Evolution of Image Captioning Models: A Systematic PRISMA Review

Abdelkrim SAOUABE[1], Khalid TIZRA[2], Doha BANOUI[3]
Akkodis Research, Paris, France[1]
LIDSI Laboratory-Faculty of Sciences-Ain Chock, University Hassan II, Casablanca, Morocco[2]
2IACS Laboratory-ENSET Mohammedia, University Hassan II, Casablanca, Morocco[3]

***Abstract*—This article presents a systematic review of image captioning approaches conducted according to the PRISMA methodology, ensuring a rigorous, transparent, and reproducible analysis of the literature. The study traces the evolution of image captioning methods, beginning with early machine learning–based techniques that rely on handcrafted visual features, object detection, and template-based or statistical language models. While these approaches established foundational concepts, they are constrained by limited scalability and semantic expressiveness. Specific challenges include difficulty in capturing complex object relationships and inability to generate diverse descriptions for the same image. Image captioning represents a key research problem at the intersection of computer vision and natural language processing, aiming to automatically generate coherent and semantically accurate textual descriptions of visual content. Due to its multimodal nature and practical relevance, it has attracted increasing attention in artificial intelligence research. The review then examines the transition toward deep learning–based models, which have become dominant due to their improved performance. Encoder–decoder architectures are analyzed, highlighting the use of convolutional neural networks for visual representation and recurrent neural networks for caption generation. Attention-based models are discussed for their ability to focus on salient image regions, followed by reinforcement learning–based methods that directly optimize evaluation metrics and semantic-driven architectures that enhance caption relevance. Finally, recent advances based on Transformer architectures and large-scale multimodal pretraining are reviewed, along with key application domains and open challenges for future research in image captioning.**

***Keywords*—Image captioning; vision-language models; semantic-based models; transformer models; attention mechanism; pre-trained models; GPT-based models**

I. Introduction

In recent years, image captioning has attracted substantial attention at the intersection of computer vision and natural language processing within the broader field of Artificial Intelligence (AI) [20]. The objective of image captioning is to automatically generate linguistically coherent and semantically meaningful textual descriptions that accurately reflect the visual content of an image. This task requires not only the recognition of objects and scenes, but also the modeling of their attributes, relationships, and contextual interactions, thereby bridging visual perception and natural language understanding.

Automatic image captioning has demonstrated its practical relevance across diverse real-world applications [30]. Notable examples include assistive technologies for visually impaired individuals, automatic medical reporting in healthcare [15], and human-machine interaction systems that enable more intuitive and accessible interfaces.

Given the rapid evolution of image captioning methods, this article provides a systematic and structured review of the field. Section II presents the methodological framework adopted for this survey, based on the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines [2], ensuring transparency, reproducibility, and comprehensive coverage of the relevant literature.

Section III reviews the evolution of image captioning techniques, tracing the transition from early traditional methods based on handcrafted features and classical machine learning to modern deep learning approaches. This chapter first discusses traditional methods, followed by simple deep learning models relying on encoder–decoder architectures. It then examines attention-mechanism-based models, reinforcement-learning-based approaches, and semantic-based architectures. More recent advances are covered through Transformer-based architectures and large-scale pre-training-based models, which have significantly improved caption quality and generalization performance.

Finally, Section IV provides a critical discussion of the reviewed approaches, highlighting current challenges, comparative strengths and limitations, and emerging research directions for future image captioning systems and Section V provides conclusion of the study.

II. Systematic Review Method Using PRISMA

In this study, the methodological framework is based on the PRISMA model, which is widely recognized for ensuring rigor and transparency in systematic literature reviews [2]. The use of PRISMA strengthens the reliability of the analysis by structuring each stage of the process from the identification of relevant studies to the final synthesis of results. This framework ensures the reproducibility of the review and supports an exhaustive selection of contributions related to Image Captioning.

The PRISMA process begins with a clear definition of the review objectives and the formulation of specific research questions, which guide the entire review procedure. Based on these elements, a research protocol was developed, detailing the consulted databases, the search strategies used, and the inclusion and exclusion criteria that determine the scientific relevance of the retained articles. Search queries were constructed using keywords and terminological combinations specific to Image Captioning and executed across several specialized academic databases.The search was conducted across major academic databases including IEEE Xplore, ACM Digital Library, Springer, ScienceDirect, and arXiv.

The results of these searches were then subjected to a multi-stage selection process, following the PRISMA flow Fig. 1. First, titles and abstracts were screened to remove studies that were clearly irrelevant. Articles deemed potentially relevant underwent full-text assessment, conducted independently by at least two reviewers to ensure consistency, quality, and reliability in the selection process.

After the final selection phase, the included studies underwent systematic data extraction, covering their methodologies, proposed approaches, main contributions, employed metrics, and authors' conclusions. The extracted information was then analyzed comparatively to identify recurrent trends, methodological limitations, and emerging research directions in the field.

The criteria applied in the selection process were defined as follows:

- Inclusion criteria: scientific articles explicitly addressing Image Captioning, presenting methodological or conceptual contributions, describing the adopted approach, and specifying the evaluation metrics used.
- Exclusion criteria: non-scientific documents (reports, non-academic book chapters, etc.), publications written in languages other than English or French, duplicate records, and articles not directly related to the research topic.

The application of the PRISMA model thus ensures a structured, reliable, and reproducible systematic review, providing a solid methodological foundation for the state of the art presented in this study.

III. Evolution of Image Captioning Models

Image captioning can be divided into two major stages. Prior to 2014, it relied on conventional techniques rooted in machine learning for retrieval and segmentation. Since 2014, as the complexity of feature extraction from images has increased, new methods based on deep learning technologies have emerged as the leading approaches, delivering state-of-the-art results. Fig. 2 illustrates the progression of automatic image captioning.

Recent research has focused on image captioning using deep learning techniques. Initial approaches utilized a Convolutional Neural Network (CNN) to extract visual representations, which were then fed into a Recurrent Neural Network (RNN) for the generation of an output sequence, following an encoder–decoder model structure. Subsequently, techniques have been enhanced by incorporating region-based features, attention mechanisms, reinforcement learning strategies, semantic attributes, and even transformer models integrating self-attention and pre-training methods such as Generative Pre-trained Transformer (GPT) designed for both vision and language tasks.

These endeavors are directed towards discovering the optimal pipeline for establishing meaningful associations between visual semantics and textual components, translating visual elements into words in a sequence while preserving their inherent significance. Within this chapter, we offer a comprehensive summary of diverse techniques and approaches employed in image captioning.

A. *Traditional Methods*

Before the advent of deep neural networks, image captioning systems relied primarily on traditional approaches combining manually defined visual descriptors and language models or text templates. These methods laid the foundations for the field while illustrating the limitations of manual engineering [20]. In retrieval-based methods, the idea is to use a database of annotated images: for a new image, similar images are searched for in the visual feature space, and then one of their captions is reused to describe the target image. Visual similarity is evaluated using descriptors such as histograms, keypoint descriptors, or texture descriptors, which are transformed into vectors that can be used by classifiers or distance measures [21]. These methods have the advantage of being simple and guaranteeing a certain grammatical correctness (since the captions are human-generated), but they lack flexibility: they do not generate new sentences, and their quality depends heavily on the visual similarity between the target image and those in the database [23].

Another family of methods first identifies objects, attributes, or relationships from images using traditional classifiers or detectors such as Support Vector Machines (SVM), Conditional Random Fields (CRF), etc., then fills in predefined linguistic templates to form sentences. For example: "An <object> is <action> in <scene>." These templates, completed with the detected visual entities, produce a textual description [21]. These systems can generate grammatically correct sentences with little data, but their expressiveness is severely limited: descriptions remain stereotypical, lacking in variety, and ill-suited to complex or original scenarios [23].

To visually encode an image, these methods used classic computer vision descriptors, such as local descriptors (e.g., Scale-Invariant Feature Transform (SIFT), Histogram of Oriented Gradients (HOG)), color histograms, texture descriptors, or representations based on "bag-of-visual-words" (BoVW) [26]. These descriptors aimed to capture robust visual information that was invariant to scale, orientation, and variations in brightness, but their descriptive power remained limited in the face of the semantic and contextual richness of natural images [24].

