3]. External knowledge graphs can also be leveraged to anchor image captions in real-world knowledge [4]. Nonetheless, such models appear to require truly massive data sets, intensive computation, and complicated models for effective training, which might make these models less useful or scalable in real-world applications [5].

The complexity of the model, the efficiency and explainability of the training, and the deployability process for the final solution. Experiments comparing CNN encoders with encoder-decoders imply that while there are some tiny improvements in accuracy, the computation and flexibility increase [6, 8]. Another stream underscores the challenge to scale vision-language pretraining to hundreds of millions of image-text pairs, as demonstrated in the LEMON model, which achieves state-of-the-art performance on benchmarks like COCO and nocaps [7]. Some research on semantic ordering, captioning for multiple languages, and improved decoding focuses on the coherence and grammatical correctness of the story. However, these problems are typically addressed through multiple, disconnected parts of the captioning process, rather than through the simultaneous increase in semantic complexity and efficiency of the solution. Existing image captioning models rely either on huge transformers and associated high computational costs or on deep learning models that lack adaptive global optimization of hyperparameters and models. Look for a framework that encompasses strong image abstraction learning, semantically powerful language modeling, adaptive decoding guided by attention mechanisms, and efficient global hyperparameter optimization. Even with significant advancements in image captioning, current methods still have to balance practical deployability, computational economy, and caption quality. Furthermore, there is still a lack of research on the effects of various metaheuristic optimization techniques on training dynamics and caption production quality in a controlled architectural environment.

The following research questions should now be clearly summarized in order to address this:

1. In order to improve semantic alignment and caption quality, exactly how may metaheuristic optimization be appropriately combined with deep image captioning structures?
2. Can TAVO produce measurable advantages over traditional metaheuristics like GA and PSO for captioning model tuning?
3. In what ways does the suggested hybrid architecture balance computational viability, performance, and generalization on the benchmark?

The following are this study’s principal contributions:

- We present AttenTAVO-Cap, a hybrid image captioning architecture that combines InceptionResNetv2-based visual encoding, RoBERTa-based semantic embeddings, and an attention-guided BiLSTM decoder optimized using the TAVO algorithm.
- We present a metaheuristic-based hyperparameter optimization method designed explicitly for caption decoding, enabling sufficient global exploration along with local refinement without using gradients.
- We will conduct a controlled comparison using the same architecture but with different optimizers (TAVO, GA, and PSO), isolating the impact of caption optimization strategies.
- On the Flickr8k and Flickr30k datasets, we continuously improve human-consensus-based metrics, including CIDEr and METEOR, indicating sounder expression and semantic relevance in captions.

In contrast to transformer-based models that require substantial training, AttenTAVO-Cap uses lightweight, expressive hybrid modeling to deliver competitive and often superior semantic performance. The suggested TAVO-based tuning technique is superior to gradient-based optimization because it prevents local optima from forming and enables the framework to achieve better generalization.

2. Literature Review

Captioning in images involves integrating natural language processing and computer vision to generate natural-language descriptions of images. In the early stages, most of the work involved using encoder-decoder networks, CNNs for image feature extraction, and an RNN, particularly an LSTM, for captioning. However, despite significant advancements, challenges remain in model parameter optimization, semantic consistency, and size and scaling. The current DL model renaissance, combined with the impact of metaheuristic optimization and the rise of transformer models, has had an extraordinary influence on captioning in the current period.

Metaheuristics have entered the mainstream as effective alternatives to gradient-based methods for fine-tuning deep neural networks for image captioning. The Attend More Times (AMT) framework was presented by Du et al. [1] to improve photo captioning performance by leveraging visual attention on a constant basis during the prediction of the next word. The goal of this model is to use a dual-LSTM decoder and a CNN-based encoder to extract image features. The model consistently focuses on the visual regions of images and shows relevance in boosting the semantic aspect of image captioning using the MS-COCO dataset. The repeated attention to visual areas is relevant for increasing the semantic accuracy of images and improving the descriptive accuracy of captioning. The model improves upon previous baselines using BLEU-4, METEOR, ROUGE-L, and CIDEr. Later, Castro [2] and Yu [4] turned to transformers and leveraged the advantages of self-attention mechanisms. The idea focused on avoiding the need for models to depend on each other or their interactions. The Multimodal Transformer model by Yu has focused on multi-view representations of images and has based its approach on enhanced deep interactions with cognitive semantics. The focus of this model has been on improved performance on the MS COCO benchmark and on establishing cognitive semantics through deeper interactions. The concept has been further enhanced by Huang [4], who focuses on external knowledge graphs for knowledge reasoning. The focus of this concept has been on strengthening semantic matching between image and text areas. Zeng [5] has further contributed to this concept by focusing on enhancing human-computer interaction. The idea has focused on improving domain shifts and the unobserved data in deep learning models.

Extensive evaluations by Xu [14] and Stefanini [6] offer crucial summaries that put the quick development in picture captioning into perspective. Stefanini's survey closely examines the development of training paradigms, language-generating models, and visual encoders. It emphasizes how critical multimodal connections and BERT-like early-fusion methods have been to recent advances. Xu's research advances this vision by exploring use beyond natural images, specifically medical image captioning, where domain-specific challenges such as semantic complexity and interpretability heavily affect the model structures that need to be converted. The scalability challenges unearthed by Hu [7], who empirically studied record-scale vision-language pretraining, also shed light on the trade-offs among model size, data size, and generalization. Hu's state-of-the-art LEMON model, which was trained on 200 million transformer-sized image-text pairs with up to 675 million parameters, also highlights several computational costs and data curation issues associated with training at this scale.

The parallel efforts by Li [8] focus on the semantic ordering and understanding of complex linguistic structures within images, investigating architectures that explicitly represent semantic coherence to enhance caption quality. More specialized studies by Mahajan [9], Manikumar [10], and Maaz [11] have examined performance variability induced by architectural modifications using different CNN backbones, such as Inception V3, ResNet, and VGG16/19, as well as encoder-decoder fusion approaches on metric performance.

They note that more complex models tend to be more accurate, but their training time, computational cost, and data size significantly impact deployment in real-world scenarios. Xia [12] sought to fill a crucial gap in low-resource, multilingual captioning by developing mechanisms for fusion attention to generate image captions in Tibetan, thereby opening the door to further integration of linguistic experiences in the development of captioning systems. Lastly, Nguyen [13] proposes a CNN-LSTM hybrid model with beam search decoding, demonstrating that decoding methods are crucial for determining the semantic richness and grammaticality of output captions, thereby increasing BLEU scores on the Flickr8k corpus.

All in all, these papers empower a multifaceted trajectory of advancement in image captioning, including architectural innovation, optimisation methods, increased data, and multilingualism. Whereas the best current practices in captioning models are deep convolutional and transformer models, other essentials driving performance frontiers include reliance on metaheuristic optimization, knowledge reasoning from the global world, and horizontally scalable training pipelines. While improving the balance between computational tractability issues, semantic consistency, and domain-invariant generalization requires further effort to build models not just that learn but also understand and generalize across a vast range of environments, this line of study forms the foundational land for new hybrid techniques that will take the strengths of existing paradigms and give up their inherent weaknesses to open the field for future research to take further toward more generalizable and human-like image captioning models.

The research makes clear that there has been a gradual evolution from traditional to more complex architectures that incorporate attention, leverage knowledge beyond current training data, pretrain on large datasets, and employ metaheuristic optimization. There are still problems with interpretation, efficiency, semantic coherence, and achievement in new domains, even though transfer interpretability of convolutional algorithms is now frequently used to provide state-of-the-art results. To make image captioning jobs more human-like and generalizable, some issues—such as hybrid models that combine powerful visual features with complex language models, adaptive decoding, and global optimization algorithms—need further investigation.

2.1. Summary of Gaps and Challenges

Despite immense advances in DL architecture-based image captioning, metaheuristic optimization, and vast corpora, numerous essential gaps and challenges remain to hinder the creation of extremely robust, generalizable, and semantically coherent captioning systems.

2.1.1. Computational Complexity and Scalability

Advanced models, particularly large transformer-based models pretrained on extremely large image-text datasets [7], [3], are computationally and memory-hungry. This is especially problematic for deployment in resource-constrained environments, such as mobile phones or real-time systems. Lightweight models and efficient training are an open research direction.

2.1.2. Semantic Consistency and Contextual Understanding

While knowledge graph integration and multi-view attention mechanisms have improved semantic alignment between generated captions and visual information [4, 8], Deep contextual comprehension and narrative coherence over longer captions are still difficult. Models tend to generate generic or somewhat relevant captions in complex scene layouts or domain transfer [5]

2.1.3. Domain Generalization and Robustness

Most caption models are trained and tested on precisely selected datasets like Flickr8k, Flickr30k, or MS COCO, and this restricts their performance on out-of-distribution, diverse, or real images. Domain shift and robustness against novel data distributions remain issues that constitute an ongoing challenge [5, 6].

2.1.4. Multilingual Support and Scarcity of Data

While large datasets can facilitate powerful learning, technical fields (e.g., medical or cultural) and low-resource languages have limited captioned image datasets. Recent studies, such as Tibetan captioning with fusion attention mechanisms [12], highlight the need for adaptive structures that can learn effectively in data-scarce scenarios.

2.1.5. Balancing Accuracy and Interpretability

Deep models, especially those involving the employment of attention and transformer modules, work very much like black boxes, and it is thus not simple to explain or interpret the captioning operation in critical usage. This interpretability slows down trust and adoption, particularly in sensitive usage.

2.1.6. Optimization Trade-Offs

Metaheuristic algorithms, such as Taylor-African Vulture Optimization (TAVO), are promising for hyperparameter tuning and evading local optima; however, they are computationally expensive and add complexity. Discovering good approaches to achieving efficient tuning and model performance is a subject of further research.

2.1.7. Decoding and Language Generation

Decoding techniques like beam search improve fluency and grammaticality at the cost of trade-offs between computational cost and possible high-frequency bias in phrases, reducing the diversity of captions [13]. The open issue is to balance decoding algorithms between quality and diversity.

Filling these gaps offers promising research directions for the future. Hybrid strategies that combine the strength of convolutional backbones, transformer semantic encoders, and metaheuristic optimization methods—along with explainability and domain adaptation—have the potential to bring image captioning nearer to human-like comprehension and generation capabilities. Table 1 presents an overview of the existing works.

Table 1. An overview of the existing works

Reference	Datasets	Model/ Methodology	Result	Limitations
[1]	MS-COCO	CNN-based encoder with multi-step attention and dual-LSTM decoder (Attend-More-Times model)	METEOR: 28.3, CIDEr: 126.1, ROUGE-L: 58.0, BLEU-4: 38.1%	Performance depends heavily on attention step selection.
[2]	MS COCO, Flickr30k	Transformer-based visual attention with ResNext-101 encoder + Adam optimizer	BLEU-4: 20.10%, Top-5 Accuracy: ~73	Trade-off between model size and computational cost
[3]	MS COCO	Multimodal Transformer with multi-view visual features	Ranked 1st on MS COCO leaderboard	High model complexity requires extensive training
[4]	MS COCO, Flickr30k	Word attention + external knowledge graph injection	State-of-the-art (SOTA) on COCO and Flickr30k	Complexity in integrating external knowledge
[5]	Corel5K, PASCAL VOC	DL with human-computing-inspired methods	Efficient on the domain-shift scenario	Limited reasoning capability of DL models
[6]	Multiple (survey)	Survey of visual encoders and text generators	Comparative analysis of SOTA	Lack of a conclusive solution, open challenges
[14]	Natural and medical image datasets	Review of DNN and GAN-based models	Qualitative & quantitative comparisons	Challenges in the medical domain adaptation
[7]	ALT200M (200M image-text pairs), COCO, nocaps	Large-scale vision-language pretraining (LEMON)	New SOTA on COCO, nocaps, Conceptual Captions	High computational resources, data noise

[8]	MS COCO	COS-Net: Semantic comprehending and ordering network	CIDEr: 141.1 (Karpathy test split)	Requires sophisticated semantic filtering
[9]	Flickr 8k	CNN + LSTM encoder-decoder	Competitive BLEU score	Limited by dataset size and complex scenes
[10]	Flickr 8k	Inception V3 + BiLSTM + GloVe embeddings	Improved BLEU and ROUGE-L scores	Generalization challenges, variable embedding sizes
[11]	Flickr 8k	VGG16 vs VGG19 CNN + LSTM	Comparable BLEU score	Sensitivity to training epochs, model convergence
[12]	Flickr8k, Flickr30k-tic Tibetan captions	CNN + LSTM with fusion attention for Tibetan captions	Improved BLEU and ROUGE-L scores	Low-resource language data scarcity
[13]	Flickr 8k	Merge model combining CNN and LSTM + beam search decoding	BLEU-1 and beam search scores > 60	Trade-off between accuracy and training memory

3. Methodology

The proposed AttenTAVO-Cap framework is a DL based pipeline for automated image caption generation, integrating visual, semantic, and sequential modeling components with optimization. After extracting and aligning the feature representations, the system operates as a singular caption generation engine by interacting with visual features and BERT-based embeddings through attention-enforced BiLSTM decoding. The final captions are produced in a sequential manner and cross-evaluated with standard linguistic metrics. Figure 1 systematically illustrates each stage of the proposed AttenTAVO-Cap framework, from image preprocessing to final caption generation.

