International Journal of Computational and Experimental Science and ENgineering (IJCESEN)

Vol. 11-No.4 (2025) pp. 10338-10349
http://www.ijcesen.com

ISSN: 2149-9144

Research Article

Advancing Multilingual Caption Generation with Multi-View Encoders and Triple-Stage Transformer Decoding

Anjali Sharma[1*], Mayank Aggarwal[2]

[1]Department of Computer Science and Engineering, Gurukula Kangri (Deemed to be) University, Haridwar, India
*** Corresponding Author Email:** 23631001@gkv.ac.in - **ORCID:** Orcid: 0009-0004-1757-4622

[2]Department of Computer Science and Engineering, Gurukula Kangri (Deemed to be) University, Haridwar, India
Email: mayank@gkv.ac.in - **ORCID:** 0000-0003-1778-2080

Article Info:

DOI: 10.22399/ijcesen.4864
Received : 05 November 2025
Revised : 15 December 2025
Accepted : 20 December 2025

Keywords

Multilingual Image Captioning, multi-View Visual Encoding, Attention Mechanism, Transformer-based Decoder

Abstract:

This work introduces a multilingual image captioning framework that leverages complementary visual repre- sentations through a multi-view encoder and a triple-stage transformer-based decoder. The encoder integrates hierarchical visual features by combining ConvNeXt, which provides strong semantic and contextual rep- resentations, with Swin Transformer, which captures fine-grained local details. A Gated Attention Fusion module unifies these views into comprehensive visual embeddings. The decoder operates in three stages: initial coarse caption generation, syntactic refinement, and final multilingual translation using a pre-trained mBART (Multilingual BART) model. This modular design enables effective multilingual captioning without requiring parallel datasets. Experiments on the MS-COCO dataset, demonstrate that the proposed system outperforms existing baselines. It achieves BLEU-4 scores of 0.53 (Hindi) and 0.52 (English), CIDEr scores of 0.94 and 0.91, and F1 scores of 0.88 and 0.95, respectively. Furthermore, the system attains Word Error Rates (WER) of 0.10 in English and 0.25 in Hindi, indicating strong fluency and semantic coherence. These results highlight the scalability and effectiveness of the approach for real-world multilingual captioning tasks

1. Introduction

Combining computer vision and natural language processing to create semantically rich and linguistically flexible captions from pictures is a cumbersome task **[1]**. Due to recent developments in this sector, numerous systems have been developed that improve the calibre and applicability of produced descriptions. These models focus on understanding and translating visual content into logical textual descriptions to handle problems like coherence, accuracy, and semantic richness.

It iteratively improves semantic alignment using a multi-step refining technique, producing comprehensive captions that capture subtle visual correlations **[2]**. Semantic Scenes Encoder (SSE) extracts scene graphs from images and integrates them into the encoding process, improving the understanding of overall scenes. SSE outperforms conventional region-based techniques in producing thorough and cohesive captions by fusing encoded visual information with learnt semantic data **[3]**. This approach creates visual embeddings for objects and their connections. Using an attention-based sequence-to-sequence paradigm has improved performance in generating captions accurately describing image content **[4]**. CLIP encodings serve as rich image features for a textual decoder, enabling the generation of meaningful captions without additional attention mechanisms. This method produces relevant captions for unseen images, demonstrating robustness **[5]**.

Using a hybrid deep learning framework combining CNNs (VGG16/ResNet-50) **[6]**. and LSTMs **[7]**. for automated image cap- tioning, enhanced by Transformer-based attention mechanisms to align visual and textual semantics have proved effective. Evaluated on Flickr **[8]**. K using BLEU, METEOR, and ROUGE metrics, the system demonstrates robust performance in generating context-aware captions, offering a scalable solution to bridge visual content with interpretable language for modern digital Picture captioning technology can enhance image retrieval systems and assist those who are visually impaired **[9]**. However, it might be difficult to ensure

that produced captions adequately depict complicated visual situations and are objective

Table 1 *Key Challenges in Multilingual Image Captioning*

Challenge	Description and Citation
Visual-Textual Alignment	Mapping high-dimensional visual features to sequential language requires precise cross-modal attention mecha- nisms. Poor alignment affects caption coherence and accuracy. [11]
Multi-Scale Feature Representation	Combining global scene context with localized object-level details is challenging, particularly for single-branch encoder models. [12]
Linguistic Diversity	Multilingual captioning must support languages with varying syntax, morphology, and word orders, demanding a flexible decoding framework. [13]
Unified End-to-End Training	Integrating vision encoding, attention-based decoding, and multilingual translation increases computational and architectural complexity. [14]

1.1 Motivation

Although encoder-decoder frameworks have advanced significantly, most current models are created for monolingual outputs and only use one visual feature, which frequently restricts their capacity to generalize across intricate scenes and languages. Picture captioning systems that are both visually comprehensive and language-agnostic are needed in today's globally interconnected society. This requires models that can accurately represent a variety of visual semantics in many languages, from fine-grained item characteristics to global scene composition. Due to their absence of cross-lingual adaptation mechanisms and multi-scale visual representation, traditional CNN-based encoders frequently fall short of meeting this dual need.

Recent advances in multi-head attention, vision transformers, and multilingual pre-trained language models (e.g., mBART) **[10]** present new opportunities to bridge these gaps. By integrating multi-view vision encoders with advanced transformer-based decoders, the work aims to make a systems that understand images more deeply and describe them accurately across different languages.

1.2 Challenges

Despite its potential, multilingual image captioning presents several technical challenges. Key challenges are shown in Table 1:

1.3 Contributions

To address these issues, this work suggests a picture captioning architecture with the following key contributions:

1. Multi-View Visual Encoding: The work employs a dual-branch encoder combining ConvNeXt for global scene under- standing and Swin Transformer for local, attention-based feature extraction. A Gated Attention Fusion module integrates these features into a unified representation.
2. Triple-Stage Transformer Decoder: The decoder consists of three stages:
   - Coarse Decoder: Generates an initial caption draft.
   - Refine Decoder: Improves grammatical coherence and structural consistency.
   - Translation Decoder: Uses a pretrained mBART model to produce fluent, accurate translations in the target language.
3. End-to-End Trainable Pipeline: The full system is implemented as a unified architecture trainable on paired image- caption datasets. The use of mBART allows for the translation into new languages without additional training. Notably, the architecture eliminates the need for manual translation of source data, enabling multilingual caption generation without relying on parallel multilingual datasets.

