Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control
Journal homepage: http://kinetik.umm.ac.id
ISSN: 2503-2267
Vol. 10, No. 4, November, Pp. 475-486

The evolution of image captioning models: trends, techniques, and future challenges

Ade Bastian[1], Abrar Wahid*[1], Zacky Hafsari[1], Ardi Mardiana[1]
Departement of Informatics, Majalengka University, Majalengka, Indonesia[1]

Article Info
Keywords:
Computational Efficiency, Image Captioning, Knowledge Integration, Systematic Literature Review, Vision-Language Models

Article history:
Received: April 15, 2025
Accepted: July 25, 2025
Published: November 01, 2025

Cite:
A. Bastian, A. Wahid, Z. Hafsari, and A. Mardiana, "The Evolution of Image Captioning Models: Trends, Techniques, and Future Challenges", *KINETIK*, vol. 10, no. 4, Nov. 2025.
https://doi.org/10.22219/kinetik.v10i4.2305

*Corresponding author.
Abrar Wahid
E-mail address:
221410088@unma.ac.id

Abstract
This study provides a comprehensive systematic literature review (SLR) of the evolution of image captioning models from 2017 to 2025, with a particular emphasis on the impending problems, methodological enhancements, and significant architectural developments. The evaluation is guided by the increasing demand for precise and contextually aware image descriptions, and it adheres to the PRISMA methodology. It selects 36 relevant papers from reputable scientific databases. The results indicate a significant transition from traditional CNN-RNN models to Transformer-based architectures, which leads to enhanced semantic coherence and contextual comprehension. Current methodologies, such as prompt engineering and GAN-based augmentation, have further facilitated generalization and diversity, while multimodal fusion solutions, which incorporate attention mechanisms and knowledge integration, have improved caption quality. Additionally, significant areas of concern include data bias, equity in model assessment, and support for low-resource languages. The study underscores the fact that modern vision-language models, such as Flamingo, GIT, and LLaVA, offer robust domain generalization through cross-modal learning and joint embedding. Furthermore, the efficacy of computing in restricted environments is improved by the development of pretraining procedures and lightweight models. This study contributes by identifying future prospects, analyzing technical trade-offs, and delineating research trends, particularly in sectors such as healthcare, construction, and inclusive AI. According to the results, in order to optimize their efficacy in real-world applications, future picture captioning models must prioritize resource efficiency, impartiality, and multilingual capabilities.

1. Introduction

Image captioning is a multidisciplinary pursuit that integrates computer vision and natural language processing (NLP), both of which have significantly advanced due to the development of deep learning and multimodal learning technologies. Research in this domain is proliferating, driven by its extensive applicability in assistive technology, content indexing, and human-computer interaction, employing large datasets and sophisticated methodologies [1]. Preliminary advancements, particularly through encoder-decoder architectures employing Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) such as CNN-LSTM, have significantly improved the extraction of visual data and the generation of descriptive text [2]. The emergence of Vision-Language Models (VLMs), integrating Transformers with language models such as GPT and BERT, has markedly enhanced the production of nuanced and contextually accurate image descriptions [3].

Recent breakthroughs include knowledge embedding for specialised tasks such as medical imaging [4], [5], [6], the application of Text Graph Convolutional Networks (TextGCN) for semantic interpretation [7], and CLIP-based region-aware methodologies for multimodal comprehension [8]. Diffusion models such as MADiffCC [9], patch-based and multi-label classifiers [10], and YOLO V5-based systems [11] have enhanced robustness and accessibility. Moreover, modern frameworks address challenges in noisy environments through the use of data augmentation, dual network designs, and consistency losses [12]. Attention-based methodologies, such Relation-Aware Selection (RAS) and Fine-grained Semantic Guidance (FSG) [13], improve model precision in complex visual tasks.

Multimodal approaches have enabled applications in visual object tracking [14], fMRI-based image reconstruction [15], and zero-shot image captioning [10], despite persistent challenges such as hallucination and phrase length [16]. Research has advanced to include low-resource languages like Bengali [17], [18], while in the medical domain, Cross-modal Augmented Transformer (CAT) models improve contextual captioning [19]. The production of effective captions remains challenging, resulting in the development of lightweight models like SCAP [20] and scene-aware frameworks such as SSE [21].

Innovative frameworks such as CapFlow [22], concept modelling [23], [24], [25], and Vision Transformer-based systems like VisualSiteDiary [26], highlight the dynamic advancement of the field. However, challenges persist in data annotation, computational costs, and the administration of domain-specific terms [27], [28], [19]. Techniques like VisualGPT and SATIC offer partial solutions via self-attention and parallel generation [27], [28]. Applications extend beyond captioning, impacting domains such as fraud detection [29], medical diagnosis [30], annotation-free captioning [31], radiological image retrieval [32], and historical image interpretation [33]. The scope now includes mathematics [34], multimodal retrieval [35], remote sensing [36], and GNN-assisted image-text comprehension [37], along with culturally significant domains such as Thangka image captioning [38] and beam search optimization [39]. Image captioning is crucial for the automation of radiography and the recording of construction [40].

Notwithstanding swift advancements, several fundamental issues persist and have grown increasingly intricate. Complex models frequently encounter obstacles due to insufficient labelled data, particularly in specialized fields like medicine, which constrains the model's capacity to generate precise descriptions. The substantial computational expenses associated with training intricate models provide a significant obstacle to broader adoption; however, methodologies like Visualgpt and Satic endeavour to provide answers via an efficient attention mechanism. These technical issues emerge within a broader context characterized by rapid and fragmented development. The proliferation of diverse architectures, including transformers and VLM, alongside specialized methodologies for distinct applications, has generated a complex research environment. Consequently, researchers frequently encounter challenges in identifying prevailing patterns, assessing trade-offs among methodologies, and ascertaining the most critical future trajectory. This scenario establishes a distinct research gap, specifically the absence of a comprehensive evaluation of the most recent systematic literature that synthesizes and delineates this progression.

This paper provides a thorough Systematic Literature Review (SLR) that examines the progression of image captioning models from 2017 to 2025. This review used the PRISMA approach to synthesize the findings of 36 pertinent papers that have undergone a peer review procedure. This paper delineates the evolution of the architectural paradigm, transitioning from the traditional CNN-RNN model to the foundational architecture of contemporary transformers and vision-language, while evaluating their comparative efficacy in generating coherent, accurate, and comprehensive descriptions. This study emphasizes the incorporation of advanced semantic enrichment methods, such as rapid engineering, knowledge graph embedding, and cross-modal attention. The latest models, including Flamingo, Git, and Llava, are highlighted for their prospective solutions via multimodal embedding and cooperative reciprocal learning. This review consolidates the existing research landscape, delineating the strengths and weaknesses of diverse methodologies while emphasizing unresolved challenges pertaining to data bias, multilingual generalization, and computational efficiency, thereby establishing a framework for future advancements in automatic image captioning.