Despite their pioneering role, these traditional methods suffered from several structural weaknesses:

- Rigidity and lack of generality: Retrieval and template-based methods did not generate new sentences, and templates limited the diversity of expressions [21].
- Poor semantic understanding: Manual visual descriptors were unable to capture complex relationships between objects, attributes, and overall context, nor were they able to model deep linguistic dependencies [28].

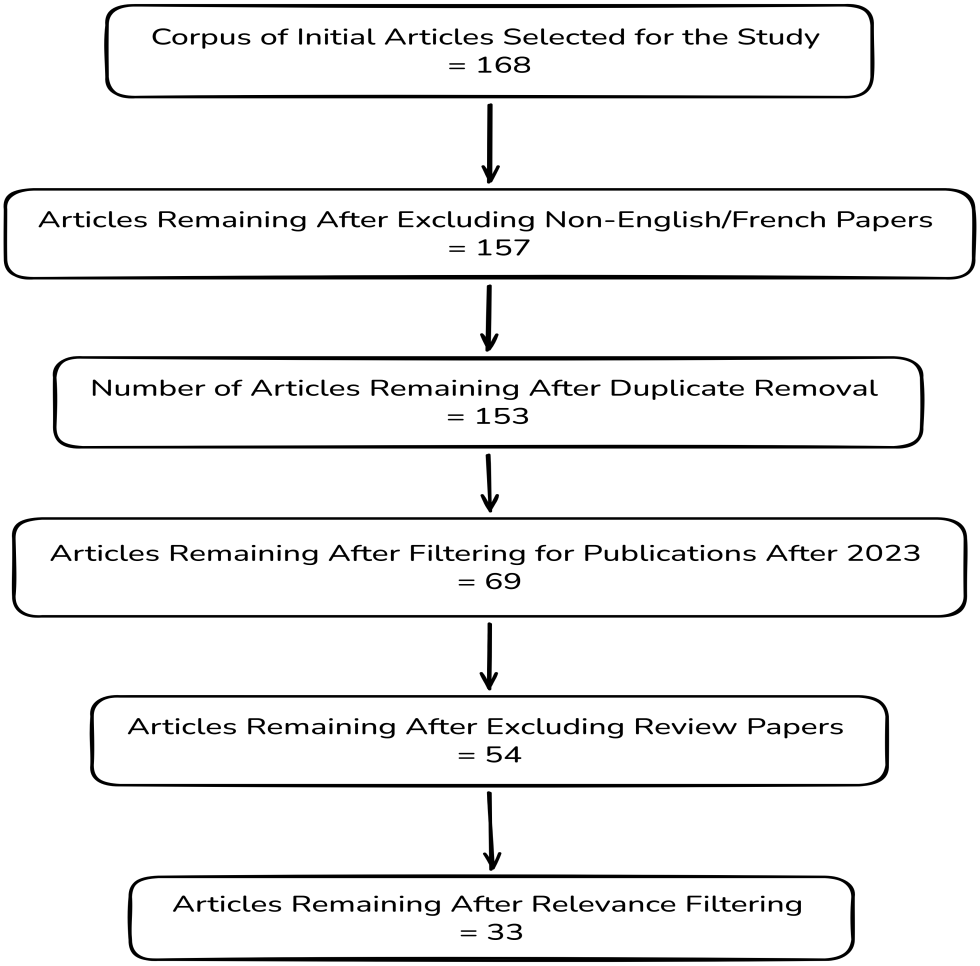
Corpus of Initial Articles Selected for the Study = 168
Articles Remaining After Excluding Non-English/French Papers = 157
Number of Articles Remaining After Duplicate Removal = 153
Articles Remaining After Filtering for Publications After 2023 = 69
Articles Remaining After Excluding Review Papers = 54
Articles Remaining After Relevance Filtering = 33

Fig. 1. Systematic filtering and selection of articles following the PRISMA methodology.

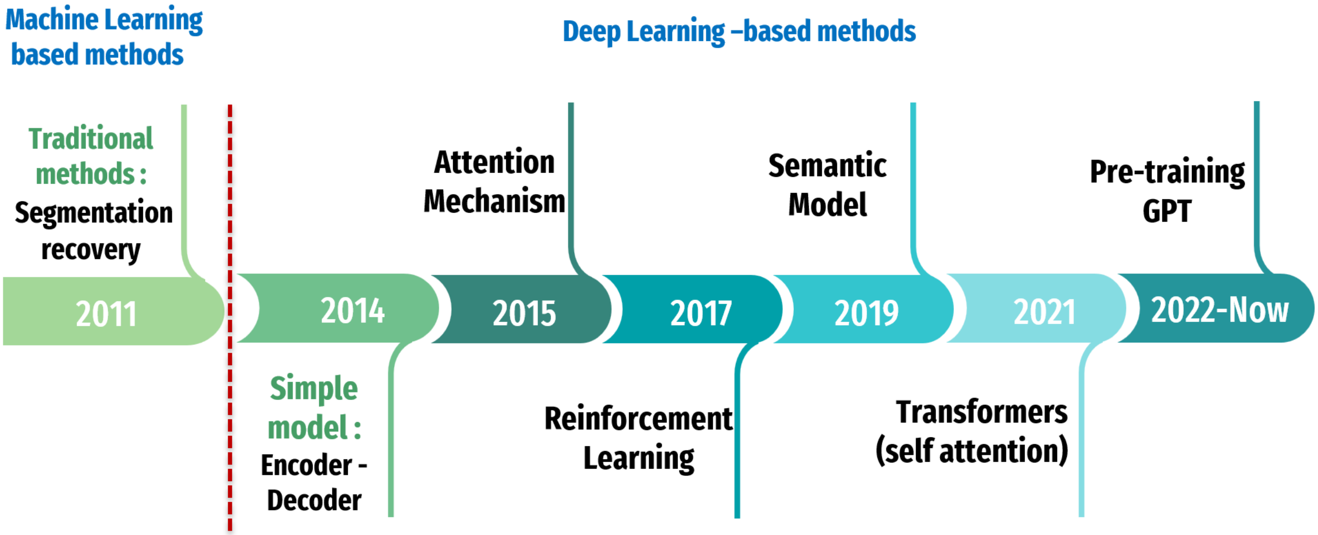
Machine Learning based methods
Deep Learning –based methods
Traditional methods : Segmentation recovery
2011
Simple model : Encoder - Decoder
2014
Attention Mechanism
2015
Reinforcement Learning
2017
Semantic Model
2019
Transformers (self attention)
2021
Pre-training GPT
2022-Now

Fig. 2. Evolution of image captioning methods.

- Disjointed vision-language pipeline: Vision (feature extraction) and text generation were separated, without joint learning, which limited adaptability and visual–linguistic consistency [23].

These observations, drawn from fundamental research on automatic visual captioning, explain why the community gradually turned to deep learning techniques in early 2014. The limitations of manual descriptors and rigid models led to the adoption of learning architectures capable of automatically acquiring joint visual and linguistic representations [29].

B. Simple Models: Encoder–Decoder-Based Architectures

First and foremost, the so-called encoder–decoder architecture for image captioning combines a vision module (encoder) and a text generation module (decoder) in an end-to-end manner: the image is first transformed into a vector representation, which is then used to generate a natural caption, word by word.

- Encoder — typically a CNN pre-trained for image classification (e.g., on ImageNet). The last classification layer is removed to use deep activations as a visual representation. This representation encodes the relevant visual features of the image: shapes, textures, objects, global context.
- Decoder — typically an RNN, often a Long Short-Term Memory (LSTM) network, which takes the visual representation from the encoder (possibly transformed) as its initial context, then generates a caption in sequence: at each time step, the decoder predicts a word based on the visual context and the history of previous words.

