International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 09 | Sept - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

A Review on Deep Learning based Image Caption Generation Methods

Poornima K.M.[1], Shifa Anjum[2], Shifa Naz[3], Shivaraj T[4], Soorya Udupa K B[5]

Department of CS&E,
JNN College of Engineering, Shivamogga, Karnataka, India.
[3]shifanaz9002@jnnce.ac.in

Abstract - This review investigates recent advancements in Deep Learning based image caption generation methodologies. The progression from early encoder-decoder models to attention-based systems and transformer frameworks is examined, alongside notable datasets and evaluation metrics. Strengths and shortcomings of leading methods are analyzed to provide researchers with a consolidated understanding of current trends. The proposed conceptual model integrates Convolutional Neural Networks for visual feature extraction with transformer mechanisms for sequence generation, aiming to produce accurate, fluent, and context-aware captions. Comparative discussions of state-of-the-art techniques are supported with evaluations based on benchmark datasets like MS- COCO and Flickr8k, emphasizing attention strategies including soft, hard, and adaptive variations.

Key Words: Image Captioning, CNN, Transformer, Attention, Deep Learning.

1. INTRODUCTION

The task of image caption generation lies at the intersection of computer vision and natural language processing by automatically generating coherent textual descriptions for visual content. Recent advances in deep learning, particularly the use of Convolutional Neural Networks (CNNs) for extracting image features and sequence models such as LSTMs and transformers for caption generation, have significantly enhanced this task. Transformer architectures, in particular, offer improved contextual understanding, better handling of long-range dependencies, as well as parallel processing benefits. This technology has applications ranging from improving accessibility for visually impaired individuals to automating content generation and improved multimedia retrieval. The literature reflects a variety of approaches, from models that combine convolutional and recurrent layers to attention-based systems that selectively focus on image regions or semantic attributes. This paper reviews the evolution of such models, evaluates their design choices, and outlines an integrated CNN-Transformer approach informed by these findings.

2. LITERATURE REVIEW

The Literature Survey reviews existing research in image captioning that have advanced through multiple innovation stages, beginning with basic encoder–decoder architectures and advancing toward models that employ complex attention mechanisms and transformer-based designs. This section outlines influential works, organized by their methodological contributions.

In [1] presented a comprehensive review of modern image captioning methods, analyzing developments in deep learning models that incorporate attention mechanisms. Their survey covered a wide range of architectures from CNN-RNN hybrids to transformer-based designs-alongside a comparison of key datasets and evaluation protocols. Special emphasis is given to attention mechanisms and their variations soft, hard, adaptive, semantic, and others which significantly improve caption quality The comprehensive review presented in [1] provides a deep dive into the evolution of modern image captioning. This highlights the paradigm shift from traditional computer vision and natural language processing techniques to more integrated deep learning models. A major focus of the review is on the critical role of attention mechanisms in enhancing captioning models. It discusses various attention types, including soft, hard, adaptive, and semantic attention, and explains how these mechanisms allow models to selectively focus on the most relevant parts of an image. The survey not only compares different architectures, such as the widely used CNN-RNN hybrids and the more recent Transformer-based designs, but also provides a valuable comparison of key datasets and the evaluation protocols used to assess model performance. This review serves as a crucial resource for understanding the historical progression and the current state-of-the-art in the field.

In [2] the Transformer architecture is introduced, which relies entirely on attention mechanisms, eliminating RNNs/CNNs for sequence modelling and becoming foundational for modern captioning models It offers advantages like parallelization, context awareness, and better long-range dependency modelling. However, it requires large datasets and is computationally intensive The Transformer architecture, as introduced in [2], marks a significant departure from traditional sequence modeling by relying exclusively on attention mechanisms. Unlike earlier models that depended on recurrent (RNN) or convolutional (CNN) layers, the Transformer processes input sequences in parallel, which greatly enhances its training efficiency. This parallelization capability allows it to capture long-range dependencies more effectively than RNNs, which process information sequentially. The core of the Transformer is its multi-head self-attention mechanism, which enables the model to weigh the importance of different words in a sequence when encoding a specific word. This self-attention mechanism, along with a positional encoding to maintain sequence order, makes the Transformer a foundational architecture for many modern language and vision-language models, including those used for image captioning.

