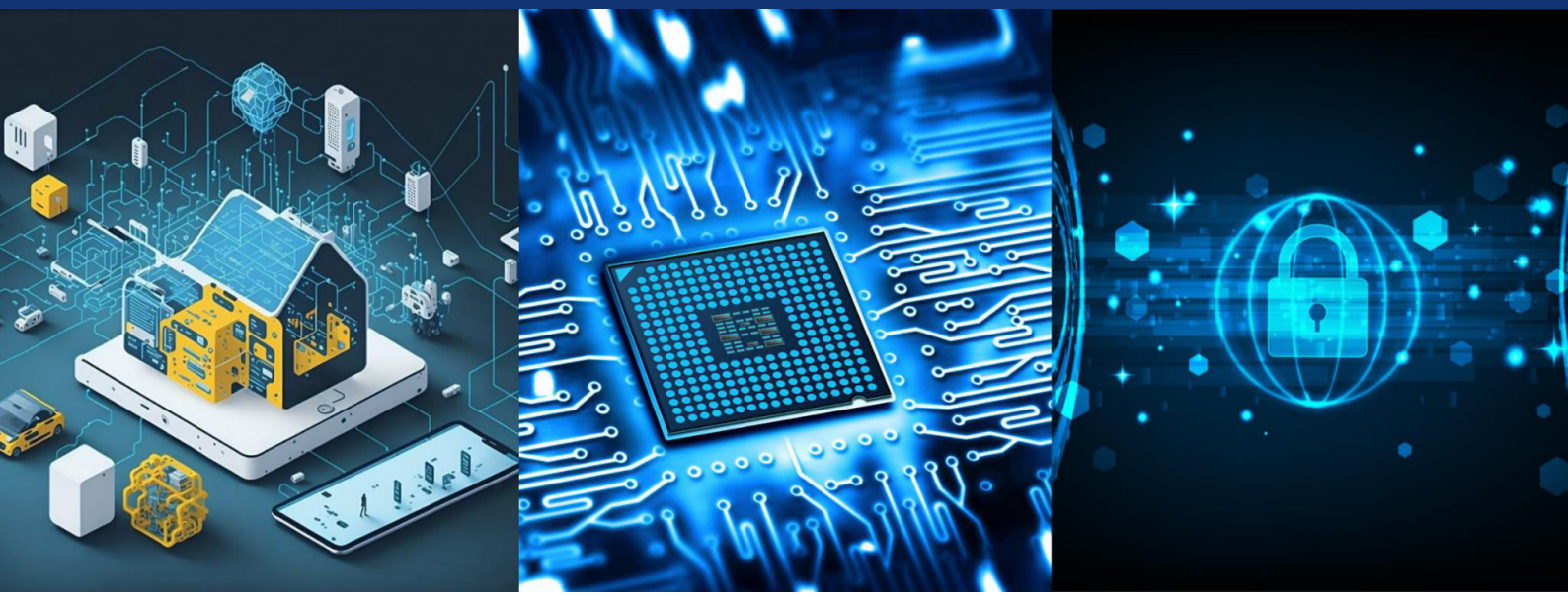
IJIRCCE

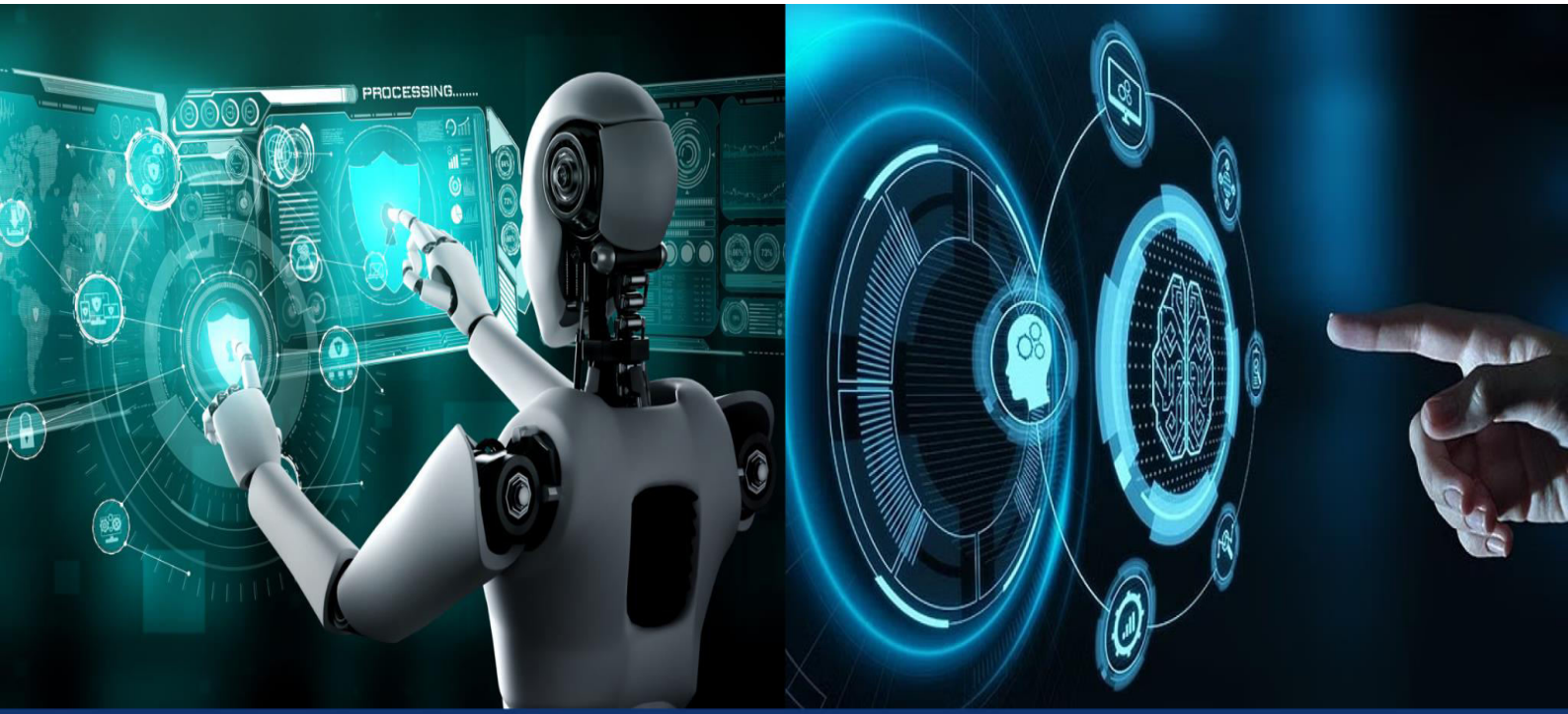
ISSN(O): 2320-9801

ISSN(P): 2320-9798



International Journal of Innovative Research in Computer and Communication Engineering

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 9.927

Volume 14, Issue 8, August 2026

© 2026 IJIRCCE | Volume 14, Issue 8, August 2026 DOI: 10.15680/IJIRCCE.2026.1408022

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

A Comprehensive Survey on Image Caption Generation Using Deep Learning: Evolution, Comparative Analysis, and Emerging Vision-Language Models

Umesh N Wagh, Prof. Dr. B.D.Phulpagar

Department of Computer Engineering, P. E. S. Modern College of Engineering, Pune, India

Department of Computer Engineering, P. E. S. Modern College of Engineering, Pune, India

ABSTRACT: Image caption generation is a multimodal artificial intelligence task that automatically converts visual content into natural-language descriptions. The problem combines computer vision for visual understanding with natural language processing for sentence generation, making it substantially more complex than conventional image classification. Deep learning has transformed this field from early convolutional neural network (CNN) and recurrent neural network (RNN) encoder-decoder models toward attention-based systems, object-aware models, graph reasoning, reinforcement learning, Transformer architectures, vision-language pretraining, and multimodal large language models (MLLMs). This survey reviews the evolution of these approaches and organizes representative studies according to their central modelling strategy. Particular attention is given to visual attention, object-level representations, scene and relationship modelling, Transformer-based cross-modal interaction, and the recent transition toward pretrained vision-language systems. The survey also compares benchmark datasets such as Flickr8k, Flickr30k, MS COCO, Conceptual Captions, and nocaps, and discusses evaluation metrics including BLEU, METEOR, ROUGE-L, CIDEr, SPICE, CLIPScore, BERTScore, and CHAIR. In addition, major limitations are examined, including object hallucination, weak visual grounding, dataset bias, long-tail recognition, computational cost, and the mismatch between lexical similarity metrics and factual correctness. The review concludes with future directions in grounded captioning, controllable generation, multilingual captioning, parameter-efficient adaptation, knowledge-grounded generation, efficient deployment, and trustworthy multimodal AI.

