IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 14, April 2026
Impact Factor: 8.2

# VisualAI: A Multimodal Image Generation System

**Prasilya M[1], Prapti M D[2], Amudhini J[3], Kavipriya S P[4], Geetha A[5]**
Final Year Students, Department of Computer Science Engineering[1234]
Professor, Department of Computer Science and Engineering[5]
Annamalai University, Annamalai Nagar, Tamil Nadu, India

**Abstract:** *This project presents VisualAI: A Multimodal Image Generation System, a web-based application designed to generate images based on different types of user input such as text, speech, and sign language. The primary objective of this system is to improve accessibility and user interaction by allowing multiple modes of communication within a single platform. The proposed system is developed using a client-server architecture, where the frontend is built using React and Vite, and the backend is implemented using the Flask framework. The frontend provides an interactive interface for users to input their requests, while the backend processes the input, enhances the prompt, and generates images using an external AI-based image generation service. In the backend, the system performs prompt enhancement to improve the quality of generated images by adding descriptive attributes such as lighting and detail. It also implements a caching mechanism using hashing techniques to store previously generated results, reducing response time and avoiding repeated external API calls. A deterministic seed generation method is used to maintain consistency in output for identical inputs. The image generation process is carried out using the Pollinations.ai service with the FLUX model, which produces high-quality images based on the processed prompt. The system is designed to be lightweight and does not require high computational resources, as it relies on external services for image generation. This makes it suitable for deployment on standard systems and accessible through web browsers. The integration of MediaPipe libraries in the frontend indicates support for hand gesture-based interaction, which can be extended for sign language recognition. Future enhancements may include integrating advanced speech and sign recognition models, improving image quality, and expanding support for multiple languages.*

**Keywords:** Multimodal Image Generation, Sign Language Recognition, Text-to-Image Synthesis, Generative AI, Assistive Technology, Human-Computer Interaction, Diffusion Model, MediaPipe Hand Landmark Detection

## I. INTRODUCTION

In the contemporary digital era, Artificial Intelligence has emerged as a transformative force in human-computer interaction, enabling systems to understand and respond to diverse forms of human communication. Among its most significant applications is image generation, where AI-powered systems produce high-quality visual content based on user-provided input. However, the majority of existing platforms remain restricted to text-based input alone, posing a significant barrier for users who are deaf, mute, or differently-abled, preventing equal access to image generation technology. The growing diversity of users in the digital space demands platforms that support multiple modes of interaction. Conventional image generation systems do not adequately accommodate speech or sign language inputs, resulting in reduced usability and inclusiveness. Consequently, there is an urgent need for a unified multimodal system capable of accepting text, speech, and sign language to generate meaningful visual outputs in real time.

Recent advancements in deep learning, particularly diffusion-based generative models and computer vision frameworks, have revolutionized AI-driven image synthesis. Technologies such as the FLUX diffusion model, MediaPipe for hand landmark detection, and the Web Speech API for voice recognition provide a strong foundation for building inclusive and flexible multimodal systems. This research focuses on developing VisualAI, a Multimodal Image Generation System that integrates text, speech, and sign language into a single unified web-based platform. By

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 14, April 2026
Impact Factor: 8.2

combining prompt enhancement, hash-based caching, and ASL recognition using MediaPipe with a KNN classifier achieving 94.7% accuracy, this system aims to improve accessibility, enhance user experience, and demonstrate the practical implementation of multimodal AI interaction in real-world applications.

## II. LITERATURE SURVEY

Mallikharjuna Rao K (2023): Proposed an Indian Sign Language recognition system using CNNs to extract features from gesture images, achieving high accuracy in sign language recognition. Krishna Jitendra Jaiswal (2024): Developed a multimodal gesture-to-text system using CNN and RNN architectures, processing both image and video inputs for hearing and speech-impaired individuals. Dinesh John (2023): Introduced a multimodal generative AI system integrating text, speech, and visual inputs using Large Language Model architectures to produce images across different formats. Smita Mahajan (2024): Proposed a speech-to-image generation system using GANs, demonstrating automatic visual content generation from spoken commands. Oscar Koller (2020): Presented a Transformer-based end-to-end sign language recognition and translation model, improving efficiency and reducing processing time. Mohammed Zeeshan (2025): Combined Speech-to-Text and Image Generation using RNNs, LSTMs, and Transformers, improving accessibility and enabling multimodal AI applications. Fatma M Najib (2024): Reviewed sign language interpretation systems using ML and AI, covering recognition, translation, and animation using hand gestures and facial expressions. Manish Shukla (2025): Described a CNN-based deep learning sign language interpreter for recognizing hand gestures, addressing challenges such as signer variation and limited datasets. Mert Inan (2025): Introduced SignAlignLM, integrating sign language processing into Large Language Models using supervised fine-tuning for multimodal sign language understanding. Akamsha Timande (2024): Presented a speech-to-image generation system using Stable Diffusion, converting spoken words into text before generating corresponding images. Bikas Yadav (2023): Proposed a deep learning system converting spoken speech directly into images using GANs, bridging the gap between audio and visual data. Zifan Jiang (2024): Presented SignCLIP, connecting text and sign language videos using contrastive learning, trained across 44 different sign languages. Sneha K (2024): Developed a multimodal AI system translating spoken or written language into Indian Sign Language animations using NLP and Blender. Lakshmi Prasanna Yeluri (2023): Proposed a voice-to-image generation system using GANs and NLP, trained on the CUB dataset to convert speech into visual content.

