IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Mathematical Modelling of Engineering Problems
Vol. 12, No. 9, September, 2025, pp. 3127-3134
Journal homepage: http://iieta.org/journals/mmep

Empirical and Iterative Analysis of Deep Learning Models for Image Captioning Using Systematic Perspective on Metrics, Architectures, and Trade-offs

Kothakonda Chandhar[1,2*], Manchala Sadanandam[1]

[1] Computer Science and Engineering, Kakatiya University, Warangal 506009, India
[2] Computer Science and Engineering (AI&ML), Kakatiya Institute of Technology & Science, Warangal 506015, India

Corresponding Author Email: chandu19024@gmail.com

https://doi.org/10.18280/mmep.120917

Received: 13 June 2025
Revised: 12 August 2025
Accepted: 18 August 2025
Available online: 30 September 2025

Keywords:
image captioning, deep learning, BLEU score, ROUGE-L, transformer, LSTM, multimodal scenarios

ABSTRACT

Image captioning integrates computer vision and natural language processing, requiring both accurate visual understanding and coherent language generation. While diverse deep learning approaches ranging from encoder–decoder models to Transformer-based architectures have emerged, few studies provide standardized, empirical comparisons across models. This work addresses that gap through a systematic and iterative evaluation, where performance insights are refined over successive analysis cycles to ensure reliability. The study benchmarks recent models using five key dimensions: latency, computational complexity, accuracy, Bilingual Evaluation Understudy (BLEU), and Recall-Oriented Understudy for Gisting Evaluation-Longest Common Subsequence (ROUGE-L). Evaluations consider architectural design (Long Short-Term Memory (LSTM), Transformer, hybrid), feature-extraction strategies (global Convolutional Neural Network (CNN) features vs. object-level detection), attention mechanisms, and training paradigms such as self-supervised learning. To improve interpretability, we introduce a multi-modal tabular and visual framework that combines comparative tables with performance plots, thereby enabling clear observation of trade-offs between accuracy and efficiency. The findings show Transformer-based architectures achieve the highest Consensus-based Image Description Evaluation (CIDEr) and BLEU scores on Microsoft Common Objects in Context (MS COCO) and Flickr datasets, while lightweight models offer competitive performance for real-time use cases. Gaps remain in handling language diversity, explainability, and domain generalization. By offering a reproducible benchmarking approach and actionable insights, this work aids researchers and practitioners in selecting and optimizing captioning models under varying operational constraints.

1. INTRODUCTION

Image captioning, the process of generating coherent and semantically accurate natural language descriptions for visual content, has emerged as a pivotal problem in Artificial Intelligence (AI). It requires a seamless integration of computer vision for visual understanding and natural language processing (NLP) for sentence generation. The ability to produce high-quality captions has far-reaching applications, including assistive technologies for the visually impaired, content-based image retrieval, human–computer interaction, and context-aware media generation.

Recent advances in deep learning have accelerated progress in this field, with architectures evolving from early Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) encoder–decoder frameworks to Transformer-based multimodal architectures capable of modeling complex cross-modal relationships. These advancements have produced a variety of approaches differing in architectural design, feature-extraction strategies, attention mechanisms, and training paradigms such as self-supervised and multitask learning.

However, as diversity increases, the challenge of effective evaluation and comparison becomes more pronounced. Current literature reviews in image captioning are predominantly descriptive, summarizing architectures without providing standardized, empirical, metric-based comparisons. Many lack reproducibility standards, making it difficult to validate findings or conduct fair cross-model comparisons. Furthermore, existing surveys often generalize categories without closely examining performance using well-established evaluation metrics such as Bilingual Evaluation Understudy (BLEU), Consensus-based Image Description Evaluation (CIDEr), Metric for Evaluation of Translation with Explicit ORdering (METEOR), and Recall-Oriented Understudy for Gisting Evaluation-Longest Common Subsequence (ROUGE-L). This limits their utility for researchers or practitioners who must select models based on operational constraints like latency, hardware limitations, or domain-specific requirements.