This architecture provides a robust, scalable solution to multilingual image captioning, balancing visual richness with linguistic adaptability.

2. Related work

2.1 CNN-RNN Architectures in Early Image Captioning

Encoder–decoder approach in image captioning was pioneered by combining CNN-based visual encoders with recurrent neural network (RNN) decoders. The Show and Tell model **[15]** used Inception-v3 with LSTMs to generate captions, while Donahue et al. **[16]** introduced Long-term Recurrent Convolutional Networks (LRCNs) for handling temporal dependencies in video captioning. Attribute-guided captioning, as in Jia et al. **[17]**, improved relevance by conditioning generation on semantic attributes detected by CNNs.

The application of attention mechanisms marked a turning point. Xu et al. **[18]** developed Show, Attend and Tell, dynamically attending to different image regions, while Anderson et al. **[19]** combined bottom- up object detection with top-down LSTM-based attention, achieving cutting edge performance on COCO. Pan et al. **[20]** further advanced this by introducing X-Linear Attention Networks with bilinear pooling.

Limitations: Despite improvements, CNN-RNN frameworks suffer from three inherent shortcomings: (1) difficulty modeling long-range dependencies in captions due to recurrent structures, (2) poor parallelization compared to attention-based architectures, and (3) limited capacity to jointly encode hierarchical visual features. These limitations motivated the shift toward Transformer-based captioning.

2.2 Transformers and Visual-Language Pretraining

Transformers **[21]** introduced parallel token processing and self-attention, making them well-suited for captioning. Initial adaptations such as AoA networks **[22]** improved cross-modal attention by refining interactions between visual and textual embeddings. Subsequent models, including OSCAR 23 and VinVL **[24]**, leveraged large-scale vision-language pretraining to align object tags and captions, significantly boosting performance. Similarly, SimVLM **[25]** and OFA **[26]** unified captioning, VQA, and translation through prefix language modeling.

Limitations: These models deliver strong monolingual performance but remain English-centric, limiting their applicability in multilingual contexts. Moreover, reliance on massive image-text datasets (hundreds of millions of pairs) creates barriers for extending to low-resource languages.

2.3 Multilingual Image Captioning

With the rise of multilingual pre-trained language models, captioning systems have begun exploring multilingual generation. mBART **[10]** enables cross-lingual text generation by pretraining across multiple languages. The UC2 framework **[27]** introduced cross- lingual contrastive learning, aligning multilingual text embeddings with visual features. Zero-shot transfer has also been demonstrated: Rusli et al. **[28]** showed that XLM-R can caption in languages like Japanese and Indonesian without direct training data. Limitations: Despite progress, most multilingual captioning approaches rely heavily on machine- translated datasets, propa-gating translation errors and bias. Additionally, the curse of multilinguality **[29]** shows that as the number of supported languages increases, model performance often degrades without proportional scaling in capacity. Finally, many methods treat captioning as a simple translation task, without integrating fine-grained visual grounding for each language.

2.4 Multi-View and Hierarchical Visual Encoding

Recent work has emphasized multi-view visual encoders to better capture complementary information. ViLBERT **[30]** and LXMERT **[31]** used dual-stream processing with cross-modal fusion, while VLDeformer **[32]** combined CLIP's global semantic

Column Descriptions: Attn: Uses attention mechanism; Multi-view: Employs multiple visual perspectives (e.g., ConvNeXt + Swin); VL Pretrain: Vision–language pretraining used; Multilingual: Supports multilingual generation.

embeddings with Swin Transformer patches, achieving strong captioning scores. Hierarchical encoders such as RSTNet **[33]** and SGAE 34 modeled object relationships and scene graphs for improved coherence.

Table 2 *Comparative Overview of Captioning Approaches Discussed in Literature (✓= Present, x = Absent)*

Model / Paper	Attn	Multi-view	VL Pretrain	Multilingual
Show, Attend and Tell[18]	✓	x	x	x
Bottom-Up Top-Down [19]	✓	x	x	x
X-Linear Attention [20]	✓	x	x	x
Meshed-Memory[35]	✓	x	x	x
OSCAR [23]	✓	x	✓	x
VinVL [24]	✓	x	✓	x
SimVLM [25]	✓	✓	✓	x
OFA[26]	✓	✓	✓	x
mBART [15]	x	x	x	✓
UC2 [27]	x	x	✓	✓
ZeroNLG [30]	x	✓	✓	✓
VLDeformer [32]	✓	✓	✓	x
RSTNet [33]	✓	x	x	x
SGAE [34]	✓	x	x	x
CLIPCap [37]	x	x	✓	x
CAVP[36]	✓	✓	x	x
Proposed Method (Ours)	✓	✓	✓	✓

Limitations: While these models improve visual grounding, they often neglect multilingual generation. Moreover, many rely on rigid fusion mechanisms (e.g., fixed early/late fusion) that cannot dynamically adapt to input complexity.

2.5 Research Gaps and Opportunities

A comparative overview of prior works is shown in Table 2. From this analysis, three major gaps emerge:

1. Monolingual Focus: Most captioning systems remain English-only, limiting accessibility in multilingual contexts.
2. Incomplete Visual Representation: Existing encoders often capture either global or local features, but rarely both in a unified manner.
3. Dependence on Parallel Corpora: Many multilingual systems require large-scale parallel datasets, which are scarce for low-resource languages.

Our Approach: To overcome these gaps, we propose a multilingual captioning model with three innovations: (1) a dual- branch multi-view encoder combining ConvNeXt (global semantics) and Swin Transformer (local details), (2) a gated fusion mechanism for adaptive integration of visual perspectives, and (3) a triple-stage decoder with mBART to refine and translate captions without relying on parallel multilingual corpora.
This design explicitly addresses the limitations of prior CNN-RNN, Transformer-based, and multilingual captioning approaches, offering a scalable solution for multilingual caption generation with rich visual grounding.

3. Methodology

We propose a multilingual captioning system that couples a multi-view visual encoder with a triple-stage transformer decoder. The encoder produces complementary token sequences from ConvNeXt (convolutional inductive bias) and Swin Transformer (hierarchical self-attention). A Gated Attention Fusion (GAF) module aligns and merges these views into a shared embedding space. Decoding proceeds in three stages: (i) coarse generation, (ii) refinement conditioned on the draft, and (iii) translation with a pretrained mBART model. Figure 1 overviews the pipeline.