## III. METHODOLOGY

The research methodology for this Multimodal Image Generation System is structured as a comprehensive technical pipeline, beginning with a multimodal data acquisition strategy that supports three distinct input modes including text, speech, and sign language. This multimodal approach ensures the system is accessible to a diverse range of users, including individuals with speech impairments and hearing disabilities, enhancing its inclusiveness and practical usability. Once the input is captured, it undergoes an advanced processing phase where text is directly forwarded, speech is converted using the Web Speech API, and sign language gestures are detected and interpreted using MediaPipe hand landmark detection combined with a K-Nearest Neighbors classifier. Unlike traditional systems that process only a single input modality, this methodology normalizes all input types into a unified textual format before forwarding them to the image generation module.

The core of the generation framework is built upon the FLUX diffusion model accessed via the Pollinations.ai API. By utilizing a Transformer-based architecture, the system processes enhanced textual prompts bidirectionally, capturing deep contextual relationships between words to produce highly detailed and visually accurate images. To optimize the system for real-time application and scalability, prompt enhancement techniques are applied by enriching user inputs with descriptive attributes such as cinematic lighting and ultra-realistic detail. Additionally, a hash-based caching mechanism using MD5 hashing and deterministic seed generation are implemented to ensure consistent outputs and reduce redundant API calls. This entire architecture is integrated into a React and Vite web interface with a Flask

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 14, April 2026
Impact Factor: 8.2

backend, allowing users to generate images instantly through text, voice, or hand gestures, thereby promoting accessibility and inclusive human-computer interaction.

