



Auto Caption AI: A Transformer-Based Vision-Language System for Automated Image Captioning, Translation and Speech Synthesis

Deepika¹, Kirti², Pushap³, Kunal Tyagi⁴, Ms.Anisha⁵

Department of CSE-AI&DS,

Panipat Institute of Engineering & Technology, Samalkha, India

Mail ID: anisha.cse.ai-ds@piet.co.in

Abstract— Auto Caption AI is an intelligent image captioning system that generates meaningful, context-aware descriptions for images using advanced vision-language transformer models. As visual content continues to grow across digital platforms, there's an increasing demand for systems that can interpret images in natural language. Manual captioning is often time-consuming and inconsistent, making automation essential. At the core of this project is the Salesforce BLIP model, a transformer-based architecture that combines a visual encoder with a text decoder to extract semantic information from images and produce fluent, human-like captions. The model is implemented using the Hugging Face Transformers library and runs on PyTorch, ensuring high performance across various image types. The user interface is built with Streamlit, allowing users to upload images and instantly receive captions through a web app. The backend runs on Google Colab, leveraging GPU acceleration for real-time inference, and public access is provided via ngrok. By integrating powerful transformer-based vision-language models with intuitive deployment tools, Auto Caption AI provides a seamless and practical solution for automatic image captioning, proving especially valuable for individuals with visual impairments who rely on textual cues for image interpretation.

Keywords: Image Captioning; BLIP; Vision-Language Transformer; Streamlit; MarianMT; Text-to-Speech; Accessibility; Deep Learning.

INTRODUCTION

Auto Caption AI is a state-of-the-art system for image captioning that utilizes deep learning techniques to automatically produce natural-language descriptions of digital images. In the current digital landscape, where immense quantities of visual content are created and disseminated across multiple platforms, the capability to understand and articulate what images depict is essential for enhancing user experiences, improving accessibility, and facilitating effective information retrieval. While manual captioning has its advantages, it is essentially time-consuming, susceptible to variability, and not practical for extensive applications[1].

The primary innovation is in merging computer vision with natural language processing (NLP) through a single transformer-based framework. Auto Caption AI implements this concept by leveraging advanced vision-language models to connect visual data with linguistic expressions. It utilizes the BLIP (Bootstrapping Language-Image Pre-training) model created by Salesforce, a leading vision-language transformer recognized for its superior performance on image captioning benchmarks. BLIP [6][7] is proficient at grasping the semantic subtleties of images and converting them into coherent text outputs, making it an excellent option for practical applications.

The application has been created using Python, utilizing the Hugging Face Transformers library as the means for loading and operating the BLIP model. PyTorch serves as the core deep learning



framework, providing both adaptability and computational efficiency during the inference process. The system includes a web-based front end developed with Streamlit, enabling users to engage with the model in an intuitive and responsive way. By incorporating ngrok, the application deployed on Google Colab can be accessed publicly through a secure and temporary URL, facilitating real-time usage without requiring a dedicated server. The purpose of the system is to use cutting-edge artificial intelligence techniques to automate the process of creating contextually relevant and meaningful textual descriptions for digital photos, thereby enhancing accessibility for visually impaired users, improving content organization and searchability, and encouraging scalable AI solutions for real-world use across journalism, social media, healthcare and education.

A. Target Users

The Auto Caption AI is designed to serve a wide range of users, offering tailored solutions for both technical and non-technical individuals. Visually impaired individuals form one of the main target user groups, since the system enables screen readers and other assistive technology to convey the content of images by translating visual content into descriptive text. Content creators and social media managers frequently have to caption a lot of photos to increase interaction, enhance SEO, or satisfy accessibility requirements, a process that Auto Caption AI automates. E-commerce and retail businesses can describe product photos automatically, boosting discoverability by search engines, while educators and e-learning platforms can provide meaningful descriptions for instructional images, charts, and diagrams. Researchers and data annotators, digital libraries and archival services, news and media professionals, and software developers or AI enthusiasts also stand to benefit from the system.