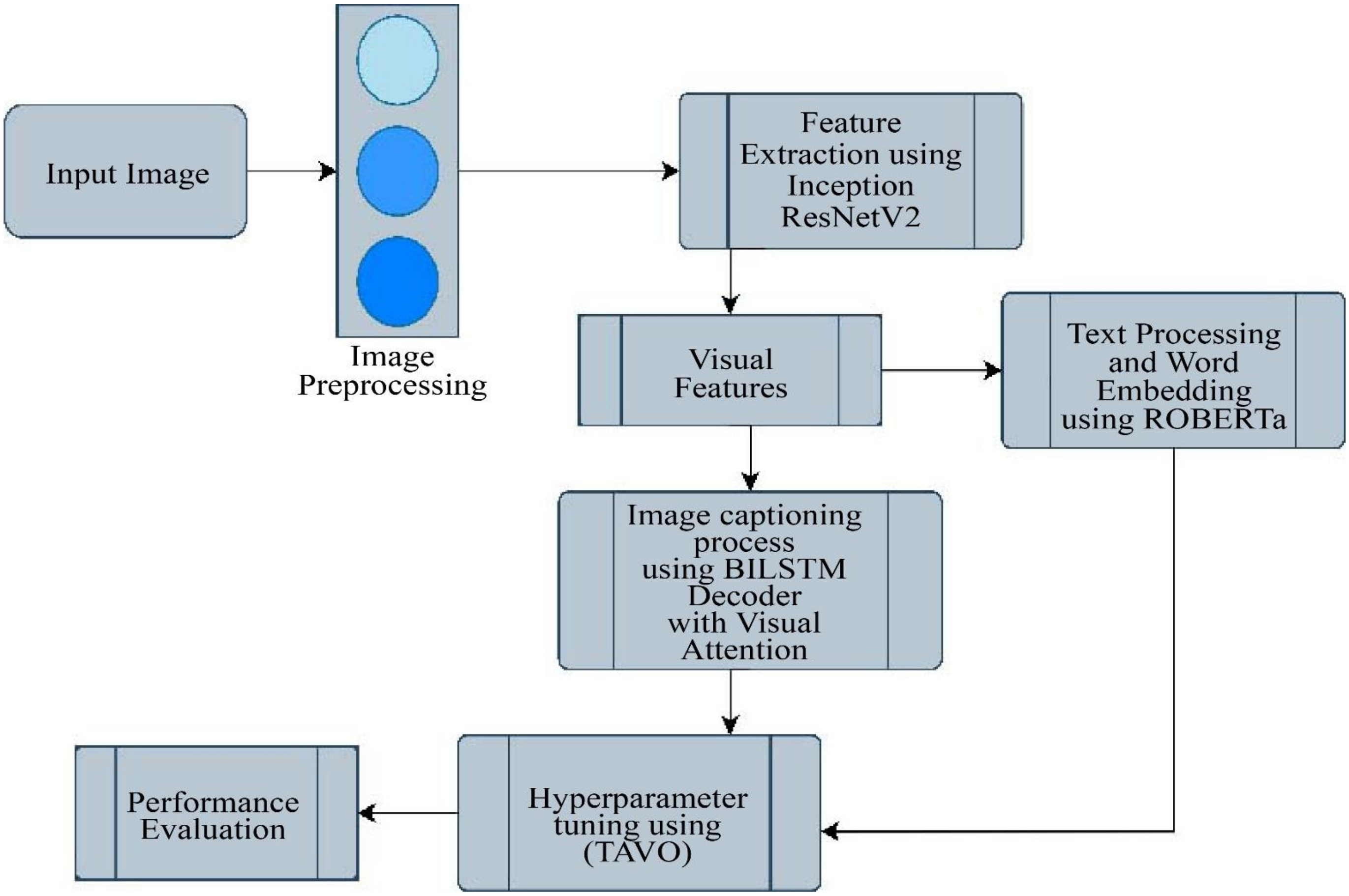

Input Image
Image Preprocessing
Feature Extraction using Inception ResNetV2
Visual Features
Text Processing and Word Embedding using ROBERTa
Image captioning process using BILSTM Decoder with Visual Attention
Hyperparameter tuning using (TAVO)
Performance Evaluation

Fig. 1 Workflow of this research

3.1. Data Collection

The Flickr8k and Flickr30k data sets are widely used training and testing corpora for image captioning algorithms. Flickr8k contains 8,000 images, while Flickr30k includes over 31,000 images, all of which are linked to five descriptive captions written by humans. The data sets contain a vast range of objects and scenes, offering linguistic structural variability that makes them excellent for training generalization-capable models. All the images are available in RGB format and vary in size and resolution. The databases are available for research and provide a realistic and challenging setting for caption generation tasks. Figure 2 depicts a sample of the dataset.

Fig. 2 Sample images from the dataset

The publicly accessible benchmark datasets (Flickr8k and Flickr30k), which are available for scholarly research, are used in this study. Images and captions were gathered with proper user authorization in accordance with the original dataset's licensing. However, inherent sociological and cultural biases, such as the overrepresentation of particular activities, objects, or demographics, may be reflected in these datasets and have an impact on the caption generation process. We examined model outputs for systematic errors and biased descriptions, especially with regard to gendered language, object attribution, and activity labeling, in order to lessen these impacts. We found no deliberate amplification beyond dataset patterns. From a societal standpoint, better picture captioning improves accessibility and human–computer interaction, but careful deployment is required to prevent perpetuating preconceptions.

3.2. Data Preprocessing

Firstly, preprocessed the images and captions from the Flickr8k and Flickr30k datasets for training. All images were resized to 299 × 299 × 3 pixels, as that is the input dimension required for the InceptionResNetV2 encoder. Pixel values were scale-normalized to [0,1] to stabilize convergence. For the caption data, first lowercase each sentence, remove punctuation and special characters, and then tokenize the text. To lower the sparsity of vocabulary, very infrequent ($\leqslant$ 5 occurrences) words were eliminated. Start and end tokens were introduced to indicate sentences, and all the sequences were then transformed into a sequence of integer indices. When dealing with the input, it is ensured that all elements have the same length by adding padding with zeros in the position of the maximum caption length. Besides, to have a more meaning-oriented input, the tokenized sentences were also embedded through a RoBERTa language model that had been pre-trained and thus generated contextualized word-level representation. At this stage, the vision and linguistic features representation becomes the reference for the following attention-guided caption generation.

3.3. Proposed Method

To address the challenge of accurate and fluent image captioning, a new hybrid DL framework is proposed called AttenTAVO-Cap. This architecture takes advantage of the discriminative representation ability of InceptionResNetv2, the embedding capability of RoBERTa, and the sequential learning capability of a Bidirectional LSTM decoder with visual attention. Hyperparameters of the decoder are tuned with the Taylor-African Vulture Optimization Algorithm (TAVO). We enable adaptive decoders and end-to-end optimization. The joint architecture enables the model to attend to meaningful parts of the image, match them with contextual semantic information from text embeddings, and produce coherent captions that are contextually correct.

3.3.1. RoBERTa Embeddings

By leveraging the RoBERTa [15], which is a robustly optimized BERT pretraining approach, we convert the preprocessed captions into high-dimensional semantic embeddings.

The handlers of each caption are tokenized and sent to the deep transformer layers of RoBERTa to encode contextual dependencies among words. These embeddings are semantically richer inputs to the decoder, thereby facilitating better alignment between linguistic and visual modalities.

Formally, let a tokenized input caption be represented as:

$$x = x_1, x_2, \ldots\ldots\ldots, x_T$$

Where$x_t x_t$ is the token at position t, and T is the length of the caption. RoBERTa applies multi-head self-attention layers and feedforward blocks over this sequence to produce context-sensitive representations:

$$H = [h_1, h_2, \ldots\ldots\ldots, h_T], h_t \epsilon \mathrm{R}^{\mathrm{d}}$$

Where$h_t h_t$ is the contextualized embedding for the token $x_t x_t$, and d is the hidden size of the transformer model.

The output matrix $h_t \epsilon \mathrm{R}^{\mathrm{T\times d}}$ $h_t \epsilon \mathrm{R}^{\mathrm{T\times d}}$ is then used by the decoder for generating captions.

The attention mechanism inside RoBERTa is defined by:

$$Attention(Q,K,V) \ = \ softmax(\frac{Qk^T}{\sqrt{d_k}})V$$

Where Q, K, V are the query, key, and value matrices derived from input embeddings, and $d_k d_k$ is the dimension of the key vectors used for scaling.

3.3.2. Feature Extraction via InceptionResNetv2

The AttenTAVO-Cap framework's feature extraction phase is adapted from InceptionResNetv2, a deep CNN hybrid structure with the power of Inception modules for representation, along with efficient learning of residual connections. InceptionResNetv2 was proposed by Szegedy et al. (2016) [16] as an extension of earlier Inception architectures, following the power to achieve faster convergence and improved feature reuse. Factorized convolutions, asymmetric kernels, and residual shortcuts in InceptionResNetv2 allow it to learn multiscale features efficiently from input images.

Suppose that the input image is in the form:

$$X \in R^{299\times 299\times 3}$$

Representing a color image of size 299×299 with three RGB channels.

The image goes through a set of Inception blocks with residual connections via the InceptionResNetv2 network, each of which gives a filtered feature map:

$$Y^{(l)} = X^{(l)} + f^{(l)}(X^{(l)})$$

Where $f^{(l)} f^{(l)}$ is the transformation applied by the l-th Inception block (e.g., 1×1, 3×3 convolutions, and pooling), and $X^{(l)}$ $X^{(l)}$ is the input to that block. The residual addition supports gradient flow preservation and accelerates learning in deep networks.

After the final convolutional layer, a Global Average Pooling (GAP) operation is done to shrink spatial features into a fixed-length vector:

$$z_k = \frac{1}{H\times W}\sum_{i=1}^{H} \ \sum_{j=1}^{H} \quad F(i,j,k), for \ k = 1,\ldots,2048$$

Where $F \in R^{H\times W\times 2048} F \in R^{H\times W\times 2048}$ is the final convolutional feature map, and $z_k z_k$ is the average activation over the spatial domain for the k-th channel.

This produces a 2048-dimensional feature vector:

$$V \in R^{2048}$$

Which is a dense, compressed representation of the image, containing both low- and high-level semantic and spatial information. These features are then fed to the decoder to generate a caption with attention.