Multimodal space-based approaches generally rely on an architecture composed of four fundamental modules: A linguistic encoder, a visual encoder, a multimodal fusion space, and a linguistic decoder. In this scheme, the visual module relies on a deep CNN to extract relevant semantic features from images, while the linguistic encoder derives compact vector representations for each word and models their temporal dynamics using recurrent layers. The multimodal space then aligns the visual and textual features in a common representation, enabling their effective combination within the generative process [30]. This functional organization naturally translates into an encoder–decoder architecture, where the encoder maps the visual characteristics of an image to an intermediate representation, similar to the processing of an input sequence in machine translation. This paradigm explicitly divides the task into two steps:

- Encoding, responsible for extracting relevant information from the input image.
- Decoding, dedicated to generating the linguistic sequence from the encoded representations.

Several studies are based on this architecture, including the model described in study [30], which uses two complementary neural networks, the first, a CNN, extracts objects from the image as well as their spatial structure. The second, generally an LSTM-type recurrent network, uses these visual representations to generate a coherent, fluent, and grammatically correct caption.

Building on this line of research, [6] presents an innovative method for automatically generating image descriptions using a Bi-directional Long Short-Term Memory (Bi-LSTM) model and an optimization method called Novel Moth Flame Optimization (NMFO). The goal of this approach is to address the challenge of automatically generating captions for images by combining computer vision and natural language processing techniques. The Bi-LSTM model is used to generate descriptions by capturing long-term dependencies and leveraging the context of the image. The NMFO optimization method, inspired by the behavior of moths attracted to flames in nature, is applied to fine-tune the model's parameters and improve its performance. A logarithmic spiral based on correlation is used to guide the optimization process. The approach is evaluated on widely used datasets, including Flickr8k and MS-COCO, using standard metrics such as Bilingual Evaluation Understudy (BLEU) and Consensus-based Image Description Evaluation (CIDEr) to assess the quality of the generated captions. The model achieves 0.5883 BLEU and 0.8303 CIDEr on Flickr8k, and 0.7988 BLEU and 0.8341 CIDEr on MS-COCO, confirming its effectiveness across different benchmarks.

To improve caption diversity and the modeling of long-term dependencies, [17] proposes a hybrid approach for automatic image captioning that integrates ResNet-50[31] for visual feature extraction, LSTM networks for sequential text generation, and Beam Search decoding to enhance caption diversity. The method addresses fundamental challenges in image captioning, including rare word handling, creative description generation, and the modeling of long-term dependencies in textual sequences. Experimental evaluation on the Flickr8k dataset demonstrates improved BLEU scores compared to baseline models, indicating enhanced caption quality. While the approach shows practical applicability across domains such as accessibility tools and visual search systems, the authors acknowledge several limitations, including the absence of attention mechanisms for fine-grained visual–textual alignment and limited generalization beyond the training dataset. Despite these constraints, the method provides a practical and computationally efficient solution for image captioning in controlled environments, demonstrating the effectiveness of combining established deep learning architectures with search-based decoding strategies.

In a complementary direction, the study [19] investigates the optimal combination of visual features and word embeddings for automatic image captioning within an encoder–decoder framework. The authors conduct a comprehensive empirical analysis evaluating ten CNN architectures for visual feature extraction paired with two types of word embeddings for textual generation. Experimental validation is performed on standard benchmarks including MSCOCO and Flickr30k, with evaluation conducted using established metrics such as BLEU, Metric for Evaluation of Translation with Explicit ORdering (METEOR), Recall-Oriented Understudy for Gisting Evaluation (ROUGE), and CIDEr to assess caption accuracy and semantic coherence. The results demonstrate that caption quality is significantly influenced by the choice and combination of visual and linguistic components, with certain architecture–embedding pairs yielding substantially superior performance. This systematic experimental approach identifies best practices for integrating visual and textual modalities, providing practical guidelines for designing more effective captioning systems. The work emphasizes the importance of modular and well calibrated architectural design in achieving high-quality captions across varied visual contexts, though computational complexity and data dependency may constrain generalization to diverse domains.

Low-data regimes are tackled in study [18], which introduces Few-shot Remote Sensing Image Captioning (FRIC), a framework for generating captions for satellite images with very few annotations (<1% of the standard dataset). Its main innovation is an optimized decoding strategy combining multi-model ensembles, auto-distillation, and parameter sharing. Annotated samples are divided into subsets to train several base models, whose predictions are then aggregated to improve robustness and generalization. FRIC does not rely on external data or pre-trained models, using pseudo-labels to exploit unannotated images. Experimental evaluations show that FRIC significantly outperforms existing methods with only 0.8% of annotations, and ablation studies confirm the importance of each component.

C. Attention-Mechanism-Based Architectures

Attention mechanisms have become a cornerstone of modern image captioning by enabling models to dynamically focus on relevant visual regions during word generation. Originally introduced in neural machine translation [44], attention was successfully adapted to image captioning through models such as Show, Attend and Tell [45]. In these frameworks, CNN encode images into spatial visual representations, while recurrent decoders generate captions by weighting visual regions according to their relevance at each time step. This process improves descriptive precision, semantic coherence, and the recognition of secondary objects compared to conventional encoder–decoder models [46].

Subsequent research has proposed more advanced attention variants, including hierarchical attention [47], semantic attention [48], and multi-head attention inspired by transformer architectures [49]. These extensions enhance the modeling of object relationships, visual attributes, and complex linguistic dependencies. Building upon these foundations, later approaches introduced adaptive, relational, and context-aware attention strategies to strengthen visual–linguistic alignment and capture fine-grained visual details. Collectively, these developments demonstrate the central role of attention-based architectures in improving the accuracy, relevance, and interpretability of image captioning systems.

To further exploit visual hierarchies, the study [9] proposes a method to automatically generate image captions using a hierarchical attention mechanism and policy gradient optimization. This approach combines an encoder–decoder model with a hierarchical attention mechanism to capture both the global and local visual features of the image. Policy gradient optimization is used to train the model using positive and negative rewards. The model is trained on the annotated MSCOCO dataset and evaluated using standard automatic metrics such as BLEU, METEOR, CIDEr, and ROUGE-L. Its objective is to enhance the quality and accuracy of the generated captions by integrating hierarchical attention with policy gradient optimization. The results show competitive performance, with 72.611 BLEU-1, 52.769 BLEU-2, 37.802 BLEU-3, 27.243 BLEU-4, 24.731 METEOR, 88.140 CIDEr, and 56.048 ROUGE-L.

Building on advances in convolutional backbones, [22] evaluates ConvNeXt variants integrated into an LSTM-based captioning pipeline with visual attention on MSCOCO. ConvNeXt-Base, trained without teacher-forcing, achieves BLEU-4 of 34.76, outperforming soft-attention (+43.04%), hard-attention (+39.04%), Vision Transformer (ViT) (+4.57%), and Data-efficient Image Transformer (DeiT) (+0.93%), while improving Top-5 accuracy by +6.68% and reducing loss by 18.72% versus MobileNetV3. The results demonstrate that ConvNeXt offers an effective balance between performance and computational efficiency, providing a compelling alternative to both traditional CNNs and transformers for image captioning tasks.

Metaheuristic optimization is explored in study [25], which introduces an Automated Image Captioning using Sparrow Search Algorithm with Improved Deep Learning (AIC-SSAIDL), a framework that integrates MobileNetV2 and an Attention-Mechanism LSTM (AM-LSTM) with Sparrow Search Algorithm (SSA) for visual hyperparameter tuning and Fruit Fly Optimization (FFO) for text generation optimization. Evaluated on Flickr8k, Flickr30k, and MSCOCO datasets, the model achieves BLEU-1 of 80.40, BLEU-4 of 38.04, METEOR of 33.58, and CIDEr of 137.45 on MSCOCO, outperforming traditional baseline methods. The approach demonstrates that metaheuristic optimization can enhance caption quality, though computational complexity and limited comparison with recent transformer architectures remain noted limitations.

Within the medical domain, [27] proposes a hybrid framework combining YOLOv4 for object detection with an attention-based LSTM model for medical image captioning. The Flamingo Search Optimization (FSO) algorithm enhances alignment between detected anatomical structures and generated descriptions. Evaluated on the PEIR Gross dataset (7,442 annotated images), the model achieves a BLEU score of 81.78%, representing a +4.42% improvement over baseline methods. While the approach demonstrates enhanced caption quality and reduced analysis time, the authors acknowledge limitations including sensitivity to image noise and computational overhead.