B. Topic Background

As images and videos are becoming more and more prevalent on digital platforms, there is a growing need for intelligent systems that can comprehend and analyze visual content. Earlier attempts at automated captioning relied on manually created features or rule-based systems, which were not flexible or generalizable. Image captioning[16] has advanced significantly with the development of deep learning, particularly transformer-based systems; vision-language models such as BLIP have been created to integrate language decoders and image encoders in a single architecture, enabling more precise, context-aware, and fluid caption production. At the same time, frameworks and tools like Hugging Face Transformers and PyTorch make these sophisticated models more accessible, while user interface libraries like Streamlit allow AI applications to be implemented as interactive web apps without extensive frontend programming.

LITERATURE REVIEW

This review investigates the developing field of image captioning by looking into transformer-based vision-language models, new developments in deep learning, and their incorporation into practical applications, while highlighting the drawbacks of previous methods that relied on manually created features or preset templates.

A. Vision-Language Modelling

Vision-language models aim to connect visual inputs with natural language outputs. A notable example is BLIP, which uses a Vision Transformer for encoding images and a Transformer-based decoder for caption generation. It demonstrates strong performance across both image captioning and visual question answering (VQA)[3] tasks and achieves significant gains (e.g., +2.8 CIDEr) by leveraging noisy web-scale image-text pairs for training.

B. CNN-LSTM Based Image Captioning

Earlier architectures such as CNN-LSTM have also shown promise in caption creation. These systems employ LSTM decoders to produce phrases and CNNs (such as VGG16) to extract visual information from images. The robustness of models trained on datasets such as Flickr8k and MSCOCO[10] has been validated by their respectable BLEU[2] scores as shown in Table I.



C. Visual Attention and Context Modelling

Contextual spatial relation extraction (CSE) and visual attention mechanisms[4] are recent developments that improve captioning by concentrating on particular areas of the picture and comprehending spatial relationships. Hybrid models that integrate CNNs, attention layers, and LSTMs have shown more accurate captions and improved semantic understanding.

C. Domain Research

Systems for automatically captioning images are highly relevant across many fields, and their applications keep growing as AI-powered technologies for accessibility and content comprehension become more popular. In education, real-time captioning can explain visual resources to enhance comprehension, particularly for students who struggle with language or visual impairments, while translated captions give multilingual students further access. In accessibility and assistive technology, such systems can read aloud descriptions of images to people who are blind or visually challenged, increasing digital inclusion. In e-commerce and social media, automatic product-image captioning[11] can increase customer trust and searchability, while smart devices such as glasses and home assistants can use captions to explain what their cameras capture.

TABLE I: SUMMARY OF RELATED WORK

Study	Technique(s) Used	Key Finding
Sasibhooshan et al. (2023)	WCNN + LSTM + attention (VAP, CSE)	Attention and spatial relation extraction yield more contextually accurate captions than CNN-LSTM baselines.
Parate & Choudhary (2022)	CNN (feature extraction) + LSTM (sequence generation)	CNN-LSTM pipelines enhance accessibility, news captioning and social-media use cases.
Li et al. (2022) - BLIP	Vision Transformer encoder + Transformer decoder, bootstrapped from noisy web data	Outperformed prior models on multiple benchmarks; +2.8 CIDEr gain and strong generalization across captioning, VQA and retrieval.
Goyal, Goyal & Virmani (2022)	VGG16 (Hybrid Places-1365) CNN encoder + LSTM decoder	Achieved BLEU-1 score of 0.735 on MSCOCO; validated CNN-LSTM effectiveness on Flickr8k and MSCOCO.

PROBLEM OBJECTIVE

A. Problem Description

Even with major advancements in computer vision, automatically creating informative captions for photos is still a difficult operation that necessitates a thorough comprehension of both natural language generation and visual semantics. Historically, manual labor has been used to annotate photographs with pertinent written descriptions, which is labour-intensive, time-consuming, and not scalable for real-time use cases or huge datasets. Current automated methods, like keyword tagging or simple object detection, only provide a portion of the answer, since they are unable to generate captions that are relevant, grammatically sound, and acceptable for the context.

B. Proposed Solution

This project offers a comprehensive Image Captioning System that combines text-to-speech, multilingual translation, and deep learning into a single, user-friendly application to overcome the difficulties posed by manual image annotation, language barriers, and accessibility restrictions.