Algorithm 1 Multilingual Captioning with Multi-View Encoding and Triple-Stage Decoding
Require: image I, target language L
1: Fg ← ConvNeXt(I); Fl ← Swin(I)
2: Hg ← FgWg; Hl ← FlWl
3: Zg ← Attn(Hg, Hl, Hl); Zl ← Attn(Hl, Hg, Hg)
4: Z ← [Zg; Zl]; E ← GAF(Z)
5: y(1) ← Dec1(E) ▷ coarse draft
6: y(2) ← Dec2([E; Emb(y(1))]) ▷ refine
7: u ← mBART(⟨L⟩, Emb(y(2))) ▷ translate
8: return u

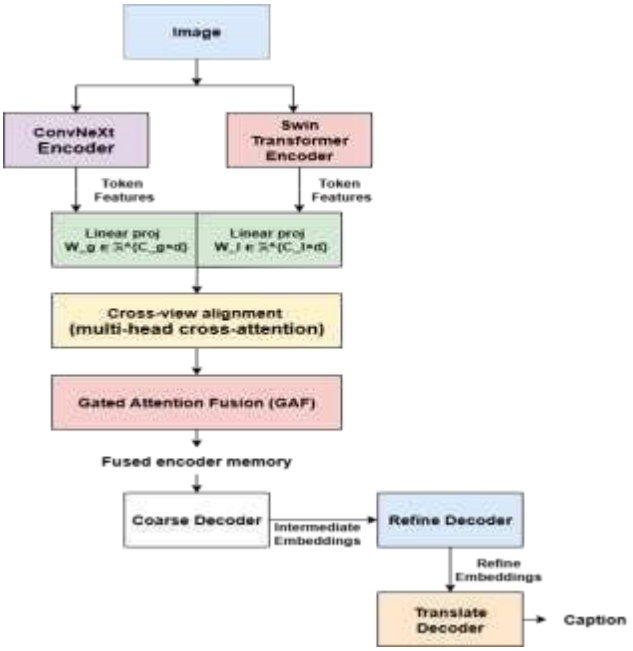

Figure 1 *Overall System Architecture*

3.1 Algorithmic Overview

Algorithm 1 displays all of the recommended stages for the systems.

Table 3 *Training Hyperparameters*

Parameter	**Value**
Optimizer	Adam
Weight Decay	1 × 10−4
Initial Learning Rate	3 × 10−4
Learning Rate Scheduler	Linear warm-up + decay
Batch Size	16
Max Epochs	20
Loss Function	Cross-Entropy
Gradient Clipping	Yes (norm = 1.0)

3.2 Multi-View Encoder

(a) Visual Feature Extraction:

1. Token sequences. ConvNeXt features are spatially pooled into a grid of tokens Fg ∈ RTg×Cg and Swin produces Fl ∈ RTl×Cl . We project both to a common model dimension d:

Hg = FgWg, Hl = FlWl, Wg ∈ RCg×d , Wl ∈ RCl×d .

2. Cross-view alignment. We contextualize each stream with the other via cross-attention:

Zg = Attn HgW(g), HlW(g), HlW(g),
q k v
Zl = Attn HlW(l), HgW(l), HgW(l)
q k v

with W(·) ∈ Rd×dh and multi-head aggregation. We then match lengths by concatenation: Z = [Zg; Zl] ∈ RT×d, T = Tg + Tl.

3. Gated Attention Fusion (GAF). Let U, V ∈ Rd×d. We compute token-wise gates and content transforms on the concatenated sequence:

G = σ(ZU), Z˜ = tanh(ZV), E = LN G ⊙ Z˜ + (1 − G) ⊙ Z ,

where G, Z˜, Z ∈ RT×d. The fused encoder memory E conditions all decoder stages.

(b) Triple-stage decoder

1. Stage 1: Coarse decoder. An autoregressive Transformer decodes a draft caption:

p1(wt | w<t, E) = Softmax FFN(Dec1(E, y<t)) .

2. Stage 2: Refine decoder. A second decoder attends to both E and the draft embeddings Y(1):

p2(wt | w<t, E, Y(1)) = Softmax FFN(Dec2([E; Y(1)], y<t)) .

3. Stage 3: Translate decoder (mBART). We prepend the target language token ⟨L⟩ and translate the refined caption:

p3(ut | u<t, ⟨L⟩, Y(2)) = mBART ⟨L⟩, Y(2), u<t ,

producing tokens ut in language L.

4 Training setup

Table 3 illustrates the training hyperparameters.

5 Experiments

5.1 Dataset

The MS COCO dataset **[39],** a well-used benchmark in image captioning research, was used in this work's tests to assess the efficacy of the suggested captioning system. There are more than 118,000 training and 5000 validation pictures, each with five human-written English subtitles. The dataset is freely available at: https://www.kaggle.com/datasets/awsaf49/coco-2017-dataset To support multilingual generation, the original English captions were augmented using machine translation (used in ablation study). Specifically, the pretrained mBART-50 **[13]** model was used to translate captions into multiple target languages, with a primary focus on Hindi. The translated captions were manually verified for consistency and grammatical correctness in a subset of samples to ensure training quality.

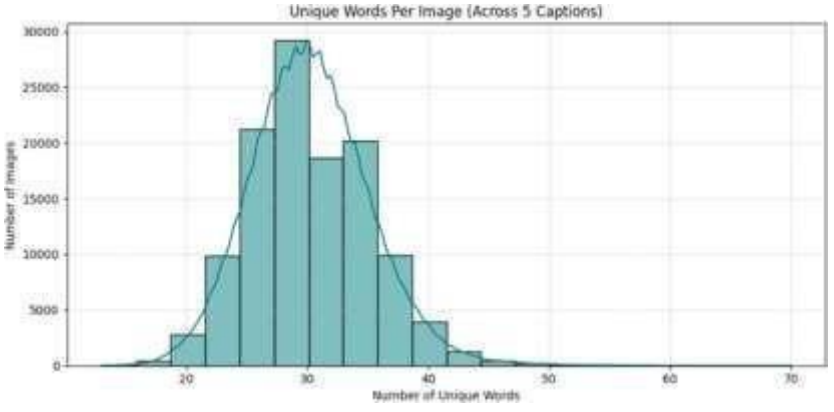

Figure 2 *Unique Words per Image*

Figure 2 visualizes the distribution of unique words per image across five captions in the dataset. Most images contain between 27 to 33 unique words, indicating moderate lexical diversity in human annotations.