The following portions of this document are structured as outlined below. Section 2 outlines the research process and the selection criteria. Section 3 delineates findings and thorough discussions obtained from ten research enquiries. Section 4 concludes the analysis and outlines potential research directions.

2. Research Method

The methodological strategy adopted in conducting a comprehensive literature review on the development of image text models is described in this section. We adhere to the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) principles to ensure a thorough and open review procedure [41]. To help authors report systematic reviews and meta-analyses with high transparency and completeness, PRISMA is an evidence-based guideline [41]. The identification, screening, eligibility evaluation, and final inclusion phases of the article selection process are all outlined in the PRISMA flowchart.

2.1 Search Strategy and Data Source

To ensure a comprehensive and representative dataset, keyword queries were carefully constructed using domain-specific terminology. The primary keyword was “Image Captioning,” complemented by related terms such as “Transformer,” “Multimodal,” “Vision-Language,” “GAN,” “Prompt Engineering,” and “Fairness.” These terms were selected based on prior studies and recurring concepts across multiple publications. Figure 1 shows the frequency of these keywords as they appeared during the literature search process, highlighting the dominance of multimodal learning and Transformer-based models in recent research.

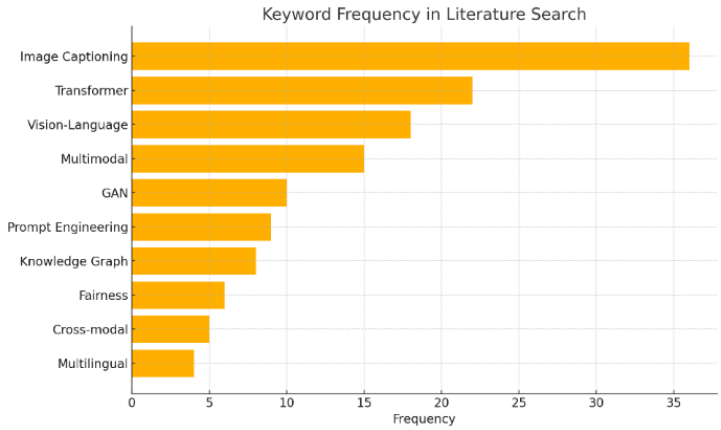

Figure 1. Frequency of Keywords used in Literature Search Across Major Databases (2017–2025)

The literature was collected from three major scientific databases: IEEE Xplore, ScienceDirect, and Scopus. Each database was queried using the same keyword combinations across the same publication range (2017–2025). The number of articles obtained from each source varied slightly depending on indexing coverage and access restrictions. Figure 2 presents the distribution of selected articles across the three databases.

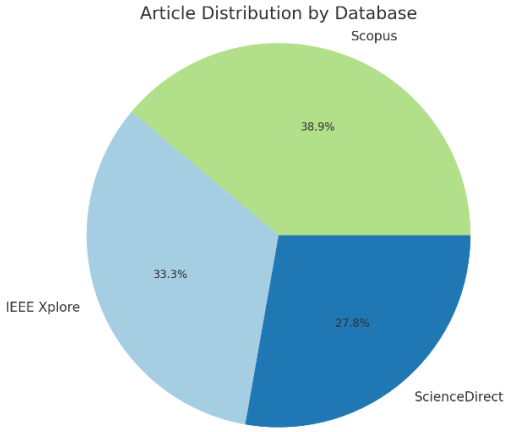

Figure 2. Distribution of Selected Articles by Database Source

2.2 Formulate Research Questions

To guide the analysis, ten research questions (RQ1–RQ10) were formulated based on recurring themes and gaps found in preliminary reading. The mapping of these questions is visualized in Figure 3.

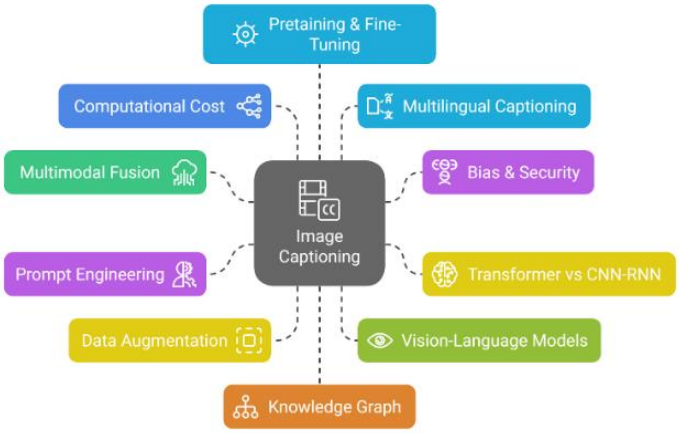

Figure 3. Mind Mapping Research Questions

The research questions are organized as follows:

- RQ1: How do current multimodal fusion techniques improve the accuracy of image captioning models compared to unimodal approaches?
- RQ2: How do image captioning models address data bias and security issues, and to what extent are fairness metrics applied in model evaluation?
- RQ3: How does the Transformer-based model perform compared to a CNN-RNN architecture in generating more accurate and contextual image descriptions?
- RQ4: How can prompt engineering strategies improve the quality of descriptions produced by vision-language-based image captioning models such as CLIP and BLIP?
- RQ5: How can the integration of knowledge graph with image captioning models enrich semantic information in image descriptions?
- RQ6: To what extent are current multilingual image captioning techniques able to accommodate linguistic and cultural variations in image description?
- RQ7: How do current image captioning models balance the trade-off between caption accuracy and computational efficiency in resource-constrained environments?
- RQ8: How does the development of vision-language foundation models such as Flamingo, GIT, and LLaVA affect accuracy and generalization in image captioning tasks?
- RQ9: How effective are newer data augmentation techniques, such as GAN-based augmentation or synthetic data generation, in improving the quality of image captioning models?
- RQ10: How do novel pretraining and fine-tuning strategies in image captioning models affect the model's generalization ability across different domains?