- **Vision-Language Model Integration:** BLIP is utilized as the final production model owing to its excellent caption fluency and ease of integration, after comparison with a ViT-GPT2 pipeline.
- **Multilingual Translation (MarianMT)[17]:** English-generated captions are automatically translated into other target languages, including Hindi, French, and Spanish, using pretrained MarianMT models from Hugging Face.
- **Text-to-Speech (gTTS):** The translated or original caption can be converted into speech for interactive use cases and visually impaired users.
- **Streamlit Web Interface:** Deploys the entire system, enabling users to upload an image, view the caption, select a translation language, and listen to the speech output.

D. Real-Life Challenges

Although Auto Caption AI uses sophisticated deep learning models to automate picture captioning, the creation and implementation of the system presented several real-world difficulties typical of AI-powered systems.

- **Handling Diverse Image Content:** No model can fully generalize to every visual circumstance without additional refinement; fine-tuning on domain-specific datasets and data augmentation can improve adaptability.
- **Limitations in Caption Precision:** Occasional errors or ambiguous descriptions occur, particularly for images with several unconnected elements; visual grounding strategies or a caption-refinement layer can help.
- **Balancing Model Size and Accessibility:** High accuracy models such as BLIP are resource-intensive; lightweight alternatives (e.g., Tiny-ViT or DistilBERT-based captioning) can widen accessibility on low-resource devices.
- **User Privacy and Data Security:** Uploaded images are processed in-session without permanent storage; secure, encrypted transmission and compliance with data-protection regulations remain important for production use.
- **Deployment Without Dedicated Servers:** Hosting via Colab and ngrok is economical but transient; migrating to Render, Hugging Face Spaces, or Streamlit Cloud offers more reliable, concurrent access.

METHODOLOGY AND SYSTEM ARCHITECTURE

The core of Auto Caption AI's methodology is the creation of an AI-powered real-time application that can comprehend visual input and provide natural language captions that are contextually relevant. Each functional component of the modular architecture functions independently while blending in perfectly with the pipeline as a whole: images are uploaded and preprocessed, captions are created using the BLIP or ViT-GPT2 model, MarianMT[17] is used for optional language translation, and gTTS is used for speech synthesis. Model evaluation and selection entailed the implementation and comparison of several image captioning architectures, namely GRU, LSTM, ViT-GPT2, and BLIP. BLIP was chosen as the final model because it continuously generated captions that were more accurate and contextually rich.

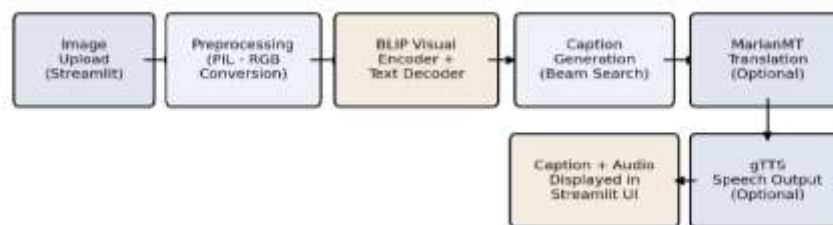


Fig. 1. High-level system architecture of Auto Caption AI.

A. Key Components

- **Input Module:** A Streamlit interface enables users to upload photos in JPEG or PNG format; PIL converts the image to RGB format.



- **Caption Generation Module:** The uploaded image is run through the BLIP transformer, which produces a fluent, contextually correct caption.
- **Translation Module:** MarianMT models translate the generated English caption into a user-selected language in real time.
- **Text-to-Speech Module:** gTTS converts the generated or translated caption into spoken audio played back within the app.
- **Deployment Module:** The application is hosted on Google Colab with GPU acceleration and exposed publicly via a temporary ngrok URL.

B. System Workflow

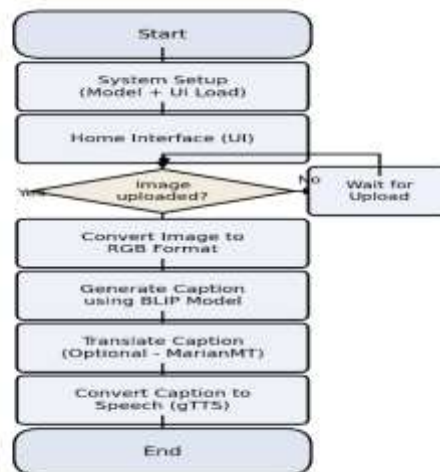


Fig. 2. Flowchart of the Auto Caption AI processing pipeline.