5.2 Training Dynamics and Model Stability

Figure 3 illustrates training and validation dynamics over ten epochs. Loss decreases steadily without signs of overfitting, as validation closely tracks training curves. Accuracy rises sharply in early epochs and stabilizes around 82–83% after the sixth

epoch, with minimal train–validation gap. These trends confirm stable convergence and effective learning of image–text representations, supporting robust multilingual caption generation.

5.3 Convergence and Sensitivity Analysis

To evaluate training stability, we examined two behaviors: convergence, measured via validation loss, and sensitivity, measured as fluctuations in validation accuracy (Figure 4). Validation loss drops sharply in the first few epochs and then plateaus, confirming stable convergence. Sensitivity decreases over time, indicating reduced variability and stronger generalization. These results highlight that the proposed Gated Fusion encoder with the triple-stage decoder (including mBART) trains reliably and supports early stopping with minimal risk of overfitting.

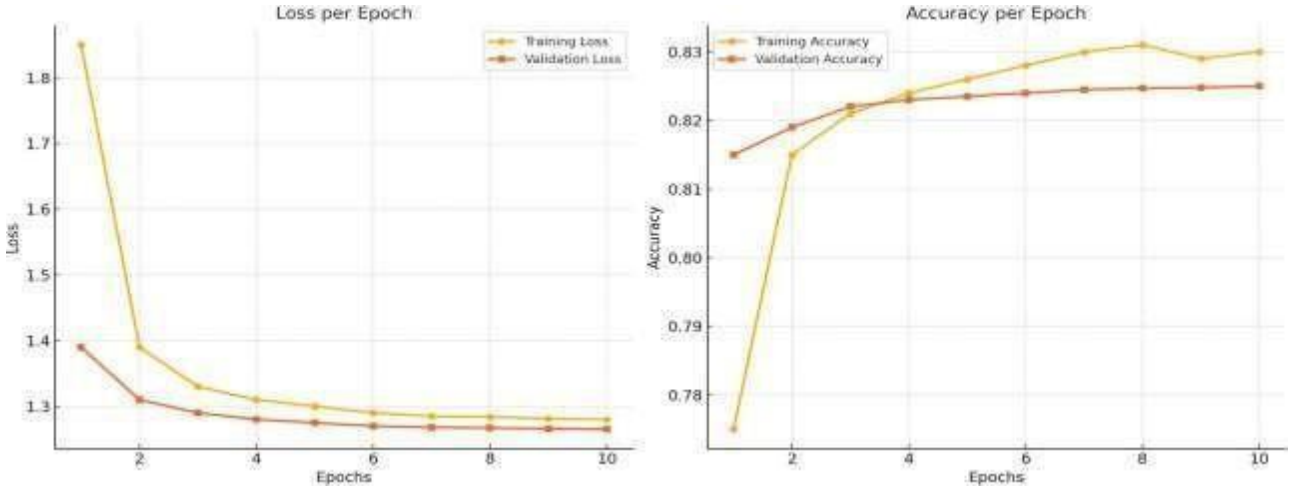

Figure 3 *Training and validation dynamics: Left – loss per epoch, Right – accuracy per epoch.*

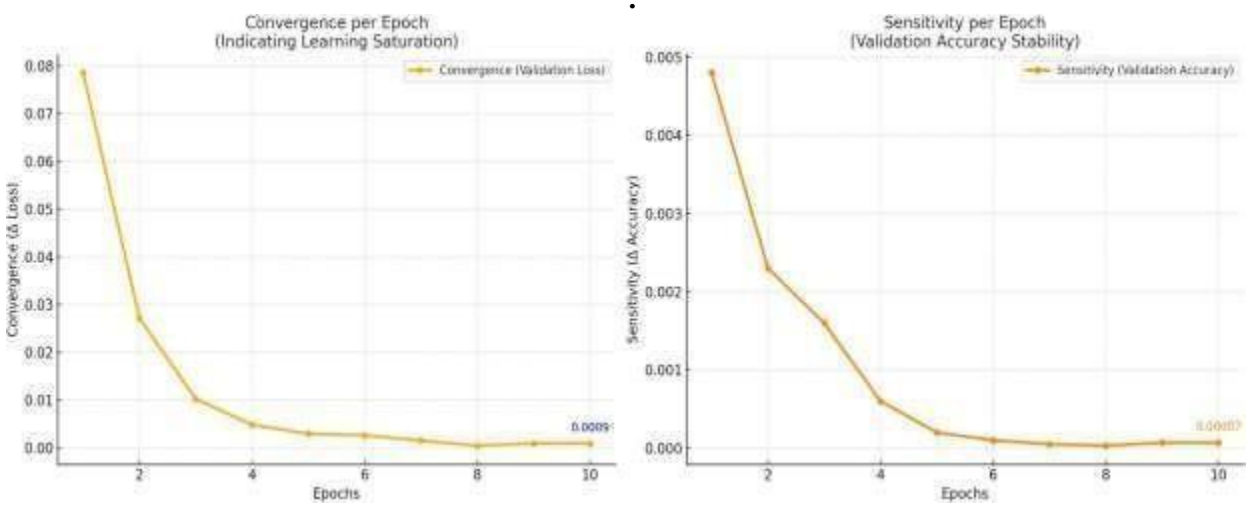

Figure 4 *Left: Convergence measured as change in validation loss between epochs. Right: Sensitivity measured as change in validation accuracy across epochs*

5.4 Evaluation Measures

The work employ a diverse set of metrics to evaluate both the fluency and fidelity of the generated captions:

- BLEU **[40]**: Evaluates n-gram precision between produced and original captions. The work report BLEU-1 through BLEU-4 to capture both lexical overlap and phrase-level correctness.
- METEOR **[41]**: Accounts for synonymy, stemming, and alignment-based scoring to assess semantic relevance.
- CIDEr **[42]**: Measures consensus between generated captions and references using TF-IDF weighted n-gram matching.
- WER **[43]** and CER **[44]**: Word Error Rate and Character Error Rate were used only for English-to- Hindi translation evaluation to assess surface-level differences in generated sequences.