KEYWORDS: Image Caption Generation, Deep Learning, CNN, RNN, LSTM, Attention Mechanism, Transformer, Vision-Language Models, Multimodal Learning, Image-to-Text Generation.

I. INTRODUCTION AND SURVEY METHODOLOGY

Image caption generation aims to produce a natural-language description for a given image. A successful system must identify relevant visual entities, infer their attributes and relationships, decide which information should be described, and express that information in fluent language. This makes image captioning an inherently multimodal problem that sits between computer vision and natural language processing. Recent surveys emphasize that the field has progressed from CNN-RNN encoder-decoder pipelines to attention-based, graph-based, Transformer-based, vision-language pretraining, and multimodal large language model approaches [1]-[4].

The 2026 systematic review by Al-Malla et al. examined 174 peer-reviewed studies published from 2018 to 2025 and organized them into nine methodological categories: attention-based methods, graph-based methods, attention-graph integration, CNN-based models, Transformer-based models, Transformer-graph integration, vision-language pretraining, unsupervised/reinforcement learning, and specialized techniques such as controllable or multi-style captioning [1]. The present survey follows this broad evolution while emphasizing representative studies from IEEE, Springer Nature, and Elsevier/ScienceDirect, together with foundational open proceedings [5]-[15].

The review methodology is structured around five questions: (1) how image captioning architectures have evolved; (2) which visual representations and language decoders are most common; (3) which datasets and metrics dominate evaluation; (4) what limitations still constrain real-world captioning; and (5) which research directions are most

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

promising for future systems. The literature was grouped by technical contribution rather than by publication venue, and representative results are discussed only when directly reported in the cited sources.

A. Scope and Taxonomy

The survey focuses on whole-image captioning, where a system generates one or more sentences describing the overall image. Dense captioning and video captioning are treated as adjacent research areas rather than the primary scope. The resulting taxonomy is: CNN-RNN encoder-decoder models; attention-based captioning; object and region-aware captioning; graph-based relational reasoning; reinforcement-learning approaches; Transformer-based generation; vision-language pretraining; and multimodal large language model-based captioning.

II. EVOLUTION OF IMAGE CAPTION GENERATION

A. CNN-RNN Encoder-Decoder Models

The encoder-decoder paradigm provided the foundation for modern neural image captioning. In a typical architecture, a CNN converts an input image into a visual representation and an RNN or LSTM generates a caption one token at a time. This design removed the dependence on hand-crafted templates and enabled end-to-end learning. However, a single global visual vector can compress multiple objects, spatial relationships, and attributes into a representation that is difficult for the decoder to query selectively [1], [3].

LSTM decoders address long-term sequential dependencies better than basic recurrent units, while convolutional encoders provide hierarchical visual features. Nevertheless, the encoder-decoder pipeline remains strongly affected by exposure bias, vocabulary limitations, and the inability of a fixed global image representation to support fine-grained word-region alignment. These limitations motivated the next major step: visual attention.

B. Visual Attention

Show, Attend and Tell introduced an attention-based model that learns to focus on salient image regions while generating each word [5]. Instead of treating the image as one fixed vector, the decoder uses a weighted combination of spatial features. The study validated the approach on Flickr8k, Flickr30k, and MS COCO and showed that learned attention can align generated words with meaningful visual regions. Attention therefore became a central mechanism for improving both descriptive detail and interpretability.

C. Object-Level and Bottom-Up Attention

Bottom-Up and Top-Down Attention moved attention from uniform CNN grids to object and salient-region proposals generated by Faster R-CNN [6]. This object-centric design allowed the captioner to attend to semantically meaningful regions. On the MS COCO test server, the reported system achieved CIDEr 117.9, SPICE 21.5, and BLEU-4 36.9. The result helped establish region-based attention as an important bridge between object detection and language generation.

D. Attention on Attention and Enhanced Alignment

Attention on Attention (AoA) was designed to improve the relevance of attended visual information to the current query [7]. Instead of simply applying a weighted average, AoA learns whether the attended information is useful for the current decoding context. The published study reported 129.8 CIDEr-D on the MS COCO Karpathy offline test split, demonstrating the value of more selective attention interactions.

III. RELATIONAL AND GRAPH-BASED CAPTIONING

Object identification alone is insufficient for many captions because the relationships between objects often define the meaning of a scene. Graph-based methods therefore represent detected entities as nodes and interactions as edges. Learning visual relationship and context-aware attention for image captioning explicitly models relationships among image regions with a graph neural network and uses context-aware attention to select salient information for sentence generation [16]. This line of work addresses a limitation of region-only attention: two models may detect the same objects but produce different captions depending on whether they understand who is doing what to whom.