Finally, a culturally specific application is explored in [32], which addresses automatic captioning of Buddhist Thangka paintings, a domain with limited annotated data and high visual complexity. The authors propose a Semantic Concept Prompt and Multimodal Feature Optimization network (SCAMF-Net), integrating two key modules: Semantic Concept Prompt (SCP), which incorporates cultural knowledge through contextual vectors, and Multimodal Feature Optimization (MFO), which enhances image–text alignment and filters noisy data. Evaluated on a dataset of 3,974 annotated Thangka images and MSCOCO, the model achieves significant improvements with BLEU-4 of 63.6% (+8.7%), METEOR of 52.0% (+7.9%), and CIDEr of 562.4 (+76.6) compared to existing methods. Despite high computational complexity and reliance on annotated data, SCAMF-Net represents a substantial advancement in processing culturally significant heritage images, demonstrating the effectiveness of integrating domain-specific semantic knowledge with multimodal feature optimization for specialized captioning tasks.

D. Reinforcement-Learning-Based Architectures

To overcome the limitations of sequence-level training, reinforcement learning (RL) has been widely explored in image captioning. In [8], Conditional Generative Adversarial Networks (cGANs) are employed to improve upon pre-established image captioning models that rely on RL. The authors show that adversarial training can complement RL-based optimization by encouraging more realistic and diverse captions. Another line of work focuses on reward design. For instance, [10] utilizes a gradient policy methodology to optimize rewards based on human evaluations, directly incorporating human preferences into the learning process. This direction highlights the importance of aligning automatic metrics with human judgment.

Human attention modeling is further explored in study [11], which emulates human attention preferences and refines attention through RL by incorporating linguistic evaluation rewards. It is worth noting that enhancing random strategies within a reasonable timeframe can be challenging. Consequently, standard image captioning models typically involve pretraining with cross-entropy or masked language models, followed by fine-tuning using RL techniques, with metrics at the sequence level serving as rewards. The study reports state-of-the-art performance using the full dataset with distractors, where the Mirrored Viewpoint-Adapted Matching (M-VAM) model improves BLEU-4, METEOR, SPICE, and CIDEr, for example achieving BLEU-4 up to 50.3, METEOR up to 37.0, SPICE up to 24.4, and CIDEr up to 114.9. On the version of the dataset without distractors, the model achieves even stronger results, with CIDEr scores of 117.4 and 119.1 and BLEU-4 scores of 45.5 and 50.1, alongside SPICE scores of 30.7 and 31.2. These larger gains highlight that, without distractor influence, the model can more accurately capture semantic changes associated with viewpoint variations.

A more explicit use of human feedback is presented in study [10], which improves image captioning by integrating offline human evaluations. Generated captions are assessed through individual and pairwise comparisons on the T2 dataset, and the collected feedback is used to refine the model via reinforcement learning. Experiments across multiple datasets show consistent performance gains, with up to +3.19% in average evaluation score and +3.4% in ranking metrics, while higher policy-gradient weighting further improves informativeness, correctness, and fluency, demonstrating the effectiveness of human in the loop optimization.

Interactive and uncertainty-aware RL is explored in study [33], which proposes a medical image captioning framework combining uncertainty estimation, keyword prediction, caption generation, and selective user feedback. Using evidential learning, the model queries users only for uncertain keywords and updates itself with weighted feedback. Evaluated on IU-Xray, PEIR Gross, and MIMIC-CXR, it outperforms strong baselines, achieving 0.851 mean Average Precision (mAP) and 0.766 F1 on PEIR, with consistent gains in BLEU, ROUGE, and METEOR, demonstrating efficient and accurate uncertainty-guided captioning.

Beyond captioning alone, [34] introduces a RL-inspired method using reinforce to fine-tune pretrained computer vision models with task-specific rewards across object detection, panoptic segmentation, image colorization, and captioning. The approach involves pretraining via Maximum Likelihood Estimation (MLE) followed by reward-based optimization. Evaluated on MSCOCO, Objects365, ImageNet, and MSCOCO Captions datasets, the method demonstrates significant improvements: mAP increases from 39.2% to 54.3%, Panoptic Quality (PQ) from 43.1% to 46.1%, CIDEr from 120.0 to 134.5, and colorfulness from 0.41 to 1.79. The method enhances performance without altering model architecture and is applicable across diverse tasks. However, the authors acknowledge heavy reliance on reward design and high computational costs. This approach provides a framework for fine-grained model tuning using human feedback and targeted optimization strategies.

RL has also been used to align model reasoning with human expectations. For example, [35] introduces an InteractiveLy RationaLizing Vision-LangUage ModEls (ILLUME), a tuning paradigm for improving alignment between vision–language model outputs and human reasoning. While pretrained language models are effective for image captioning and visual question-answering (VQA), their outputs often fail to reflect expected commonsense rationales. ILLUME incorporates human feedback into the training loop: given an image–question–answer prompt, the model generates multiple candidate rationales, from which a human critic selects preferred ones used for fine-tuning. This iterative process enriches training data and shapes the model's ability to produce explanations aligned with human intent. Remarkably, ILLUME achieves performance comparable to standard supervised fine-tuning while requiring significantly less annotated data and minimal human interaction. The framework demonstrates that integrating human preferences into model training can enhance interpretability and trustworthiness in multimodal AI systems.

Finally, diversity-focused RL is investigated in [36], which introduces Unlikely Negative Knowledge Training (UNKT), an approach for reducing generic descriptions in image captioning by teaching the model what to avoid alongside what to generate. The system employs a negative teacher trained to produce bland captions and a student model that learns to avoid them through negative knowledge transfer. Evaluated on MSCOCO, UNKT outperforms traditional methods in both quality metrics (BLEU, METEOR, CIDEr) and diversity metrics (SelfCIDEr, Div-n). Ablation studies confirm the importance of each model component, while human evaluations find UNKT captions more informative and fluent. Easily integrable into other captioning architectures, UNKT enhances caption richness and relevance while reducing redundancy, paving the way for more expressive systems without requiring additional annotations.

E. Semantic-Based Architectures

Semantic modeling constitutes another important direction in image captioning. Initial exploration in this direction was undertaken by study [12], followed by subsequent work such as [39], which proposed the idea of utilizing a Graph Convolutional Network (GCN) to integrate semantic and spatial relations among objects. The semantic relationship model is derived by applying a pre-trained classifier to Visual Genome, predicting actions or interactions between pairs of objects. Retrieval-based techniques then generate descriptions for a provided image by selecting the one that shares the closest semantic meaning, phrases, or expressions from a database. As a result, the generated captions tend to exhibit grammatical accuracy, fluency in expression, and a resemblance to natural language.

In study [12] the authors present an innovative approach using a hierarchical parsing model to generate descriptive captions for images. The approach divides the image into distinct regions to understand its structure and content, capturing the relationships between these regions to produce detailed captions. The process through which the model is trained involves supervised optimization, where sets of images and their corresponding captions are combined for use. The approach is evaluated on the MSCOCO dataset using standard metrics such as BLEU, METEOR, ROUGE-L, CIDEr-D, and SPICE. Despite strong quantitative results : 95.9 BLEU-1, 90.4 BLEU-2, 81.6 BLEU-3, 71.0 BLEU-4, 38.1 METEOR, 74.1 ROUGE-L, and 130.2 CIDEr-D, the method faces several limitations. Its performance depends heavily on accurate object detection, and the hierarchical structure increases model complexity. In addition, potential errors in hierarchical analysis and the difficulty of handling complex scenes may negatively impact the quality of the generated captions.