The bar chart in Figure 5 compares English and Hindi captioning performance. English captions achieve higher scores on most metrics, particularly METEOR and F1, reflecting strong lexical and structural alignment. Hindi captions, however, remain

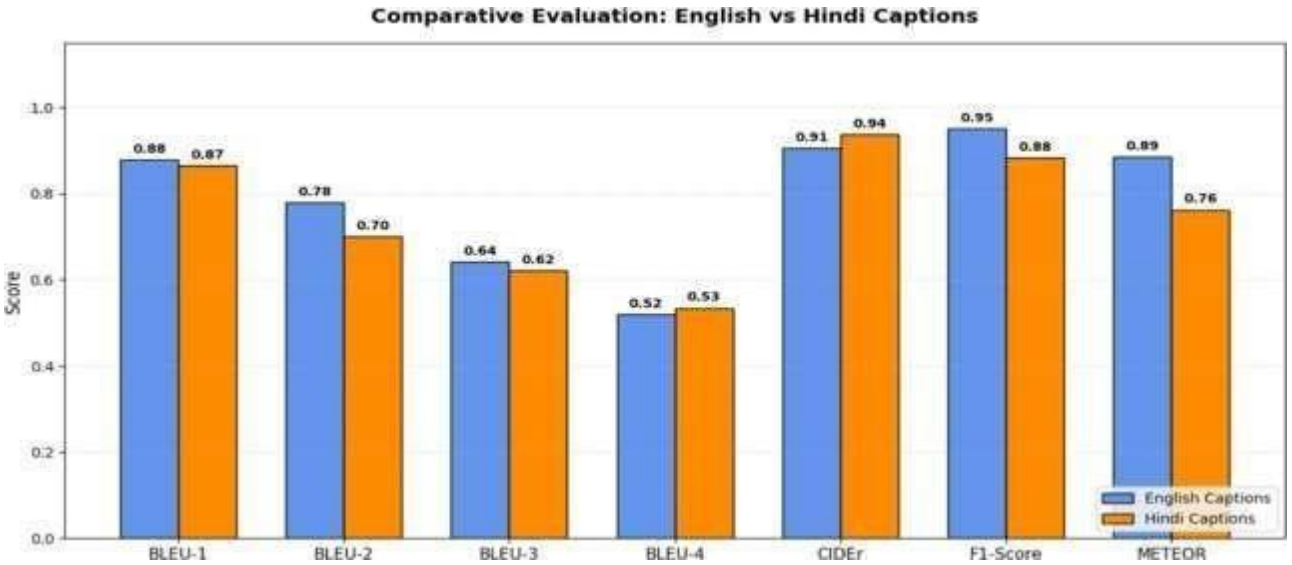

Figure 5 *Comparison Evaluation of English vs. Hindi Image Captions Across Multiple Metrics*

Table 4 *Comparative Metrics: English vs Hindi Captions*

Metric	English Score	Hindi Score
BLEU-1	0.88	0.87
BLEU-2	0.78	0.70
BLEU-3	0.64	0.62
BLEU-4	0.52	0.53
CIDEr	0.91	0.94
F1-Score	0.95	0.88
METEOR	0.89	0.76

Table 5 *WER and CER Comparison: English vs Hindi Captions*

Metric	**English Score**	**Hindi Score**
WER	0.10	0.25
CER	0.12	0.26

competitive, with BLEU-4 (0.53) slightly exceeding English (0.52) and a higher CIDEr score (0.94 vs. 0.91), indicating effective semantic preservation across languages.

Figure 6 and Table 5 show that Hindi captions have higher WER (0.25) and CER (0.26) than English (0.10, 0.12), reflecting greater lexical and morphological complexity. Despite this, the model still produces semantically coherent captions, indicating robustness with scope for further fine-tuning.

5.5 Ablation Study

The work conducts a thorough ablation investigation on the validation set in order to comprehend the contribution of each component inside the suggested design. Three important architectural decisions are assessed:

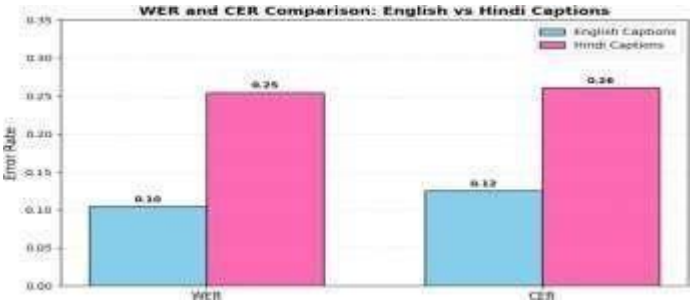

Figure 6 *WER and CER Error Rates for English and Hindi Captions*

5.5.1 Single vs. Multi-View Encoding:

The work compares a baseline model using a single visual encoder (ConvNeXt) against the suggested multi-view encoder which fuses ConvNeXt and Swin outputs. The multi-view setup consistently improves semantic richness and localization precision, leading to improved caption fluency and accuracy. In this configuration, the model is trained on the translated Hindi captions of the COCO dataset.

5.5.2 Dual vs. Triple-Stage Decoding:

Next, the study compare a dual-stage decoder (coarse + refinement) against the suggested triple-stage pipeline (coarse, refinement, and mBART-

based translation). Results show that the addition of the translation stage not only facilitates multilingual output but also improves fluency, particularly in morphologically rich languages like Hindi. The dual-stage model is trained using the Hindi captions from the COCO dataset.

5.5.3 Effect of Multilingual Translation:

To isolate the effect of the translation module, we compare model outputs with and without the mBART decoder. The translated captions are evaluated using both machine metrics (BLEU, METEOR) and human inspection. The results confirm that the mBART module substantially raises the model's capacity to generalize across languages without retraining on target-language data. able 6 and Figure 7 illustrates the outcome of the ablation study.

5.6 Comparisions with state-of-art Models

Table 7 shows that the proposed model outperforms prior English and Hindi captioning methods, achieving the best BLEU scores across all levels. Notably, it reaches 88.1 (BLEU-1) in English and 53.0 (BLEU-4) in Hindi, demonstrating strong generalization across both high- and low-resource languages.