The initial research question (RQ1) is to examine whether image captioning models achieve enhanced performance through the integration of multimodal information, specifically the combination of textual and visual inputs. This topic aims to assess the efficacy of multimodal and unimodal strategies in generating accurate visual descriptions. The second research question (RQ2) is motivated by the necessity to understand the technical and ethical dimensions of image captioning, specifically about how models address biased input while ensuring security and fairness. This includes the analysis of fairness standards in model evaluation. The objective of RQ3 is to evaluate the performance of Transformer-based models in comparison to traditional CNN-RNN architectures. This investigation seeks to determine the model architecture that generates captions with greater contextual and semantic richness. RQ4 examines how prompt engineering, especially with advanced vision-language models such as CLIP and BLIP, may enhance the expressiveness and relevance of image captions. This study examines the growing significance of human-in-the-loop strategies in improving output quality. The objective of RQ5 is to investigate whether the integration of external semantic knowledge, such as knowledge graphs, enhances the informativeness and depth of image captions produced by AI models.

The objective of RQ6 is to assess the multilingual picture captioning model's robustness in addressing diverse linguistic and cultural contexts. This involves evaluating its capacity to adapt and preserve semantic nuances amid linguistic variations. The objective of RQ7 is to investigate how image captioning models reconcile high accuracy with computational efficiency, particularly in resource-constrained environments such as embedded or mobile systems. RQ8 aims to investigate the impact of novel vision-language foundation models such as Flamingo, GIT, and LLaVA on generalization and overall efficacy in image captioning tasks across diverse datasets. RQ9 aims to determine if advanced data augmentation techniques, including synthetic data and GAN-generated samples, can enhance the accuracy and robustness of image captioning models. Finally, RQ10 examines how different pretraining and fine-tuning methodologies improve a model's ability to generalize to unfamiliar tasks or domains, hence increasing transferability and adaptability in real-world applications.

2.3 Data Eligibility and Analysis of the Literature

Figure 4 displays the PRISMA flowchart [41], which delineates the systematic procedure for article selection, including the phases of searching, inclusion, and exclusion. The flowchart comprises three primary phases: identification, screening, and inclusion. During the identification step, an extensive literature review was conducted across prominent academic databases such as IEEE Xplore, ScienceDirect, and Scopus, encompassing the publishing period from 2017 to 2025. The year 2017 was selected as the first reference to document the progression and recent developments in the domain of Image Captioning during the past eight years. The principal search phrase utilized in the identifying procedure was “Image Captioning”.

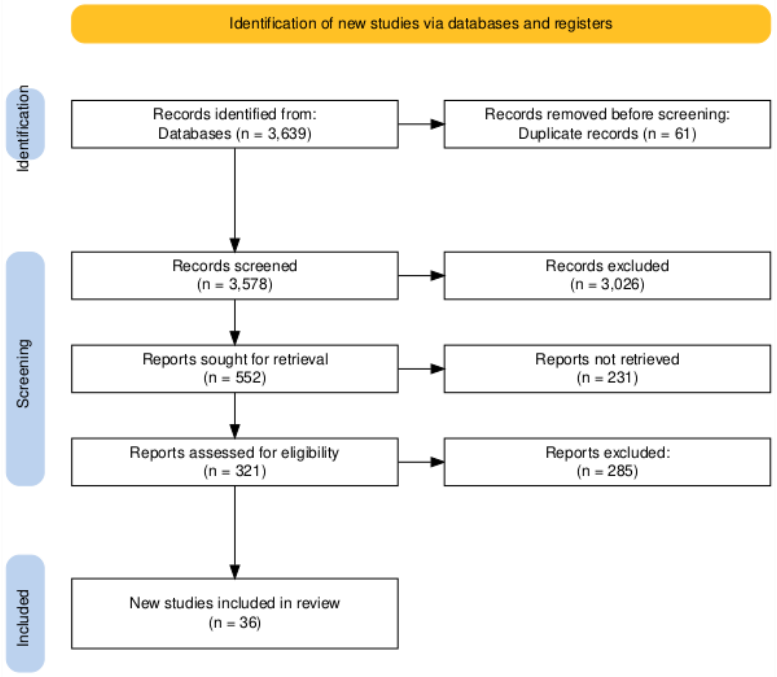

Figure 4. Study Selection Using the PRISMA Flow Diagram Method, Consisting of the Identification, Screening, and Inclusion Steps

The systematic literature descriptive collection process in this study follows the PRISMA 2020 (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) Guidelines, which provide a transparent framework for selecting and reporting study results. PRISMA helps researchers organize each stage from introduction to final inclusion of relevant studies [41]. The first stage is identification, where 3,639 articles were retrieved from various scientific databases. Of these, 61 articles were removed because they were duplicates. This step is important to avoid double counting and maintain the integrity of data selection [42]. The next stage is screening, which begins by screening 3,578 articles based on title and abstract. At this stage, 3,026 articles were removed because they did not meet the initial inclusion criteria, such as topic relevance, publication type, or did not meet the image captioning study domain. After that, 552 articles were reviewed for full-text report retrieval. However, 231 of them could not be obtained, either due to limited access or incomplete documents.

Next, the eligibility stage was carried out on 321 complete articles that were successfully obtained. This evaluation was based on stricter criteria such as methodological contribution, focus on image captioning models, and the existence of experiments and results that could be evaluated. A total of 285 articles were then excluded because they did not meet these criteria. In the final stage, namely inclusion, a total of 36 articles that met all the criteria were selected for further analysis in systematic insight. This final number is the result of a strict and transparent selection process in accordance with the PRISMA reporting principles, which are widely considered the gold standard in systematic literature review [43].

3. Results and Discussion

The main conclusions of the systematic literature review are presented in this section, which also provides a thorough synopsis of image caption models and debates that support the study's research topics. The discussion is divided into two subsections: the first discusses the development, advantages, and disadvantages of the various model groups found in the 36 chosen studies; the second answers each research question by combining knowledge from the examined literature.

3.1 Summary of Image Captioning Models

The evolution of image captioning models in the past decade indicates a transition from conventional CNN-RNN encoder-decoder frameworks to Transformer-based architectures and Vision-Language models (VLMs). Initial models like CNN-LSTM were proficient at fundamental visual-to-text conversions but exhibited constraints in semantic depth and adaptability when confronted with intricate sceneries or specialized images [44], [45]. These models generally attain BLEU-4 scores under 30 on benchmarks like MSCOCO, indicating their limited contextual capability. Table 1 provides a comparative overview of the primary categories of image captioning models examined in this work, emphasizing their structural characteristics, benchmark performance, and trade-offs.