Residual connections are a crucial part of the ResNet architecture, enabling signals to flow smoothly both forward and backward through the layers. These connections play a significant role in reducing the vanishing gradient issue that often arises during the training of deep networks, helping the model to reach convergence more quickly. The User can think of the signal movement within a residual unit as something that can be expressed mathematically:

$$F(x_l) = \ W \times x_l \ + \alpha \qquad (1)$$

$$R(F) + h(x_l) \qquad (2)$$

$$y_l = R\, x_l \ + \ 1 \ = \ R(y_l) \qquad (3)$$

The last output of the residual model has the label $x_l x_l$ +1. The value represents the offset. $x_l x_l$denotes the input; w indicates the weight; R shows the Relu function; $y_l y_l$ denotes the sum of two branches; h($x_l x_l$) shows a simple transformation for input; F$(x_l)(x_l)$ signifies the convolution function; and Relu shows an activation function, which can be advantageous to the spread of the ladder and prevents the divergence of the ladder from becoming significantly attenuated late in the multi-layer convolution.

3.3.3. BiLSTM Decoder with Visual Attention

The system generates natural language descriptions that represent the input image through this component. The system performs translation of visual data to linguistic sentences, which allows visual understanding to become a linguistic representation.A Bidirectional LSTM decoder with visual attention mechanisms performs the entire caption generation task. The attention module at every decoding step generates a weighted context vector by identifying which visual areas matter most for current word generation. The BiLSTM receives context vectors to model forward and backward dependencies before generating output word probabilities through softmax.

Mathematically, the attention mechanism is defined as:

1. Attention Score Calculation

$$e_{ti} = W_a . tanh(W_v V_i + W_h h_{t-1})$$

2. Softmax Attention Weights

$$\alpha_{ti} = \frac{exp(e_{ti})}{\sum_j exp\,(e_{ti})}$$

3. Context Vector (Weighted Sum)

$$c_t = \sum_j e_{ti} v_i$$

Where V_i V_i is the i-th visual feature from the image, $h_{t-1} h_{t-1}$ is the previous hidden state from the BiLSTM decoder, $W_a W_v W_h$ $W_a W_v W_h$ is the learnable weight matrices, $tanh$ $tanh$ is the non-linearity to mix the image and decoder states. This computes a score $e_{ti} e_{ti}$ representing how relevant the i-th image feature $v_i v_i$ is at the t-th decoding step, α_{ti} α_{ti} converts the raw scores $e_{ti} e_{ti}$ into attention weights α_{ti} α_{ti} that sum to 1, and c_t c_t is the context vector, which is a weighted average of image features that is weighted by the attention scores. Each decoding step receives the context vector c_t c_t together with the current word embedding $e_t e_t$ for their processing through the BiLSTM:

4. BiLSTM Input:

$$[h_t^{\rightarrow}, h_t^{\leftarrow}] = BiLSTM([e_t ; c_t])$$

5. Word Prediction

$$o_t = Softmax(W_o[h_t^{\rightarrow}, h_t^{\leftarrow}] + b_o)$$

Where $h_t^{\rightarrow} h_t^{\rightarrow}$ is the forward LSTM output and $h_t^{\leftarrow} h_t^{\leftarrow}$ backward LSTM output. A complex meaning of previous and upcoming time step information exists because these elements combine at time step t, $e_t e_t$ is the embedding of the current input word (e.g., “a”, “man”, etc.), and c_t is c_t is the context vector from the attention mechanism summarizes what parts of the image to focus on at this step.

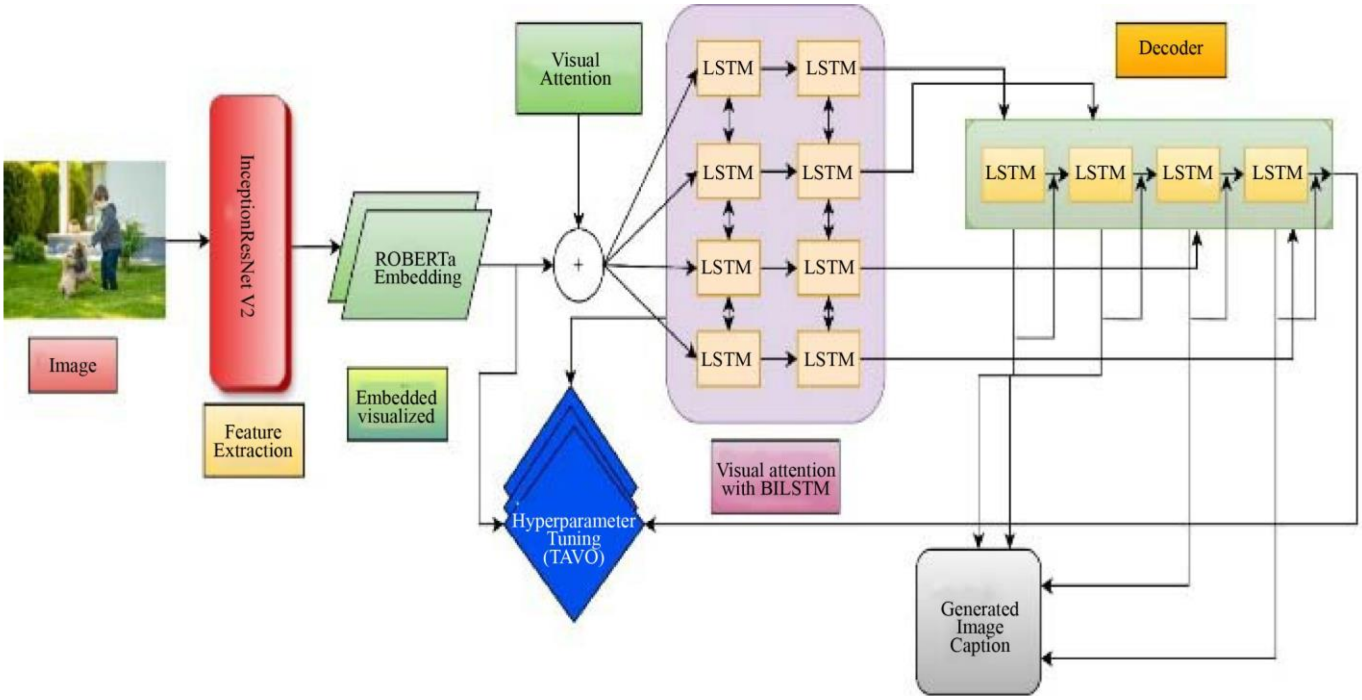


Fig. 3 Schematic representation of the proposed AttenTAVO-Cap framework image captioning

Figure 3 shows an architecture that integrates Inception ResNetv2 as the visual feature extractor with RoBERTa as the semantic text embedding module. The features are then merged and fed into a BiLSTM-based visual attention layer. The decoder, which is trained with the Taylor-African Vulture Optimization Algorithm (TAVO), generates context-informed image captions.

3.3.4. Metaheuristic Optimization (TAVO, GA, PSO)

To further enhance the performance of the captioning task, we add metaheuristic optimization techniques to adapt critical decoder hyperparameters. The first optimizer explored here is the Taylor-African Vulture Optimization Algorithm (TAVO), a hybrid algorithm that couples the global search capability of the African Vulture Optimization Algorithm

(AVOA) and Taylor series-based local refinement accuracy. TAVO enhances convergence rate and local exploitation accuracy with gradient-free, derivative-inspired search behavior. The optimization process begins by randomly initializing a population of candidate solutions, each corresponding to a unique set of decoder hyperparameters (such as hidden units and learning rate). Figure 4 illustrates the Flowchart of the TAVO hyperparameter optimization process.

TAVO then Iteratively Updates Each Candidate based on Two Stages,

- Global exploration: Inspired by the flight behavior of a vulture, candidate solutions are scattered in the search space.
- Local exploitation: Taylor series approximations are utilized for locally refining promising areas around the current best solutions for fine-tuning adjustment.

Two classic metaheuristics are also employed for performance benchmarking:

- Genetic Algorithm (GA): This GA [17] replicates natural selection through crossover and mutation operators applied to hyperparameter chromosomes. Selection occurs based on fitness scores computed against validation BLEU-4 or CIDEr.
- Particle Swarm Optimization (PSO): PSO is an algorithm [18] that emulates swarm behavior by allowing candidate solutions (particles) to modify their position and velocity according to global best and personal best solutions. It is highly effective for continuous hyperparameter optimization.

3.3.5. Hyperparameter Search Strategy and Reproducibility

A structured hyperparameter optimization technique is used since the decoder configuration affects the AttenTAVO-Cap framework's performance.

The vector of decoder hyperparameters is defined as:

$$\theta = \{H, \eta, D, \lambda\}$$

Where H defines the number of hidden units in the BiLSTM, η is the learning rate, D denotes the dropout ratio, and λ is the L2 regularization.

The following are the fixed search ranges:

$$H \in [256, 1024]$$

$$\eta \in [10^{-5},\ 10^{-3}]$$

$$D \in\ [0.2, 0.6]$$

$$\lambda \in [10^{-6},\ 10^{-3}]$$

The validation BLEU-4 score is employed as the primary fitness function for each candidate solution, and CIDEr is used for secondary validation. Every optimization run is performed for a maximum of N = 50 repetitions with a population of P = 20. All experiments are carried out utilizing fixed random seeds for data shuffling, optimizer population creation, and weight initialization to guarantee reproducibility. Every optimization algorithm uses the same batch size, early stopping conditions, and training/validation splits.

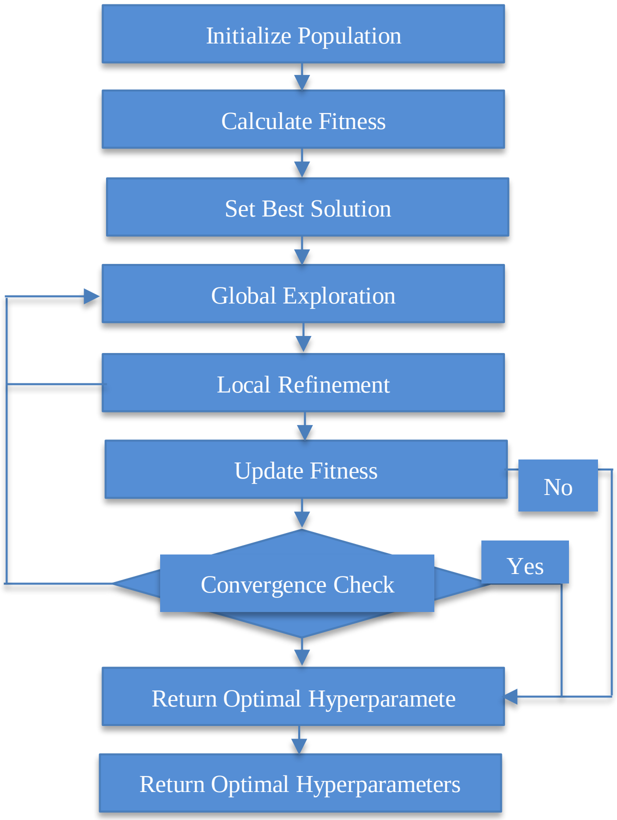


Fig. 4 Flowchart of the TAVO hyperparameter optimization process

3.3.6. Convergence Behavior of Metaheuristic Optimizers

TAVO, GA, and PSO convergence is examined by monitoring the optimal fitness value throughout iterations. Because of its hybrid search strategy, which combines local refining based on Taylor series with global exploration to prevent premature standstill and expedite exploitation close to ideal regions, TAVO exhibits faster convergence.

The process of optimization ends when either:

- The maximum iteration limit is reached, or
- The relative advancements in fitness stay below a predefined threshold $\in\ 10^{-4}$ for five consecutive iterations.

Figure 5 depicts the detailed attention-based BiLSTM decoding process with visual feature weighting.