System setup initializes the processor, Streamlit interface, and BLIP transformer model. The user is then shown a home interface where an image can be uploaded; the system waits idle if no image is submitted. Once uploaded, the image is converted to RGB format and preprocessed by the BLIP processor into tensor representations suitable for transformer-based analysis. The BLIP model receives the processed image and generates a descriptive caption, which is decoded and shown next to the uploaded picture in the Streamlit interface. Finally, a secure ngrok tunnel is created from the Google Colab environment, making the entire program publicly available for real-time AI-powered image captioning.

C. Development Methodology

Each functional module - caption creation, translation, text-to-speech, and user interface - was developed and tested independently before being incorporated into the entire pipeline, following an iterative and modular development process spanning requirement analysis, system design, implementation, testing, deployment, and iteration and maintenance. BLIP was chosen as the main model after testing and output comparison; the translation module was subsequently updated to handle more languages and the speed of the caption-generation process was improved.

D. Data Collection and Dataset Preparation

The approach relies on datasets where each image is accompanied by a relevant, human-annotated caption to guarantee dependable output from the captioning models; these datasets are used as ground truth to compare model performance and test caption quality. Users can upload their own JPEG or PNG images through the Streamlit interface as dynamic input, while additional sample images were retrieved via URLs from sources such as Pexels for consistent testing, comprising scenes involving people, animals, landscapes, and other elements. Uploaded images are converted



to RGB format using the PIL library, and images were purposefully varied in lighting, complexity, and subject matter to assess how well the models produce precise and contextually rich captions.

E. Implementation Highlights

The final application utilizes modern transformer models (BLIP, ViT-GPT2)[12][13] for accurate, rich captioning, with captions generated and displayed instantly upon image upload. Multilingual support allows captions to be translated into multiple languages using MarianMT, and speech output through gTTS enables text-to-speech playback of the caption. The Streamlit interface makes the tool accessible to both technical and non-technical users: users open the app via the ngrok URL, upload an image file, view the automatically generated caption, optionally select a target language from a dropdown for translation, and click a "Listen to Caption" button to activate gTTS, after which the original image, generated caption, and audio output (if enabled) are all displayed in real time.

RESULTS AND DISCUSSION

A. Hardware and Software Requirements

Real-time caption generation, translation and speech playback were supported by an Intel Core i5 (or higher) processor, a minimum of 8 GB RAM, 256 GB SSD storage, and an optional NVIDIA GTX/RTX GPU for accelerated inference. On the software side, the system runs on Python 3.7+ with Hugging Face Transformers, PyTorch, gTTS, Streamlit, OpenCV and NumPy, developed and tested in Jupyter Notebook / VS Code and deployed through Google Colab with Chrome or Firefox as the target browser.

B. Comparison with Initial Project Goals

The initial goals of automated image captioning, accessibility and multilingual support, and interactive cloud deployment were all achieved: the system automatically generates captions on image upload using the BLIP model and supports multiple image types with coherent real-time outputs; captions can be translated into several languages using MarianMT and converted into speech using gTTS; and the application is deployed via Streamlit and accessible using ngrok[14] links, supporting image upload, caption display, language selection, and audio playback. The primary challenges identified were occasional misinterpretation of complex or abstract scenes and slight inference delays on Colab when a GPU was unavailable.

C. Model Result

The Auto Caption AI project successfully generates descriptive captions for a wide variety of images using transformer-based vision-language models. The BLIP model demonstrated superior performance in terms of accuracy, contextual relevance, and fluency. In real-time usage, it consistently produced captions that correctly described the primary subjects and their environment with minimal hallucination or generic phrasing. Though the model performs well in standard cases, occasional inaccuracies may occur in visually complex scenes or low-resolution images[8][9][10].