To better combine visual and semantic information, [7] presents an approach that integrates attention networks to semantically represent the objects in the image and the words in the caption. This approach combines visual and semantic features of objects to generate more accurate and meaningful captions. The proposed model skillfully merges visual and semantic information using a transformer-based attention block. This fusion enhances visual attention and establishes richer semantic associations during caption generation. The evaluation is conducted on the MSCOCO dataset, the model obtains 78.6 BLEU-1, 36.0 BLEU-4, 27.6 METEOR, 57.7 ROUGE-L, and 120.98 CIDEr-D, confirming its overall effectiveness.

Human–human interactions are specifically addressed in study [37], which focuses on learning interactions in images using weak textual supervision. The proposed approach combines scene graphs (SG), GCNs, transformer decoders, and cross-attention mechanisms to align semantic and syntactic structures between images and text. Evaluated on diverse datasets including UT-Interaction, AVA, NTU RGB+D, imSitu, and Who's Waldo using BLEURT and Natural Language Inference (NLI)-based factuality scores, the CLIP Captioning with Conceptual Captions plus pseudo Human–Human Interactions (CLIPCap CC+pHHI) model achieves a BLEURT score of 0.46, outperforming state-of-the-art captioning and action recognition models. While the method significantly enhances understanding of human interactions, the authors acknowledge challenges with data bias and coherence in complex scenarios. The work represents a promising advancement in modeling human behavior from images with implications for intelligent visual analysis systems.

Entity-aware captioning is investigated in study [38], which proposes a multimodal knowledge graph that explicitly links visual objects with named entities while capturing their semantic relations, enriched with external sources such as Wikipedia. This approach tackles the challenge of generating captions that identify named entities and specific events requiring knowledge beyond object recognition. Evaluated on two large-scale news datasets GoodNews (445k images) and NYTimes800k (763k images), the method significantly outperforms state-of-the-art baselines, achieving higher scores on BLEU-4 and CIDEr metrics, with F1 scores for entity recognition surpassing all competitors across both datasets. The key finding demonstrates that integrating multimodal external knowledge enables captions that are descriptively accurate and semantically informative about depicted entities and events. This work advances entity-aware captioning toward human-like reasoning with practical implications for news media, accessibility, and intelligent information retrieval applications.

Visual scene graphs are central in study [39], which addresses image captioning by leveraging such graphs to enhance descriptive quality. The proposed to TransForm Scene Graphs into more descriptive Captions (TFSGC) model is built on a homogeneous transformer architecture with a Mixture of Experts (MoE) decoder. The method encodes objects, attributes, and relations via Multi-Head Attention Graph Neural Networks (MHA-GNN), then dynamically selects relevant experts to generate captions. Evaluated on MSCOCO and Visual Genome datasets using CIDEr-D, Recall@50, and cross-entropy loss metrics, TFSGC outperforms existing models by producing richer and more accurate captions with notable CIDEr improvements. The approach represents a theoretical advancement by effectively integrating graph structure into language generation. However, the authors acknowledge that performance heavily relies on scene graph and visual feature quality, with model complexity potentially hindering practical adoption. TFSGC offers a meaningful contribution to image captioning while raising challenges for generalization and real-world implementation.

Unsupervised learning is explored in study [40], which introduces a framework for unsupervised image captioning that eliminates the need for paired image–text datasets by using object relationships as a bridge between images and text. The method employs relational distant supervision to align visual content with external textual knowledge. It consists of three modules: relationship learning, relationship-to-sentence generation to create pseudo caption pairs, and an image captioning module trained on these pairs. Experiments on benchmark datasets show that the approach outperforms existing unsupervised captioning methods, producing more semantically aligned and context-aware captions.

F. Transformer-Based Architectures

The introduction of the transformer architecture by the authors of [49] marked a major paradigm shift in sequence modeling, replacing recurrent structures with fully attention-based mechanisms capable of capturing long-range dependencies more efficiently. Unlike RNN- or LSTM-based encoder–decoder models, transformers rely entirely on multi-head self-attention to process input sequences in parallel, significantly improving both computational efficiency and representational capacity.

In image captioning, early adaptations of transformers demonstrated substantial gains by enabling richer interactions between visual features and linguistic tokens. Works such as

Meshed-Memory Transformer (MMT) [50] and Object Relation Transformer (ORT) [51] showed that transformer decoders can effectively integrate object-level embeddings, spatial relations, and global contextual information, outperforming traditional attention models. Transformer-based encoders also allow images to be represented not only as global CNN features but as structured sets of patches or regions, facilitating fine-grained attention and relational reasoning [52].

Following the introduction of the transformer architecture, numerous researchers have proposed transformer-based models to further improve image captioning performance. These works aim to better model long-range dependencies, capture complex visual–semantic relationships, and enhance the alignment between image regions and generated words. By leveraging self-attention and cross-attention mechanisms, transformer-based approaches have demonstrated superior capability in handling object interactions, contextual reasoning, and global scene understanding. Among these research efforts, several representative models have been introduced, each proposing architectural innovations or learning strategies that contribute to improved caption accuracy, coherence, and semantic richness.

The introduction of the reinforced attention model in [13] marked a significant shift in the landscape of language generation. Shortly thereafter, the transformer model emerged as a pivotal building block, leading to further advancements in Natural Language Processing (NLP), exemplified by innovations like BERT [53] and GPT [54]. Currently, it serves as the conventional architecture for numerous natural language understanding tasks. Since the captioning of images can be considered as a sequence-to-sequence challenge, the transformer architecture has also been adapted for this purpose. One such model is evaluated on the MSCOCO dataset using standard captioning metrics and achieves 94.4 BLEU-1, 87.8 BLEU-2, 78.2 BLEU-3, 67.4 BLEU-4, 36.8 METEOR, 72.8 ROUGE-L, and 122.6 CIDEr-D, reflecting solid performance in both accuracy and descriptive quality.

The transformer decoder typically operates through masked self-attention over the generated words, followed by cross-attention where linguistic queries attend to visual representations produced by the encoder, and a final feed-forward network, ensuring unidirectional caption generation through causal masking. In image captioning, several works adopt this decoder with limited architectural modifications, while others extend it to better integrate visual attributes or directly process image patches without relying on convolutional networks. Among these approaches, the ORT exemplifies an effective adaptation, combining object detection with a transformer-based framework that incorporates relative geometric relationships into the attention mechanism [3]. Trained on the MSCOCO dataset, ORT achieves strong performance across standard metrics, including 128.3 CIDEr-D, 22.6 SPICE, 80.5 BLEU-1, 38.6 BLEU-4, 28.7 METEOR, and 58.4 ROUGE-L, demonstrating the capacity of transformer-based architectures to generate accurate and semantically rich image captions.

To jointly exploit complementary visual representations, [4] proposes the Dual-Level Collaborative Transformer (DLCT), an image-captioning model designed to automatically generate coherent and descriptive captions. The approach exploits both grid-level and region-level visual features to benefit from their complementarity and introduces a Comprehensive Relation Attention (CRA) mechanism that captures complex visual and spatial relationships by combining absolute and relative positional information. The model further distinguishes itself through a dual form of collaboration between visual perception and caption generation modules and across the attention layers within the generator, enabling more accurate and contextually aligned captioning. The method is evaluated on the MSCOCO dataset using standard metrics, achieving 133.8 CIDEr, 82.4 BLEU-1, 67.4 BLEU-2, 83.8 BLEU-3, 74.0 BLEU-4, 29.5 METEOR, and 59.1 ROUGE, confirming the effectiveness of the dual-level collaborative design.

Redundancy in visual inputs is addressed by study [41], which introduces CropCap, a transformer-based approach with a Cross-Partition Dependency (CPD) module that refines visual feature interactions. By reducing visual information redundancy and optimizing dependencies between image regions, the model captures fine-grained spatial and temporal relationships more effectively. Evaluated on MSCOCO, the approach achieves BLEU-4 of 41.7 and CIDEr of 138.8, outperforming existing models. The results demonstrate that modeling cross-partition dependencies significantly improves caption relevance. CropCap nonetheless represents a theoretical and practical advance in image captioning, paving the way for more accurate and less redundant models.