This paper seeks to bridge these gaps by introducing a reproducible, multi-metric benchmarking framework for

recent image captioning models. Our approach is empirical, and iterative performance insights are refined through successive evaluation cycles, ensuring robustness and reliability. Models are assessed across five dimensions: latency, computational complexity, accuracy, BLEU, and ROUGE-L, with results presented in both raw and normalized forms. The framework includes multi-modal tabular and visual representations, enabling intuitive observation of trade-offs between performance and efficiency.

The main contributions of this work are:

- Comprehensive empirical benchmarking of recent image captioning models, integrating both quantitative metrics and qualitative insights.
- A unified, reproducible evaluation protocol that ensures transparency and facilitates fair cross-model comparison.
- Categorization of models by methodological features, including attention mechanisms (e.g., dual self-attention, cross-modal alignment) and feature-extraction strategies (e.g., global CNN features, object-level detection).
- Identification of research gaps in language-specific captioning, emotion-aware caption generation, and explainability.

Through this structured synthesis, we aim to provide a consolidated reference point and decision-making guide for researchers, practitioners, and system designers, supporting the development of next-generation image captioning systems tailored to diverse application contexts.

2. REVIEW OF EXISTING MODELS USED FOR IMAGE CAPTIONING ANALYSIS

Image captioning, an interdisciplinary domain bridging computer vision and natural language processing, has evolved significantly with deep learning. The core objective remains producing semantically accurate and linguistically coherent descriptions of visual content, a task requiring precise object recognition and fluent sentence generation. Existing models can be systematically categorized into four broad groups:

(i) LSTM-based encoder–decoder architectures,

(ii) Transformer-based models,

(iii) Self-supervised and semi-supervised frameworks,

(iv) Lightweight or multi-task systems. This taxonomy provides a structured basis for evaluating design trade-offs, computational efficiency, and domain adaptability.

Table 1 presents the model's empirical review analysis.

Table 1. Model's empirical review analysis

References	Method Used	Findings	Strengths	Limitations
[1]	Hierarchical Clustering + LSTM	Examines clustering for data reduction and compares LSTM variants	Efficient in reducing data redundancy; improved performance on MS-COCO	Limited to LSTM-based architectures; lacks attention mechanisms
[2]	Multitask DenseNet201 Encoder-Decoder	Demonstrates transfer learning benefits across tasks	Robust and adaptable across tasks; strong regularization	High complexity; possible overfitting without tuning
[3]	Systematic Literature Review	Aggregates trends across 548 studies; identifies core models and metrics	Comprehensive overview; identifies gaps	Non-empirical; lacks new model proposal
[4]	Recurrent Fusion Transformer (RCT)	Combines recurrent attention and feature fusion in Transformer	Enhanced semantic understanding; competitive performance	Transformer complexity may limit deployment
[5]	SMOT: Self-supervised Modal Optimization Transformer	Leverages self-supervision for cross-modal optimization	Performs well with limited data; robust semantic alignment	Relies on high-quality pretext tasks
[6]	Survey on Automatic Image Captioning	Highlights attention-based models and challenges like language diversity	Covers emerging research directions	Descriptive; lacks empirical validation
[7]	ETransCap: Lightweight Transformer	Emphasizes linear complexity for real-time captioning	High efficiency; real-time potential	Trade-off between speed and expressiveness
[8]	V16HP1365 Encoder + Dual Self-Attention	Combines spatial encoding with GRU decoding	Captures diverse visual semantics	Limited validation across diverse datasets
[9]	Neuraltalk+ with Context-Aware Fusion	Introduces real-time captioning with similarity comparison	Fast training; supports assistive tech	Less tested on large-scale benchmarks
[10]	SCAP: Lightweight Sifting + Hierarchical Decoding	Hierarchical decoding aligns visual and textual semantics	Effective for low-resource settings	Simplistic modeling of high-level semantics
[11]	Next-LSTM (ResNeXt + LSTM)	Improves LSTM captioning with advanced visual features	Better generalization on Flickr8k	Relies heavily on image encoder quality
[12]	FeiM: Grid Features + Transformer	Explores learnable feature queries for better alignment	Strong local-global contextual modeling	Grid features may be computationally demanding
[13]	Dilated ResNet + Attention + SE Module	Improves receptive field and feature selection	Enhanced contextual capture	Complex integration of modules
[14]	BMFNet: Bidirectional Multimodal Fusion	Dual-path cross-attention with multimodal fusion	Improved CIDEr; deep feature interaction	Increased model complexity
[15]	Weakly Supervised Grounded Captioning	Estimates region-word alignment without annotations	Reduces annotation cost; robust alignment	Semantic matching sensitive to noise
[16]	DVAT: Dual Visual Align-Cross Attention	Integrates region/grid with cross attention	High accuracy and speed; strong visual fusion	Requires optimal region segmentation
[17]	BIANet: Bidirectional Interactive Alignment	Cross-feature alignment between grid and region paths	Improved semantic alignment	Relatively high training cost
[18]	Emotion-Aware GAN (ResNet + Capsule Net)	Generates sentiment-rich captions	Effective emotional expression	Emotion classification challenges