Table 1. Comparative Summary of Image Captioning Model Categories

Model Category	Example Models	Architecture Highlights	Key Metrics (MSCOCO)	Strengths	Limitations
CNN-RNN (Early)	CNN-LSTM [44]	Encoder-Decoder, Sequential Text Gen	BLEU-4: ~28	Simple, Fast Training	Weak in context, limited semantics
Attention-based	Show-Attend-Tell [46]	Visual Attention + RNN	BLEU-4: ~31, METEOR: ~25	Focus on salient regions	Still sequential, limited long dependencies
Hybrid / Knowledge	TSFNet, VisualGPT [47], [27]	Multistream, Knowledge Infusion	CIDEr: ~113	Rich semantics, prompt-aware	Complex structure
Reinforcement	RL-Rank [48], SCST [49]	Reward-based Fine-tuning	CIDEr: 106+, BLEU-4: ~33	Optimized for evaluation metrics	Training instability
Cross-lingual	CALM, CLIDCap [50]	Multilingual Transformer, Attention Fusion	BLEU-4: ~30+ (multi-lang)	Language adaptability	Domain vocabulary mismatch
Lightweight	SCAP [20]	Efficient Attention	BLEU-4: ~29	Low computation	Trade-off with semantic depth
Transformer	X-Transformer [51], TSFNet [47]	Self-Attention, Parallel Tokens	BLEU-4: 35+, CIDEr: 110+	Strong context modeling	High training cost

The implementation of attention methods enhanced model concentration on prominent visual areas, resulting in greater descriptive detail, while simultaneously increasing architectural complexity [51], [46]. Transformer-based systems utilizing Embedded Heterogeneous Attention or Cross on Cross Attention [52], [51] have markedly enhanced long-range dependency modelling and semantic alignment. These models attain superior performance, frequently surpassing BLEU-4 scores of 35 and CIDEr scores exceeding 110 on regular datasets, establishing them as state-of-the-art for general-purpose captioning [53].

Hybrid models enhance this potential by including semantic knowledge, hierarchical encoding, or retrieval-based assistance. The Triple-Stream Fusion Network (TSFNet) integrates vision, text, and external knowledge to provide more precise and contextually aware captions [47]. Concurrently, models based on reinforcement learning enhance task-specific measures such as CIDEr but encounter difficulties in training stability [48], [49].

Recent advancements emphasize cross-domain generalization and multilingual adaption. Cross-lingual models utilizing consensus-aware learning and heterogeneous attention [50] demonstrate promising outcomes across many linguistic contexts, whilst lightweight designs like SCAP [20] seek to minimize computational demands without substantially sacrificing accuracy.

The progression of model development illustrates a compromise between performance improvements and computational intricacy. Transformer-based and knowledge-enhanced models consistently surpass previous CNN-RNN methodologies, particularly in terms of accuracy and contextual fluency. Nonetheless, practical implementation is limited by computing requirements, necessitating additional investigation into efficiency-focused captioning methods.

3.2 Discussion Based on Research Questions

This subsection addresses the research questions provided in this study by utilizing the conclusions derived from the reviewed literature. The data from the 36 selected papers facilitates a deeper understanding of the evolution, trends, and persistent challenges in image captioning research, addressing each inquiry. The responses aim to highlight both the achievements and the aspects that require additional investigation.

RQ1: How do current multimodal fusion techniques improve the accuracy of image captioning models compared to unimodal approaches?

Multimodal fusion approaches enhance caption accuracy by allowing the model to grasp more nuanced contextual meanings through several input modalities. The TSFNet approach amalgamates visual features, hierarchical scene comprehension, and semantic attention over three simultaneous streams, facilitating more thorough caption production [47]. This architecture yields a significant enhancement in CIDEr and BLEU scores relative to unimodal CNN-

RNN models [44], which generally analyze solely visual variables without accounting for semantic depth. The competitive advantage resides in TSFNet's capacity to synchronize image regions with textual semantics, hence minimizing generic or misaligned captions. Figure 5 explains that to create a truly good image description, an AI model cannot simply “see” the image (unimodal), but must also “understand” the meanings, concepts, and relationships within it (multimodal).

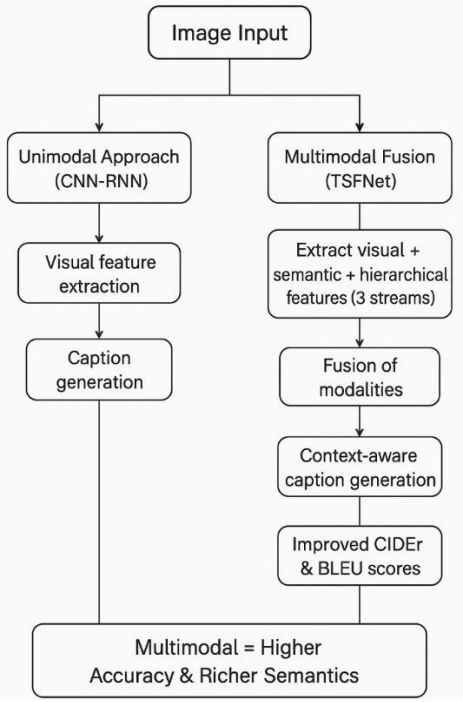

Figure 5. Workflow Comparison between CNN-RNN and Transformer Architectures for Image Captioning Tasks

RQ2: How do image captioning models address data bias and security issues, and to what extent are fairness metrics applied in model evaluation?

Many models prioritize enhancing accuracy, neglecting the examination of fairness and bias. Bengali captioning models utilizing EfficientNetV2S and Inception ResNetV2 address low-resource language bias by employing domain-specific datasets [17], [18]. Nonetheless, these initiatives are devoid of standardized fairness indicators or assessments for demographic representativeness. Current models seldom explicitly address security problems, including adversary image disturbances and hallucination threats. The emphasis persists on metric optimization (e.g., BLEU, METEOR), highlighting a distinct research deficiency in fairness-aware image captioning.

RQ3: How does the Transformer-based model perform compared to a CNN-RNN architecture in generating more accurate and contextual image descriptions?

Transformer-based models exhibit enhanced efficacy in semantic alignment and contextual fluency. Models employing Cross on Cross Attention [51] or Embedded Heterogeneous Attention [52] consistently surpass CNN-LSTM models [44], [45] on conventional benchmarks. Transformer models can simultaneously focus on numerous visual regions and preserve long-range dependencies in text production, facilitating more cohesive and informative captions. Conversely, CNN-RNN systems frequently generate more simplistic and less informative captions owing to their sequential nature and restricted context modelling. The performance disparity is apparent in BLEU-4 score enhancements exceeding 5–7 points for Transformer-based models. Figure 6 is a comparative illustration showing the evolution and superiority of AI technologies for image understanding, where Transformer is the superior approach.