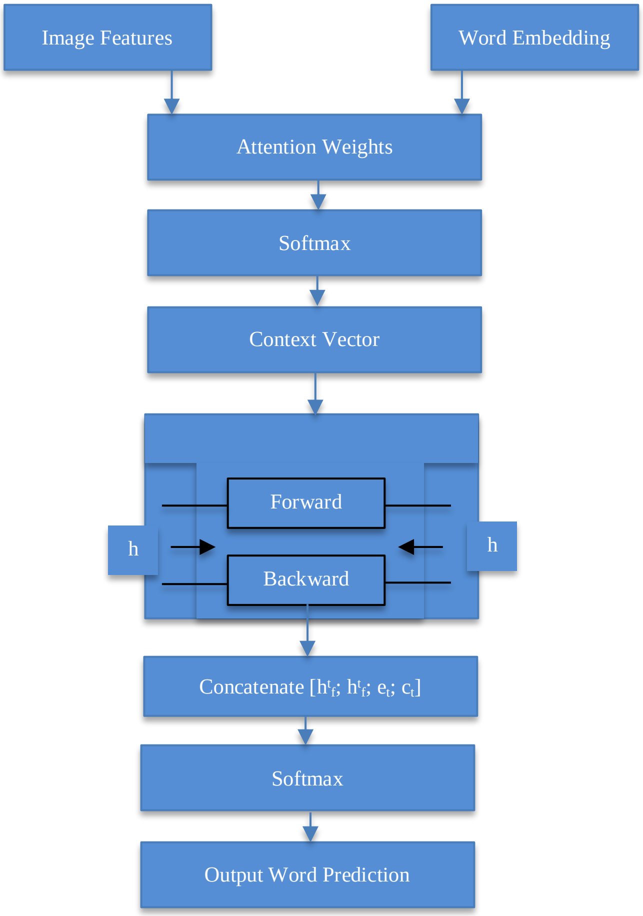


Fig. 5 Detailed attention-based BiLSTM decoding process with visual feature weighting

Algorithm 1: Taylor-African Vulture Optimization Algorithm (TAVO)

```
Input:
● $L_b U_b L_b U_b$: Search space boundaries
  P: Population size
● T: Maximum iterations
● f(·): Fitness function
Output:
● $X^* X^*$: Optimal solution
Step 1: Initialize a population $\{X_1, X_2, \ldots\ldots, X_p$ $X_1, X_2, \ldots\ldots, X_p\}$ of candidate hyperparameter vectors within $L_b U_b L_b U_b$
Step 2: Evaluate the fitness $f(X_i) f(X_i)$ of each individual in the population
Step 3: Identify the current best solution $X^* X^*$ with the highest fitness
Step 4: For iteration t = 1 to T:
● For each individual $X_i X_i$:
  ○ Exploration phase (African Vulture flight behavior):
  ○ $X_i^{t+1} X_i^{t+1} = X^* + r_1.(r_2.X^* - r_3.X_i^t)$
  $X^* + r_1.(r_2.X^* - r_3.X_i^t)$
  Where $r_1\, r_1, r_2 r_2, r_3 r_3$~U(0,1)
  ○ Exploitation phase (Taylor-based local refinement):
    Apply Taylor series approximation:
  ○ $f(X+\Delta) \approx f(X)\Delta + \frac{f''(X)}{2!}\Delta^2$
  $f(X+\Delta) \approx f(X)\Delta + \frac{f''(X)}{2!}\Delta^2$
    Use it to guide small local updates around $X^* X^*$
  ○ Selection:
  $X_i^{t+1} = argmin(f(X_i^t), f(X_i^{t+1}))$
Step 5: Update global best $X^* X^*$ if a better solution is found
Step 6: Repeat until convergence or T iterations are complete
Step 7: Return $X^* X^*$ as the optimized hyperparameter configuration
```

Algorithm 2: Genetic Algorithm (GA) for Hyperparameter Selection

```
Input:
● $D_{img}, D_{cap} D_{img}, D_{cap}$: Data input
● P: population size
● $C_r C_r$: crossover rate
● $M_r M_r$ : mutation rate
● G: max generations
Step 1: Population initializations from the subset according to certain constraints.
Randomly initialize the population of hyperparameter vectors $\{H_1, H_2, \ldots\ldots, H_P\} H_1, H_2, \ldots\ldots, H_P\}$
Step 2: For generation $g = 1$ to $G g = 1$ to $G$:
● Compute fitness $F_j\ F_j$ of each individuals $G_j G_j$ based on the BLEU-4/CIDEr from BiLSTM + Attention caption results
● Choose elite individuals (ranches) by using tournament/roulette selection
● Perform crossover on chosen parents, resulting in children
● Mutate the offspring to diversify them.
● Create new population $H_{new} H_{new}$ from offspring and elites
Step 3: After G generations, return best $H^* H^*$ and train final baseline model $M_{GA} M_{GA}$
```

Algorithm 3: PSO for Hyperparameter Tuning

```
Input:
• $D_{img}, D_{cap}D_{img}, D_{cap}$: input data,
• P: Swarm size
• $\omega\omega$: Inertia weight
• $c_1 c_1$: Cognitive coefficient
• $c_2 c_2$: Social coefficient
• T: Max iteration
Step 1: Particle initialization, each particle is initialized with a random hyperparameter vector $X_i X_i$, velocity $V_i V_i$, and best position $P_i P_i$.
Step 2: Assess the fitness function $f(X_i) f(X_i)$ of each particle by BLEU-4/CIDEr.
Step 3: Finding the global best position $G^* G^*$
Step 4: For iteration $t = 1 to Tt = 1 to T$:
    • For each particle:
    • Update velocity:
      $V_i = \omega.V_i + c_1.rand().(P_i - X_i) + c_2.rand().G^* - X_i$
      $V_i = \omega.V_i + c_1.rand().(P_i - X_i) + c_2.rand().G^* - X_i$
    • Update position: $X_i = X_i + V_i X_i = X_i + V_i$
    • Compute new fitness and update $P_i P_i$ and $G_i G_i$ if it is better suited
Step 5: Then Return $G^*$ $G^*$and train $M_{PSO} M_{PSO}$.
```

3.4. Baseline Models

To benchmark the performance of the AttenTAVO-Cap model, selected several baseline models for comparison. These models share common architecture, but differ in optimization regimes, or philosophies of design:

HABGRU + AVOA: Habituation-aware BiLSTM with an African Vulture Optimization Algorithm is a hybrid GRU model previously listed in captioning literature. This model serves as a good benchmark model, as it is biologically inspired and designed with metaheuristic tuning.

AttenTAVO-Cap (GA) is a version of attending with a vulture for a cape, where hyperparameters are tuned using a Genetic Algorithm for comparison between one optimization method versus the proposed Taylor-based vulture optimization.

AttenTAVO-Cap (PSO) is a different version tuned using Particle Swarm Optimization, representing an acceptable baseline under the swarm intelligence taxonomy. Given that all baseline models use the same backbone architecture, InceptionResNetv2 for visual features, RoBERTa for semantic embedding, and a visual attending BiLSTM decoder, this creates a level playing field between performance measures and architecture. All baseline models were also tested using the exact same performance measures as the proposed model, e.g., BLEU-1 to BLEU-4, METEOR, ROUGE-L, and CIDEr were calculated from the Flickr8k and Flickr30k datasets, and HABGRU + AVOA used Flickr8k.

Using the same testing framework is essential for a fair performance comparison, as it isolates optimization from architectural variables.

To ensure the outcomes are accurate and reliable, a rigorous evaluation method was adopted, including split-data experiments. For the experiments conducted using data from the Flickr8k and Flickr30k datasets, the pattern was 80% training, 10% testing, and 10% validation.

The images were shuffled before division, and all the captions of an image were in the same group. To compare the performance scores of AttenTAVO-Cap models with those of other models on test images, the models' scores were compared using a paired t-test with a 95% confidence level ($p < .05$) for each test image. The CIDEr and METEOR scores for the TAVO-optimized models show significant improvements on test images across both datasets, indicating that the improvements are not random variations in performance. In addition to average-score improvements, we conducted an exploratory analysis of predictive failures.

The predictive failures were categorized into the following groups: (i) omission of the object, (ii) assignment of the attribute of color/number, and (iii) assignment of the relation of the semantic aspect. In both scenarios, our approach performed better than the others in terms of both the omission of semantics and the semantics of the relation.

4. Results and Discussion

4.1. Experimental Setup and Implementation

The image captioning architecture, named AttenTAVO-Cap, has also been assembled for accelerated training environments using GPU resources to achieve faster convergence and better DL performance. In building this architecture utilizing Python 3.9, it depended heavily on libraries such as PyTorch, Hugging Face Transformers, and NumPy.

The Google Colab environment with Tesla T4 GPU resources has been heavily relied upon for training. First, we would like to highlight the hardware, software, and training environment used in our experiment as follows: In addition to the hardware components already mentioned, we used the Hugging Face RoBERTa caption tokenizer and the Torchvision image transform for data preprocessing. For image feature extraction, InceptionResNetv2 was used, pre-trained on the ImageNet dataset.

Table 2. Experimental setup for AttenTAVO-cap image captioning

Training Parameter	**Value**
Batch Size	32
Optimizer	Adam (metaheuristic fully tuned)
Epochs	50
Learning Rate Scheduler	ReduceLROnPlateau
Dropout Rate	0.5
Learning Rate	0.0001
BERT Embedding	RoBERTa-base
Decoder Units	BiLSTM, 128–512 hidden units
Attention Vector	Size 64–256

For the entire training process, the forward and backward passes, calculation of loss, and optimization routines using the TAVO, GA, and PSO algorithms, were performed on the GPU. As explained in the sections that follow, we employed a well-optimized training strategy, backed by powerful computational resources, to achieve remarkable captioning performance and reliable convergence within 50 epochs. The model was trained and tested on two benchmark datasets for picture captioning: Flickr8k and Flickr30k. Flickr8k contains 8,000 photographs, each with five human-written descriptions, whereas Flickr30k includes 30,000 images. Both datasets are common standards for assessing the generalizability and linguistic fluency of image captioning models**.** Table 2 presents the experimental setup for AttenTAVO-Cap Image Captioning.

4.2. Evaluation Metrics

In this section, we have evaluated AttenTAVO-Cap with TAVO, GA, and PSO using standard image captioning metrics on both datasets, Flickr8k and Flickr30k. The performance of the proposed model was analyzed based on image captioning metrics: BLEU, METEOR, ROUGE-L, and CIDEr. Examples of precision-based captioning metrics, such as BLEU and its subunits, BLEU-1 to BLEU-4, which are unigrams up to 4-grams, capture basic oversight and structural interconnection and overall fluency and syntactical similarity. METEOR is a synonymy-based extension of BLEU, which does better in correlating to real-life evaluations by humans due to factors of synonymy, stemming, and penalization based on the order of words. As for ROUGE-L, it computes the Longest Common Subsequence (LCS) and is referred to as structural similarity and fluency. CIDEr is customized for captioning due to its assessment of consensus among multiple references pertaining to human citation references through TF-IDF N-grams, highlighting merit-based content capture and precision detailing. Tables 3 and 4 showcase the metric values for the models.