[19]	SEA: Self-Enhanced Attention	Refines attention weights for better feature focus	Improves CIDEr; emphasizes key regions	Limited novelty beyond attention tuning
[20]	TAVOHDL-ICS: Hybrid DL + Optimization	Bio-inspired hyperparameter tuning with hybrid encoder-decoder	Outperforms on small datasets; robust optimization	Complex and less interpretable architecture

(a) LSTM-based encoder–decoder architectures

Early deep learning approaches used CNNs (e.g., Visual Geometry Group (VGG), ResNet) for image encoding, followed by RNNs, particularly LSTMs—for sequence generation. Rahman et al. [1] proposed a method that improves LSTM-based captioning via hierarchical clustering to reduce feature redundancy, thereby lowering computational load. Empirical analysis shows that stacked LSTMs slightly improve BLEU scores on MS-COCO but at the cost of increased inference latency. Such architectures remain effective for moderate-sized datasets but struggle with long-range dependency modeling compared to modern attention-based systems.

(b) Transformer-based models

Attention-based architectures have transformed captioning by capturing global context without sequential bottlenecks. The Recurrent Fusion Transformer in the study [4] combines recurrent gating with multi-head self-attention, improving semantic coherence by modeling fine-grained feature interactions. While these models outperform LSTM baselines in accuracy and CIDEr scores, they are computationally heavier, making them less suitable for resource-constrained environments. Their strength lies in complex relational reasoning, but they require careful regularization to avoid overfitting on small datasets.

(c) Self-supervised and semi-supervised frameworks

To reduce reliance on large annotated datasets, the self-supervised modal optimization transformer (SMOT) [5] synchronizes cross-modal embeddings using contrastive objectives. This enables competitive performance in low-data regimes, addressing a major limitation of fully supervised captioning. The trade-off is that performance still lags behind supervised transformers on high-resource datasets, but these methods are highly promising for domain adaptation and low-resource languages.

(d) Lightweight and multi-task systems

Multi-task learning extends captioning models to serve multiple vision tasks with shared encoders. For example, Bayisa et al. [2] introduced a tensor-based DenseNet201 backbone supporting classification, detection, and captioning, with task-specific decoders. This approach improves generalizability and reduces model duplication, but sharing representations can introduce task interference, where optimizing one task harms another. Such systems are particularly attractive for edge deployment due to reduced model size and unified maintenance.

Critical Observations: Prior surveys [3, 6] provide valuable historical context, datasets (MS COCO, Flickr8k/30k), and evaluation metrics (BLEU, METEOR, ROUGE-L, CIDEr, SPICE), but often lack empirical cross-comparisons under standardized conditions. Our analysis highlights clear trade-offs:

- LSTM-based models: computationally lighter, weaker at long dependencies.
- Transformers: highest accuracy, higher computational demand.
- Self-supervised: data-efficient, slightly lower peak performance.
- Multi-task: efficient deployment, potential task interference.

This categorization enables a clearer comparison framework and sets the stage for the empirical evaluation in the following sections.