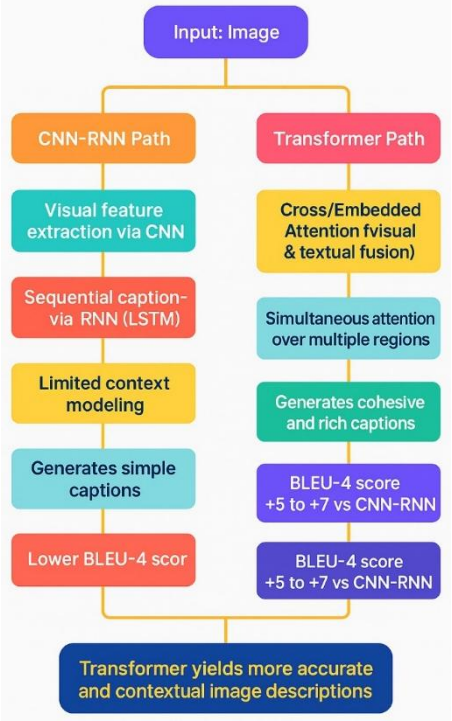

Figure 6. Architectural and Performance Comparison between CNN-RNN and Transformer Paths

RQ4: How can prompt engineering strategies improve the quality of descriptions produced by vision-language-based image captioning models such as CLIP and BLIP?

Prompt Engineering markedly enhances the quality of descriptions by offering structured directives to the language-vision model. This strategy is executed using several methodologies in the studied literature. For instance, Knowledge Distillation Prompting [53] use distilled knowledge to enhance the model's comprehension, hence strengthening the recognition of linkages between things. Cascade Semantic Prompt Alignment employs a hierarchical prompt structure to comprehend multi-level contexts, proving highly effective in deconstructing intricate visual sceneries. These methods, functioning as advanced promotions, have been demonstrated to enhance the relevance and detail of descriptions generated by models such as CLIP and BLIP.

RQ5: How can the integration of knowledge graph with image captioning models enrich semantic information in image descriptions?

The incorporation of knowledge graphs enables models to transcend mere object recognition by integrating external semantic linkages. The analyzed study indicates that Axiom captioning employs graph neural networks to facilitate advanced reasoning in mathematical picture production [34], but semantic scene encoders such as SSE [21] more successfully depict spatial and relational context. These methodologies result in captions that encompass inferred concepts rather than solely observed content, so enhancing coherence and depth of meaning.

RQ6: To what extent are current multilingual image captioning techniques able to accommodate linguistic and cultural variations in image description?

Multilingual capabilities are starting to develop, especially in models evaluated on Bengali datasets [17], [18]. These models demonstrate basic fluency in low-resource languages; nevertheless, they have yet to integrate cultural context or stylistic adaptation. The captions are often literal translations instead of culturally relevant storytelling. This suggests that although multilingual captioning is technically achievable, the semantic quality across cultures poses a barrier that necessitates additional investigation.

RQ7: How do current image captioning models balance the trade-off between caption accuracy and computational efficiency in resource-constrained environments?

Efficiency-oriented models such as SCAP [20] implement streamlined attention methods to minimize computational time and resource requirements. These models attain satisfactory accuracy (BLEU-4 around 29), although remain inferior to larger Transformer-based models in terms of richness and fluency. The trade-off prioritizes real-time or mobile deployment, albeit at the expense of detailed captioning, particularly in intricate scenarios. This underscores a persistent performance-efficiency disparity in the domain.

RQ8: How does the development of vision-language foundation models such as Flamingo, GIT, and LLaVA affect accuracy and generalization in image captioning tasks?

The emergence of foundational language-vision models (VLM) like Flamingo, GIT, and Llava has profoundly transformed the methodology of image captioning. Even though not all are included among the 36 publications meticulously selected through PRISMA, a comprehensive analysis of contemporary trends would be incomplete without addressing their influence. The choice of these three models as representative pillars is grounded in their distinct contributions, Flamingo is distinguished by its effective few-shot learning capabilities, enabling swift adaptability with limited data [54]. Git from Microsoft demonstrates the efficacy of a singular generative architecture that attains state-of-the-art performance across multiple benchmarks following extensive pre-training [55]. Simultaneously, Llava advanced an architectural framework that integrates a vision encoder with a Large Language Model (LLM) for activities necessitating reasoning and comprehension of directives [56]. Their combined effect is a notable enhancement in zero-shot generalization, cross-domain accuracy, and the capacity to generate more sophisticated descriptions, establishing new benchmarks for future models. Figure 7 summarizes and illustrates the significant impacts of Vision-Language Foundation Models on the field of Image Captioning.

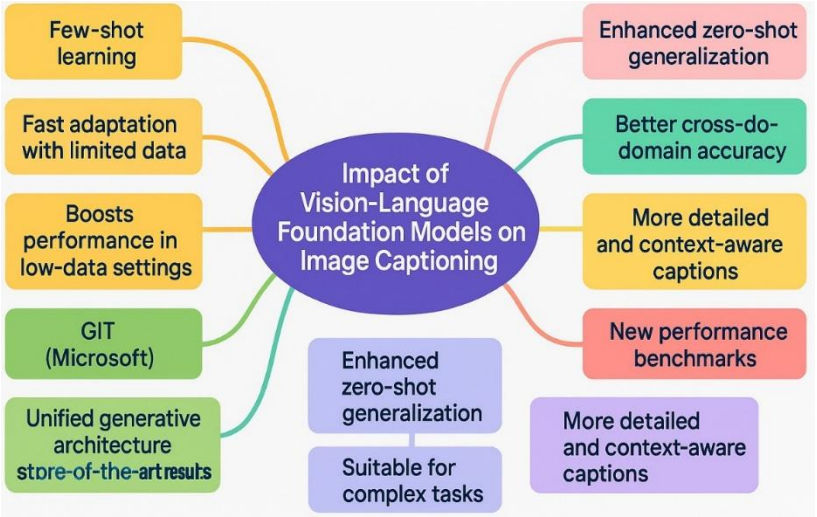

Figure 7. A concept Map Summarizing the Impact of Vision-Language Foundation Models (VLFM) on the Field of Image Captioning

RQ9: How effective are newer data augmentation techniques, such as GAN-based augmentation or synthetic data generation, in improving the quality of image captioning models?

Synthetic augmentation techniques have demonstrated the ability to improve model robustness, especially in areas with scarce data. In medical and remote sensing applications, methods such as patch-based masking [8] and noise injection [12] have been employed to replicate real-world conditions. These augmentations enhance caption stability and precision across diverse input qualities; however, the intricacy of synthetic pipeline integration may prolong training duration.