Cross-lingual modeling is further expanded in study [42] which image captioning approach that generates coherent multilingual captions from a single image. The Embedded Heterogeneous Attention Transformer (EHAT) models global and local image–text correspondences across languages using heterogeneous attention mechanisms. Evaluated on the MSCOCO English–Chinese dataset with BLEU, METEOR, ROUGE, and CIDEr, EHAT outperforms strong monolingual baselines (BLEU-4/CIDEr: 40.1/133.5 in English and 32.9/111.6 in Chinese), demonstrating effective multilingual caption generation with semantic consistency.

Finally, The role of visual feature encoding is revisited in [43], which examines how simplified grid-based visual representations can improve image captioning compared to complex region-based features from object detectors. The authors introduce a FEature Interaction Module (FeiM), a transformer-based model with two innovations: learnable feature queries that probe global visual grids as local signals, and a feature interaction module that refines representations through spatial and channel-wise relations before linguistic integration. Tested on MSCOCO, FeiM achieves state-of-the-art results with CIDEr of 135.2, outperforming Meshed-Memory Transformer (M2 Transformer) by over 4%, alongside high BLEU-4 (40.5), METEOR (29.9), and SPICE (23.7) scores. The study demonstrates that grid features can rival and surpass region-based representations when paired with targeted architectural enhancements, indicating that image captioning systems can achieve top-tier performance with lighter, more transferable models, enabling broader and more efficient deployment across vision–language tasks.

G. Pre-Training-Based Models for Image Captioning

With the rapid progress of deep learning, pre-training strategies have emerged as a powerful paradigm for image captioning, enabling models to leverage large-scale vision-language data before task-specific fine-tuning. Unlike earlier approaches trained from scratch on limited caption datasets, pre-trained models learn general multimodal representations that can be efficiently adapted to downstream tasks, resulting in improved generalization and robustness. This paradigm shift has been largely driven by advances in large language models and Transformer-based architectures.

Among these, GPT-based models have played a central role in advancing caption generation. Initially introduced by Radford in [54], GPT models are autoregressive transformer decoders pre-trained on massive text corpora to capture rich syntactic and semantic language patterns. When adapted to image captioning, GPT-style decoders are typically combined with visual encoders (e.g., CNNs or ViT), allowing the model to generate captions conditioned on visual embeddings while benefiting from strong linguistic priors [5], [55].

Pre-training is often performed using large-scale image-text pairs through objectives such as masked language modeling, contrastive learning, or conditional text generation. Representative models, including ViLBERT [56], UNITER [57], OSCAR [55], and BLIP [58], demonstrate that jointly pre-training on multimodal data significantly enhances caption fluency, semantic alignment, and performance on benchmarks such as MSCOCO. These approaches effectively bridge the gap between visual perception and natural language generation, establishing pre-trained transformer and GPT-based models as a cornerstone of modern image captioning systems.

A prominent example of vision-language pre-training is CLIP, introduced in study [5]. Contrastive Language-Image Pre-training (CLIP) is a dual-encoder model composed of an image encoder and a text encoder jointly trained to project images and their corresponding captions into a shared latent semantic space. This unified representation enables CLIP to perform effectively across a wide range of computer vision and multimodal tasks, including image classification and text-driven image generation or editing. In parallel, other works adapt GPT, a powerful pre-trained language model that functions as both a decoder and a transformer, widely used in tasks such as text summarization and neural machine translation due to its superiority over static word embeddings. Experimental evaluations show that CLIP achieves strong zero-shot image classification performance, obtaining 98.4 on Yahoo, 76.2 on ImageNet, and 58.5 on SUN, highlighting its substantial improvement over previous zero-shot transfer approaches.

Adversarial pre-training ideas are revisited in study [8], which presents an approach to enhance image caption generation using Conditional Generative Adversarial Networks (cGANs). The cGAN model consists of a generator and a discriminator, enabling the generation of realistic and coherent captions that align with the visual content of the image. It is trained iteratively using a loss function that encourages the generator to produce realistic captions and the discriminator to correctly distinguish between real and generated captions. Once trained, the generator is utilized for crafting descriptions for new images, thereby improving the quality and relevance of the produced captions. The proposed approach is evaluated on the MS-COCO dataset using standard automatic evaluation metrics to measure caption quality. The experimental results demonstrate strong performance, with scores of 95.6 BLEU-1, 90.1 BLEU-2, 81.7 BLEU-3, 71.5 BLEU-4, 38.2 METEOR, 74.4 ROUGE-L, and 124.3 CIDEr, indicating the effectiveness of the model in generating high-quality image captions. The utilization of cGANs thus enables the generation of more realistic and coherent captions, offering a significant enhancement in image caption generation.

Shifting from static images to procedural reasoning, [14] proposes a language-first framework for procedure planning that leverages the reasoning capabilities of pre-trained language models over vision-based approaches. Rather than relying on visual latent representations, the method converts visual observations into textual descriptions and employs language models to generate intermediate procedural steps between initial and goal states. This approach exploits the abstraction and generalization strengths inherent in natural language processing. Evaluated on two large-scale instructional video datasets COIN and CrossTask, the framework achieves a success rate of 98.9%, representing a 19.2% improvement over existing state-of-the-art methods. The results indicate that language-based reasoning provides superior coherence and accuracy in step prediction compared to purely vision-centric models. While the approach demonstrates significant potential for structured task execution in AI systems, the authors acknowledge limitations with long-sequence planning and domain-specific scenarios, suggesting directions for future research toward real-time adaptability.

In the clinical domain, [15] proposes an automated radiology report generation system that integrates transformer-based architectures with contrast-based image enhancement to address radiologists workload and diagnostic complexity. The framework adopts an encoder–decoder architecture using CheXNet for visual feature extraction and BERT for textual encoding [1], while Multi-Head Attention (MHA) enables effective fusion of visual and semantic information. Four contrast enhancement techniques are applied during preprocessing to reduce noise and improve diagnostic feature recognition in chest X-ray images. Experimental results show that incorporating image enhancement leads to a 15% performance improvement over baseline models in report generation quality. These findings demonstrate that combining transformer-based multimodal modeling with contrast-enhanced preprocessing improves the relevance and informativeness of generated radiology reports, highlighting its potential to support clinical decision-making.

Knowledge-enriched pre-training is investigated in [16] to address the limitation of image captioning models that generate generic or inaccurate descriptions due to weak integration of real-world knowledge. The authors propose Knowledge-guided Replay (K-Replay), a framework that preserves and refines knowledge embedded in vision–language pre-trained models while reducing hallucinations and improving caption fidelity. K-Replay relies on a dual-objective learning strategy, combining a knowledge prediction task to maintain knowledge retention and a knowledge distillation constraint to ensure consistency with original pre-trained representations. To assess this approach, the authors introduce KnowCap, a benchmark encompassing multiple knowledge domains, including landmarks, brands, foods, and movie characters. Experimental evaluations show significant gains, with a +20.9 CIDEr improvement (from 78.7 to 99.6) and a +20.5% increase in knowledge recognition accuracy (from 34.0% to 54.5%). These results demonstrate that K-Replay produces more informative and faithful captions while effectively preserving pre-trained knowledge. Overall, this work establishes a scalable paradigm for knowledge-enhanced image captioning, with strong applicability to domains requiring precise and context-aware visual descriptions, such as accessibility systems, cultural heritage analysis, and content-based image retrieval.

Finally, Extending pre-training to structured documents, [59] addresses the challenge of jointly modeling intertwined text and visual information in documents, infographics, and user interfaces, where traditional methods often separate modalities or rely heavily on OCR. The authors propose Pix2Struct, a pre-training model that learns to reconstruct simplified HTML from web screenshots, encoding both textual and structural layout information. Built on a ViT encoder and text decoder, the model was trained on 80 million screenshots from the C4 corpus, producing Base (282M parameters) and Large (1.3B parameters) versions. The approach enables robust handling of diverse formats without domain-specific engineering. Evaluated across nine benchmarks covering illustrations, UI, natural images, and documents, Pix2Struct-Large outperformed prior visual models on 8 of 9 tasks, achieving new state-of-the-art results on six tasks including significant gains in Screen2Words (64.3 to 109.4 CIDEr) and Widget Captioning (127.4 to 136.7 CIDEr). While slightly behind specialized Optical Character Recognition (OCR) pipelines on some tasks, it demonstrated strong generalization and independence from external OCR. This work highlights screenshot parsing as a scalable pre-training paradigm, advancing flexible models for accessibility, document processing, and interface analysis.