Table 6. *Ablation Study Results on COCO Validation Set*

Configuration	BLEU-4	CIDEr	METEOR	F1-Score
English Captions Single Encoder + Dual Decoder	0.4789	0.8712	0.8123	0.6971
Multi-View Encoder + Dual Decoder	0.5071	0.9125	0.8454	0.7246
Multi-View + Triple Decoder	0.5189	0.9065	0.8860	0.7423
Full Model (Base Language)	**0.5244**	**0.9135**	**0.8903**	**0.9523**
Hindi Captions Single Encoder + Dual Decoder	0.4700	0.8500	0.7200	0.6800
Multi-View Encoder + Dual Decoder	0.5000	0.8800	0.7400	0.7000
Multi-View + Triple Decoder	0.5100	0.9000	0.7500	0.7200
Full Model (with Translation)	**0.5300**	**0.9400**	**0.7600**	**0.8850**

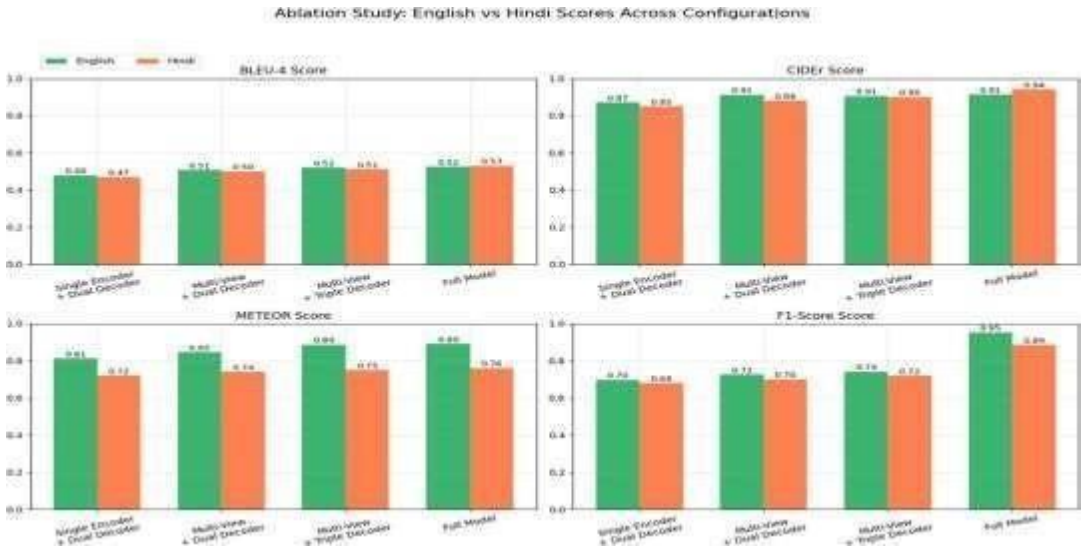

Figure 7 *Ablation Study: English vs Hindi Scores Across Configurations*

5.7 Computational Efficiency

Table 8 summarizes the model's efficiency. With 628M parameters and ∼54 GFLOPs per forward pass, the system balances capacity and cost (Params/FLOPs ratio = 0.67). Inference is efficient (19.36 ms, 2.77 GB GPU memory), making the 2.5 GB model scalable and practical for real-time multilingual captioning.

5.8 Qualitative Analysis

Figure 8 shows examples of reference captions, generated English outputs, and their Hindi translations. The model preserves semantic content well, with close alignment to references and fluent cross-lingual rendering. Minor lexical artifacts and syntactic deviations occur in complex scenes, but overall the outputs remain coherent, demonstrating the system's effectiveness for multilingual captioning.

Table 7 *Comparison of English and Hindi Image Captioning Models*

Model/Authors	B1	B2	B3	B4
English Image Captioning				

INJECT-TAG [45]	76.1	59.1	44.4	33.5
SCST [46]	80.2	—	—	38.0
GAT [47]	81.1	66.1	51.8	39.9
BT [48]	80.5	65.2	50.6	39.7
Neural_Talk [49]	74.0	56.7	43.3	31.3
Ens Caption [50]	81.7	65.3	51.1	39.2
PAG Net [51]	83.2	62.8	46.3	40.8
TRANSKG [52]	76.2	—	—	34.4
Proposed Model (Ours)	**88.1**	**78.21**	**64.34**	**52.44**
Hindi Image Captioning				
Mishra et al. [53]	62.9	43.3	29.1	19.0
P. Singh et al. [54]	51.3	30.4	16.7	12.4
Rijul Dhir [55]	57.0	39.0	26.4	17.3
Ankit Rathi [56]	58.0	47.0	39.0	35.0
Akash Ram Singh [57]	58.0	40.0	27.0	12.0
Virendra Kumar Meghwal [58]	62.5	45.8	32.8	23.2
S. Singh et al. [59]	64.29	48.09	36.55	21.91
S.K. Mishra et al. [60]	35.79	19.97	11.68	6.76
Proposed Model (Ours)	**87.24**	**70.00**	**62.36**	**53.00**

Note: B1–B4 represent BLEU-1 to BLEU-4 scores respectively.

Table 8 *Model Performance Metrics*

Attribute	**Value**
Model Size	2515.12 MB
Total Parameters	628,457,200
Trainable Parameters	628,457,200
Non-trainable Parameters	0
Number of Layers	397
FLOPs (approx.)	54 GFLOPs
Params/FLOPs Ratio	0.67
Inference Time	19.362 ± 1.121 ms
GPU Memory Usage	2769.77 MB

6. Limitations and future work

Despite strong multilingual performance, several limitations remain. First, transfer to truly low-resource languages (e.g., Bengali, Marathi) is hindered by limited pretraining coverage and tokenization mismatches in the multilingual LM, yielding inflectional errors and syntactic drift—a data-coverage, not merely decoding, issue. Second, the coarse→refine→translate cascade can propagate early mistakes; auxiliary losses help, but residual lexical artifacts persist on long, cluttered scenes. Third, semantic grounding weakens under domain shift (e.g., industrial imagery, graphics, or dense text-in-image), increasing hallucination risk. Fourth, dual encoders plus a three-stage decoder increase compute cost during training and impose moderateinference overhead, constraining real-time use on edge devices. Finally, standard metrics (BLEU/METEOR/CIDEr) emphasize lexical overlap and consensus rather than faithfulness and grounding, particularly in morphologically rich targets.