Table 3. Full evaluation metrics on Flickr8k (test set)

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr
Proposed: AttenTAVO-Cap (TAVO)	0.70	0.54	0.45	0.29	38	67	194
AttenTAVO-Cap (GA)	0.66	0.49	0.41	0.34	35	61	185
AttenTAVO-Cap (PSO)	0.65	0.47	0.39	0.33	37	63	189
HABGRU + AVOA	0.64	0.45	0.36	0.31	33	60	183

Table 4. Full evaluation metrics on Flickr30k (test set)

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr
Proposed: AttenTAVO-Cap (TAVO)	0.68	0.52	0.43	0.29	35	63	191
AttenTAVO-Cap (GA)	0.64	0.47	0.39	0.32	34	59	188
AttenTAVO-Cap (PSO)	0.63	0.45	0.37	0.30	33	61	186
HABGRU + AVOA	0.60	0.43	0.34	0.31	31	57	183

Across both datasets, the AttenTAVO-Cap (TAVO) model demonstrated a superior advantage over the remaining variants. The TAVO model achieved the highest scores on the Flickr8k data with BLEU-1 at 0.70, BLEU-2 at 0.54, BLEU-3 at 0.45, BLEU-4 at 0.29, METEOR at 38, ROUGE-L at 67, and CIDEr at 194, as shown above. These metrics suggest that both the syntactic and semantic relationships with the ground truth captions are very strong. The GA-tuned AttenTAVO-Cap scores slightly lower than AttenTAVO-Cap across all metrics (for example, BLEU-4: 0.34, METEOR: 35, CIDEr: 185), and the PSO-tuned version with scores of BLEU-4: 0.33 and CIDEr: 189, respectively. However, the baseline HABGRU + AVOA model produces the lowest scores across all metrics (for example, BLEU-4: 31, METEOR: 60, and CIDEr: 183), strongly indicating the advantages of the AttenTAVO architecture. A similar runnable context can also be observed across the Flickr30k dataset, as all models exhibited slightly lower scores due to the additional complexity and variability of the dataset. Also, reusable metrics indicate that AttenTAVO-Cap (TAVO) achieved relatively high baseline scores of BLEU-4: 0.29, METEOR: 35, ROUGE-L: 63, and CIDEr: 191 again. The GA and PSO models all exhibited a modest drop from TAVO's scores, with PSO scoring BLEU-4: 0.30 and CIDEr: 186. Meanwhile, HABGRU + AVOA, again, scored the lowest across the base metrics evaluated with BLEU-4: 0.31, METEOR: 31, and CIDEr: 183. Overall, these comparisons reveal the overall effectiveness of the TAVO optimization strategy and highlight the strengths of the AttenTAVO-Cap model based on the size of the datasets and the complexity of the captioning task.

4.2.1. Comparison with AVOA-Based and Transformer-Based Models

In contrast to other AVOA-based approaches, such as the HABGRU + AVOA model, the proposed TAVO technique includes an additional Taylor Series Local Refinement step that runs parallel to the global search conducted through the African Vulture Optimization algorithm. The hybrid technique enables the African Vulture Optimization algorithm to exploit the strongest indicative signals of successful hyperparameters, resulting in faster convergence, improved stability, and better overall performance in the semantic metrics CIDEr, METEOR, and ROUGE-L for both Flickr8k and Flickr30k. AVOA uses random-driven and population-level updates, whereas in TAVO, refinements inspired by the derivative components do not require gradient computations. However, transformer-based solutions to image captioning tasks usually perform in the hundreds, thanks to heavy reliance on the self-attention mechanism and pre-training over massive datasets, perhaps measured in the hundreds of millions. In the AttenTAVO-Cap framework, the best use of the capabilities is aimed at, not the size. This is because transformer-based models often impose extremely demanding computational requirements and require enormous, curated datasets, which may not be feasible in less-than-real-time settings. From our work, we show that, through expertly designed hybrid models, semantic alignment can be achieved reasonably with less massive transformers. To showcase these differences, we have made an equal comparison in Table 4 between AttenTAVO-Cap and other AVOA- and transformer-based models for captioning. Table 5 presents a comparative study of our model with AVOA-based and transformer-based image captioning models.

Table 5. Comparative analysis of AttenTAVO-Cap with AVOA-based and transformer-based image captioning models

Aspect	AVOA-based Models	Transformer-based Models	Proposed TAVO
Optimization Strategy	AVOA	Gradient-based (Adam/SGD)	Hybrid metaheuristic (AVOA + Taylor refinement)
Local Search Capability	Limited	Implicit through gradients	Explicit Taylor-based exploitation
Model Scale	Moderate	Very large	Moderate
Data Requirement	Medium	Very high	Medium
Computational Cost	High	Very high	Controlled
Semantic Consistency	Moderate	High	High
Generalization on Small Datasets	Limited	Often weak	Strong
Deployment Feasibility	Moderate	Low	High

Table 6. Comparison of evaluation metrics with others

Ref(s)	Model / Methodology	BLEU-4 (%)	(METEOR / ROUGE-L / CIDEr)
[1]	Attention-based Two-LSTM Image Captioning Model (CNN + Multi-Attention + LSTM)	38.1	28.3 / 58 / 126.1
[3]	Transformer + ResNeXt-101 + Adam Optimizer (MS COCO, Flickr30k)	20.10	— / — / —
[19]	Linear-Time Sequence Model (Flickr30k)	59	79/73/130
[20]	Advanced Context-Aware Object Relational Model	44.39	41.58/64/-
Ours	AttenTAVO-Cap	29.00	38 / 67 / 194

4.3. Comparative Analysis and Discussion

In this section, we interpret the evaluation results and explain why our model performs the best. Among all models, AttenTAVO-Cap optimized using Taylor-African Vulture Optimization (TAVO) showed the best results, as shown in Table 6. With a CIDEr score of 194 and a BLEU-4 score of 0.29 or 29%, it outperformed all other versions by a wide margin. The next-best model was an Attention-based two LSTM image captioning model (CNN+Multi-attention+LSTM) [1], which reached BLEU-4 = 38.1% and CIDEr = 126.1. Conversely, the model (Transformer + ResNeXt-101 + Adam Optimizer (MS COCO, Flickr30k) [3] employed in past research had a somewhat reduced BLEU-4 of 20.10 %. Recently, research in image captioning has focused on applying novel architectures to improve efficiency and semantic quality. In study [19], a linear-time sequence model architecture is used on Flickr30k and achieves very encouraging results, with BLEU-4 = 59%, METEOR = 79, and ROUGE-L ≈ 73. However, compared with other methods, the CIDEr score is relatively low, around 130, indicating that, although improved in terms of efficiency and semantic quality, the current caption outcome may still relate less accurately to human-written captions than the optimization-based hybrid outcome. In [20], focusing on how object relationships and contextual dependencies are implemented in images, the findings are more pertinent and semantically significant: ROUGE-L ≈ 64, BLEU-4 = 44.39%, and METEOR ≈ 41.58. These results indicate higher logical and semantic accuracy and, consequently, better comprehension of object relationships in the caption. But because it lacks values for CIDEr metrics, assessing performance on human-consensus metrics becomes difficult, given CIDEr's importance in determining semantic relevance and accuracy across a variety of captions.

Compared to the above arches, the proposed AttenTAVO-Cap model yields well-rounded performances. Moreover, in the context of high CIDEr scores and METEOR and ROUGE-L evaluations, optimization-driven models have been a prominent choice for efficiency and relational understanding.

4.4. Training and Testing Performance

To assess the caption-generating performance of the suggested AttenTAVO-Cap Framework, extensive comparisons against several baseline models using the same training arrangements were undertaken. Using the same 80/20 training-validation-test split, the trials were carried out on the Flickr8k and Flickr30k datasets. Evaluation of caption quality was based on five major metrics: BLEU (1-4), METEOR, ROUGE-L, and CIDEr. Figure 6 performance curves show the training curve of the proposed AttenTAVO-Cap for its 50 epochs of training. The right side presents the loss curve that depicts all the loss types, thus showing the behavior of both training loss and validation loss of the model. Initially, the training loss is quite high, and is about 0.85, and then it is reduced very fast to nearly 0.03 in the first 20 epochs. Therefore, the rate of descent is similar to a logarithm; it never gets nearly straight, and the end is reached with this situation at the end of the 50th epoch. This is the constant diminishing of the loss that tells the model was never standing still and was learning the best attributes internally to make predictions as accurately as possible in each epoch.

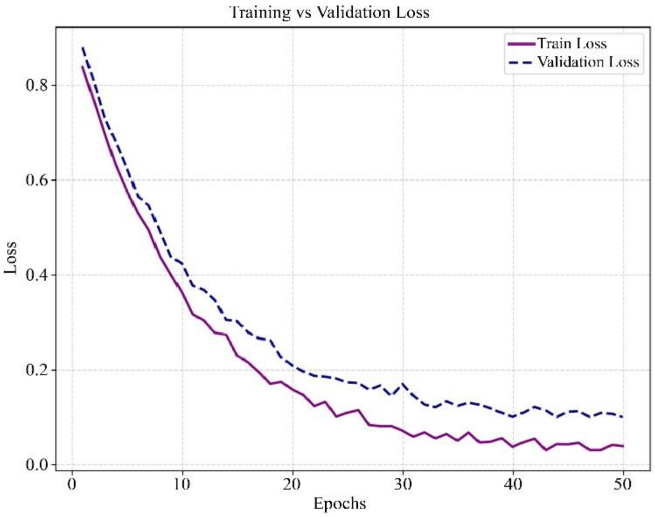


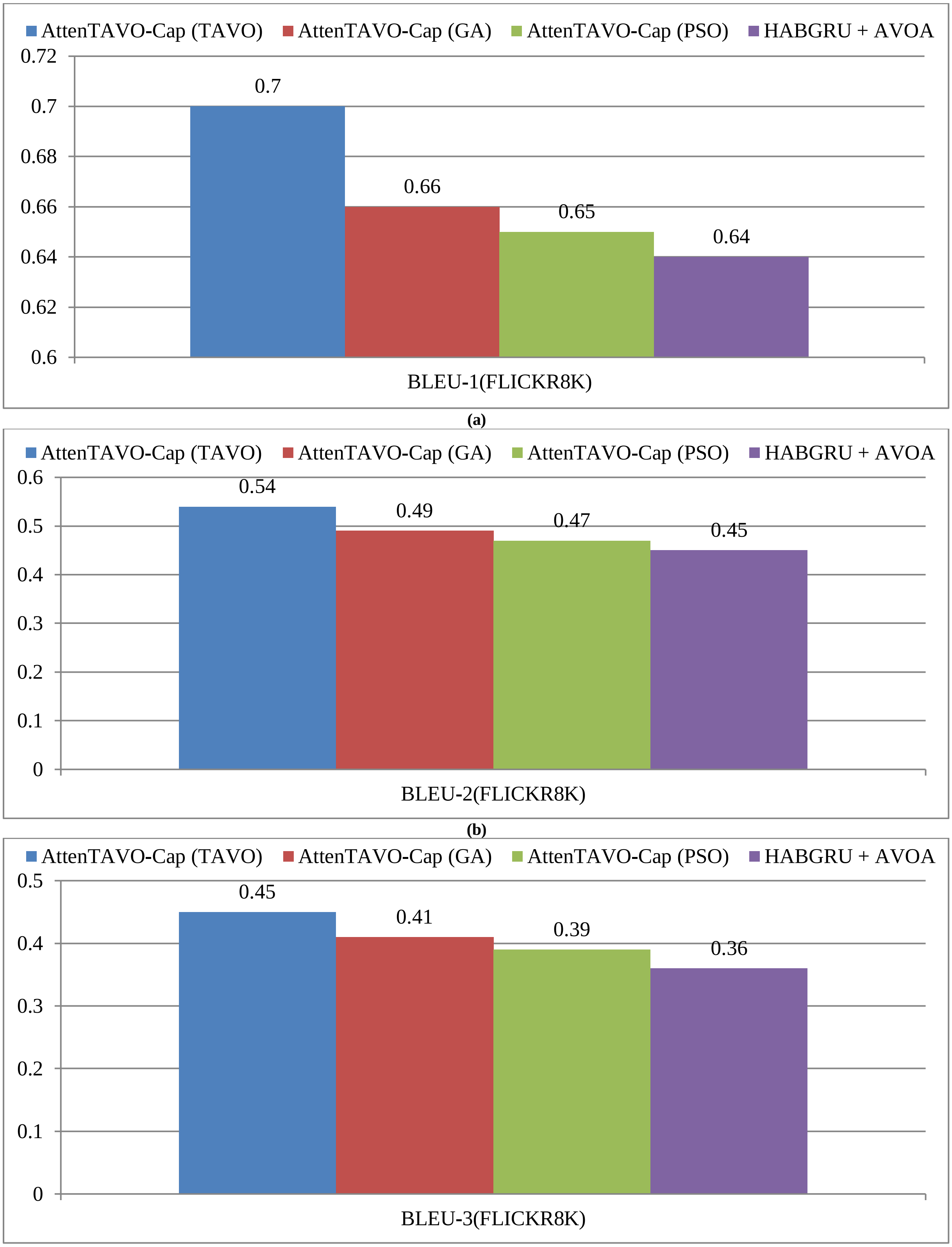
Fig. 6 Training vs validation loss curve of the AttenTAVO-Cap