BLIP generated context-rich, fluent captions with an estimated qualitative accuracy of approximately 72.15% based on human evaluation of relevance and fluency, outperforming other models in describing fine details and scene context. Transformer-based models demonstrated significantly higher fluency than RNN-based models, while captions sometimes lacked precision for abstract or cluttered images, with accuracy decreasing slightly under poor lighting or uncommon subjects.

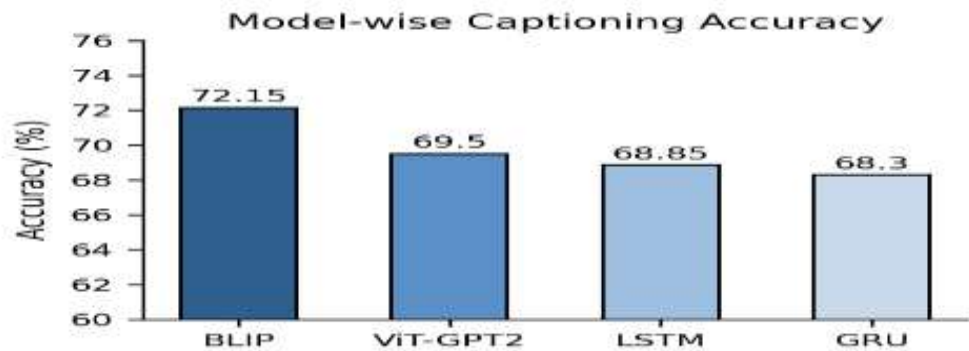


Fig. 3. Comparative captioning accuracy across the four evaluated models.

As summarized in Table II, BLIP offered the best balance between fluency, detail, and contextual accuracy among the four evaluated architectures, with its transformer architecture allowing for better generalization and performance than the CNN-LSTM and CNN-GRU baselines.

TABLE II. MODEL COMPARISON METRICS

Aspect	BLIP	ViT-GPT2	GRU-based	LSTM-based
Architecture	ViT encoder + Transformer decoder, joint training	ViT encoder[5] + GPT-2 decoder (loosely coupled)	CNN encoder + GRU decoder	CNN encoder + LSTM decoder
Accuracy (%)	72.15 - highest	69.50 - good, lower than BLIP	68.30 - baseline	68.85 - slightly above GRU
Caption Fluency	High - rich, coherent captions	Moderate - decent, less nuanced	Moderate fragmented - generic	Moderate-High - better context retention
Inference Speed	Efficient - optimized for GPU	Slightly slower - larger model size	Fast on CPU - low compute	Slightly slower than GRU
Best Use Case	Real-time, high-quality captioning (accessibility, smart assistants)	Moderate-performance educational/archival use	Low-resource environments	Time-dependent sequence modeling



CONCLUSION AND FUTURE SCOPE

Using advanced transformer-based vision-language models, the Auto Caption AI system effectively illustrates how to automate the creation of image captions. The method addresses the laborious and repetitive process of captioning visual content by combining deep learning techniques with contemporary web technology to give users precise, natural-language descriptions of images. The BLIP model, which integrates verbal and visual data into a single transformer framework, produces captions that are cohesive and contextually rich compared to previous RNN-based models and simpler architectures.

However, image quality variations, complex scenes with overlapping objects, and unusual viewpoints or occlusions can affect the accuracy and richness of the generated captions. Notwithstanding these drawbacks, the system continues to exhibit remarkable scalability and robustness, producing reliable results for a broad range of images, while the user-friendly Streamlit interface ensures that individuals with little to no technical experience may run it with ease.

A. Applications and Implications

The system has wide-ranging implications across sectors. Educational institutions can deliver inclusive, multilingual learning materials and interactive, narrated visual content. Corporate environments can use automated captioning for accessibility-compliant documentation, training material, and global communication. Government and law-enforcement bodies can apply it to surveillance review, public communication, and evidence documentation, while the healthcare industry can use AI-generated descriptions to support (not replace) clinical documentation and patient education. Retailers can automate product-image captioning and multilingual listings, and public event venues can provide accessible displays and live interpretation for global audiences.