Despite their strong performance, GPT-based and large-scale pre-trained captioning models face several critical limitations: 1) *Computational cost*: training and fine-tuning require substantial GPU resources; 2) *Bias propagation*: models may inherit and amplify biases present in large-scale pre-training corpora, leading to unfair or stereotypical descriptions; 3) *Hallucination risk*: models may generate plausible but factually incorrect descriptions.

IV. Discussion

This study provides a comprehensive and systematic overview of the evolution of image captioning methods, ranging from early traditional machine learning approaches to recent deep learning models based on large-scale pre-training and GPT. By adopting a PRISMA-based systematic review methodology, this work ensures a structured, transparent, and reproducible analysis of the literature, thereby offering a reliable snapshot of the current state of the art.

One of the main contributions of this article lies in its chronological and conceptual analysis of image captioning techniques. The review highlights the paradigm shift from handcrafted feature-based and rule-driven models toward deep learning architectures, which have demonstrated superior performance in capturing complex visual semantics and generating fluent natural language descriptions. Encoder–decoder frameworks established the foundation for this transition, while attention mechanisms further enhanced model interpretability and performance by enabling selective focus on salient image regions. Subsequently, reinforcement learning approaches addressed metric misalignment issues, and semantic-based models introduced higher-level reasoning capabilities. More recently, Transformer-based architectures and pre-training strategies have emerged as dominant paradigms, significantly improving generalization and caption quality across diverse datasets.

Another important contribution of this work is the systematic comparison of representative approaches across multiple dimensions, including architectural categories, core techniques, evaluation datasets, and performance metrics. The comprehensive summary presented in Table I synthesizes these elements and provides a unified view of how different methods relate to one another. This table not only facilitates an objective comparison of existing approaches but also highlights common evaluation practices, such as the widespread use of MSCOCO data base and metrics like BLEU, METEOR, CIDEr, ROUGE, and SPICE.

Moreover, this review emphasizes the diversity of datasets and evaluation protocols employed in the literature, revealing challenges related to fair comparison and reproducibility. The inclusion of human feedback, semantic reasoning, and multimodal pre-training in recent works suggests a clear trend toward more human-aligned.

Overall, this article contributes added value by combining a rigorous systematic review methodology with a structured synthesis of image captioning approaches. It provides researchers and practitioners with a clear and up-to-date understanding of methodological trends, strengths, and limitations, while also identifying promising directions for future research in multimodal learning and vision–language modeling.

V. Conclusion and Perspectives

This review synthesized the evolution of image captioning methodologies, from early machine learning approaches to state-of-the-art deep learning architectures, including encoder–decoder models, attention mechanisms, reinforcement learning strategies, semantic-aware frameworks, Transformer-based models, and large-scale pre-trained systems. These developments have led to substantial improvements in caption accuracy, semantic coherence, and multimodal representation learning, underscoring the rapid progress of vision–language models.

Nevertheless, important challenges persist, particularly in describing complex visual scenes, capturing fine-grained inter-object relationships, generalizing to domain-specific or out-of-distribution data, and reducing the high computational cost of large pre-trained models. These limitations hinder both robustness and practical deployment in constrained environments.

Future research directions emphasize the growing role of prompt engineering and task-conditioned learning to enhance visual–linguistic alignment and adaptability with minimal supervision. When combined with vision–language pre-training, prompt-based approaches show strong potential for improving zero-shot and few-shot generalization, especially in specialized domains such as medical imaging, autonomous systems, and assistive technologies.

Furthermore, the exploration of hybrid multimodal architectures that integrate attention mechanisms, reasoning capa-

TABLE I. SUMMARY OF IMAGE CAPTIONING METHODS, DATASETS, AND EVALUATION METRICS

	Approaches cited in	[6]	[17]	[19]	[22]	[9]	[27]	[11]	[10]	[34]	[12]	[7]	[37]	[3]	[4]	[41]	[42]	[43]	[5]	[8]	[16]	[15]	[14]
Categories	Encoder–Decoder	×	×	×																			
	Attention				×	×	×																
	Reinforcement Learning							×	×	×													
	Semantic Models										×	×	×										
	Transformer													×	×	×	×	×					
	Pre-training / GPT																		×	×	×	×	×
Techniques	CNN			×														×					
	LSTM/BiLSTM	×	×		×		×																
	ResNet		×	×																			
	YOLOv4						×																
	GCN/GNN												×										
	Attention Mech.				×	×	×					×	×				×					×	
	Transformer											×	×	×	×	×	×	×				×	×
	GAN																			×			
	Policy Gradient					×		×	×	×													
	Human Feedback							×	×														
	Scene Graphs												×										
	CLIP																		×				
	BERT																					×	
	CheXNet																					×	
	Knowledge Distillation																				×		
	Feature Queries																	×					
	LLM/GPT																						×
Dataset	MS-COCO	×		×	×	×				×	×	×		×	×	×	×	×		×			
	Flickr8k	×	×																				
	Flickr30k			×																			
	ImageNet									×									×				
	PEIR Gross						×																
	Objects365									×													
	COCO Captions									×													
	Caption Ratings								×														
	UT-Interaction												×										
	AVA												×										
	NTU RGB+D												×										
	Yahoo																		×				
	SUN																		×				
	COIN																						×
	CrossTask																						×
	KnowCap																				×		
	Chest X-ray																					×	
Metrics	BLEU	×	×	×	×	×	×	×			×	×		×	×		×	×		×		×	
	METEOR			×		×		×			×	×		×	×	×	×	×		×			
	CIDEr	×		×		×		×		×	×			×	×	×	×	×		×	×		
	ROUGE			×		×					×	×		×	×	×	×			×			
	SPICE							×			×			×				×		×			
	BLEURT												×										
	Accuracy																		×		×		
	mAP									×													
	PQ									×													
	Success Rate																						×

bilities, and external knowledge sources may further advance the semantic understanding and interpretability of generated captions. Overall, image captioning is expected to continue evolving through innovations in multimodal learning, offering broad scientific, industrial, and societal impact.

REFERENCES

[1] Saouabe, Abdelkrim, Hicham Oualla, and Imad Mourtaji. ”Data Encoding with Generative AI: Towards Improved Machine Learning Performance.” International Journal of Advanced Computer Science and Applications 15.10 (2024).

[2] Saouabe, Abdelkrim, et al. ”Large-Scale Image Indexing and Retrieval Methods: A PRISMA-Based Review.” International Journal of Advanced Computer Science and Applications 15.7 (2024).

[3] Herdade, Simão et al. “Image Captioning: Transforming Objects into Words.” Neural Information Processing Systems (2019).

[4] Luo, Yunpeng et al. “Dual-Level Collaborative Transformer for Image Captioning.” ArXiv abs/2101.06462 (2021): n. pag.

[5] Radford, Alec et al. “Learning Transferable Visual Models From Natural Language Supervision.” International Conference on Machine Learning (2021).

[6] Ravulaplli, Lakshmi Tulasi. ”A Novel Bi-LSTM Based Automatic Image Description Generation.” Ingenierie des Systemes d’Information 28.2 (2023).

[7] Hafeth, D. A., Kollias, S., and Ghafoor, M., “Semantic representations with attention networks for boosting image captioning,” *IEEE Access*, vol. 11, pp. 40230–40239, 2023.

[8] Chen, Chen, et al. ”Improving image captioning with conditional generative adversarial nets.” Proceedings of the AAAI conference on artificial intelligence. Vol. 33. No. 01. 2019.

[9] Yan, Shiyang, et al. ”Image captioning via hierarchical attention mechanism and policy gradient optimization.” Signal Processing 167 (2020): 107329.

[10] Seo, Paul Hongsuck, et al. ”Reinforcing an image caption generator using off-line human feedback.” Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 34. No. 03. 2020.