In [3] introduced the Neural Image Caption (NIC) framework, which treats caption generation as a translation task from visual features to natural language. Using a Convolutional Neural Network(CNN) to encode input visual data and an LSTM to decode it into a sentence, the model demonstrated the effectiveness of end-to-end learning for integrating visual as well as linguistic understanding within a single trainable system The Neural Image Caption (NIC) framework, as introduced in [3], models image captioning as a translation problem. It uses a Convolutional Neural Network (CNN) to encode the input image into a compact representation, or visual features. This is analogous to a human "seeing" an image and processing its contents. Then, a Long Short-Term Memory (LSTM), a type of recurrent neural network (RNN), acts as a decoder, taking these visual features and translating them into a natural language sentence, word by word. This end-to-end learning approach was groundbreaking because it allowed the entire system from visual understanding to linguistic generation to be trained as a single, cohesive unit. This integrated framework was highly effective and demonstrated the power of combining deep learning models

from computer vision and natural language processing to solve complex, multimodal tasks.

In [4] presented the Long-term Recurrent Convolutional Network (LRCN) architecture, which being a versatile model that is capable of combining convolutional layers for visual feature extraction along with recurrent layers for sequential output generation. It supports variable-length inputs/outputs and learns spatial-temporal dynamics. It is dependent on the selection of the visual encoder and it lacks newer attention mechanisms It is designed to be end-to-end trainable and has been shown to provide advantages over models that are defined and optimized separately. LRCN models are also considered "doubly deep" because they are compositional in both spatial and temporal layers. The model can directly map variable-length inputs, such as video frames, to variable-length outputs, like natural language text. This flexibility allows it to be applied to a variety of sequential learning tasks, including video activity recognition, image description, and video description. The model's success is rooted in its ability to jointly learn temporal dynamics and convolutional perceptual representations.

Attention-based model is proposed in [5], that integrates visual attention mechanisms, using a Convolutional Neural Network (CNN) as well as an RNN with LSTM units. It introduced probabilistic(soft) & deterministic(hard) attention, enhancing interpretability and performance. Hard attention requires stochastic sampling and reinforcement learning, while soft attention is computationally heavier the soft attention model is differentiable and can be trained using standard backpropagation. In contrast, the hard attention model is a stochastic process that is trained using methods like reinforcement learning. An important advantage of this attention framework is its ability to be visualized, which provides insight into which parts of the image the model is "seeing" as it generates the output.

In [6] introduced a semantic attention strategy that augments visual attention with high-level semantic cues derived from detected image attributes. This leads to more accurate and semantically rich captions, as a result it achieves superior performance on benchmark datasets. Butthis may lead to increased complexity & risk of overfitting The semantic attention strategy integrates both top-down and bottom-up approaches to image captioning. The model learns to selectively focus on proposed semantic concepts and merges them with the hidden states and outputs of a recurrent neural network. This process creates a feedback loop that connects the two distinct computation paradigms. The semantic attention mechanism is able to identify semantically important concepts within an image, weigh the relative strength of attention across multiple concepts, and dynamically shift its focus based on the current task. By detecting visual attributes from a bottom-up perspective and guiding the process with a top-down visual feature, the algorithm can more accurately predict new words. This novel approach has demonstrated superior performance on public benchmarks like Microsoft COCO and Flickr30K, consistently surpassing other state-of-the-art methods across various evaluation metrics.

In [7] introduced an adaptive attention mechanism with a visual sentinel, allowing the model to decide between visual information and internal linguistic knowledge. It offers improved interpretability and State-of-the-Art performance. Drawbacks include higher computational complexity and dependency on attention accuracy This is a novel approach because most attention-based models force the network to actively attend to the image for every word it generates, even for non-visual words like "the" and "of". By incorporating the visual sentinel, the model gains a "fallback" option, which improves efficiency and can prevent misdirection from gradients of non-visual words. This architecture achieved state-of-the-art results on benchmark datasets such as Microsoft COCO and Flickr30K. The model's ability to decide when and where to attend also enhances interpretability, as it can be visualized to show which words are grounded in visual information versus linguistic context.

In [8] proposes a combined bottom-up component that identifies specific object regions using detection network, while top-down component guides language model to attend to key image segments at each step of caption generation, enhancing overall precision & explainability. Disadvantages include architectural complexity & sensitivity to hyper-parameter tuning. The model introduces new approach to image captioning by combining Bottom-up & Top-down attention. The bottom-up component, powered by a Faster R-CNN detection network, automatically identifies a set of salient object regions in the image. This is a significant improvement over previous grid-based attention methods that treated all regions equally, regardless of their content. The top-down component then uses this information to selectively attend to these detected regions as it generates each word of the caption. This combined strategy allows model to better focus on relevant visual information, leading to more accurate & detailed captions. This achieved state-of-the-art results on benchmark datasets like Microsoft COCO.