RQ10: How do novel pretraining and fine-tuning strategies in image captioning models affect the model's generalization ability across different domains?

Pretraining on varied vision-language tasks followed by fine-tuning on certain domains markedly improves generalization. For example, models such as SSE [57] and SATIC [28] utilize semantic guidance or self-attention mechanisms during fine-tuning to adjust to new domains, including medical imaging. These methodologies regularly

produce superior accuracy and contextual awareness relative to models developed from the ground up, demonstrating that intentional pretraining is a crucial facilitator of domain transferability.

3.3 Summary of Key Findings

This review delineates key trends and deficiencies in the advancement of image captioning models. Transformer-based and vision-language foundation models have markedly surpassed conventional CNN-RNN architectures in producing precise and contextually nuanced descriptions, while presenting new trade-offs in computational expense. Prompt engineering and knowledge integration serve as effective ways to improve semantic alignment and domain-specific reasoning, especially in zero-shot and specialized scenarios. Notwithstanding these advancements, issues remain in guaranteeing equity, addressing linguistic and cultural subtleties, and sustaining efficiency in resource-constrained settings. Present methods exhibit constrained efficacy in bias reduction and multilingual adaptation, underscoring the necessity for more inclusive and socially responsible captioning frameworks. The efficacy of pretraining, fine-tuning, and data augmentation techniques is apparent in enhancing model generalization, while their implementation is still disjointed. These findings collectively emphasize the need for a more cohesive, efficient, and egalitarian strategy in the future development of image captioning systems.

4. Conclusion

This paper provides a thorough systematic literature evaluation that charts the development of image captioning algorithms from 2017 to 2025. The analysis of 36 selected articles confirmed a substantial paradigm change from traditional CNN-RNN architectures to more advanced Transformer-based models and Vision-Language Models (VLM). The primary findings indicate that this contemporary method consistently yields a description that excels in accuracy, contextual coherence, and semantic comprehension. This advancement is bolstered by the incorporation of fresh methodologies, including rapid engineering, knowledge graphs, and intricate multimodal attention procedures. The advent of VLMs such as Flamingo, Git, and Llava has significantly expanded the boundaries of generalization capabilities and zero-shot learning applications.

The primary contribution of this work consists of three layers. Initially, we present a systematic historical framework that delineates the chronological evolution of architecture and highlights the principal technological milestones within this domain. Secondly, we consolidate the efficacy of diverse advanced methodologies, providing researchers with a cohesive framework to comprehend the trade-offs of different approaches. Third, by precisely delineating the unresolved issues, we pinpoint the most pressing research gaps, particularly with bias, computing efficiency, and multilingual capabilities.

Future research should concentrate on advancing efficient architectural computers for enhanced accessibility, establishing standardized benchmarks for assessing justice and prejudice mitigation, and augmenting multilingual models with profound cultural contexts rather than mere literal translations. By addressing these challenges, the domain of image captioning can progress towards developing models that not only see but also comprehend the visual world with accuracy, equity, and utility.

Acknowledgement

The researcher would like to sincerely thank the Faculty of Engineering, Universitas Majalengka, for their assistance and the favorable research atmosphere they provided during the preparation of this work. Without the institution's infrastructure and academic support, this research would not have been feasible. Additionally, the researcher would like to express gratitude to everyone who indirectly contributed to shaping the direction and focus of this literature review through discussions and inputs. It is intended that this work will significantly advance computer vision research and stimulate more investigation into the subject of image captioning, particularly in addressing practical issues through a multidisciplinary approach.

References

[1] H. T. Ho *et al.*, “A Review on Vision-Language-Based Approaches: Challenges and Applications,” *Comput. Mater. Contin.*, vol. 82, no. 2, pp. 1733–1756, 2025. https://doi.org/10.32604/cmc.2025.060363

[2] N. M. Khassaf and N. H. M. Ali, “Improving Pre-trained CNN-LSTM Models for Image Captioning with Hyper-Parameter Optimization,” *Eng. Technol. Appl. Sci. Res.*, vol. 14, no. 5, pp. 17337–17343, 2024. https://doi.org/10.48084/etasr.8455

[3] S. Tyagi *et al.*, “Novel Advance Image Caption Generation Utilizing Vision Transformer and Generative Adversarial Networks,” *Computers*, vol. 13, no. 12, 2024. https://doi.org/10.3390/computers13120305

[4] H. B. Duy *et al.*, “A dental intraoral image dataset of gingivitis for image captioning,” *Data Br.*, vol. 57, p. 110960, 2024. https://doi.org/10.1016/j.dib.2024.110960

[5] Y. Li, X. Zhang, T. Zhang, G. Wang, X. Wang, and S. Li, “A Patch-Level Region-Aware Module with a Multi-Label Framework for Remote Sensing Image Captioning,” *Remote Sens.*, vol. 16, no. 21, pp. 1–20, 2024. https://doi.org/10.3390/rs16213987

[6] K. Cheng, E. Cambria, J. Liu, Y. Chen, and Z. Wu, “KE-RSIC: Remote Sensing Image Captioning Based on Knowledge Embedding,” *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 18, pp. 4286–4304, 2024. https://doi.org/10.1109/JSTARS.2024.3523944

[7] S. Das and R. Sharma, “A TextGCN-Based Decoding Approach for Improving Remote Sensing Image Captioning,” *IEEE Geosci. Remote Sens. Lett.*, pp. 1–6, 2024. https://doi.org/10.1109/LGRS.2024.3523134