Iteratively in Table 1, efficiency and lightweight architectures have also been an area for research, especially for real-time or resource-oriented deployment. ETransCap in the study [7] is a transformer model characterized by linear complexity and is optimized for computational effectiveness, while SCAP in the study [10] proposes a sifting attention mechanism alongside a hierarchical decoding approach for proper yet computationally cheap captioning. Also, Neuraltalk+ [9], combining dual context-aware fusion and a lightweight self-attention decoder, exhibits faster convergence with real-time assistive scope. Furthermore, the architectural novelty is being extended in the study [8] by merging the V16HP1365 encoder with a dual self-attention network and GRU-based decoder, hence capturing spatial diversities among visual features accompanied by context refinements via attention. The BMFNet [14] adopts a bidirectional multimodal fusion strategy to enhance visual-semantic representation through cross-attention mechanisms and channel-level fusion. This would allow deeper interactivity between image regions and caption tokens and, thus, was reported to gain an extra 2.8% in CIDEr sets. Attention mechanisms are still the backbone of modern image captioning systems. SEA [19] refines classical self-attention by re-weighting attention based on internal distributions to focus on salient features.

DVAT [16] and BIANet [17] propose dual-path and bidirectional alignment architectures, respectively, to facilitate deep interaction between grid and region features thereby reinforcing semantic alignment. These two models show a dominant performance on the MS COCO benchmark, thereby marking the importance of visual-textual co-adaptations.

Grounded captioning, which aligns text components with their respective image regions, is the focus of the study conducted by Rashied and Jeribi [21]. This approach, presented by Du et al. [15], uses weakly supervised semantic matching loss and region-word matching to avoid completely relying on exhaustive annotations. Likewise, TAVOHDL-ICS [20] uses a bio-inspired optimization strategy when tuning hyperparameters as part of a deeper hybrid framework combining Inception ResNetv2, BERT embeddings, and bidirectional GRUs. In constrained datasets like Flickr400 sets, the system showed improvement as generalization and captioning accuracy improved. The sentiment-aware generation mechanism introduced by Yang et al. [18] under a GAN-based approach for fine-grained captioning captures the positive and negative emotional tones separately. Providing a better lens to assess emotional alignment in captions is a capsule-based discriminator. Further, dilated convolutions have been explored by Li et al. [13] for devising larger receptive fields in feature maps, whereby contextual feature extraction is enhanced in ResNet. Finally, in combination with attention and squeeze-and-excitation modules, a further major improvement in caption accuracy and semantic richness for the process can be remarked. In this way, the recently proposed FeiM model conducted by Yan et al. [12] integrates grid feature representations with a state-of-the-art feature

interaction module to enhance local-global context integration. It also allows learnable feature queries to be imposed in a transformer set-up, thus pushing the boundaries of caption generation in fine-grained visual understanding. This research trajectory in image captioning appears to have continued its progressive shift from traditional LSTM-based approaches to increasingly hybrid, more sophisticated transformer models. Bidirectional attention, multimodal fusion, and self-supervised learning are promising innovations of this new form that are set to revolutionize the caption quality, efficiency, and generalizability. Empirical evidence from studies [1-20] generally supports the idea that task-specific feature extraction, modality alignment, and contextual reasoning should be the critical pillars for next-generation image captioning systems.

Figure 1. Model's integrated result analysis

3. COMPARATIVE RESULT ANALYSIS

To objectively and empirically compare image captions that are current, a synthesis aligned to PRISMA was conducted based on the most recent peer-reviewed studies in process. It also lists model-specific design choices, performance across widely accepted benchmarks, and the observed trade-offs in terms of accuracy, scalability, and resource demands. Commonly, studies adopted evaluation metrics such as BLEU, CIDEr, METEOR, and ROUGE-L, with the MS COCO dataset as the common evaluation ground set. Where exact performance metrics were not reported, approximate values were inferred, generally based on architectural complexity and benchmark norms for the process. An overview of these comparative findings is depicted in Figure 1, and a detailed analysis is in Table 2 as follows:

Table 2. Model's statistical review analysis

Reference	Method Used	Dataset Used	Performance Metrics	Key Findings	Strengths	Limitations
[1]	Hierarchical Clustering + LSTM	MS COCO	BLEU, CIDEr	Stacked LSTM improves accuracy over single LSTM	Reduced computational load, high LSTM interpretability	Does not use attention; scalability limited
[2]	DenseNet201 Multitask Encoder-Decoder	MS COCO, ImageNet	BLEU, METEOR	Performs competitively across tasks; strong feature reuse	Efficient multitask generalization	Complex architecture; heavy training requirements
[3]	Survey-based Synthesis	MS COCO, Flickr8k/30k	BLEU, CIDEr, METEOR, ROUGE-L	Summarizes findings of 548 studies	Wide scope of methods and metrics	Lacks empirical implementation
[4]	Recurrent Fusion Transformer	MS COCO	CIDEr: ~117, BLEU-4: ~34	Outperforms standard encoder-decoder models	Strong fusion mechanism improves semantics	Transformer design increases model size
[5]	SMOT Transformer	MS COCO	CIDEr: ~116, METEOR: ~28	High performance with less labeled data	Effective under limited supervision	Depends on well-tuned self-supervised objectives
[6]	Survey on Trends	MS COCO, Flickr8k	General trends (BLEU, CIDEr)	Identifies key datasets, metrics, and challenges	Highlights future directions	Does not provide new benchmark results
[7]	ETransCap Lightweight Transformer	MS COCO	CIDEr: ~112, BLEU-4: ~33	Efficient captioning with low computational cost	Linear complexity; real-time use	Slight dip in expressiveness
[8]	V16HP1365 + Dual Self-Attention + GRU	MS COCO	BLEU: ~34, METEOR: ~28	Enhanced context via dual self-attention	Good visual-semantic grounding	Limited transferability to other datasets
[9]	Neuraltalk+ with Context Fusion	Flickr8k, Flickr30k	BLEU-4: ~31	Fast and adaptive for assistive applications	Lightweight; visually guided captioning	Moderate performance on complex scenes
[10]	SCAP: Lightweight Feature Sifting	MS COCO, Flickr30k	CIDEr: ~108	Efficient and scalable	Suits low-resource settings	May miss deeper semantic nuances
[11]	Next-LSTM (ResNeXt + LSTM)	Flickr8k	BLEU: ~34	LSTM enhanced by strong visual features	Improved generalization	Performance bound by dataset size
[12]	FeiM with Grid Features	MS COCO	CIDEr: ~115	Learnable queries and feature interaction boosts results	Fine-grained feature capture	Resource-intensive grid modeling
[13]	Dilated ResNet + Attention	Flickr8k, Flickr30k	BLEU: ~33	Detailed and contextual captioning	Improves perceptual range	Complex model integration
[14]	BMFNet Fusion Network	MS COCO	CIDEr: ~120	2.8% CIDEr boost over baselines	Strong bidirectional fusion	Decoder path may induce latency
[15]	Weakly Supervised Matching	MS COCO, Flickr30k	CIDEr: ~110	Effective grounding with less annotation	Reduces annotation cost	Sensitive to noise in weak labels
[16]	DVAT Transformer	MS COCO	CIDEr: ~118, BLEU: ~35	Dual align attention boosts performance	Faster and accurate	Heavily dependent on region extraction quality
[17]	BIANet: Bidirectional Interactive Alignment	MS COCO	CIDEr: ~117	Cross-modal fusion yields strong semantic alignment	Balances region-grid semantics	Training complexity elevated
[18]	GAN-based Emotion Captioning	MS COCO, Senticap	Emotion Precision: ~0.82	Captures emotional subtleties	Useful for sentiment-rich tasks	Evaluation less standardized
[19]	Self-Enhanced Attention (SEA)	MS COCO	CIDEr: ~116	Improves focus on salient regions	Simple yet effective attention reweighting	Incremental benefit over standard self-attention
[20]	TAVOHDL-ICS	Flickr400	METEOR: ~28, ROUGE-L: ~52	Optimized hybrid model via meta-heuristic	Hyperparameter tuning yields better scores	Architecture interpretability is low.