7. Conclusion

We introduced a multilingual image captioning framework that combines a multi-view visual encoder with triple-stage decoding. On the vision side, ConvNeXt and Swin are treated as complementary encoders whose token sequences are cross-aligned and fused via a gated attention mechanism into a shared memory E. On the language side, decoding proceeds from coarse drafts to refinement and culminates in multilingual generation via mBART conditioned on a target language token. A clarified training objective supervises the final translation stage and optionally adds auxiliary losses to stabilize earlier stages.

Empirically, the architecture delivers consistent gains over strong baselines on COCO (English) and transfers effectively to Hindi without using parallel image–Hindi data, indicating that language conditioning at Stage-3 can substitute for costly multilingual caption corpora. Ablations isolate the benefits of (i) multi-view encoding, (ii) cross-view alignment + gated fusion,

(iii) staged decoding, and (iv) auxiliary supervision. Efficiency measurements show the approach is deployable with modest beams and amenable to compression.

Figure 8 *Multilingual image captioning examples with English and Hindi outputs using mBART*

Beyond incremental improvements, the contribution is conceptual: decoupling visual grounding from multilingual surface realization through a staged interface, while enforcing alignment and information flow with an explicit fusion design. Future work will focus on low-resource adaptation, controllable style, stronger grounding under domain shift, and model compression— advancing towards inclusive, reliable, and efficient multilingual captioning.

Author Statements:

- **Ethical approval:** The conducted research is not related to either human or animal use.
- **Conflict of interest:** The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper
- **Acknowledgement:** The authors declare that they have nobody or no-company to acknowledge.
- **Author contributions:** This study was jointly developed by Dr. Mayank Aggarwal and Anjali Sharma through collaborative discussions and refinement. Dr. Aggarwal contributed to the overall research direction and methodological rigor, while Anjali Sharma carried out the primary investigation, analysis, and manuscript preparation. Both authors reviewed and approved the final manuscript.
- **Funding information:** The authors declare that there is no funding to be acknowledged.
- **Data availability statement:** The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.
- **Use of AI Tools:** The author(s) declare that no generative AI or AI-assisted technologies were used in the writing process of this manuscript.

References

[1] Megahed FM, Chen YJ, Colosimo BM, et al. Adapting OpenAI's CLIP Model for Few-Shot Image Inspection in Manufacturing Quality Control: An Expository Case Study with Multiple Application Examples. *arXiv preprint arXiv:2501.12596.* 2025.

[2] Li H, Wang H, Zhang Y, Li L, Ren P. Underwater image captioning: Challenges, models, and datasets. *ISPRS Journal of Photogrammetry and Remote Sensing.* 2025;220:440–453.

[3] Wei H, Li Z, Huang F, Zhang C, Ma H, Shi Z. Integrating scene semantic knowledge into image captioning. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM).* 2021;17(2):1–22.

[4] Khan R, Islam MS, Kanwal K, Iqbal M, Hossain MI, Ye Z. Attention based sequence-to-sequence framework for auto image caption generation. *Journal of Intelligent & Fuzzy Systems.* 2022;43(1):159–170.

[5] Lv G, Sun Y, Nian F, Zhu M, Tang W, Hu Z. COME: Clip-OCR and Master ObjEct for text image captioning. *Image and Vision Computing.* 2023;136:104751.

[6] Theckedath D, Sedamkar R. Detecting affect states using VGG16, ResNet50 and SE-ResNet50 networks. *SN Computer Science.* 2020;1(2):79.

[7] Yu Y, Si X, Hu C, Zhang J. A review of recurrent neural networks: LSTM cells and network architectures. *Neural computation.* 2019;31(7):1235– 1270.

[8] Xu L, Tang Q, Lv J, Zheng B, Zeng X, Li W. Deep image captioning: A review of methods, trends and future challenges. *Neurocomputing.* 2023;546:126287.

[9] Safiya K, Pandian R. A real-time image captioning framework using computer vision to help the visually impaired. *Multimedia Tools and Applications.* 2024;83(20):59413–59438.

[10] Chipman HA, George EI, McCulloch RE, Shively TS. mBART: multidimensional monotone BART. *Bayesian Analysis.* 2022;17(2):515–544.

[11] Suresh KR, Jarapala A, Sudeep P. Image captioning encoder–decoder models using CNN-RNN architectures: A comparative study. *Circuits, Systems, and Signal Processing.* 2022;41(10):5719–5742.

[12] Pan Y, Li Y, Yao T, Mei T. Bottom-up and top-down object inference networks for image captioning. *ACM Transactions on Multimedia Computing, Communications and Applications.* 2023;19(5):1–18.

[13] Liu Y, Gu J, Goyal N, et al. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics.* 2020;8:726–742.

[14] Osman AA, Shalaby MAW, Soliman MM, Elsayed KM. A survey on attention-based models for image captioning. *International Journal of Advanced Computer Science and Applications.* 2023;14(2).

[15] Vinyals O, Toshev A, Bengio S, Erhan D. Show and tell: A neural image caption generator. In: 2015:3156–3164.

[16] Donahue J, Anne Hendricks L, Guadarrama S, et al. Long-term recurrent convolutional networks for visual recognition and description. In: 2015:2625–2634.

[17] Jia X, Gavves E, Fernando B, Tuytelaars T. Guiding the long-short term memory model for image caption generation. In: 2015:2407–2415.

[18] Liu H, Brailsford T. Reproducing "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention". In: . 2589. IOP Publishing. 2023:012012.

[19] Anderson P, He X, Buehler C, et al. Bottom-up and top-down attention for image captioning and visual question answering. In: 2018:6077–6086.

[20] Pan Y, Yao T, Li Y, Mei T. X-linear attention networks for image captioning. In: 2020:10971–10980.

[21] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. *Advances in neural information processing systems.* 2017;30.

[22] Ahsan H, Bhalla N, Bhatt D, Shah K. Multi-modal image captioning for the visually impaired. *arXiv preprint arXiv:2105.08106.* 2021.

[23] Nguyen N, Bi J, Vosoughi A, Tian Y, Fazli P, Xu C. Oscar: Object state captioning and state change representation. *arXiv preprint arXiv:2402.17128.* 2024.