In [9] introduced VLP, a unified Transformer-based model handling image captioning and VQA through bidirectional and sequence-to-sequence training objectives. It enables efficient fine- tuning. It requires massive datasets and has high computational cost Unified Vision-Language Pre-training (VLP) model built on a Transformer architecture. Its key innovation is a shared multi-layer Transformer network that is used for both encoding and decoding, which sets it apart from other models that use separate encoders and decoders. VLP is trained on a massive dataset of image-text pairs using two unsupervised learning objectives: a bidirectional masked vision-language prediction and a sequence-to-sequence masked vision-language prediction. This unified approach allows the model to be efficiently fine-tuned for a variety of tasks, including both vision-language generation (e.g., image captioning) and understanding (e.g., visual question answering). The paper demonstrates that VLP achieves state-of-the-art results on both types of tasks, which shows the effectiveness of its single, comprehensive architecture.

In [10] proposes CLIP, a generative model using CLIP embeddings for generating high-fidelity images from text prompts. While primarily for text-to-image generation, it offers insights into cross- modal learning. It is not directly applicable to captioning and has high GPU/memory requirement This approach explicitly separates the tasks of understanding the text and generating the image, which enhances image diversity while maintaining photorealism and semantic alignment. Although primarily developed for text-to-image generation, it provides valuable insights into cross-modal learning by demonstrating how a model can learn powerful representations that connect visual and linguistic information. The paper also notes that using diffusion models for the prior is computationally more efficient and leads to better quality outputs.

In [11] introduces Dense Captioning, which generates descriptions for multiple regions within an image by jointly performing region localization and caption generation. It provides fine-grained, region-specific descriptions. Challenges include managing overlapping regions & risk of semantic

inconsistency. It is an end-to-end dense captioning model that jointly localizes and describes image regions. It uses CNN backbone with Region Proposal Network (RPN) to generate candidate boxes, followed by a second "localization & captioning" network that extracts features for each region & feeds them into LSTM decoders. They introduce *joint inference* modules linking bounding-box regression & caption generation, & *context fusion* module that injects global image context into LSTM. In practice the model's LSTMs output both next word & refined box offsets together, which improves alignment between language & visual detections. This 2-stage network achieves large gain (≈73% relative) over prior baselines on Visual Genome, demonstrating that combining detection & description in one model gives more coherent region-level captions.

In [12] introduces SCST, a reinforcement learning technique that optimizes non-differentiable metrics like CIDEr by using model's inference output as a baseline. This reduces exposure bias. Training can be computationally intensive & less stable introduce Self-Critical Sequence Training (SCST) to directly optimize a captioning model for evaluation metrics. They use a standard CNN encoder + LSTM decoder but train it with REINFORCE, where the model's own inference-time greedy output serves as the baseline. In other words, the network is rewarded for generating captions that score higher than its greedy prediction under metrics like CIDEr. This policy-gradient training dramatically boosts performance: their SCST model on MSCOCO improves the CIDEr score from 104.9 to 114.7. Using the model's own output as the "self-critical" baseline also stabilizes training (reducing exposure bias and variance) compared to naive RL, yielding more natural and fluent captions at the expense of higher computational cost.

In [13] proposes an attribute-enhanced framework that incorporates high-level semantic attributes the caption generation process. This enhances recognition of fine-grained content. It relies on accurate attribute detection & increases preprocessing/model complexity. By fusing these attributes with standard visual features, their model is able to produce captions that exhibit greater descriptive detail & lexical variety propose LSTM-A, an attribute-augmented captioning architecture. Here, a CNN extracts visual features & a separate attribute predictor produces a vector of semantic attributes from the image. These high-level attributes are fused into the LSTM decoder along with the usual visual features. The authors implement five variants (LSTM-A1 through LSTM-A5) that differ in *where* (initial step vs. each step) and *when* (before or after feeding visual features) the attribute vector is input. For example, LSTM-A5 injects predicted attributes at every time step of the decoder, while LSTM-A1 uses only attributes at the start. The best variant substantially improves scores on COCO (CIDEr ≈ 100.2% versus ≈95% for a baseline model). This attribute fusion lets the model mention finer details & boosts lexical diversity, though it adds overhead for attribute extraction.