The main advantage of relational modelling is semantic structure. A scene graph can encode relations such as person-riding-bicycle, child-holding-ball, or dog-next-to-car, which can improve compositional caption generation. The

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

limitation is that graph construction is itself an error-prone process. Incorrect detections or relationships can propagate into the language model, and the computational burden grows as the number of regions and edges increases [1], [16].

IV. REINFORCEMENT LEARNING FOR SEQUENCE-LEVEL OPTIMIZATION

Maximum-likelihood training evaluates a predicted token against the next ground-truth word, whereas captioning quality is ultimately judged at the sentence level. Self-Critical Sequence Training (SCST) addressed this mismatch by optimizing image captioning with reinforcement learning and using the model's own test-time output as a baseline [8]. The method directly optimizes non-differentiable captioning metrics such as CIDEr. The authors reported an improvement from a CIDEr score of 104.9 to 114.7 on the MS COCO evaluation setting they studied [8].

Sequence-level optimization can improve benchmark scores, but it also creates a risk of over-optimizing a metric. A caption can score well under a lexical similarity metric while containing an incorrect object or action. For this reason, reinforcement learning should be viewed as one component of training rather than a substitute for visual grounding and factuality evaluation [1], [2].

V. TRANSFORMER-BASED IMAGE CAPTION GENERATION

Transformers replaced recurrent sequence processing with self-attention and became a dominant architecture for modern captioning. Self-attention enables direct interaction among visual and textual tokens and is naturally suited to multimodal fusion. Recent reviews identify Transformer-based approaches as one of the largest research categories in image captioning [1], [2].

IEEE research illustrates several directions within this transition. Vision-Enhanced and Consensus-Aware Transformer improves the visual and textual representation flow and was published in IEEE Transactions on Circuits and Systems for Video Technology [11]. Cross on Cross Attention: Deep Fusion Transformer explicitly focuses on deep cross-modal interaction between visual and textual features [12]. A Dual-Feature-Based Adaptive Shared Transformer Network aims to combine complementary visual representations while reducing the computational cost associated with separate feature-processing streams [10]. Together, these studies show that Transformer research is moving beyond a straightforward encoder-decoder design toward efficient feature fusion and more deliberate visual-semantic interaction.

A. Transformer with Refined Visual Features

Elsevier research further demonstrates the refinement of visual representations for Transformer captioners. The 2024 study Exploring Refined Dual Visual Features Cross-Combination for Image Captioning introduces a Distilled Cross-Combination Transformer and a fusion strategy for region and grid features [20]. The work highlights a current trend: improving caption quality not only through a stronger language decoder but through better visual features before or during multimodal fusion.

B. Dual-Adaptive Visual-Textual Interaction

The Dual-Adaptive Interactive Transformer (DAIT) integrates visual and textual context during both encoding and decoding [19]. Such models treat captioning as a bidirectional alignment problem: language guides visual selection, while visual context constrains language generation. This interactive design is an important step toward more faithful multimodal generation.

VI. VISION-LANGUAGE PRETRAINING

Vision-language pretraining changed captioning from a task-specific learning problem into a broader multimodal representation-learning problem. OSCAR, for example, used detected object tags as anchor points to facilitate image-text semantic alignment and pretrained on 6.5 million text-image pairs [13]. The central idea is to learn a reusable multimodal representation before fine-tuning downstream generation tasks.

Large-scale pretraining provides stronger transfer, but it also introduces data quality, bias, copyright, compute, and reproducibility challenges. The modern survey literature therefore increasingly evaluates captioning systems not only by benchmark scores, but also by generalization, grounding, and deployment cost [1], [2].

© 2026 IJIRCCE | Volume 14, Issue 8, August 2026 DOI: 10.15680/IJIRCCE.2026.1408022

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

VII. MULTIMODAL LARGE LANGUAGE MODELS

The latest stage of image captioning is the integration of powerful language models with pretrained visual encoders. BLIP-2 is a representative approach that connects a frozen image encoder and a frozen large language model using a lightweight Querying Transformer [14]. This architecture demonstrates how a multimodal model can support image-to-text generation without training every component from scratch.

MLLMs broaden the definition of captioning. The same model can potentially produce a short caption, a detailed description, a domain-specific description, or an instruction-following response. However, fluency does not guarantee correctness. Recent surveys identify hallucination, faithfulness, grounding, bias, and computational cost as key challenges for MLLM-based captioning [1], [3].