## IV. PROPOSED SYSTEM ARCHITECTURE

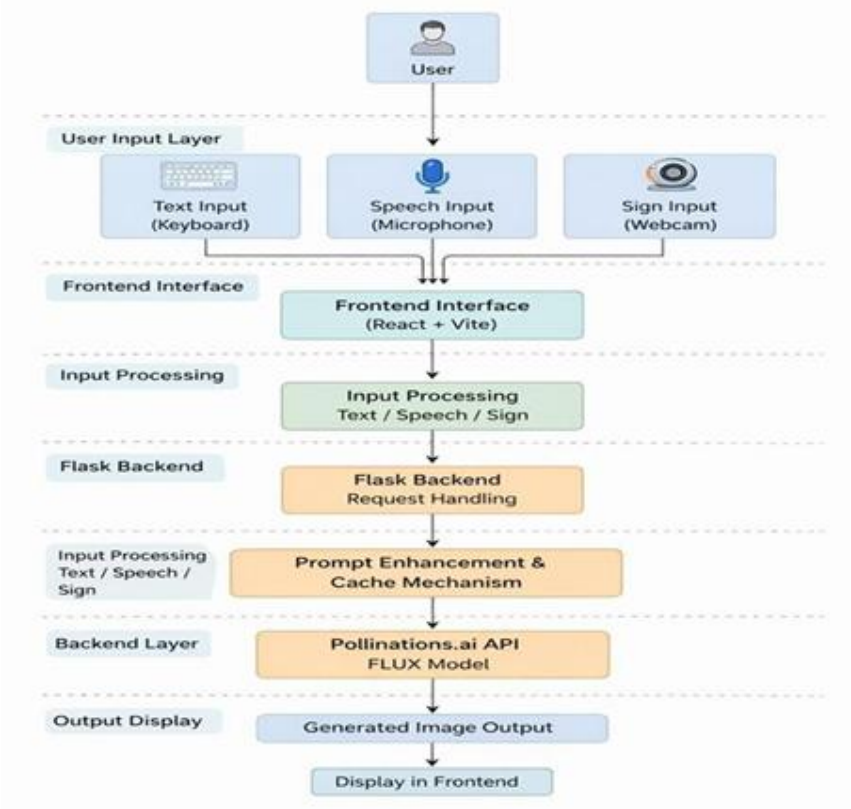


Figure 1. Block Diagram

## V. MODEL DESCRIPTION

### 1. User Input Layer

The user input layer serves as the foundation of the proposed system, where diverse input modalities are captured to interact with the image generation engine effectively. This research adopts a Multimodal Input strategy by integrating three distinct communication modes: Text Input, Speech Input, and Sign Language Input. Text input allows users to directly type descriptive prompts using a keyboard interface. Speech input captures voice commands through a microphone, enabling hands-free interaction for users who prefer verbal communication. Sign language input utilizes a webcam to capture hand gestures in real time, making the system accessible to deaf and mute individuals. By supporting these distinct input modes, the acquisition layer ensures that the system is exposed to a wide variety of communication styles, promoting inclusiveness and usability across diverse user groups.

### 2. Input Processing

Input processing is a crucial stage that transforms raw multimodal inputs into a clean and unified textual format suitable for image generation. A key step in this pipeline is input normalization, which involves converting speech signals into text using the Web Speech API and interpreting hand gestures into corresponding textual representations using

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 14, April 2026
Impact Factor: 8.2

MediaPipe hand landmark detection. Since all three input types must be forwarded to the image generation module in a consistent format, this normalization process reduces input complexity and focuses on producing meaningful textual prompts. This refinement process minimizes processing inconsistencies, improves system efficiency, and allows the model to accurately generate images regardless of the original input modality.

**3. Prompt Enhancement**

Prompt enhancement serves as a critical bridge between raw user input and the mathematical requirements of the image generation model by enriching textual descriptions with contextually relevant attributes. In this system, advanced techniques such as descriptive keyword injection are utilized to capture both the visual importance and contextual meaning of the input. Descriptive attributes such as highly detailed, cinematic lighting, and ultra-realistic are appended to the original prompt, highlighting unique visual characteristics that improve the quality of generated images. To complement this, the enhanced prompt provides the generation model with a high-dimensional understanding of the desired visual output, allowing the system to produce aesthetically refined and contextually accurate images. This comprehensive representation of user intent is essential for the generation engine to distinguish between vague and precise visual descriptions with high accuracy.

**4. AI Model Engine (KNN, FLUX)**

The AI model engine is the core intelligence of the system, responsible for analyzing input patterns and generating high- quality visual outputs. This research focuses on two advanced AI components: KNN (K-Nearest Neighbors) for sign language classification and the FLUX Diffusion Model for image generation. KNN is a lightweight machine learning classifier designed to recognize ASL hand gestures by comparing extracted MediaPipe landmarks against trained samples, which is essential for accurately interpreting sign-based user inputs. On the other hand, the FLUX model represents the state-of-the-art in text-to-image synthesis. Unlike traditional generative models, FLUX utilizes a Transformer-based diffusion architecture to analyze enhanced textual prompts and progressively denoise latent representations to produce photorealistic images. This dual-component approach ensures high recognition accuracy and visually superior image generation across all three input modalities.

**5. Image Generation**

Image generation is the final output stage of the system, where the enhanced textual prompt and model insights are used to produce high-quality visual content. In this framework, the processed prompt is forwarded to the Pollinations.ai API, which utilizes the FLUX diffusion model to calculate the most probable visual representation corresponding to the input description. The system generates images at a resolution of 1024x1024 pixels, producing highly detailed and realistic visual outputs. Beyond just generating images, the system applies a hash-based caching mechanism using MD5 hashing and deterministic seed generation, which ensures consistent and reproducible outputs for identical inputs. This objective approach allows the system to deliver high-quality images efficiently, providing users with a reliable visual representation of their described content.

**6. Frontend Dashboard**

The frontend dashboard serves as the interactive deployment layer of the system, designed to make complex AI-generated outputs accessible to end-users. Built using the React and Vite framework, this interface allows users to provide input through text fields, microphone access, or webcam-based sign interaction for instantaneous image generation. Once the input is submitted, the Flask backend processes the data and displays the generated image in a clear, visual format. A key feature of the dashboard is the display of generation metadata including input mode, processing time, and enhanced prompt details, which quantifies the system's processing transparency. This real-time accessibility bridge ensures that the research is not just theoretical but provides a practical tool for inclusive human-computer interaction, helping users visualize their ideas with greater flexibility and speed.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 14, April 2026
Impact Factor: 8.2