In [14] introduces a large-scale dataset harvested & cleaned from web images & alt-text, emphasizing open-domain & varied content. It minimizes noisy annotations & is useful for pretraining large vision-language models. Some captions may still carry noise, & there is less focus on spatial relationships introduce the Conceptual Captions dataset (~3.3M images from web alt-text) & analyse model architectures on it. They find that a hybrid CNN–Transformer model performs best using an Inception-ResNet-v2 image encoder combined with a Transformer decoder yields the highest accuracy. This architecture benefits from the rich residual-inception visual features & self-attention sequence modelling, which handles large, noisy caption data better than traditional RNN decoders. Their experiments show Transformer-based captioners achieve notably higher output quality on Conceptual Captions, indicating that modern multi-head attention models generalize well to open-domain, web-scale image descriptions.

In [15] applies the Reinforce algorithm within SCST framework to optimize metrics directly, using inference predictions as a baseline. It generates more natural and diverse captions. Training may be unstable & complex. SCST framework is further refined in [15] by similarly applying REINFORCE-based training to an attention-driven captioner. In this work, the model (again a CNN+LSTM encoder–decoder) uses its own decoded caption as the reward baseline, directly optimizing metrics during training. The authors observe that this self-critical training encourages more natural & diverse sentence generation, since it rewards correct but novel word choices. They also note that using the model's greedy output as baseline significantly lowers gradient variance (and avoids training a separate critic). However, as with SCST, the RL-trained model can be sensitive to reward weighting, careful tuning to ensure stable convergence.

In [16] proposes Transformer-based model that replaces recurrent layers with self-attention. It benefits from parallelization & long-range dependency modelling, offering faster training & improved fluency. It requires large datasets & is computationally intensive during pre-training propose a fully-transformer image captioning model that uses a Vision Transformer (ViT) as encoder & a GPT-2 based Transformer as decoder. They train this model with both image & text augmentations, reporting strong results on COCO (e.g. BLEU-4 ≈34.3, CIDEr ≈104.2) that surpass standard CNN–LSTM baselines. Each augmentation strategy alone improves performance, though combining them yields no further gain, and transfer-learning does not significantly help.

In [17] presents a unified Transformer framework that formulates tasks like image captioning, VQA, and visual grounding as text generation problems. This allows for cross-task generalization & shared parameters. It demands massive training corpus introduce a unified vision-language pretraining framework (VLP) that employs one shared multi-layer Transformer for both encoding and decoding. The model is pre-trained on large-scale image–text corpora with two masked objectives - a bidirectional prediction and a sequence-to-sequence (causal) prediction-controlled by special self-attention masks. This single-model approach can be fine-tuned for caption generation or VQA, achieving state-of-the-art results on diverse benchmarks (COCO Captions, Flickr30k, VQA 2.0) for both captioning and VQA tasks.

In [18] introduces a two-stage model that detects objects first and then uses them to guide caption generation via template filling. This results in grounded and interpretable captions. Its accuracy depends heavily on object detection quality tackle dense image captioning by detecting and describing multiple regions per image. Their end-to-end architecture is a two-stage pipeline (region proposal followed by a localization-and-captioning network) where an LSTM decoder jointly predicts region descriptions and bounding-box offsets, incorporating global visual context. The model's novel *joint inference* and *context fusion* modules produce a compact, efficient system that achieves a roughly 73% relative mAP gain on Visual Genome dense captioning compared to prior methods.

In [19] presented CLIP, a comparative objective method for developing unified visual–textual representations on large-scale internet data. The learned multimodal embeddings are versatile, enabling zero-shot transfer to a wide range of vision–language tasks, including caption generation, with no

further retraining required. It provides high-quality embeddings usable as features for caption generation, though it's not a dedicated captioning model. It requires adaptation or pairing with a decoder for captioning further detail the implementation of the dense-captioning system. They use a Faster R-CNN backbone to generate candidate regions and an LSTM-based head to generate each region's caption and refine its box. The authors experiment with various design choices (e.g. different LSTM structures and context-fusion strategies) and confirm that the integrated joint-inference design yields the best performance, aligning with the large mAP improvements reported in [18].