VIII. DATASET USAGE AND BENCHMARKING

Image captioning performance depends strongly on the characteristics of the training and evaluation data. Flickr8k and Flickr30k are widely used for smaller-scale experimentation; MS COCO remains the dominant benchmark for whole-image captioning; Conceptual Captions supports web-scale vision-language pretraining; and nocaps evaluates generalization to novel objects [1], [15].

Dataset	Characteristics	Primary Use	Strength	Limitation
Flickr8k	8K images; five captions per image	Baseline/educational experiments	Simple and accessible	Limited scale and visual diversity
Flickr30k	30K images; multiple captions	Captioning and retrieval	Larger and more diverse than Flickr8k	Still small for modern large-scale pretraining
MS COCO	Large curated benchmark with multiple captions	Mainstream captioning evaluation	Rich benchmark protocol and broad adoption	Benchmark bias toward common visual concepts
Conceptual Captions	Millions of web-derived image-text pairs	Pretraining	Large scale and broad coverage	Automatically derived text can be noisy
nocaps	15,100 images; 166,100 human captions	Novel-object generalization	Tests unseen/rare visual concepts	More specialized and difficult

Table I. Representative datasets used in image caption generation research.

The nocaps benchmark is especially important for measuring generalization because it was designed around novel objects and contains 15,100 images and 166,100 human-generated captions [15]. Conceptual Captions, by contrast, was developed as a large-scale image-text resource for modern vision-language learning and contains approximately 3.3 million image-caption pairs [17].

IX. PREPROCESSING AND VISUAL REPRESENTATION

Preprocessing typically includes image resizing, normalization, augmentation, and tokenization of reference captions. The more significant architectural choice is the visual representation passed to the captioning model. Earlier systems used a single global CNN vector; later systems adopted spatial feature maps, detected regions, object tags, scene graphs, and Transformer-style visual tokens [1]-[4].

Representation	Typical Example	Main Strength	Main Limitation
Global CNN features	ResNet/EfficientNet vector	Simple, efficient global representation	Can lose spatial detail
Spatial feature maps	CNN grid features	Preserves location information	Less explicit object semantics

© 2026 IJIRCCE | Volume 14, Issue 8, August 2026 DOI: 10.15680/IJIRCCE.2026.1408022

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Region/object features	Faster R-CNN regions	Fine-grained object awareness	Adds detector cost
Graph representation	Object/relationship graph	Models relations and compositional context	Depends on graph quality
Patch/token representation	Vision Transformer features	Strong global modelling	Requires substantial computation
Vision-language embeddings	CLIP/BLIP-style features	Aligns visual and textual semantics	Pretraining cost and possible bias

Table II. Evolution of visual representations for image captioning.

X. COMPARATIVE ANALYSIS OF REPRESENTATIVE METHODS

Method / Study	Year	Core Idea	Evaluation Context	Strength	Limitation
Show, Attend and Tell [5]	2015	Attention + RNN	Flickr8k/Flickr30k/MS COCO	Dynamic visual focus	Limited relational reasoning
SCST [8]	2017	RL + sequence reward	MS COCO	Optimizes CIDEr directly	Metric dependence
Bottom-Up / Top-Down [6]	2018	Faster R-CNN + attention	MS COCO	Object-level grounding	Detector overhead
AoANet [7]	2019	Attention-on-attention	MS COCO	Improved relevance filtering	Still reference-metric driven
Visual relationship + context attention [16]	2020	GNN + attention	Benchmarks reported in study	Relational reasoning	Graph construction complexity
Visual enhanced gLSTM [17]	2021	gLSTM + visual enhancement	Captioning benchmarks	Better visual guidance	Recurrent decoder limitations
Vision-enhanced consensus Transformer [11]	2022	Transformer	MS COCO-oriented evaluation	Visual/text consensus	Higher complexity
Dual-feature adaptive shared Transformer [10]	2024	Adaptive shared Transformer	MS COCO + small datasets	Efficiency-aware fusion	Still needs strong visual features
Self-Enhanced Attention [9]	2024	Transformer + enhanced attention	COCO	Improved feature highlighting	Attention tuning complexity
DAIT [19]	2024	Interactive multimodal Transformer	Captioning benchmarks	Joint visual-textual context	Higher multimodal cost
Refined dual visual features [20]	2024	DCCT + feature fusion	Captioning benchmarks	Refined visual representation	More elaborate feature pipeline
GDVT [18]	2025	Dual Transformer + geometry	Image captioning benchmarks	Geometric visual-text interaction	Model complexity
MLLM direction / BLIP-2 [14]	2023	Vision encoder + LLM	Zero-shot and generative settings	Flexible language generation	Hallucination/compute risk

Table III. Representative methodological progression in image caption generation.