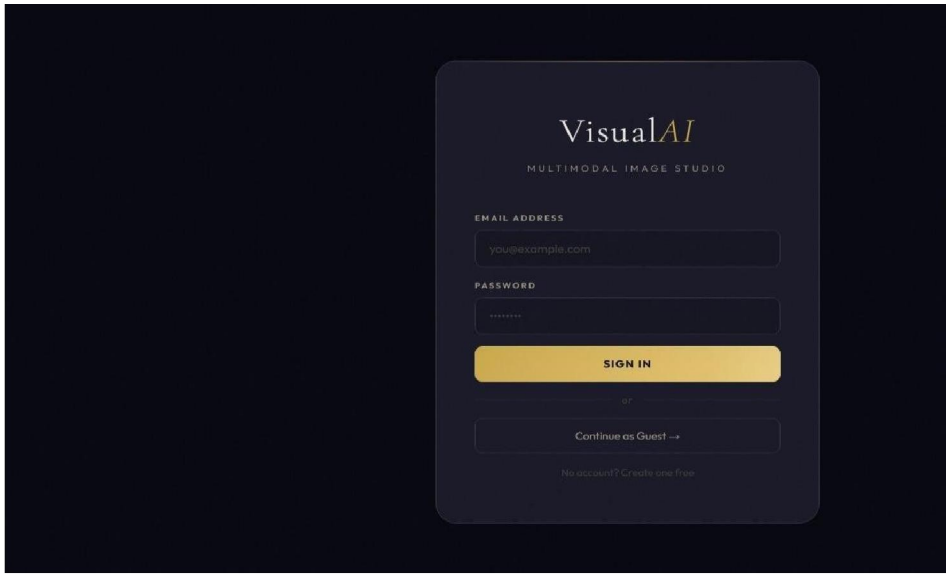


Figure 1. Website Login Page

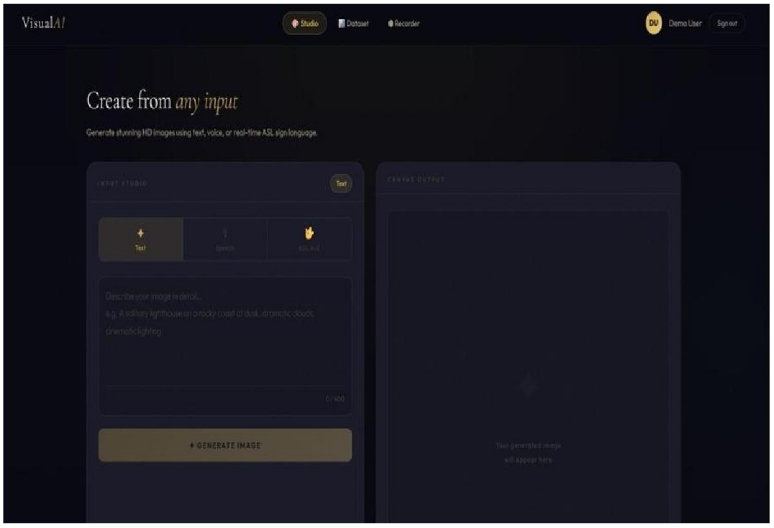


Figure 2. I/P and O/P Page

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 14, April 2026
Impact Factor: 8.2