[24] Zhang P, Li X, Hu X, et al. VinVL: making visual representations matter in vision-language models. CoRR abs/2101.00529 (2021). *arXiv preprint arXiv:2101.00529.* 2021.

[25] Wang Z, Yu J, Yu AW, Dai Z, Tsvetkov Y, Cao Y. Simvlm: Simple visual language model pretraining with weak supervision. *arXiv preprint arXiv:2108.10904.* 2021.

[26] Wang P, Yang A, Men R, et al. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In: PMLR. 2022:23318–23340.

[27] Zhou M, Zhou L, Wang S, et al. Uc2: Universal cross-lingual cross-modal vision-and-language pre-training. In: 2021:4155–4165.

[28] Rusli A, Shishido M. On the Applicability of Zero-Shot Cross-Lingual Transfer Learning for Sentiment Classification in Distant Language Pairs. *arXiv preprint arXiv:2412.18188.* 2024.

[29] Pfeiffer J, Goyal N, Lin XV, et al. Lifting the curse of multilinguality by pre-training modular transformers. *arXiv preprint arXiv:2205.06266.* 2022.

[30] Lu J, Batra D, Parikh D, Lee S. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems.* 2019;32.

[31] Tan H, Bansal M. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490.* 2019.

[32] Zhang L, Wu H, Chen Q, et al. VLDeformer: Vision–language decomposed transformer for fast cross-modal retrieval. *Knowledge-Based Systems.* 2022;252:109316.

[33] Tan Y, Wang B, Yan Z, Liu H, Zhang H. RST-Net: a spatio-temporal residual network based on Region-reConStruction algorithm for shared bike prediction. *Complex & Intelligent Systems.* 2023;9(1):81–97.

[34] Yang X, Tang K, Zhang H, Cai J. Auto-encoding scene graphs for image captioning. In: 2019:10685–10694.

[35] Cornia M, Stefanini M, Baraldi L, Cucchiara R. Meshed-memory transformer for image captioning. In: 2020:10578–10587.

[36] Yang B, Liu F, Zou Y, Wu X, Wang Y, Clifton DA. Zeronlg: Aligning and autoencoding domains for zero-shot multimodal and multilingual natural language generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence.* 2024.

[37] Mokady R, Hertz A, Bermano AH. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734.* 2021.

[38] Zha ZJ, Liu D, Zhang H, Zhang Y, Wu F. Context-aware visual policy network for fine-grained image captioning. *IEEE transactions on pattern analysis and machine intelligence.* 2019;44(2):710–722.

[39] Chen X, Fang H, Lin TY, et al. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325.* 2015.

[40] Ghassemiazghandi M. An Evaluation of ChatGPT's Translation Accuracy Using BLEU Score. *Theory and Practice in Language Studies.* 2024;14(4):985–994.

[41] Lavie A, Denkowski MJ. The METEOR metric for automatic evaluation of machine translation. *Machine translation.* 2009;23:105–115.

[42] Young JC, Arthur R, Williams HT. CIDER: Context-sensitive polarity measurement for short-form text. *Plos one.* 2024;19(4):e0299490.
[43] Klakow D, Peters J. Testing the correlation of word error rate and perplexity. *Speech Communication.* 2002;38(1-2):19–28.
[44] James J, Gopinath DP, others . Advocating character error rate for multilingual asr evaluation. *arXiv preprint arXiv:2410.07400.* 2024.
[45] Zhang J, Mei K, Zheng Y, Fan J. Integrating part of speech guidance for image captioning. *IEEE Transactions on Multimedia.* 2020;23:92–104.
[46] Zhou Y, Wang M, Liu D, Hu Z, Zhang H. More grounded image captioning by distilling image-text matching model. In: 2020:4777–4786.
[47] Wang C, Gu X. Dynamic-balanced double-attention fusion for image captioning. *Engineering Applications of Artificial Intelligence.* 2022;114:105194. doi: https://doi.org/10.1016/j.engappai.2022.105194
[48] Cui W, He X, Yao M, et al. Landslide image captioning method based on semantic gate and bi-temporal LSTM. *ISPRS International Journal of Geo-Information.* 2020;9(4):194.
[49] Xiao X, Wang L, Ding K, Xiang S, Pan C. Dense semantic embedding network for image captioning. *Pattern Recognition.* 2019;90:285-296. doi: https://doi.org/10.1016/j.patcog.2019.01.028
[50] Yang M, Liu J, Shen Y, et al. An Ensemble of Generation- and Retrieval-Based Image Captioning With Dual Generator Generative Adversarial Network. *IEEE Transactions on Image Processing.* 2020;29:9627-9640. doi: 10.1109/TIP.2020.3028651
[51] Zhang Z, Cheng B, Wang Z, et al. PAGNet: Pluggable Adaptive Generative Networks for Information Completion in Multi-Agent Communication. *arXiv preprint arXiv:2502.03845.* 2025.
[52] Zhang Y, Shi X, Mi S, Yang X. Image captioning with transformer and knowledge graph. *Pattern Recognition Letters.* 2021;143:43–49.
[53] Mishra SK, Dhir R, Saha S, Bhattacharyya P, Singh AK. Image captioning in Hindi language using transformer networks. *Computers & Electrical Engineering.* 2021;92:107114.
[54] Singh P, Raja F, Sharma H. Generating Image Captions in Hindi Based on Encoder-Decoder Based Deep Learning Techniques. In: , , Springer, 2024:81–94.
[55] Dhir R, Mishra SK, Saha S, Bhattacharyya P. A deep attention based framework for image caption generation in hindi language. *Computación y Sistemas.* 2019;23(3):693–701.
[56] Rathi A. Deep learning apporach for image captioning in Hindi language. In: IEEE. 2020:1–8.
[57] Singh AR. *Generating Semantically Correct Hindi Captions Using Deep Neural Network*. PhD thesis. Dublin, National College of Ireland, 2021.
[58] Meghwal VK, Mittal N, Singh G. A Multiheaded Attention-Based Model for Generating Hindi Captions. In: Springer. 2023:677–684.
[59] Singh J, Garg KK, Panwar A. Effective Image Captioning Using Multi-layer LSTM with Attention Mechanism. In: Springer. 2023:65–73.
[60] Mishra SK, Sinha S, Saha S, Bhattacharyya P. A deep learning based framework for image paragraph generation in hindi. In: 2022:792–800.