In [20] argues that captioning pretraining alone can produce vision encoders competitive with contrastively pretrained ones. It introduces a standard encoder-decoder transformer trained only on captioning tasks. It offers simplified pretraining and competitive performance. Disadvantages include less inference efficiency and training complexity study the algebraic-geometric problem of uniform K-stability in families of Q-Fano varieties (proving it is a Zariski-open condition).

The summary of Literature Survey is given in table-1.

3. CONCLUSION

Over the past decade, image captioning research has progressed from relatively simple CNN–RNN architectures to advanced transformer-based designs. Attention mechanisms-whether spatial, semantic, or adaptive-have proven essential in improving the quality, interpretability, and fluency of generated captions. The integration of transformers has greatly enhanced the contextual modelling & computational efficiency. Despite these gains, challenges remain, including the need for more diverse outputs, adaptation to smaller datasets, and reduction of computational demands. Future work is expected to leverage large-scale multimodal pretraining, knowledge integration, and fine-tuning for specific application domains. This review consolidates major developments, offering insights for researchers aiming to build next-generation captioning models that achieve high precision while being accurate as well as resource-efficient.

REFERENCES

1. Haoran Wang, Yue Zhang, Xiaosheng Yu, "*An Overview of Image Caption Generation Methods", Journal of Visual Communication and Image Representation*, Vol. 71, 2020, pp. 102-117.
2. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, et al., *Attention Is All You Need*, Proc. of NIPS, Vol. 30, 2017,pp.5998-6008.
3. Oriol Vinyals, Alexander Toshev, Samy Bengio, Dumitru Erhan, *Show and Tell: A Neural Image Caption Generator*, Proc. of IEEE CVPR, 2015, pp. 3156- 3164.
4. Jeff Donahue, Lisa Anne Hendricks, Marcus Rohrbach, et al., *Long-term Recurrent Convolutional Networks for Visual Recognition and Description*, Proc. of IEEE CVPR, 2015, pp. 2625- 2634.
5. Kelvin Xu, Jimmy Lei Ba, Ryan Kiros, et al., *Show, Attend and Tell: Neural Image Caption Generation with Visual Attention*, Proc. of ICML, Vol. 37, 2015, pp. 2048 2057.
6. Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, Jiebo Luo, *Image Captioning with Semantic Attention*, Proc. of IEEE CVPR, 2016, pp. 4651- 4659.
7. Jiasen Lu, Caiming Xiong, Devi Parikh, Richard Socher, *Knowing When to Look: Adaptive Attention via A Visual Sentinel for Image Captioning*, Proc. of IEEE CVPR, 2017, pp. 375-383
8. Peter Anderson, Xiaodong He, Chris Buehler, et al., *Bottom- Up and Top-Down Attention for Image Captioning and Visual Question Answering*, Proc. of IEEE CVPR, 2018, pp. 6077- 6086.
9. Luowei Zhou, Hamid Palangi, Lei Zhang, et al., *Unified Vision-Language Pre Training for Image Captioning and VQA*, Proc. of AAAI, Vol. 34, 2020, pp. 13041 13049.
10. Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, et al., *Hierarchical Text Conditional Image Generation with CLIP Latents,* arXiv preprint arXiv:2204.06125,2022.
11. Justin Johnson, Andrej Karpathy, Li Fei-Fei, *Dense Captioning with Joint Inference*, Proc. of IEEE CVPR, 2016, pp. 4565-4574.
12. Steven Rennie, Etienne Marcheret, Yousef Mroueh, Jerret Ross, Vaibhava Goel, *Self-Critical Sequence Training for Image Captioning*, Proc. of IEEE CVPR, 2017, pp. 7008-7024.
13. Ting Yao, Yingwei Pan, Yehao Li, Tao Mei, *Boosting Image Captioning with Attributes*, Proc. of IEEE ICCV, 2017, pp. 4894-4902.
14. Peter Young, Alice Lai, Micah Hodosh, Julia Hockenmaier, *Conceptual Captioning: A Clean Training Dataset for Image Captioning*, Proc. of ACL, 2018, pp. 2287 2297.
15. Yash Goyal et al., *SCST: Reinforcement Learning for Image Captioning*, arXiv preprint arXiv:1706.03850, 2017.
16. Yang Wang et al., *Visual Captioning Transformer*, Proc. of IEEE Transactions on Multimedia, Vol. 24, 2022, pp. 1420-1432.
17. Jiasen Lu, Dhruv Batra, Devi Parikh, et al., *Unifying Vision- and-Language Tasks via Text Generation*, Proc. of ICML, 2021, pp. 9934-9945.
18. Michael Tschannen, Manoj Kumar, Andreas Steiner, et al., *Image Captioners Are Scalable Vision Learners Too*, Proc. of NeurIPS, Vol. 36, 2023.
19. Alec Radford, Jong Wook Kim, Chris Hallacy, et al, CLIP: *Learning Transferable Visual Models from Natural Language Supervision*, Proc. of ICML, 2021, pp. 8748 8763.
20. Michael Tschannen, Manoj Kumar, Andreas Steiner, et al., *Image Captioners Are Scalable Vision Learners Too*, Proc. of NeurIPS, Vol. 36, 2023, pp. 12743–12755