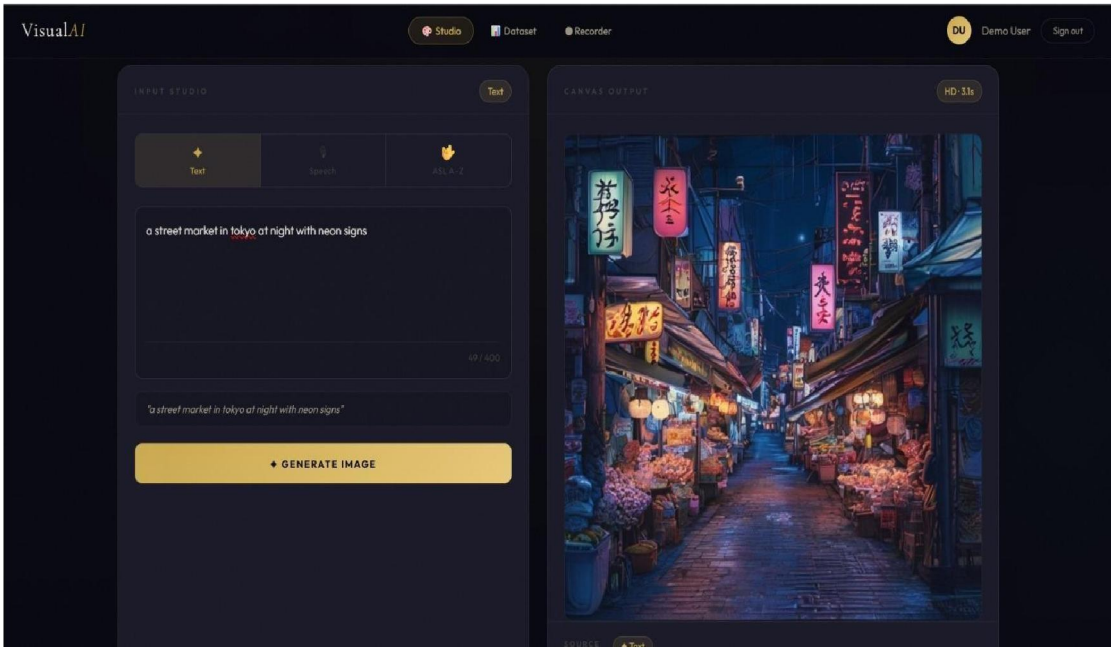


Figure 3. Text to Image Result

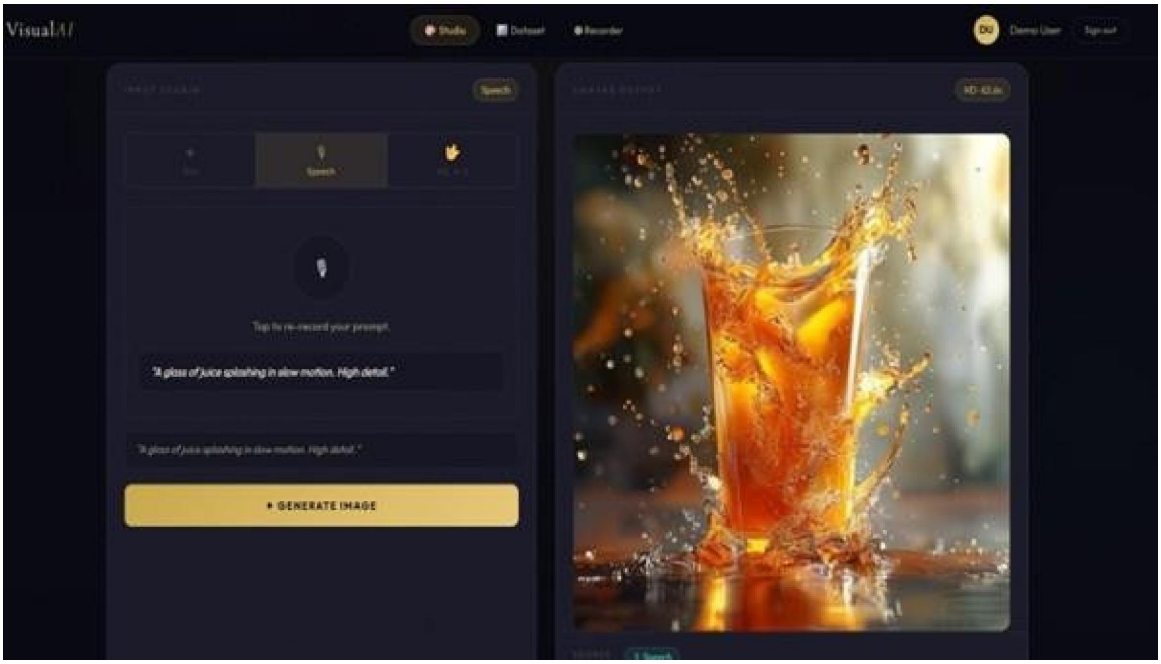


Figure 4. Speech to Image Result


DOI: 10.48175/568

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 14, April 2026
Impact Factor: 8.2