Table 7. Performance comparison of captioning models based on BLEU-4 and CIDEr (Flickr8k and CIDEr (Flickr8k)

Optimization Method	Model Architecture	BLEU-4	CIDEr
Proposed: AttenTAVO-Cap (TAVO)	InceptionResNetv2 + RoBERTa + BiLSTM + Attention	0.29	194
GA[Baseline]	InceptionResNetv2 + RoBERTa + BiLSTM + Attention	0.34	185
PSO[Baseline]	InceptionResNetv2 + RoBERTa + BiLSTM + Attention	0.33	189
AVOA + HABGRU [Baseline]	InceptionResNet + HABGRU + AVOA	0.31	183

For the validation set, one notices a similar pattern - the losses started off slightly higher, with the initial loss around 0.88, and were then gradually reduced to about 0.10. The most important thing, however, is that the gap between the two loss curves (training and validation) is always kept to a minimum, displaying no signs of overfitting. In fact, parallelism in the direction of both lines indicates the model's good generalization ability in relation to unseen data. The curves of both the training loss and the validation loss, which are convex downward and free from any sudden inclines, accompanied by up and down steps (because of random gradient updates), are clear-cut evidence of the steady performance the model is exhibiting. The trajectory of convergence of the trend assures that the model indeed extracted the essential features from its input and subsequently retained all these features well during the test phase, proving the correctness of the training strategy and the model itself. Extra indicators, including BLEU-1 to BLEU-3, METEOR, and ROUGE-L, were evaluated to provide a more complete picture of linguistic fluency and semantic coherence. Once again, ranking top among all of these criteria was the TAVO-optimized model. These findings highlight how resilient AttenTAVO-Cap is across various datasets, showcasing its knack for producing high-quality, contextually relevant captions, even when dealing with larger and more varied image collections like Flickr30k. The experimental results reveal that the AttenTAVO-Cap model not only achieves superior n-gram precision (BLEU) but also shines in generating captions that are both semantically rich and fluent, as shown by METEOR, ROUGE-L, and CIDEr. The model's impressive performance is attributed to the powerful combination of InceptionResNetv2's visual features, RoBERTa's contextual embeddings, attention-guided BiLSTM decoding, and effective hyperparameter optimization through TAVO. These results were consistently observed across both the Flickr8k and Flickr30k datasets, reinforcing the generalization capability of the proposed framework.

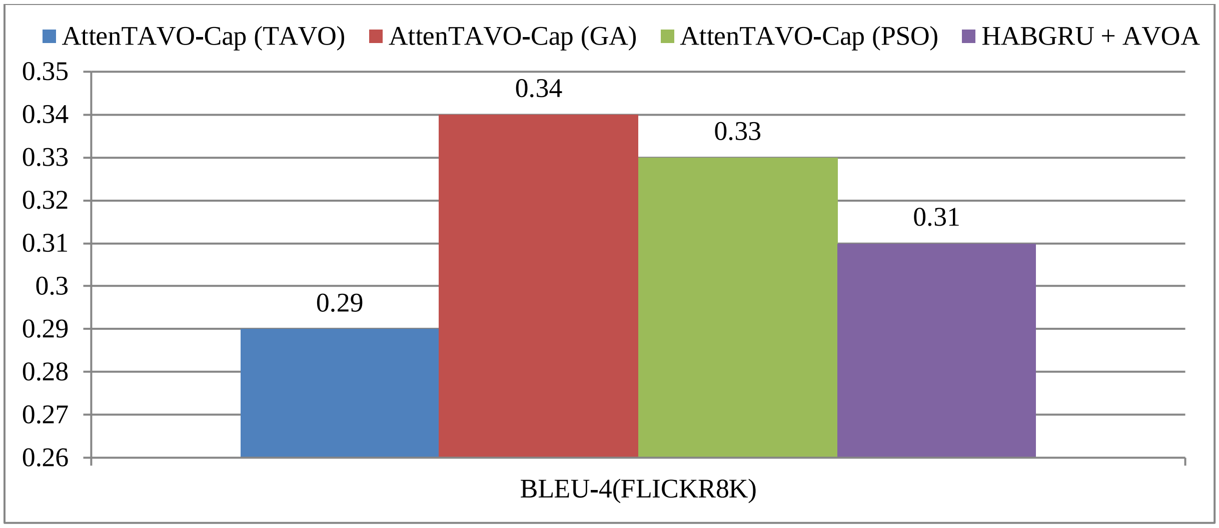
In Table 7, the Performance Comparison of Captioning Models Based on BLEU-4 and CIDEr is presented for both datasets. Figures 7 and 8 show the comparative BLEU inspection of the AttenTAVO-Cap methodology.



(a)

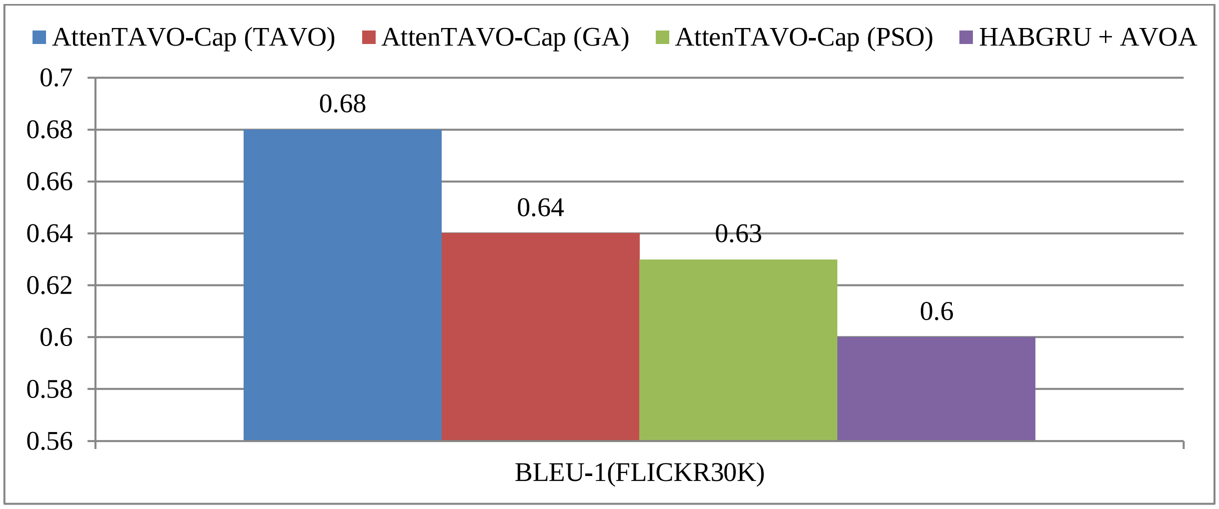
(b)

(c)

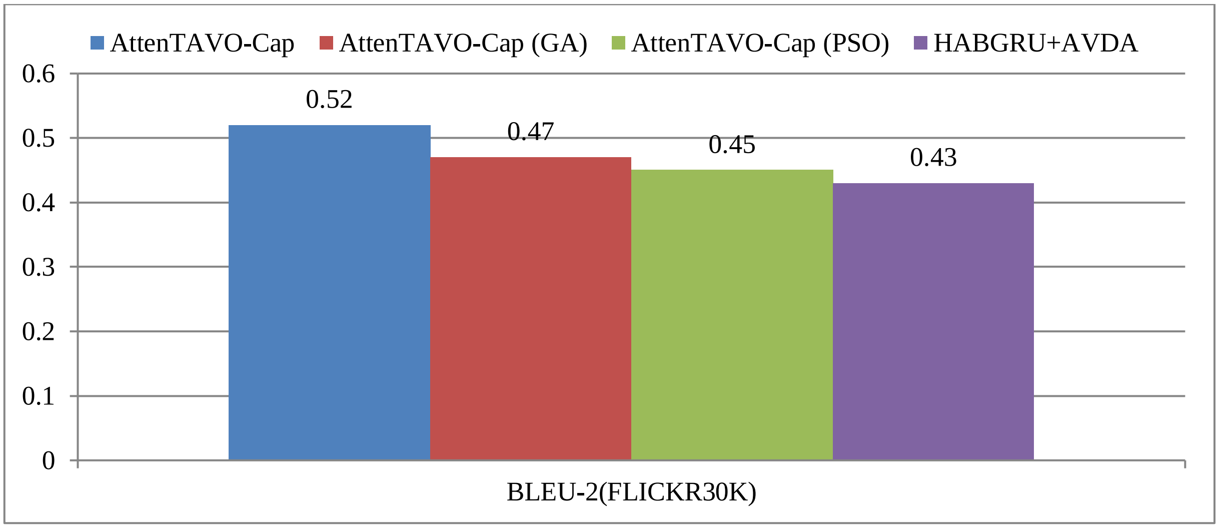


(d)

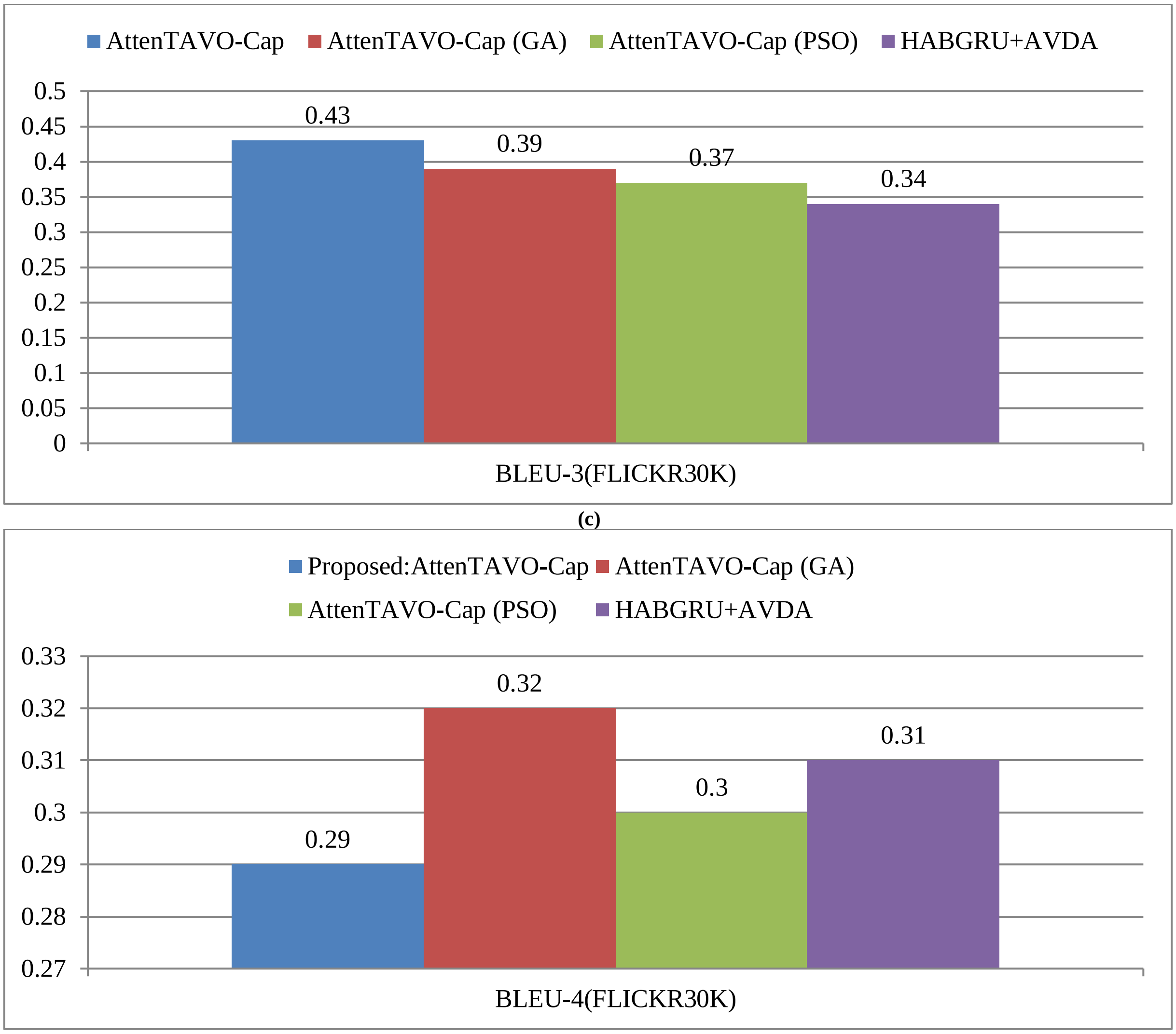
Fig. 7 Comparative analysis of AttenTAVO-Cap approach with other systems (a) BLEU-1, (b) BLEU-2, (c) BLEU-3, and (d) BLEU-4 using Flickr8k dataset.