[8] Q. Lin, S. Wang, X. Ye, R. Wang, R. Yang, and L. Jiao, "CLIP-based Grid Features and Masking for Remote Sensing Image Captioning," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 18, pp. 2631–2642, 2024. https://doi.org/10.1109/JSTARS.2024.3510414
[9] Y. Yang, T. Liu, Y. Pu, L. Liu, Q. Zhao, and Q. Wan, "Multi-Attentive Network with Diffusion Model," pp. 1–18, 2024. https://doi.org/10.3390/rs16214083
[10] X. Zhang, J. Shen, Y. Wang, J. Xiao, and J. Li, "Zero-Shot Image Caption Inference System Based on Pretrained Models," *Electron.*, vol. 13, no. 19, 2024. https://doi.org/10.3390/electronics13193854
[11] P. S. Sherly and P. Velvizhy, "'Idol talks!' AI-driven image to text to speech: illustrated by an application to images of deities," *Herit. Sci.*, vol. 12, no. 1, pp. 1–21, 2024. https://doi.org/10.1186/s40494-024-01490-0
[12] L. Yu, M. Nikandrou, J. Jin, and V. Rieser, "Quality-agnostic Image Captioning to Safely Assist People with Vision Impairment," *IJCAI Int. Jt. Conf. Artif. Intell.*, vol. 2023-Augus, pp. 6281–6289, 2023. https://doi.org/10.24963/ijcai.2023/697
[13] Y. Li, Y. Ma, Y. Zhou, and X. Yu, "Semantic-Guided Selective Representation for Image Captioning," *IEEE Access*, vol. 11, no. December 2022, pp. 14500–14510, 2023. https://doi.org/10.1109/ACCESS.2023.3243952
[14] M. Alansari, K. Alnuaimi, S. Alansari, and S. Javed, "ELTrack: Events-Language Description for Visual Object Tracking," *IEEE Access*, vol. 13, no. December 2024, pp. 31351–31367, 2025. https://doi.org/10.1109/ACCESS.2025.3540445
[15] F. Kalantari, K. Faez, H. Amindavar, and S. Nazari, "Improved image reconstruction from brain activity through automatic image captioning," *Sci. Rep.*, vol. 15, no. 1, pp. 1–17, 2025. https://doi.org/10.1038/s41598-025-89242-3
[16] Y. Qin, S. Ding, and H. Xie, "Advancements in Large-Scale Image and Text Representation Learning: A Comprehensive Review and Outlook," *IEEE Access*, vol. PP, p. 1, 2025. https://doi.org/10.1109/ACCESS.2025.3541194
[17] A. Masud, M. B. Hosen, M. Habibullah, M. Anannya, and M. S. Kaiser, "Image captioning in Bengali language using visual attention," *PLoS One*, vol. 20, no. 2 February, pp. 1–15, 2025. https://doi.org/10.1371/journal.pone.0309364
[18] B. Patra and D. R. Kisku, "Exploring Bengali Image Descriptions through the combination of diverse CNN Architectures and Transformer Decoders," *Turkish J. Eng.*, vol. 9, no. 1, pp. 64–78, 2025. https://doi.org/10.31127/tuje.1507442
[19] Y. Tang, Y. Yuan, F. Tao, and M. Tang, "Cross-modal Augmented Transformer for Automated Medical Report Generation," *IEEE J. Transl. Eng. Heal. Med.*, vol. 13, no. December 2024, pp. 33–48, 2025. https://doi.org/10.1109/JTEHM.2025.3536441
[20] Y. Zhang, J. Tong, and H. Liu, "SCAP: enhancing image captioning through lightweight feature sifting and hierarchical decoding," *Vis. Comput.*, pp. 0–26, 2025. https://doi.org/10.1007/s00371-025-03824-w
[21] F. Zhao, Z. Yu, T. Wang, and Y. Lv, "Image Captioning Based on Semantic Scenes," *Entropy*, vol. 26, no. 10, pp. 1–20, 2024. https://doi.org/10.3390/e26100876
[22] N. Shetty and Y. Li, "Detailed Image Captioning and Hashtag Generation," *Futur. Internet*, vol. 16, no. 12, 2024. https://doi.org/10.3390/fi16120444
[23] A. A. E. Osman, M. A. W. Shalaby, M. M. Soliman, and K. M. Elsayed, "Novel concept-based image captioning models using LSTM and multi-encoder transformer architecture," *Sci. Rep.*, vol. 14, no. 1, pp. 1–15, 2024. https://doi.org/10.1038/s41598-024-69664-1
[24] A. Zheng, S. Zheng, C. Bai, and D. Chen, "Triple-level relationship enhanced transformer for image captioning," *Multimed. Syst.*, vol. 29, no. 4, pp. 1955–1966, 2023. https://doi.org/10.1007/s00530-023-01073-2
[25] Y. Pan, T. Yao, Y. Li, and T. Mei, "X-Linear Attention Networks for Image Captioning," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pp. 10968–10977, 2020. https://doi.org/10.1109/CVPR42600.2020.01098
[26] Y. Jung, I. Cho, S. H. Hsu, and M. Golparvar-Fard, "VISUALSITEDIARY: A detector-free Vision-Language Transformer model for captioning photologs for daily construction reporting and image retrievals," *Autom. Constr.*, vol. 165, no. May, p. 105483, 2024. https://doi.org/10.1016/j.autcon.2024.105483
[27] J. Chen, H. Guo, K. Yi, B. Li, and M. Elhoseiny, "VisualGPT: Data-efficient Adaptation of Pretrained Language Models for Image Captioning," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2022-June, pp. 18009–18019, 2022. https://doi.org/10.1109/CVPR52688.2022.01750
[28] Y. Zhou, Y. Zhang, Z. Hu, and M. Wang, "Semi-Autoregressive Transformer for Image Captioning," *Proc. IEEE Int. Conf. Comput. Vis.*, vol. 2021-Octob, pp. 3132–3136, 2021. https://doi.org/10.1109/ICCVW54120.2021.00350
[29] J. H. Wang, M. Norouzi, and S. M. Tsai, "Augmenting Multimodal Content Representation with Transformers for Misinformation Detection †," *Big Data Cogn. Comput.*, vol. 8, no. 10, 2024. https://doi.org/10.3390/bdcc8100134
[30] S. Gautam *et al.*, "Kvasir-VQA: A Text-Image Pair GI Tract Dataset," *arXiv Prepr. arXiv2409.01437*, 2024. https://doi.org/10.1145/3689096.3689458
[31] Z. Li, D. Liu, H. Wang, C. Zhang, and W. Cai, "Exploring Annotation-free Image Captioning with Retrieval-augmented Pseudo Sentence Generation," no. VI, 2023. https://doi.org/10.1145/3696409.3700223
[32] K. Y. Cheng, M. Lange-Hegermann, J. B. Hövener, and B. Schreiweis, "Instance-level medical image classification for text-based retrieval in a medical data integration center," *Comput. Struct. Biotechnol. J.*, vol. 24, no. February, pp. 434–450, 2024. https://doi.org/10.1016/j.csbj.2024.06.006
[33] X. Guo, X. Di Liu, and J. Jiang, "A Scene Graph Generation Method for Historical District Street-view Imagery: A Case Study in Beijing, China," *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci. - ISPRS Arch.*, vol. 48, no. 3, pp. 209–216, 2024. https://doi.org/10.5194/isprs-archives-XLVIII-3-2024-209-2024
[34] E. K. Holden and K. Korovin, "Graph sequence learning for premise selection," *J. Symb. Comput.*, vol. 128, p. 102376, 2025. https://doi.org/10.1016/j.jsc.2024.102376
[35] S. Fayou, H. C. Ngo, Y. W. Sek, and Z. Meng, "Clustering swap prediction for image-text pre-training," *Sci. Rep.*, vol. 14, no. 1, pp. 1–16, 2024. https://doi.org/10.1038/s41598-024-60832-x
[36] A. Sebaq and M. ElHelw, "RSDiff: remote sensing image generation from text using diffusion model," *Neural Comput. Appl.*, vol. 36, no. 36, pp. 23103–23111, 2024. https://doi.org/10.1007/s00521-024-10363-3
[37] H. Senior, G. Slabaugh, S. Yuan, and L. Rossi, "Graph neural networks in vision-language image understanding: a survey," *Vis. Comput.*, vol. 41, no. 1, pp. 491–516, 2024. https://doi.org/10.1007/s00371-024-03343-0
[38] W. Hu, F. Zhang, and Y. Zhao, "Thangka image captioning model with Salient Attention and Local Interaction Aggregator," *Herit. Sci.*, vol. 12, no. 1, pp. 1–21, 2024. https://doi.org/10.1186/s40494-024-01518-5
[39] F. Zhao, Z. Yu, T. Wang, and H. Zhao, "Meshed Context-Aware Beam Search for Image Captioning," *Entropy*, vol. 26, no. 10, pp. 1–22, 2024. https://doi.org/10.3390/e26100866
[40] P. Sloan, P. Clatworthy, E. Simpson, and M. Mirmehdi, "Automated Radiology Report Generation: A Review of Recent Advances," *IEEE Rev. Biomed. Eng.*, vol. XX, no. Xx, pp. 1–24, 2024. https://doi.org/10.1109/RBME.2024.3408456
[41] M. J. Page *et al.*, "The PRISMA 2020 statement: An updated guideline for reporting systematic reviews," *BMJ*, vol. 372, 2021. https://doi.org/10.1136/bmj.n71