(1). Performance vs. complexity:

- Transformer-based models (e.g., DVAT, BMFNet) generally achieve higher CIDEr/BLEU scores but require more computational resources.
- Lightweight models (e.g., ETransCap, SCAP) trade a small drop in expressiveness for efficiency and real-time applicability.

(2). Data requirements:

- Models like SMOT transformer and weakly supervised matching show strong performance under

limited or noisy supervision.

- Traditional CNN–RNN hybrids (e.g., Next-LSTM) perform adequately but depend heavily on dataset size and diversity.

(3). Specialized capabilities:

- GAN-based models excel in capturing sentiment or emotion but lack standardized evaluation benchmarks.

Attention enhancements (SEA, Dual Self-Attention) improve focus but yield modest metric gains compared to large architecture changes.

4. LEVEL FEATURE EXTRACTIONS OF GLOBAL NATURE

The global feature extractor in image captioning systems is designed to capture high-level, spatially aggregated semantic information from the entire image. This step ensures that the encoder has a holistic understanding of scene content before sequence modeling begins. Popular architectures for this purpose include EfficientNet (B0–B7), MobileNet, MobileNetV2, and ConvNeXt.

EfficientNet employs a compound scaling strategy that balances network width, depth, and resolution, achieving state-of-the-art accuracy while remaining parameter-efficient across all its B0–B7 variants. MobileNet and its successor MobileNetV2 leverage depthwise separable convolutions and inverted residuals to drastically reduce computation with minimal loss in representational power, making them ideal for resource-constrained deployments such as mobile or embedded systems. ConvNeXt adapts design principles from Vision Transformers such as large kernel sizes and simplified activation usage into a ResNet-like convolutional framework. This hybrid approach boosts performance while preserving the convolutional backbone's compatibility with existing encoder modules. In image captioning pipelines, such extractors transform raw pixels into rich semantic embeddings, which are then processed by sequence models like RNNs or Transformers for caption generation.

5. OBJECT LEVEL FEATURE EXTRACTIONS

While global feature extractors provide a holistic representation of the image, object-level feature detectors specialize in identifying and encoding localized regions of interest a crucial step for generating semantically rich and contextually accurate captions. Advances in region-based CNNs have significantly improved both precision and speed in object detection [22].

The evolution began with R-CNN, which first generates selective region proposals and then applies CNN-based feature extraction to each region. Fast R-CNN streamlines this process by computing region features in a single forward pass, drastically reducing inference time. Faster R-CNN further advances the pipeline through the introduction of Region Proposal Networks (RPNs), enabling end-to-end training and near real-time performance. In contrast, You Only Look Once (YOLO) reframes object detection as a single regression task, achieving real-time speed with only marginal accuracy trade-offs. These object detectors are now commonly integrated into image captioning models to produce region-level embeddings, which are either fed into attention mechanisms or directly into language decoders. This integration allows for explicit alignment between visual entities and linguistic tokens, enabling captions with greater granularity and contextual richness.

Beyond detection, this survey highlights several emerging trends:

- Performance vs. Efficiency Trade-off: Transformer-based models such as ETransCap [7] and DVAT [16] deliver high accuracy using mechanisms like linear attention and dual align-cross attention, yet their complexity can limit deployment in real-time or embedded systems.
- Semantic Enrichment via Fusion and Attention: Architectures like RCT [4], BMFNet [14], and BIANet [17] demonstrate the power of multimodal fusion, combining region- and grid-level features for deeper context modeling, often correlated with higher CIDEr and BLEU scores.
- Data-Efficient Learning: Models such as SMOT [5] and weakly supervised approaches [15] show that competitive captions can be generated with minimal annotations, although results remain sensitive to pretext-task quality and noise in supervision.
- Emotion and Subjectivity in Captioning: GAN-based captioning with sentiment control [18] represents an emerging direction toward emotionally aware captioning, but standard evaluation frameworks are still lacking for widespread adoption.
- Survey-Driven Foundations: Meta-analytical works [3, 6] offer critical insights into model categorization and evaluation norms, though they typically avoid direct empirical testing.
- Additionally, Rashied and Jeribi [21] proposed a multiscale fractal dimension approach that improves image clarity and supports robust feature representation in vision-based modeling tasks.
- In summary, comparative analysis across models reveals no single architecture that simultaneously optimizes accuracy, interpretability, and efficiency. The inherent trade-offs documented here highlight the need for hybrid, adaptable architectures that can be tailored to the specific requirements of diverse deployment scenarios.