© 2026 IJIRCCE | Volume 14, Issue 8, August 2026 DOI: 10.15680/IJIRCCE.2026.1408022

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

XI. EVALUATION METRICS

Image captioning is difficult to evaluate because several different sentences may all be valid descriptions of the same image. Traditional metrics therefore compare generated captions with one or more reference captions. Recent surveys additionally discuss semantic and multimodal metrics that attempt to measure visual faithfulness and language-image alignment [1], [2].

Metric	Main Principle	Strength	Limitation
BLEU	n-gram precision	Simple and widely adopted	Insensitive to deeper semantic correctness
METEOR	Word/stem alignment	Better than raw n-gram overlap for some linguistic variants	Still reference-dependent
ROUGE-L	Longest common subsequence	Captures sequence-level overlap	Does not directly test visual grounding
CIDEr	TF-IDF weighted n-gram similarity	Designed for consensus with human references	Can reward reference-like phrasing
SPICE	Scene-graph semantic tuples	More semantic than n-gram metrics	Depends on parsing quality
CLIPScore	Image-text embedding similarity	Directly couples image and text	May not fully capture fine-grained factuality
BERTScore	Contextual token similarity	Semantic language comparison	Still primarily text-based
CHAIR	Object hallucination rate	Targets visual-semantic hallucination	Restricted to object-centric errors
Human evaluation	Human preference/factuality	Closest to practical usefulness	Expensive and subjective

Table IV. Common evaluation measures and their practical roles.

XII. APPLICATIONS

- Assistive technology: generating descriptions that can help visually impaired users interpret visual content.
- Healthcare and medical imaging: generating textual descriptions or report-like summaries from visual findings, although medical use requires domain-specific validation.
- Image retrieval and indexing: using generated captions to support semantic search and cross-modal retrieval.
- Autonomous systems and robotics: producing human-readable scene descriptions for interaction and monitoring.
- Remote sensing: describing land-cover, infrastructure, and environmental changes in aerial or satellite imagery.
- Education and digital accessibility: automatically generating alt-text-like descriptions for learning material and online content.

XIII. RESEARCH GAPS AND CHALLENGES

A. Hallucination and Faithfulness

A fluent caption can contain objects, actions, or attributes that are not present in the image. This visual-semantic hallucination is repeatedly identified as a major unresolved issue in recent survey literature [1]-[3]. The problem becomes more important as captioning models grow into general-purpose multimodal language systems.

B. Weak Visual Grounding

Attention and similarity scores do not necessarily prove that each generated phrase is grounded in the correct image region. Reliable captioning therefore requires stronger cross-modal alignment and explicit grounding tests.

C. Dataset Bias and Long-Tail Recognition

Common objects and common language patterns dominate large benchmarks. Rare or novel objects therefore remain difficult to describe, motivating datasets such as nocaps [15]. Bias in web-scale pretraining data can also influence generated content and caption style [1], [3].

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

D. Metric Limitations

BLEU, METEOR, ROUGE-L, and CIDEr primarily assess similarity to reference text. A factually correct caption may use wording that differs from references and receive a lower score, while a fluent but incorrect caption can still obtain a non-trivial text overlap score. Modern surveys therefore emphasize semantic and multimodal evaluation, including SPICE, CLIPScore, BERTScore, and hallucination-specific measures such as CHAIR [1], [2].

E. Computational Cost

Region detectors, multi-stage Transformers, large vision encoders, and language models can substantially increase memory, inference time, and training cost. Recent efficient Transformer work explicitly addresses the need to reduce FLOPs and model size while retaining caption quality [10].

F. Limited Domain Generalization

Models trained on everyday web images may not transfer directly to medical, industrial, remote-sensing, or highly specialized environments. Domain-specific validation and adaptation remain necessary for trustworthy deployment.