[11] Shi, Xiangxi, et al. ”Finding it at another side: A viewpoint-adapted matching encoder for change captioning.” European conference on computer vision. Cham: Springer International Publishing, 2020.

[12] Yao, Ting, et al. ”Hierarchy parsing for image captioning.” Proceedings of the IEEE/CVF international conference on computer vision. 2019.

[13] Guo, Longteng, et al. ”Fast sequence generation with multi-agent reinforcement learning.” arXiv preprint arXiv:2101.09698 (2021).

[14] Liu, Jiateng, et al. ”A language-first approach for procedure planning.” Findings of the Association for Computational Linguistics: ACL 2023. 2023.

[15] Tsaniya, Hilya, Chastine Fatichah, and Nanik Suciati. ”Automatic radiology report generator using transformer with contrast-based image enhancement.” IEEE Access 12 (2024): 25429-25442.

[16] Cheng, Kanzhi, et al. ”Beyond generic: Enhancing image captioning with real-world knowledge using vision-language pre-training model.” Proceedings of the 31st ACM International Conference on Multimedia. 2023.

[17] Dixit, B., et al. ”Challenges and a novel approach for image captioning using neural network and searching techniques.” International Journal of Intelligent Systems and Applications in Engineering 11.3 (2023): 712-720.

[18] Zhou, Haonan, et al. ”FRIC: A framework for few-shot remote sensing image captioning.” International Journal of Digital Earth 17.1 (2024): 2337240.

[19] Bartosiewicz, Mateusz, et al. ”On combining image features and word embeddings for image captioning.” 2023 18th Conference on Computer Science and Intelligence Systems (FedCSIS). IEEE, 2023.

[20] Sharma, Dhruv, Chhavi Dhiman, and Dinesh Kumar. ”Evolution of visual data captioning Methods, Datasets, and evaluation Metrics: A comprehensive survey.” Expert Systems with Applications 221 (2023): 119773.

[21] Gómez Martínez, Mario. ”Deep learning for image captioning: an encoder-decoder architecture with soft attention.” (2019).

[22] Ramos, Leo, et al. ”A study of convnext architectures for enhanced image captioning.” IEEE Access 12 (2024): 13711-13728.

[23] Ming, Yue, et al. ”Visuals to text: A comprehensive review on automatic image captioning.” IEEE/CAA Journal of Automatica Sinica 9.8 (2022).

[24] Anwar, Qazi, and Ch VS Satyamurty. ”An Analysis on Recent Approaches for Image Captioning.” CVR Journal of Science and Technology 26.1 (2024): 87-92.

[25] Arasi, Munya A., et al. ”Automated image captioning using sparrow search algorithm with improved deep learning model.” IEEE Access 11 (2023): 104633-104642.

[26] Saouabe, Abdelkrim, et al. ”Image Indexing Approaches for Enhanced Content-Based Image Retrieval: An Overview.” 2024 International Conference on Ubiquitous Networking (UNet). Vol. 10. IEEE, 2024.

[27] Ravinder, Paspula, and Saravanan Srinivasan. ”Automated medical image captioning with soft attention-based LSTM model utilizing YOLOv4 algorithm.” Journal of Computer Science 20.1 (2024): 52-68.

[28] G. Kalaiarasi, M. Sravya Sree, B. Sai Geetha, A. Yasaswi, I. Aravind, G. Nagavenkata Sreeja, Year: 2025, Image Captioning: Enhance Visual Understanding, ICITSM PART I, EAI, DOI: 10.4108/eai.28-4-2025.2357851

[29] Saouabe, Abdelkrim, et al. ”Evolution of Image Captioning Models: An Overview.” 2023 10th International Conference on Wireless Networks and Mobile Communications (WINCOM). IEEE, 2023.

[30] Bai, Shuang, and Shan An. ”A survey on automatic image caption generation.” Neurocomputing 311 (2018): 291-304.

[31] Sheng, Shou-Jun, and Zi-Wei Zhou. ”Revolutionizing Image Captioning: Integrating Attention Mechanisms with Adaptive Fusion Gates.” IAENG International Journal of Computer Science 51.3 (2024).

[32] Hu, Wenjin, et al. ”Thangka image captioning based on semantic concept prompt and multimodal feature optimization.” Journal of Imaging 9.8 (2023): 162.

[33] Zheng, Ervine, and Qi Yu. ”Evidential interactive learning for medical image captioning.” International Conference on Machine Learning. PMLR, 2023.

[34] Pinto, André Susano, et al. ”Tuning computer vision models with task rewards.” International Conference on Machine Learning. PMLR, 2023.

[35] Brack, Manuel, et al. ”Illume: Rationalizing vision-language models through human interactions.” International Conference on Machine Learning. PMLR, 2023.

[36] Fei, Zhengcong, and Junshi Huang. ”Incorporating Unlikely Negative Cues for Distinctive Image Captioning.” IJCAI. 2023.

[37] Alper, Morris, and Hadar Averbuch-Elor. ”Learning human-human interactions in images from weak textual supervision.” Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023.

[38] Zhao, Wentian, and Xinxiao Wu. ”Boosting entity-aware image captioning with multi-modal knowledge graph.” IEEE Transactions on Multimedia 26 (2023): 2659-2670.

[39] Yang, Xu, et al. ”Transforming visual scene graphs to image captions.” arXiv preprint arXiv:2305.02177 (2023).

[40] Qi, Yayun, Wentian Zhao, and Xinxiao Wu. ”Relational distant supervision for image captioning without image-text pairs.” Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 38. No. 5. 2024.

[41] Wang, Bo, et al. ”CropCap: Embedding visual cross-partition dependency for image captioning.” Proceedings of the 31st ACM International Conference on Multimedia. 2023.

[42] Song, Zijie, et al. ”Embedded heterogeneous attention transformer for cross-lingual image captioning.” IEEE Transactions on Multimedia 26 (2024): 9008-9020.

[43] Yan, Jie, et al. ”Exploring better image captioning with grid features.” Complex & Intelligent Systems 10.3 (2024): 3541-3556.

[44] Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. ”Neural machine translation by jointly learning to align and translate.” arXiv preprint arXiv:1409.0473 (2014).

[45] Xu, Kelvin, et al. ”Show, attend and tell: Neural image caption generation with visual attention.” International conference on machine learning. PMLR, 2015.

[46] Lu, Jiasen, et al. ”Knowing when to look: Adaptive attention via a visual sentinel for image captioning.” Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.

[47] Yang, Zichao, et al. ”Hierarchical attention networks for document classification.” Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies. 2016.

[48] You, Quanzeng, et al. ”Image captioning with semantic attention.” Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[49] Vaswani, Ashish, et al. ”Attention is all you need [J].” Advances in neural information processing systems 30.1 (2017): 261-272.

[50] Cornia, Marcella, et al. ”Meshed-memory transformer for image captioning.” Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020.

[51] Herdade, Simao, et al. ”Image captioning: Transforming objects into words.” Advances in neural information processing systems 32 (2019).

[52] Dosovitskiy, Alexey. ”An image is worth 16x16 words: Transformers for image recognition at scale.” arXiv preprint arXiv:2010.11929 (2020).

[53] Devlin, Jacob, et al. ”Bert: Pre-training of deep bidirectional transformers for language understanding.” Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers). 2019.

[54] Radford, Alec, et al. ”Improving language understanding by generative pre-training.” (2018): 3.

[55] Li, Xiujun, et al. ”Oscar: Object-semantics aligned pre-training for vision-language tasks.” European conference on computer vision. Cham: Springer International Publishing, 2020.

[56] Lu, Jiasen, et al. ”Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks.” Advances in neural information processing systems 32 (2019).

[57] Chen, Yen-Chun, et al. ”Uniter: Universal image-text representation learning.” European conference on computer vision. Cham: Springer International Publishing, 2020.

[58] Li, Junnan, et al. ”Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation.” International conference on machine learning. PMLR, 2022.

[59] Lee, Kenton, et al. ”Pix2struct: Screenshot parsing as pretraining for visual language understanding.” International Conference on Machine Learning. PMLR, 2023.