(a)



(b)



(c)

(d)

Fig. 8 Comparative analysis of AttenTAVO-Cap approach with other systems (a) BLEU-1, (b) BLEU-2, (c) BLEU-3, and (d) BLEU-4 using Flickr30k dataset.

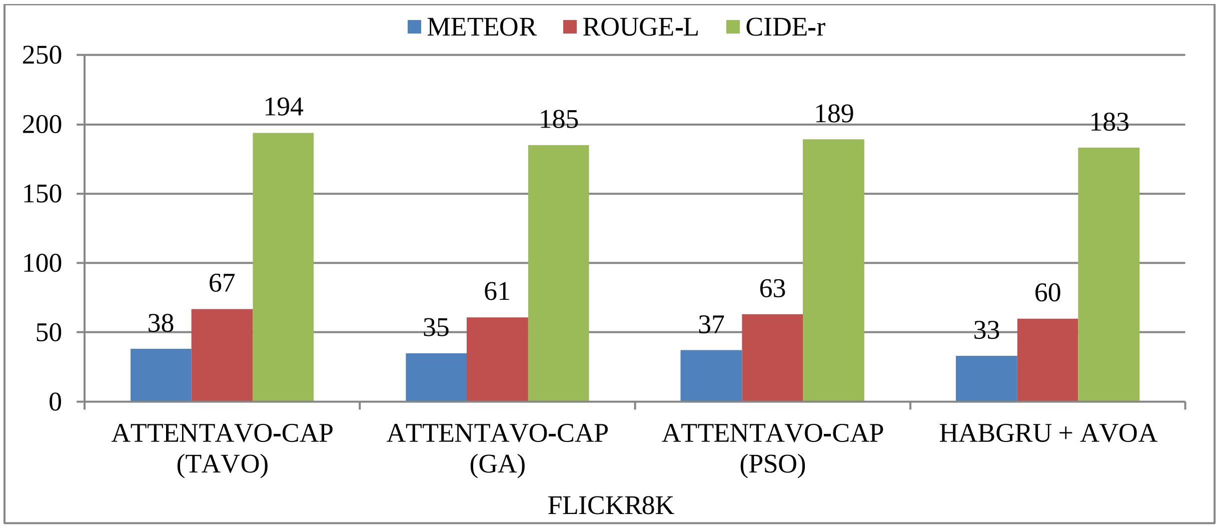


Fig. 9 Comparative analysis of AttenTAVO-Cap approach with other systems METEOR, CIDEr, and Rouge-L using Flickr8K dataset

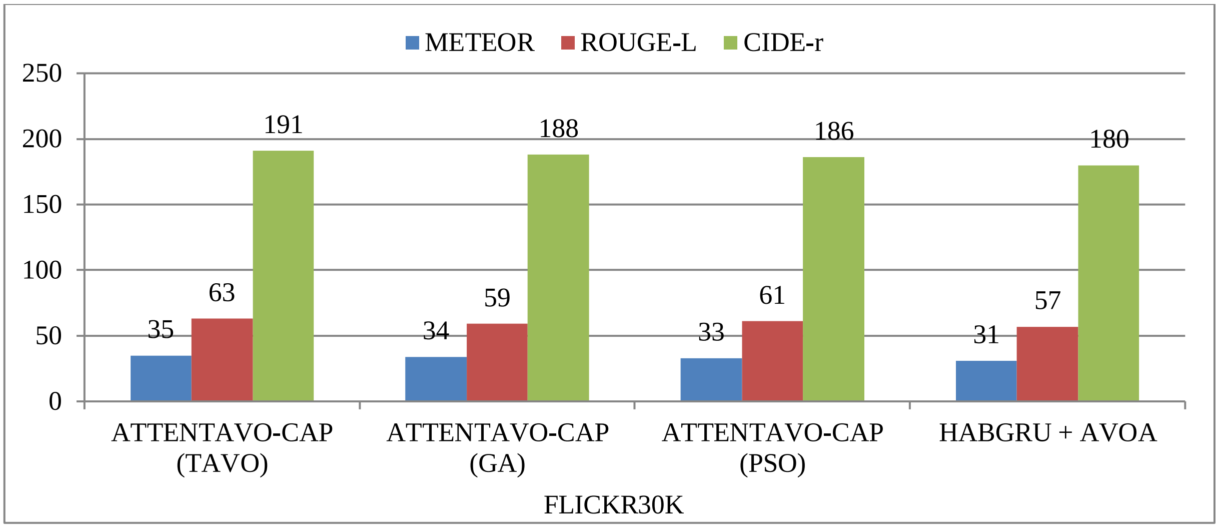


Fig. 10 Comparative analysis of AttenTAVO-Cap approach with other systems METEOR, CIDEr, and Rouge-L using the Flickr30K dataset

Figures 9 and 10 show the Comparative analysis of the AttenTAVO-Cap approach with other systems, METEOR, CIDEr, and Rouge-L, using both datasets. The comparison results unequivocally show that the suggested AttenTAVO-Cap model outperforms a previously developed hybrid model (HABGRU + AVOA) as well as baseline variants (GA and PSO).

For both the Flickr8k and Flickr30k datasets, AttenTAVO-Cap (TAVO) consistently yields the highest scores across BLEU-4, METEOR, ROUGE-L, and CIDEr metrics, as indicated in the Figures. It outperforms GA by 5 points and PSO by 6 points, respectively, with a BLEU-4 score of 0.29 on Flickr8k. It continues to Flickr30k with 0.29, while GA and PSO trail at 0.32 and 0.30. GA leads BLEU-4, but TAVO leads CIDEr (191 vs 188/186).

According to the CIDEr metric, the model demonstrates superior relevance and expressiveness compared to GA (185), PSO (189), and the baseline HABGRU + AVOA (183), achieving a higher score of 194. Indeed, the model beats all other approaches on semantic agreement and linguistic fluency, with METEOR and ROUGE-L scores of 38 and 67, respectively. The BLEU-4 for AttenTAVO-Cap on Flickr30k is 0.29, compared to 0.32 and 0.30 for GA and PSO, respectively. Also, the CIDEr score achieved by the model surpasses those of all the compared models, reaching a value of 191. The generated model produces syntactically correct, semantically dense captions that align with human references, achieving METEOR scores of 35 and ROUGE-L scores of 63.

4.5. Computational Complexity and Runtime Analysis

This section evaluates the computational overhead on training efficiency and the practicability of the newly proposed AttenTAVO-Cap framework, particularly regarding the performance of TAVO, GA, and PSO optimization techniques.

- Training Time Comparison: The three different versions of AttenTAVO-Cap were trained and tested in strictly the same hardware and software environments on a Tesla T4 GPU. There was a noticeable difference between the optimizers, even though metaheuristics have more overhead than fixed hyperparameters. Because of the constant crossover and mutation between generations, the GA optimizer was the slowest. Due to velocity modifications, the training time in PSO was moderate; however, swarm characteristics could affect it. At the same time, TAVO reached optimal points more quickly by avoiding unnecessary population-level updates through Taylor expansion-based local search. At the same time, TAVO was found to have a training time equal to or lower than PSO and superior to GA.
- GPU Utilization and Convergence Properties: GPU usage remained steady across all experiments, as most computations involved extracting CNN features and performing BiLSTM decoding. The effect of TAVO on total training costs was negligible. However, most importantly, TAVO-optimised models required fewer effective epochs and exhibited smoother convergence than GA- and PSO-optimised models.
- Feasibility for Real-World Deployment: As far as the feasibility for deployment is concerned, AttenTAVO-Cap provides a very positive impression in terms of balance between performance and computational complexity. Unlike the transformer-based captioning models, the proposed approach does not require large-scale training. Once the optimization is achieved, the complexity at the inference phase remains the same for the TAVO, GA, and PSO models. This makes the TAVO-optimized approach particularly suitable for environments with constrained computational resources.
- Effect of Optimization Strategy on Caption Quality: To gain better insights into the effect of optimization strategy in captioning tasks, we performed an ablation study by

varying just the hyperparameter optimization strategy while holding the architecture constant. In this work, all model variations share the same architecture, including the InceptionResNetv2 encoder, RoBERTa embeddings, and an attention-based Bi-LSTM decoder. The only difference between them is how each optimizer handles hyperparameter optimization.

CIDEr and METEOR reward semantic relevance and human agreement over superficial n-gram matching. TAVO's careful balance between exploration and exploitation helps the decoder reach conclusions about the choice of hyperparameters, naturally aligning visual features and contextual linguistic representations.

Additionally, the use of the Taylor series approximation in the local refinement enables the careful control of the rates of learning, hidden units, and dimensions of the attention vector to result in captions that convey semantic relationships between objects and the scene more accurately.

This results in higher CIDEr and METEOR scores than those of the GA and PSO algorithms. GA performance would sometimes fluctuate across generations due to the mutation component. In some cases, the PSO would get trapped in a local optimum.

TAVO ensures good convergence with lower variability. In conclusion, this analysis shows that it is not only architectural trends that have led to improvements in performance, but that the proposed TAVO optimization method is more significantly instrumental in ensuring the efficiency of high-quality image captioning.

4.6. Ablation Study: Effect of Optimization Strategy on Caption Generation

An ablation test was conducted by maintaining the core architecture and varying only the hyperparameter optimization algorithm so as to see the effectiveness of the selected optimization algorithm on the performance of the captioning model. In all the tests carried out, the attention-based BiLSTM decoder architecture, semantic embedder model (RoBERTa), and visual encoder architecture were maintained as InceptionResNetv2. The TAVO, GA, and PSO optimizers are the three metaheuristics involved in this ablation and compared to a previous combination of a hybrid baseline model and AVOA. The ablation test outcome is clearer on the fact that the selected optimization algorithm plays a pivotal part in the determination of a captioning model, regardless of changes in the backbone network architecture, as explained in Table 8. The captioning outcome generated by our new embedder model with TAVO-based optimization strategy achieves semantically truer and more cohesive and human-evaluated captions.

Table 8. Ablation analysis highlighting semantic and fluency gains from different optimization strategies

Optimization Strategy	Semantic Consistency (CIDEr ↑)	Linguistic Alignment (METEOR ↑)	Fluency (ROUGE-L ↑)	Overall Caption Quality
TAVO (Proposed)	Highest	Highest	Highest	Very Strong
GA	Moderate	Moderate	Lower	Medium
PSO	High	Moderate–High	Medium	Medium–High
AVOA (Baseline)	Lowest	Lowest	Lowest	Weak

In the case of CIDEr, what we notice is that the benefit brought by TAVO is the ability to get the captions aligned to the human annotations. This is brought by an effective combination of exploration and exploitation brought by TAVO because the search inspired by the African Vulture Search helps the search avoid being stuck prematurely.