[42] M. L. Rethlefsen *et al.*, "PRISMA-S: an extension to the PRISMA Statement for Reporting Literature Searches in Systematic Reviews," *Syst. Rev.*, vol. 10, no. 1, pp. 1–19, 2021. https://doi.org/10.1186/s13643-020-01542-z
[43] N. R. Haddaway, M. J. Page, C. C. Pritchard, and L. A. McGuinness, "PRISMA2020: An R package and Shiny app for producing PRISMA 2020-compliant flow diagrams, with interactivity for optimised digital transparency and Open Synthesis," *Campbell Syst. Rev.*, vol. 18, no. 2, pp. 1–12, 2022. https://doi.org/10.1002/cl2.1230
[44] S. A. Ghosal. K, Rana. A, "Aesthetic image captioning from weakly-labelled photographs," *Proc. - 2019 Int. Conf. Comput. Vis. Work. ICCVW 2019*, pp. 4550–4560, 2019. https://doi.org/10.1109/ICCVW.2019.00556
[45] C. T.-S. Zhang. M, Yang. Y, Zhang. H, Ji. Y, Shen. H.T, "More is better: Precise and detailed image captioning using online positive recall and missing concepts mining," *IEEE Trans. Image Process.*, vol. 28, no. 1, pp. 32–44, 2019. https://doi.org/10.1109/TIP.2018.2855415
[46] C. T.-S. Chen. L, Zhang. H, Xiao. J, Nie. L, Shao. J, Liu. W, "SCA-CNN: Spatial and channel-wise attention in convolutional networks for image captioning," *Proc. - 30th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR 2017*, vol. 2017-Janua, pp. 6298–6306, 2017. https://doi.org/10.1109/CVPR.2017.667
[47] L. B. Hu. N, Ming. Y, Fan. C, Feng. F, "TSFNet: Triple-Steam Image Captioning," *IEEE Trans. Multimed.*, vol. 25, pp. 6904–6916, 2023. https://doi.org/10.1109/TMM.2022.3215861
[48] Z. Y. Xu. N, Zhang. H, Liu. A.-A, Nie. W, Su. Y, Nie. J, "Multi-Level Policy and Reward-Based Deep Reinforcement Learning Framework for Image Captioning," *IEEE Trans. Multimed.*, vol. 22, no. 5, pp. 1372–1383, 2020. https://doi.org/10.1109/TMM.2019.2941820
[49] G. . . Rennie. S.J, Marcheret. E, Mroueh. Y, Ross. J, "Self-critical sequence training for image captioning," *Proc. - 30th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR 2017*, vol. 2017-Janua, pp. 1179–1195, 2017. https://doi.org/10.1109/CVPR.2017.131
[50] W. Z. Cao. S, An. G, Zheng. Z, "Vision-Enhanced and Consensus-Aware Transformer for Image Captioning," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 10, pp. 7005–7018, 2022. https://doi.org/10.1109/TCSVT.2022.3178844
[51] W. Z. Zhang. J, Xie, Y, Ding. W, "Cross on Cross Attention: Deep Fusion Transformer for Image Captioning," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 8, pp. 4257–4268, 2023. https://doi.org/10.1109/TCSVT.2023.3243725
[52] W. M. Song. Z, Hu. Z, Zhou. Y, Zhao. Y, Hong. R, "Embedded Heterogeneous Attention Transformer for Cross-Lingual Image Captioning," *IEEE Trans. Multimed.*, vol. 26, pp. 9008–9020, 2024. https://doi.org/10.1109/TMM.2024.3384678
[53] W. L. Liu A. A, Wu Q, Xu N, Tian H, "Enriched Image Captioning based on Knowledge Divergence and Focus," *IEEE Trans. Circuits Syst. Video Technol.*, 2025. https://doi.org/10.1109/TCSVT.2024.3525158
[54] J. B. Alayrac *et al.*, "Flamingo: a Visual Language Model for Few-Shot Learning," *Adv. Neural Inf. Process. Syst.*, vol. 35, no. NeurIPS, 2022.
[55] J. Wang *et al.*, "GIT: A Generative Image-to-text Transformer for Vision and Language," vol. 2, pp. 1–49, 2022, [Online]. Available: http://arxiv.org/abs/2205.14100
[56] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual Instruction Tuning," *Adv. Neural Inf. Process. Syst.*, vol. 36, no. NeurIPS, pp. 1–25, 2023.
[57] E. J. Bassey, J. H. Cheng, and D. W. Sun, "Enhancing infrared drying of red dragon fruit by novel and innovative thermoultrasound and microwave-mediated freeze-thaw pretreatments," *Lwt*, vol. 202, no. March, p. 116225, 2024. https://doi.org//10.1016/j.lwt.2024.116225