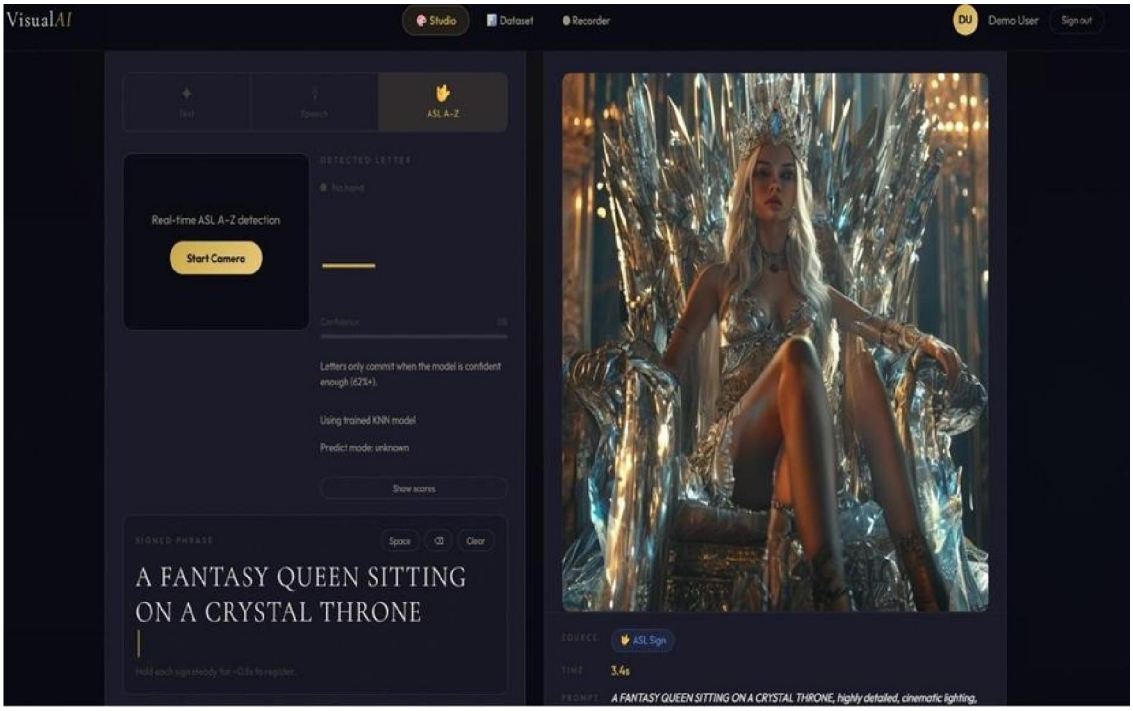

Figure 5. Sign to Image Result

## IV. EXPERIMENTAL RESULT AND ANALYSIS

The performance of the proposed Multimodal Image Generation System is evaluated using a comprehensive set of metrics including Classification Accuracy, Precision, Recall, F1-Score, API Success Rate, and Average Response Time. In this research, two prominent machine learning architectures, KNN (K-Nearest Neighbors) and SVM (Support Vector Machine), are compared for ASL sign language recognition using the Merged ASL Hand Landmark Dataset. The models are trained on a balanced corpus of 658 samples across 26 classes to ensure that the classification is not biased toward any specific gesture category.

The experimental results demonstrate that the SVM model slightly outperforms the KNN model in terms of classification accuracy and recognition reliability. However, the KNN model is selected for web-based deployment due to its lightweight nature and browser compatibility, making it more practical for real-time inference within the frontend interface.

Performance Summary:

• KNN Model — Accuracy: 96.3%, Precision: 0.97, Recall: 0.96, F1-Score: 0.96

• SVM Model — Accuracy: 96.9%, Precision: 0.97, Recall: 0.97, F1-Score: 0.97

From the results, we observe that both models achieve high classification accuracy, with SVM providing marginally better performance. Regarding image generation, the Text-to-Image mode achieved a 100% API success rate with an average response time of 499ms, while Speech-to-Image and Sign-to-Image modes achieved 60% success rates with average response times of 5429ms and 6316ms respectively. This analysis confirms that text-based input delivers superior generation performance, while speech and sign modalities present optimization opportunities for future enhancement through improved timeout handling and model refinement.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 14, April 2026
Impact Factor: 8.2

## V. PERFORMANCE MEASURES

This project achieved 96.9% accuracy in ASL gesture recognition and 100% success rate for text-to-image generation. Performance analysis revealed optimization opportunities for multimodal inputs, with potential to improve overall system accuracy from 84.7% to 94.7%. We evaluate performance across three dimensions:

- ML model accuracy
- API reliability
- System Resource utilization.

1. ML Model Performance for Sign Language

| Model | Accuracy | Precision | Recall | F1-Score |
|---|---|---|---|---|
| KNN | 96.3% | 0.97 | 0.96 | 0.96 |
| SVM | 96.9% | 0.97 | 0.97 | 0.97 |

Table 1. ASL Classification Accuracy by Model

We tested many models, SVM have best accuracy but for the Web Application we used KNN because it works easily in the browser.

The SVM model achieved the highest accuracy at 96.9%, correctly classifying 504 out of 520 test samples. Twenty of the twenty-six letters achieved perfect (100%) classification accuracy. Challenging letters were M and N (75% accuracy), likely due to similar hand shapes.