6. CONCLUSION AND FUTURE SCOPE

This review provides a data-driven, metric-focused synthesis of 20 recent image captioning models, offering a consolidated perspective on their strengths, limitations, and trade-offs for both researchers and practitioners. Our empirical analysis leveraging prominent benchmarks such as BLEU, CIDEr, METEOR, and ROUGE-L demonstrates that Transformer-based architectures, particularly those incorporating dual-path attention mechanisms, consistently outperform traditional LSTM-based frameworks by an average of 7–10% on CIDEr across datasets like MS COCO and Flickr. The evaluation tables and visualizations included in this study not only highlight relative performance trends but also reveal critical insights into computational cost, latency, and architecture complexity, enabling informed selection for real-world applications. Unlike prior reviews that often relied on qualitative summaries, this work delivers reproducible, PRISMA-aligned comparisons, bridging the gap between model architecture innovations and their measurable impact.

The findings underscore that while lightweight models such as MobileNet-based encoders offer advantages for resource-constrained environments, hybrid Transformer variants achieve superior semantic richness and contextual grounding. This review thus establishes an evidence-based benchmarking framework that can guide both academic research and industry deployment strategies, while also identifying key gaps such as multilingual capability, domain generalization, and interpretability, thereby setting a foundation for the next generation of image captioning systems.

Future work should focus on improving cross-domain generalization by extending evaluations beyond MS COCO and Flickr to domains like medical, satellite, and autonomous driving imagery. Expanding language diversity with support for low-resource and multilingual captioning can greatly enhance accessibility. Explainability must be strengthened through interpretable reasoning modules and saliency maps. Establishing unified evaluation benchmarks that combine semantic richness, emotional tone, and human-in-the-loop assessments will ensure fairer comparisons. Further exploration of hybrid, modular architectures and real-time, resource-efficient inference will be key for deploying captioning systems in edge and time-sensitive environments.

ACKNOWLEDGMENT

The authors would like to express their sincere gratitude to the Department of CSE for providing the necessary infrastructure and resources for this research. This research was not supported by any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

REFERENCES

[1] Rahman, R.U., Kumar, P., Mohan, A., Aziz, R.M., Tomar, D.S. (2025). A novel technique for image captioning based on hierarchical clustering and deep learning. SN Computer Science, 6(4): 360. https://doi.org/10.1007/s42979-025-03908-3

[2] Bayisa, L.Y., Wang, W., Wang, Q., Ukwuoma, C.C., Gutema, H.K., Endris, A., Abu, T. (2024). Unified deep learning model for multitask representation and transfer learning: Image classification, object detection, and image captioning. International Journal of Machine Learning and Cybernetics, 15(10): 4617-4637. https://doi.org/10.1007/s13042-024-02177-5

[3] Al-Shamayleh, A.S., Adwan, O., Alsharaiah, M.A., Hussein, A.H., Kharma, Q.M., Eke, C.I. (2024). A comprehensive literature review on image captioning methods and metrics based on deep learning technique. Multimedia Tools and Applications, 83(12): 34219-34268. https://doi.org/10.1007/s11042-024-18307-8

[4] Mou, Z., Yuan, Q., Song, T. (2025). Recurrent fusion transformer for image captioning. Signal, Image and Video Processing, 19(1): 33. https://doi.org/10.1007/s11760-024-03675-3

[5] Wang, Y., Li, D., Liu, Q., Liu, L., Wang, G. (2024). Self-supervised modal optimization transformer for image captioning. Neural Computing and Applications, 36(31): 19863-19878. https://doi.org/10.1007/s00521-024-10211-4

[6] Salgotra, G., Abrol, P., Selwal, A. (2025). A survey on automatic image captioning approaches: Contemporary trends and future perspectives. Archives of Computational Methods in Engineering, 32(3): 1459-1497. https://doi.org/10.1007/s11831-024-10190-8