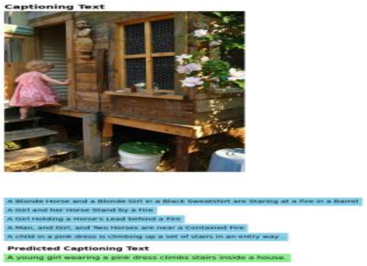
In other words, the search inspired by the Taylor series helps the adjustment of the decoder parameters. On the flip side, GA has ensured good n-gram precision, although with less convergence of hyperparameters because of its stochastic process involving mutation and crossover. It has ensured that the captured images have aesthetically pleasing results concerning their BLEU scores, although less semantic and with lower CIDEr and METEOR scores, which has been established in the results section. In comparison with GA, there is faster convergence for PSO, although its initialization method sometimes leads to less optimal exploration of its solution space. The baseline system of HABGRU + AVOA is consistently beaten on lexical and fluency measures.

This emphasizes the advantage of the combination of AttenTAVO-Cap architecture and the Taylor-opt search procedure over previous biologically inspired architectures. This ablation experiment clearly justifies that the performance benefit achieved through the AttenTAVO-Cap is not just an issue in the model architecture. A key contributing factor for such improvements is the TAVO optimization method since it incorporates the capability for converging toward the informed choice of hyperparameters through the attainment of improvements in the expressiveness and human alignment qualities.

4.7. Qualitative Results and Analysis

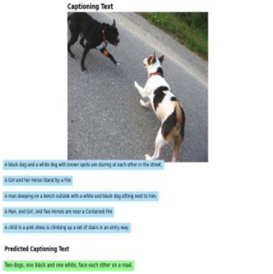
In this section, we can see the quantitative evaluation metrics, and the captions produced by the AttenTAVO-Cap model were scrutinized for qualitative evaluation by thumbnail visual examination. To demonstrate the model's competency in captioning, these samples were drawn from both Flickr8k and Flickr30k test datasets presented in Tables 9, 10, and 11.

Table 9. Ablation analysis demonstrating the improvements in fluency and semantics from various optimization techniques

Image	Ground Truth	Generated Caption (AttenTAVO-Cap)
 Captioning Text A Blonde Horse and a Blonde Girl in a Black Sweatshirt are Staring at a Fire in a Barrel A Girl and her Horse Stand by a Fire A Girl Holding a Horse's Lead behind a Fire A Man, and Girl, and Two Horses are near a Contained Fire A child in a pink dress is climbing up a set of stairs in an entry way . Predicted Captioning Text A young girl wearing a pink dress climbs stairs inside a house.	A child in a pink dress is climbing up a set of stairs in an entryway.	A young girl wearing a pink dress climbs stairs inside a building.

Analysis: The term "child in the pink dress climbing the stairs" is captured by the model, which lexically simplifies a human-created caption. The model's decoding components can spatially focus on contextually relevant regions, such as clothing and stairs, as demonstrated by its semantic correctness.

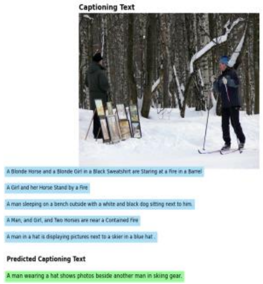
Table 10. Semantic and fluency improvements from various optimization techniques are highlighted via ablation analysis

Image	Ground Truth	Generated Caption (AttenTAVO-Cap)
 Captioning Text A black dog and a white dog with brown spots are staring at each other in the street. A Girl and her Horse Stand by a Fire A man sleeping on a bench outside with a white and black dog sitting next to him. A Man, and Girl, and Two Horses are near a Contained Fire A child in a pink dress is climbing up a set of stairs in an entry way. Predicted Captioning Text Two dogs, one black and one white, face each other on a road.	A black dog and a white dog with brown spots are staring at each other in the street.	Two dogs, one black and one white, face each other on a road.

Analysis: The caption upholds several focal aspects: two dogs, their colors, and the described interaction. The model's language is accurate and demonstrates a grammatically sound sentence, although it is explained in simple terms.

The description emphasized how visual concentration, effective decoding, and expressing reasoning drive object-to-object interactions.

Table 11. Ablation analysis and semantic and fluency gains from different optimization strategies

Image	Ground Truth	Generated Caption (AttenTAVO-Cap)
 Captioning Text A Blonde Horse and a Blonde Girl in a Black Sweatshirt are Staring at a Fire in a Barrel A Girl and her Horse Stand by a Fire A man sleeping on a bench outside with a white and black dog sitting next to him. A Man, and Girl, and Two Horses are near a Contained Fire A man in a hat is displaying pictures next to a skier in a blue hat . Predicted Captioning Text A man wearing a hat shows photos beside another man in skiing gear.	A man in a hat is displaying pictures next to a skier in a blue hat.	A man wearing a hat shows photos beside another man in skiing gear.

Analysis: Contextually, the utterance conveys a man, presumably showing off pictures, and another person dressed in skiing attire. Even while details such as "blue hat" tend towards generalizations in "skiing gear," the core of the caption remains in line with the visuals that can be detected. It picks steam and posture and has details in place.

The qualitative examples demonstrate the effectiveness of the proposed AttenTAVO-Cap (TAVO) model in understanding the scene, identifying objects, and creating natural language responses. The attention mechanism plays an important role in linking vocabulary with the relevant image, and the phrase identification is improved through the use of TAVO hyperparameters.

The minor differences between the proposed and ground truth captions do not impact the contextual correctness of the proposed captions and make them equally good as captions in a practical scenario.

5. Conclusion

This work proposes a general solution for image captioning with AttenTAVO-Cap, a novel hybrid approach that combines the attention-driven CNN-BiGRU framework with the TAVO metaheuristic. The optimization strategy not only improves the convergence rate but also enhances the capability to learn more discriminative correlations between images and texts. The critical analysis performed for the evaluation process, using the Flickr8k and Flickr30k benchmarks, determines that AttenTAVO-Cap (TAVO) obtains SOTA results in captioning in terms of the entire set of evaluation criteria, thoroughly surpassing the results obtained with the standard optimization approaches, namely PSO & GA, as well as HABGRU + AVOA in the SOTA approaches. That is, the model obtained a CIDEr score of 194 in the Flickr8k database and a score of 191 in the Flickr30k database, which are signs that indicate a strong consistency with the human-annotated captions. The BLEU & METEOR scores also confirm the semantic & syntactical aptness. The experiment succeeded in determining that metaheuristic approaches proposed in bio-inspired optimization, like the proposed TAVO, are sturdy alternatives to the traditional learning paradigms, namely in the case of captioning, which corresponds to the search in high-dimensional spaces. Although the proposed approach obtained outstanding results, more studies may investigate the extension to the transformer framework, cross-lingual captioning, and real-time captioning in embedded devices.

The proposed AttenTAVO-Cap model performs well; however, it has only been validated on medium-scale datasets and based on the CNN–BiGRU backbone architectures. Concerning Q-I, the results confirm that the integration of deep captioning architectures and metaheuristics leads to a substantial improvement in semantic alignment and captioning. Moving on to Q-II, TAVO outperforms GA and PSO on all parameters of evaluation in each case. Finally, the model achieves a competent trade-off concerning Q-III but requires further investigation on real-time feasibility concerning transformer-scale scenarios.

In the future, we will extend the AttenTAVO-Cap model using Transformer-based decoders to implement the generation of the sequences in the model efficiently. In addition, there are plans to utilize multimodal pretraining and larger and more diverse datasets, such as the MS-COCO dataset, to achieve higher levels of generalization in the model. There will also be the inclusion of explanation modules, such as Grad-CAM and attention maps, to improve the interpretability of the model.

References

[1] Jiajun Du et al., “Attend More Times for Image Captioning,” *arXiv Preprint*, pp. 1-8, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[2] Roberto Castro et al., “Deep Learning Approaches based on Transformer Architectures for Image Captioning Tasks,” *IEEE Access*, vol. 10, pp. 33679-33694, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[3] Jun Yu et al., “Multimodal Transformer with Multi-View Visual Representation for Image Captioning,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 12, pp. 4467-4480, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[4] Feicheng Huang et al., “Boost Image Captioning with Knowledge Reasoning,” *Machine Learning*, vol. 109, no. 12, pp. 2313-2332, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[5] Zhihong Zeng, and Xiaowen Li, “Application of Human Computing in Image Captioning Under Deep Learning,” *Microsystem Technologies*, vol. 27, no. 4, pp. 1687-1692, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[6] Matteo Stefanini et al., “From Show to Tell: A Survey on Deep Learning-based Image Captioning,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 1, pp. 539-559, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[7] Xiaowei Hu et al., “Scaling Up Vision-Language Pre-Training for Image Captioning,” *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, New Orleans, LA, USA, pp. 17980-17989, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[8] Yehao Li et al., “Comprehending and Ordering Semantics for Image Captioning,” *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, New Orleans, LA, USA, pp. 17969-17978, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[9] Mithilesh Mahajan et al., “Image Captioning-A Comprehensive Encoder-Decoder Approach on Flickr8K,” *2025 International Conference on Automation and Computation (AUTOCOM)*, Dehradun, India, pp. 1310-1315, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[10] Valavala S.S.S.R. Manikumar, and G. Bharathi Mohan, “Comparative Study of Deep Learning Algorithms for Image Caption Generation,” *Proceedings of the International Conference on Advances and Applications in Artificial Intelligence (ICAAAI 2025)*, pp. 160-178, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[11] Ahmad Maaz et al., “VGG Models in Image Captioning: Which Architecture Delivers Better Descriptions?,” *2024 18th International Conference on Open Source Systems and Technologies (ICOSST)*, Lahore, Pakistan, pp. 1-6, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[12] Jianjun Xia et al., “Research on Image Tibetan Caption Generation Method Fusion Attention Mechanism,” *2023 IEEE 4th International Conference on Pattern Recognition and Machine Learning (PRML)*, Urumqi, China, pp. 193-198, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[13] Duy Thuy Thi Nguyen, and Hai Thanh Nguyen, *Image Caption Generator with a Combination Between Convolutional Neural Network and Long Short-Term Memory*, Biomedical and Other Applications of Soft Computing, Springer, Cham, pp. 225-238, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[14] Liming Xu et al., “Deep Image Captioning: A Review of Methods, Trends and Future Challenges,” *Neurocomputing*, vol. 546, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[15] Khang Nhut Lam et al., “Vision Transformer and Bidirectional RoBERTa: A Hybrid Image Captioning Model Between VirTex and CPTR,” *International Advanced Computing Conference*, Kolhapur, India, vol. 1781, pp. 124-137, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[16] Christian Szegedy et al., “Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning,” *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31, no. 1, pp. 4278-4284, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[17] P. Hemashree et al., “Recuperating Image Captioning with Genetic Algorithm and Red Deer Optimization: A Comparative Study,” *International Conference on Data Science and Applications*, Jaipur, India, vol. 821, pp. 375-385, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[18] James Kennedy, and Russell C. Eberhart, “Particle Swarm Optimization,” *Proceedings of ICNN'95 - International Conference on Neural Networks*, Perth, WA, Australia, vol. 4, pp. 1942-1948, 1995. [CrossRef] [Google Scholar] [Publisher Link]

[19] Tariq Shahzad et al., “Mamba-Caption: Long-Range Sequence Modelling for Efficient and Accurate Image Captioning,” *Array*, vol. 28, pp. 1-13, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[20] Madhvi Patel et al., “Enhanced Image Captioning with Advanced Context-Aware Object Relational Model,” *Discover Computing*, vol. 28, no. 1, pp.1-27, 2025. [CrossRef] [Google Scholar] [Publisher Link]