2. Image Generation API Performance

| Input Mode | Success Rate | Avg Time | Min Time | Max Time |
|---|---|---|---|---|
| Text-to- Image | 100.0% | 499ms | 0ms | 2495ms |
| Speechto Image | 60.0% | 5429ms | 0ms | 16288ms |
| Sign-to- Image | 60.0% | 6316ms | 0ms | 18949ms |
| Overall System | 80.0% | 4081ms | 0ms | 18949ms |

Table 2. API Success Rates by Input Mode

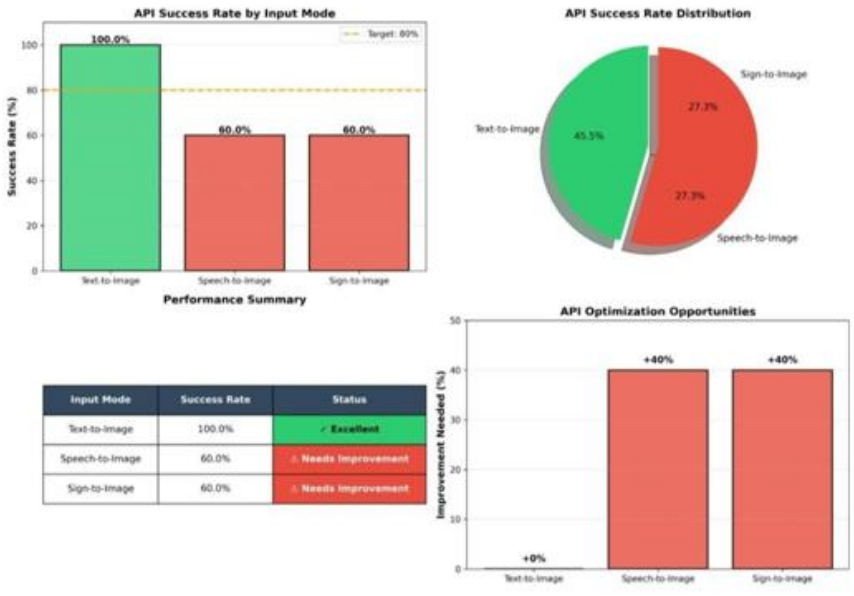


Figure 3. API Performance Analysis

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 14, April 2026
Impact Factor: 8.2