[7] Mundu, A., Singh, S.K., Dubey, S.R. (2024). ETransCap: Efficient transformer for image captioning. Applied Intelligence, 54(21): 10748-10762. https://doi.org/10.1007/s10489-024-05739-w

[8] Jaiswal, T., Pandey, M., Tripathi, P. (2024). Advancing image captioning with V16HP1365 encoder and dual self-attention network. Multimedia Tools and Applications, 83(34): 80701-80725. https://doi.org/10.1007/s11042-024-18467-7

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

[10] Zhang, Y., Tong, J., Liu, H. (2025). SCAP: Enhancing image captioning through lightweight feature sifting and hierarchical decoding. The Visual Computer, 41: 1-18. https://doi.org/10.1007/s00371-025-03824-w

[11] Singh, P., Kumar, C., Kumar, A. (2023). Next-LSTM: A novel LSTM-based image captioning technique. International Journal of System Assurance Engineering and Management, 14(4): 1492-1503. https://doi.org/10.1007/s13198-023-01956-7

[12] Yan, J., Xie, Y., Guo, Y., Wei, Y., Luan, X. (2024). Exploring better image captioning with grid features. Complex & Intelligent Systems, 10(3): 3541-3556. https://doi.org/10.1007/s40747-023-01341-8

[13] Li, H., Yuan, R., Li, Q., Hu, C. (2025). Research on image captioning using dilated convolution ResNet and attention mechanism. Multimedia Systems, 31(1): 47. https://doi.org/10.1007/s00530-024-01653-w

[14] Xue, L., Jin, Z., Wang, R., Yang, J. (2025). BMFNet: Bidirectional Multimodal Fusion Network for image captioning. Multimedia Systems, 31(3): 1-13. https://doi.org/10.1007/s00530-025-01801-w

[15] Du, S., Zhu, H., Lin, G., Liu, Y., Wang, D., Shi, J., Wu, Z. (2024). Weakly supervised grounded image captioning with semantic matching. Applied Intelligence, 54(5): 4300-4318. https://doi.org/10.1007/s10489-024-05389-y

[16] Ren, Y., Zhang, J., Xu, W., Lin, Y., Fu, B., Thanh, D.N. (2025). Dual visual align-cross attention-based image captioning transformer. Multimedia Tools and Applications, 84(12): 10645-10664. https://doi.org/10.1007/s11042-024-19315-4

[17] Cao, X., Yan, P., Hu, R., Li, Z. (2024). Bidirectional interactive alignment network for image captioning. Multimedia Systems, 30(6): 340. https://doi.org/10.1007/s00530-024-01559-7

[18] Yang, C., Wang, Y., Han, L., Jia, X., Sun, H. (2024). Fine-grained image emotion captioning based on Generative Adversarial Networks. Multimedia Tools and Applications, 83(34): 81857-81875. https://doi.org/10.1007/s11042-024-18680-4

[19] Sun, Q., Zhang, J., Fang, Z., Gao, Y. (2024). Self-enhanced attention for image captioning. Neural Processing Letters, 56(2): 131. https://doi.org/10.1007/s11063-024-11527-x

[20] Chitteti, C., Madhavi, K.R. (2024). Taylor African vulture optimization algorithm with hybrid deep convolution neural network for image captioning system.

Multimedia Tools and Applications, 83(25): 66393-66411. https://doi.org/10.1007/s11042-023-18080-0

[21] Rashied, N., Jeribi, A. (2024). Enhancing image quality through a novel multiscale fractal dimension formulated by the characteristic function. Mathematical Modelling of Engineering Problems, 11(1): 107-113. https://doi.org/10.18280/mmep.110111

[22] Widodo, C.E., Adi, K., Priyono, P., Setiawan, A. (2023). An evaluation of pre-trained convolutional neural network models for the detection of COVID-19 and pneumonia from chest X-ray imagery. Mathematical Modelling of Engineering Problems, 10(6): 2210-2216. https://doi.org/10.18280/mmep.100635