B. Future Scope

- **Fine-Tuning:** Fine-tuning BLIP on domain-specific datasets (e.g., medical, industrial, or tourism images) to increase adaptability to different image types.
- **Quantitative Evaluation:** Incorporating BLEU, CIDEr[15], or METEOR scores to enable quantitative comparison with reference captions.
- **Scalable Deployment:** Moving the system from Colab/ngrok to a hosted environment such as Render, Hugging Face Spaces, or Streamlit Cloud for permanent, concurrent access.
- **Video and Real-Time Captioning:** Extending image captioning into temporal, multi-frame video captioning for dynamic content.
- **Expressive Multilingual TTS:** Adopting more expressive speech generators (e.g., Tacotron 2, FastSpeech) for emotional, accent-aware audio captions.

REFERENCES

- [1] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and Tell: A Neural Image Caption Generator," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2015, pp. 3156–3164.
- [2] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio, "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention," in Proc. 32nd Int. Conf. Mach. Learn. (ICML), vol. 37, 2015, pp. 2048–2057.
- [3] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2018.



[4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention Is All You Need,” in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017.

[5] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, “An Image Is Worth 16×16 Words: Transformers for Image Recognition at Scale,” in *Proc. Int. Conf. Learn. Representations (ICLR)*, 2021.

[6] J. Li, D. Li, C. Xiong, and S. C. H. Hoi, “BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation,” in *Proc. 39th Int. Conf. Mach. Learn. (ICML)*, vol. 162, 2022, pp. 12888–12900.

This should be one of the most frequently cited references in your paper, because BLIP is the central model in your proposed system.

[7] J. Li, D. Li, S. Savarese, and S. Hoi, “BLIP-2: Bootstrapping Language-Image Pre-Training with Frozen Image Encoders and Large Language Models,” in *Proc. 40th Int. Conf. Mach. Learn. (ICML)*, vol. 202, 2023, pp. 19730–19742.

[8] J. Wang, Z. Yang, X. Hu, L. Li, K. Lin, Z. Gan, Z. Liu, C. Liu, and L. Wang, “GIT: A Generative Image-to-text Transformer for Vision and Language,” *Trans. Mach. Learn. Res.*, 2022.

This is useful when comparing your BLIP-based system with other Transformer-based image-to-text systems.

[9] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel, “Self-Critical Sequence Training for Image Captioning,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 7008–7024.

[10] X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollár, and C. L. Zitnick, “Microsoft COCO Captions: Data Collection and Evaluation Server,” *arXiv preprint arXiv:1504.00325*, 2015.

The paper describes the COCO caption dataset and its evaluation server, including BLEU, METEOR, ROUGE, and CIDEr.

[11] P. Sharma, N. Ding, S. Goodman, and R. Soricut, “Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset for Automatic Image Captioning,” in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics (ACL)*, 2018.

[12] M. Hodosh, P. Young, and J. Hockenmaier, “Framing Image Description as a Ranking Task: Data, Models and Evaluation Metrics,” *J. Artif. Intell. Res.*, vol. 47, pp. 853–899, 2013.

[13] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, “BLEU: A Method for Automatic Evaluation of Machine Translation,” in *Proc. 40th Annu. Meeting Assoc. Comput. Linguistics (ACL)*, 2002, pp. 311–318.

[14] R. Vedantam, C. L. Zitnick, and D. Parikh, “CIDEr: Consensus-Based Image Description Evaluation,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2015, pp. 4566–4575.

[15] S. Banerjee and A. Lavie, “METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments,” in *Proc. ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, 2005, pp. 65–72.

[16] P. Anderson, B. Fernando, M. Johnson, and S. Gould, “SPICE: Semantic Propositional Image Caption Evaluation,” in *Proc. European Conf. Comput. Vis. (ECCV)*, 2016.

[17] M. Junczys-Dowmunt, R. Grundkiewicz, T. Dwojak, H. Hoang, K. Heafield, T. Neckermann, F. Seide, U. Germann, A. F. Aji, N. Bogoychev, A. F. T. Martins, and A. Birch, “Marian: Fast Neural Machine Translation in C++,” in *Proc. ACL 2018, System Demonstrations*, Melbourne, Australia, 2018, pp. 116–121, doi: 10.18653/v1/P18-4020.