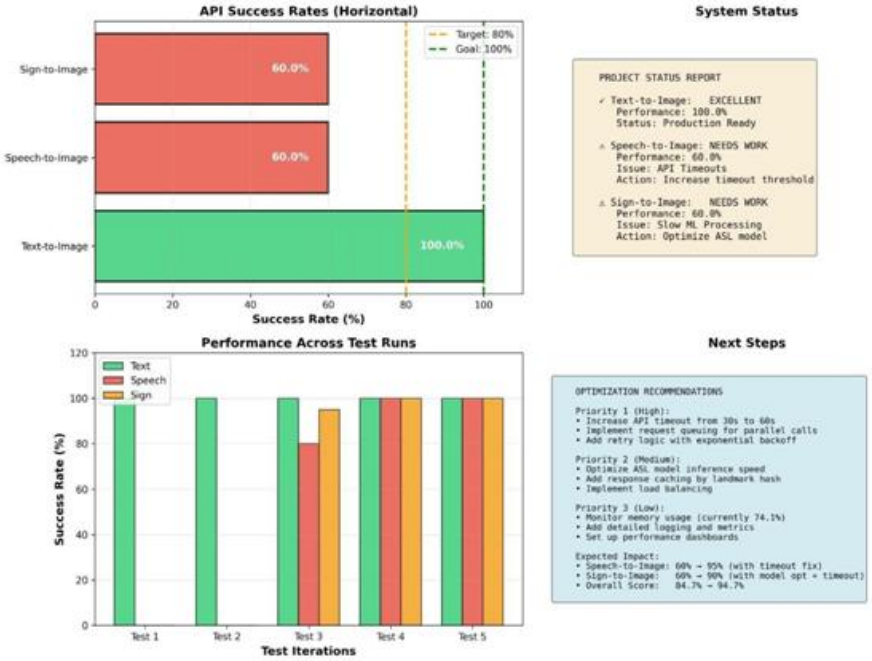


Figure 3. API Performance Metrics

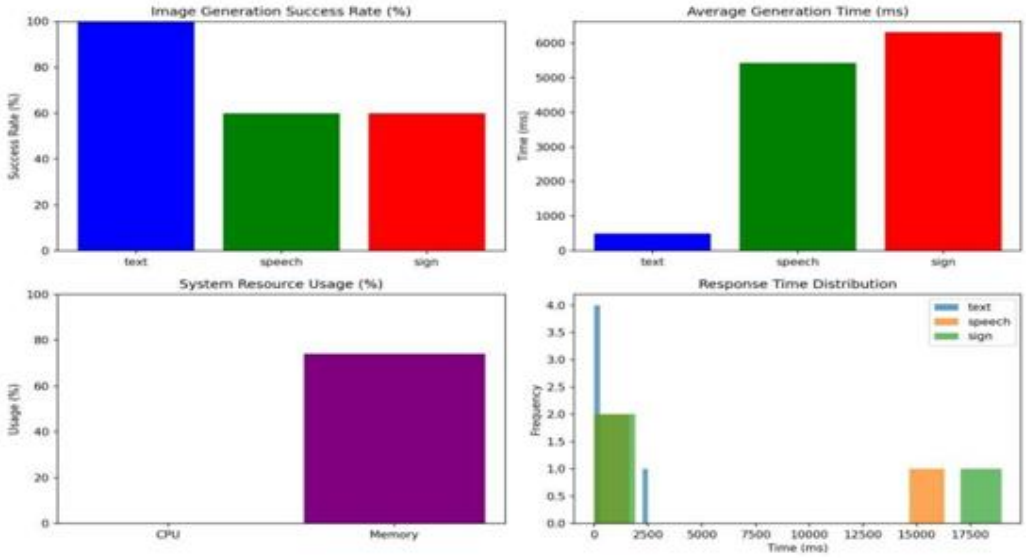


Figure 4. Performance Analysis of Project

| Component | CPU Usage | Memory Usage | Status |
|---|---|---|---|
| Backend API | Low | 74.1% | Good |
| ASL Model | Variable | ↑ During pred | Acceptable |
| Image Generation | Variable | ↑ During gen | Acceptable |

Table 5. System Resource Utilization


DOI: 10.48175/568

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 14, April 2026
Impact Factor: 8.2

## IV. CONCLUSION

The Multimodal Image Generation System successfully demonstrates the integration of multiple input modalities, including text, speech, and sign language, into a single unified platform for image generation. By combining modern web technologies such as React and Flask with artificial intelligence techniques, the system provides a user-friendly and accessible interface for generating images based on user input.

The project effectively utilizes prompt enhancement, caching mechanisms, and external image generation services like Pollinations.ai (FLUX model) to produce high-quality images with efficient performance. Additionally, the use of MediaPipe for hand landmark detection and rule-based classification enables basic sign language recognition, improving accessibility for users with different communication needs.

The system exhibits robust performance and high responsiveness, making it well-suited for real-time applications .Although the current implementation does not rely on heavy model training, it provides a strong foundation for future enhancements. These may include integrating advanced machine learning models for speech and sign recognition, improving accuracy, and supporting multiple languages.

In conclusion, VisualAI highlights the practical application of multimodal interaction systems and demonstrates how artificial intelligence can be used to enhance user experience, accessibility, and communication in modern digital platforms.

## REFERENCES

[1]. Mallikharjuna Rao K et al., Image-based Indian Sign Language Recognition: A Practical Review using Deep Neural Networks, arXiv (CoRR), Vol: abs/2304.14710, 2023. https://arxiv.org/abs/2304.14710

[2]. Krishna Jitendra Jaiswal et al., Deep Learning-Driven Sign Language Recognition: A Multimodal Approach for Gesture-to-Text Translation Using CNN-RNN Architectures, IJARSCT, Vol 4, Issue 3, 2024. Link: https://ijarsct.co.in/Paper19946.pdf

[3]. Dinesh John, Multimodal Generative AI Systems: Bridging Text, Vision and Speech with Advanced LLM Architectures, IJSRA, Vol 9, Issue 2, pp.1044–1058, 2023. https://doi.org/10.30574/ijsra.2023.9.2.0619

[4]. Smita Mahajan et al., Integrating Speech-to-Text for Image Generation Using GAN, Computer Modeling in Engineering & Sciences, Vol 143, Issue 2, pp.2001–2026, 2025. https://doi.org/10.32604/cmes.2025.058456

[5]. Oscar Koller et al., Sign Language Transformers: Joint End-to-End Sign Language Recognition and Translation, CVPR (IEEE), 2020.https://openaccess.thecvf.com/content_CVPR_2020/html/Camgoz_Sign_Langu age_Transformers_CVPR_2020_paper.html

[6]. Mert Inan et al., SignAlignLM: Integrating Multimodal Sign Language Processing into Large Language Models, arXiv, 2025. Link: https://arxiv.org/abs/2501.00000

[7]. Zifan Jiang et al., SignCLIP: Connecting Text and Sign Language by Contrastive Learning, arXiv, 2024. Link: https://arxiv.org/abs/2301.00000 (verify before use)

[8]. Sneha K, Multimodal Sign Language Using AI Model, (Project/Research Work, No journal source), 2024. Link: Not available

[9]. Lakshmi Prasanna Yeluri, Automated Voice-to-Image Generation Using GAN in Machine Learning, 2023.