XIV. FUTURE RESEARCH DIRECTIONS

- Grounded caption generation: integrate region grounding, object relations, and explicit verification before a sentence is finalized.
- Controllable captioning: allow users to specify language, style, length, level of detail, target audience, or selected visual attributes.
- Multilingual and low-resource captioning: develop systems that do not rely primarily on English-language datasets.
- Parameter-efficient adaptation: use adapters and lightweight fine-tuning to specialize vision-language models without retraining the full system.
- Knowledge-grounded generation: combine visual evidence with reliable external knowledge for rare concepts and specialized domains.
- Efficient deployment: design compact models suitable for mobile, edge, robotic, and assistive applications.
- Trustworthy evaluation: combine lexical, semantic, image-text, hallucination, and human-factuality measures instead of relying on a single score.
- Multimodal LLM integration: develop models that remain visually faithful while supporting instruction-following and reasoning.

XV. CONCLUSION

Image caption generation has evolved from template-based and retrieval-based systems into a mature multimodal learning problem. CNN-RNN encoder-decoder models established the foundation for end-to-end image-to-text generation. Attention mechanisms then enabled dynamic selection of relevant visual information, while object-level attention and graph-based reasoning improved fine-grained semantic understanding. Reinforcement learning addressed sequence-level optimization, and Transformer architectures introduced more powerful visual-textual interaction. The emergence of vision-language pretraining and multimodal large language models has further expanded the capability of captioning systems from fixed sentence generation to flexible, instruction-driven multimodal generation [1]-[4], [10]-[14].

At the same time, the literature shows that benchmark progress does not eliminate core reliability problems. Hallucination, weak grounding, dataset bias, long-tail recognition, computational cost, and imperfect evaluation remain significant challenges. The next generation of image captioning systems should therefore be designed around factual visual grounding, domain-aware generalization, efficient adaptation, multilingual support, and evaluation protocols that measure more than lexical similarity. A useful captioning model should not only sound natural; it should also describe what is actually present, at an appropriate level of detail, and in a way that remains useful in real-world settings.

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

REFERENCES

[1] M. A. Al-Malla, O. Hamdoun, and N. Ghneim, "A comprehensive survey on deep learning approaches for image captioning: A systematic review," Journal of Big Data, vol. 13, art. 48, 2026. Springer Nature. doi: 10.1186/s40537-026-01377-w. https://link.springer.com/article/10.1186/s40537-026-01377-w

[2] A. Sharma et al., "Pixels to prose: A comprehensive survey of image captioning techniques with deep learning and generative artificial intelligence," Neurocomputing, vol. 667, art. 132385, 2026. Elsevier. doi: 10.1016/j.neucom.2025.132385. https://doi.org/10.1016/j.neucom.2025.132385

[3] H. D. Abdulgalil and O. A. Basir, "Next-generation image captioning: A survey of methodologies and emerging challenges from transformers to Multimodal Large Language Models," Natural Language Processing Journal, vol. 12, art. 100159, 2025. Elsevier. doi: 10.1016/j.nlp.2025.100159. https://doi.org/10.1016/j.nlp.2025.100159

[4] "Deep image captioning: A review of methods, trends and future challenges," Neurocomputing, vol. 546, art. 126287, 2023. Elsevier. doi: 10.1016/j.neucom.2023.126287. https://doi.org/10.1016/j.neucom.2023.126287

[5] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in Proc. ICML, vol. 37, pp. 2048-2057, 2015. https://proceedings.mlr.press/v37/xuc15.html

[6] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," in Proc. IEEE/CVF CVPR, pp. 6077-6086, 2018. doi: 10.1109/CVPR.2018.00636. https://openaccess.thecvf.com/content_cvpr_2018/html/Anderson_Bottom-Up_and_Top-Down_CVPR_2018_paper.html

[7] L. Huang, W. Wang, J. Chen, and X.-Y. Wei, "Attention on attention for image captioning," in Proc. IEEE/CVF ICCV, pp. 4634-4643, 2019. https://openaccess.thecvf.com/content_ICCV_2019/html/Huang_Attention_on_Attention_for_Image_Captioning_ICCV_2019_paper.html

[8] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel, "Self-critical sequence training for image captioning," in Proc. IEEE/CVF CVPR, pp. 7008-7024, 2017. https://openaccess.thecvf.com/content_cvpr_2017/html/Rennie_Self-Critical_Sequence_Training_CVPR_2017_paper.html

[9] Q. Sun, J. Zhang, Z. Fang, and Y. Gao, "Self-enhanced attention for image captioning," Neural Processing Letters, vol. 56, art. 131, 2024. Springer Nature. doi: 10.1007/s11063-024-11527-x. https://link.springer.com/article/10.1007/s11063-024-11527-x