TABLE-1: LITERATURE SURVEY SUMMARY

Authors	Title	Methodology	Remarks
Haoran Wang, Yue Zhang, Xiaosheng Yu[1], 2020	An Overview of Image Caption Generation Methods	Survey of deep learning-based image captioning; attention mechanisms	Comprehensive review of models, attention variants, datasets, metrics, challenges
Ashish Vaswani et al. [2], 2017	Attention Is All You Need	Transformer with self-attention	Transformer was introduced; RNNs/CNNs eliminated; foundational for modern captioning models

Oriol Vinyals et al. [3], 2015	Show and Tell: A Neural Image Caption Generator	CNN + LSTM encoder-decoder	Vision & language features integrated model
Jeff Donahue et al. [4], 2015	Long-term Recurrent Convolutional Networks	CNN + LSTM with variable-length input/output	Visual recognition & video description
Kelvin Xu et al. [5], 2015	Show, Attend and Tell	Visual attention (soft and hard); CNN + RNN	Image captioning with attention, enhancing interpretability
Quanzeng You et al. [6], 2016	Image Captioning with Semantic Attention	Semantic attention with visual attributes	Dual attention using visual features and semantic concepts
Jiasen Lu et al. [7], 2017	Knowing When to Look: Adaptive Attention	Visual sentinel with adaptive attention	Chooses between visual inputs and internal language context dynamically
Peter Anderson et al. [8], 2018	Bottom-Up and Top-Down Attention	Faster R-CNN + top-down attention	Object-level attention; state-of-art performance on captioning benchmarks
Luowei Zhou et al. [9], 2020	Unified Vision-Language Pre-Training (VLP)	Unified Transformer model; pretraining on large- scale image-text dataset	Joint model for captioning and VQA; efficient fine-tuning
Aditya Ramesh et al. [10], 2022	Hierarchical Text-Conditional Image Generation with CLIP Latents	CLIP embeddings + diffusion for image generation	Cross-modal latent space modeling advancement.

Justin Johnson et al. [11], 2016	Dense Captioning with Joint Inference	CNN + RNN; region-wise captioning	Multiple image regions; supports dense descriptions
Steven Rennie et al. [12], 2017	Self-Critical Sequence Training for Image Captioning	Reinforcement Learning (SCST); CIDEr optimization	Direct optimization of evaluation metrics; reduces exposure bias
Ting Yao et al. [13], 2017	Boosting Image Captioning with Attributes	Attribute-aware image captioning	Semantic attributes to enhance diversity and accuracy

Peter Young et al. [14], 2018	Conceptual Captioning	Dataset construction with linguistic and semantic filtering	Provides a large-scale, clean caption dataset for pretraining
Yash Goyal et al. [15], 2018	SCST: Reinforcement Learning for Image Captioning	SCST + REINFORCE for metric optimization	Enhanced reinforcement training stability and performance
Yang Wang et al l. [16], 2022	Visual Captioning Transformer	Transformer-only model for captioning	Captures long-term dependencies; faster training; high fluency
Jiasen Lu et al. [17], 2021	Unified Transformer for captioning, VQA, grounding	Unified Transformer for captioning, VQA, grounding	Multitask learning with shared architecture
Michael Tschannen et al. [18], 2023	Neural Baby Talk	Two-stage: object detection + template-based generation	Grounded captions with interpretable templates
Alec Radford et al. [19], 2021	CLIP: Learning Transferable Visual Models from Natural Language Supervision	Contrastive learning with 400M image-text pairs	Foundation model for multimodal understanding; decoding module needed
Michael Tschannen et al. [20], 2023	Image Captioners Are Scalable Vision Learners Too	Encoder-decoder trained on captioning only	Demonstrates captioning as competitive pretraining alternative to contrastive models