PicLingo  
An AI-Driven Image Caption Generator

Project Report Submitted  
In Partial Fulfillment of the Requirement for the Award of the Degree of  
BACHELOR OF TECHNOLOGY  
IN  
COMPUTER SCIENCE

Submitted By:  
PRINCE SINGH (21142)  
PRATHAM HARSH (21141)  
TABISH JAVED (21160)

Under the Supervision of  
AVADESH YADAV  
Assistant Professor, Department of CSE

Institute of Engineering & Technology  
Dr. Ram Manohar Lohia Avadh University, Ayodhya  
Uttar Pradesh, India – 224001  
(Session 2024-2025)

# Certificate

This is to certify that the project report titled “PicLingo - An AI Driven Image Caption Generator” submitted by PRINCE SINGH (21142), PRATHAM HARSH (21141), and TABISH JAVED (21160) is a bonafide record of the work carried out under my supervision in partial fulfillment of the requirements for the award of the degree of Bachelor of Technology in Computer Science and Engineering.

Signature of Supervisor: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Signature of External Examiner: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Signature of H.O.D: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

# Acknowledgment

We would like to express our sincere gratitude to Er. Avadesh Kumar Yadav, Assistant Professor, Department of Computer Science and Engineering, IET Ayodhya, for his constant guidance, support, and encouragement throughout the development of this project.

We are also thankful to Prof. S.S. Mishra, Director, IET Ayodhya, for providing the necessary infrastructure and support that facilitated our research and development work.

Lastly, we are grateful to our family, friends, and peers who supported us directly or indirectly during the course of this project.

PRINCE SINGH (21142)  
PRATHAM HARSH (21141)  
TABISH JAVED (21160)

# Abstract

PicLingo is an AI-powered image caption recommendation system built to transform how textual information is derived from visual inputs. The project employs OpenAI’s CLIP (Contrastive Language–Image Pre-training) model to understand and match image content with the most contextually appropriate captions using cosine similarity. By bridging computer vision and natural language processing, PicLingo addresses the limitations of traditional captioning methods.

The project stands out for its use of a retrieval-based approach instead of generative models, making it more efficient and accurate for real-world applications such as content management, accessibility for visually impaired users, social media automation, and SEO enhancement. The backend is powered by Flask (Python), and the frontend is built using modern web technologies, deployed seamlessly via Vercel.

This report outlines the technical architecture, methodology, and testing strategies used in developing the project, offering a comprehensive understanding of the system’s capabilities and potential extensions.

# Table of Contents

1. Introduction

2. Literature Survey / Related Work

3. Technologies and Tools Used

4. System Analysis

5. System Design

6. Working of the Project

7. Implementation

8. Algorithms and Models

9. Dataset Used

10. Evaluation and Results

11. Testing

12. User Manual

13. Deployment

14. SWOT Analysis

15. Applications and Use Cases

16. Limitations and Challenges

17. Future Enhancements

18. Conclusion

19. References

20. Appendices

# 1. Introduction

In today’s digital landscape, images have become central to communication across platforms like social media, e-commerce, education, and digital marketing. However, the extraction of descriptive textual information from images remains a complex and time-consuming task. This process is crucial for improving accessibility, enhancing SEO, and automating content curation.

Manual image captioning is often inconsistent and lacks context-aware precision. The evolution of artificial intelligence, specifically the integration of computer vision with natural language processing, has opened new avenues to automate this task efficiently.

PicLingo is an AI-driven image caption recommendation system that leverages OpenAI’s CLIP model to bridge this gap between visual and textual understanding. It uses a retrieval-based approach that selects the most contextually accurate captions by comparing image and text embeddings using cosine similarity. Unlike generative models, which may produce grammatically correct but semantically off-topic captions, PicLingo’s approach ensures relevance, speed, and scalability.

This project is designed not only to demonstrate the power of CLIP in real-world applications but also to improve accessibility for visually impaired users, automate social media captions, enhance digital content workflows, and streamline SEO-based metadata generation.