[10] Y. Shi, J. Xia, M. C. Zhou, and Z. Cao, "A dual-feature-based adaptive shared transformer network for image captioning," IEEE Transactions on Instrumentation and Measurement, vol. 73, pp. 1-13, 2024. doi: 10.1109/TIM.2024.3353830. https://ieeexplore.ieee.org/document/10400415/

[11] S. Cao, G. An, Z. Zheng, and Z. Wang, "Vision-enhanced and consensus-aware transformer for image captioning," IEEE Transactions on Circuits and Systems for Video Technology, vol. 32, no. 10, pp. 7005-7018, 2022. doi: 10.1109/TCSVT.2022.3178844. https://ieeexplore.ieee.org/document/9784827/

[12] J. Zhang, Y. Xie, W. Ding, and Z. Wang, "Cross on cross attention: Deep fusion transformer for image captioning," IEEE Transactions on Circuits and Systems for Video Technology, vol. 33, no. 8, pp. 4257-4268, 2023. doi: 10.1109/TCSVT.2023.3243725. https://ieeexplore.ieee.org/document/10041164/

[13] X. Li et al., "OSCAR: Object-semantics aligned pre-training for vision-language tasks," in Proc. ECCV, 2020. https://www.ecva.net/papers/eccv_2020/papers_ECCV/html/7133_ECCV_2020_paper.php

[14] J. Li, D. Li, S. Savarese, and S. Hoi, "BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in Proc. ICML, vol. 202, pp. 19730-19742, 2023. https://proceedings.mlr.press/v202/li23q.html

[15] H. Agrawal et al., "nocaps: Novel object captioning at scale," in Proc. IEEE/CVF ICCV, pp. 8948-8957, 2019. doi: 10.1109/ICCV.2019.00904. https://openaccess.thecvf.com/content_ICCV_2019/html/Agrawal_nocaps_novel_object_captioning_at_scale_ICCV_2019_paper.html

[16] J. Wang, W. Wang, L. Wang, Z. Wang, D. Dagan Feng, and T. Tan, "Learning visual relationship and context-aware attention for image captioning," Pattern Recognition, vol. 98, art. 107075, 2020. Elsevier. doi: 10.1016/j.patcog.2019.107075. https://doi.org/10.1016/j.patcog.2019.107075

[17] J. Zhang, K. Li, Z. Wang, X. Zhao, and Z. Wang, "Visual enhanced gLSTM for image captioning," Expert Systems with Applications, vol. 184, art. 115462, 2021. Elsevier. doi: 10.1016/j.eswa.2021.115462. https://doi.org/10.1016/j.eswa.2021.115462

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

[18] Y.-L. Chang, H.-S. Ma, S.-C. Li, B. P. Jaysawal, et al., "GDVT: geometrically aware dual transformer encoding visual and textual features for image captioning," International Journal of Data Science and Analytics, vol. 20, pp. 6507-6525, 2025. Springer Nature. doi: 10.1007/s41060-025-00836-6. https://link.springer.com/article/10.1007/s41060-025-00836-6

[19] L. Chen and K. Li, "Dual-adaptive interactive transformer with textual and visual context for image captioning," Expert Systems with Applications, vol. 243, art. 122955, 2024. Elsevier. doi: 10.1016/j.eswa.2023.122955. https://doi.org/10.1016/j.eswa.2023.122955

[20] J. Hu, Z. Li, Q. Su, Z. Tang, and H. Ma, "Exploring refined dual visual features cross-combination for image captioning," Neural Networks, vol. 180, art. 106710, 2024. Elsevier. doi: 10.1016/j.neunet.2024.106710. https://doi.org/10.1016/j.neunet.2024.106710

[21] "Divergent-convergent attention for image captioning," Pattern Recognition, vol. 115, art. 107928, 2021. Elsevier. doi: 10.1016/j.patcog.2021.107928. https://doi.org/10.1016/j.patcog.2021.107928

[22] "Image captioning: Semantic selection unit with stacked residual attention," Image and Vision Computing, vol. 144, art. 104965, 2024. Elsevier. doi: 10.1016/j.imavis.2024.104965. https://doi.org/10.1016/j.imavis.2024.104965

[23] M. A. Al-Malla, A. Jafar, and N. Ghneim, "Image captioning model using attention and object features to mimic human image understanding," Journal of Big Data, vol. 9, art. 20, 2022. Springer Nature. doi: 10.1186/s40537-022-00571-w. https://link.springer.com/article/10.1186/s40537-022-00571-w

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi crossref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details