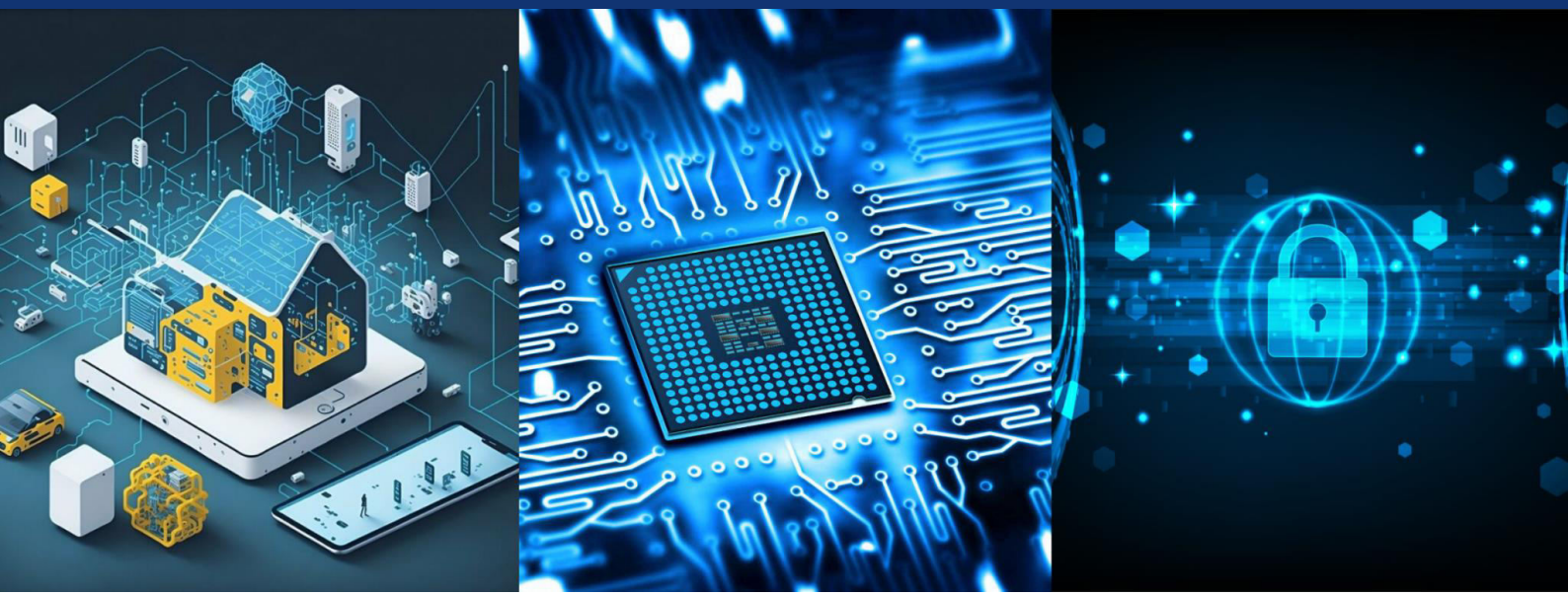
IJIRCCE

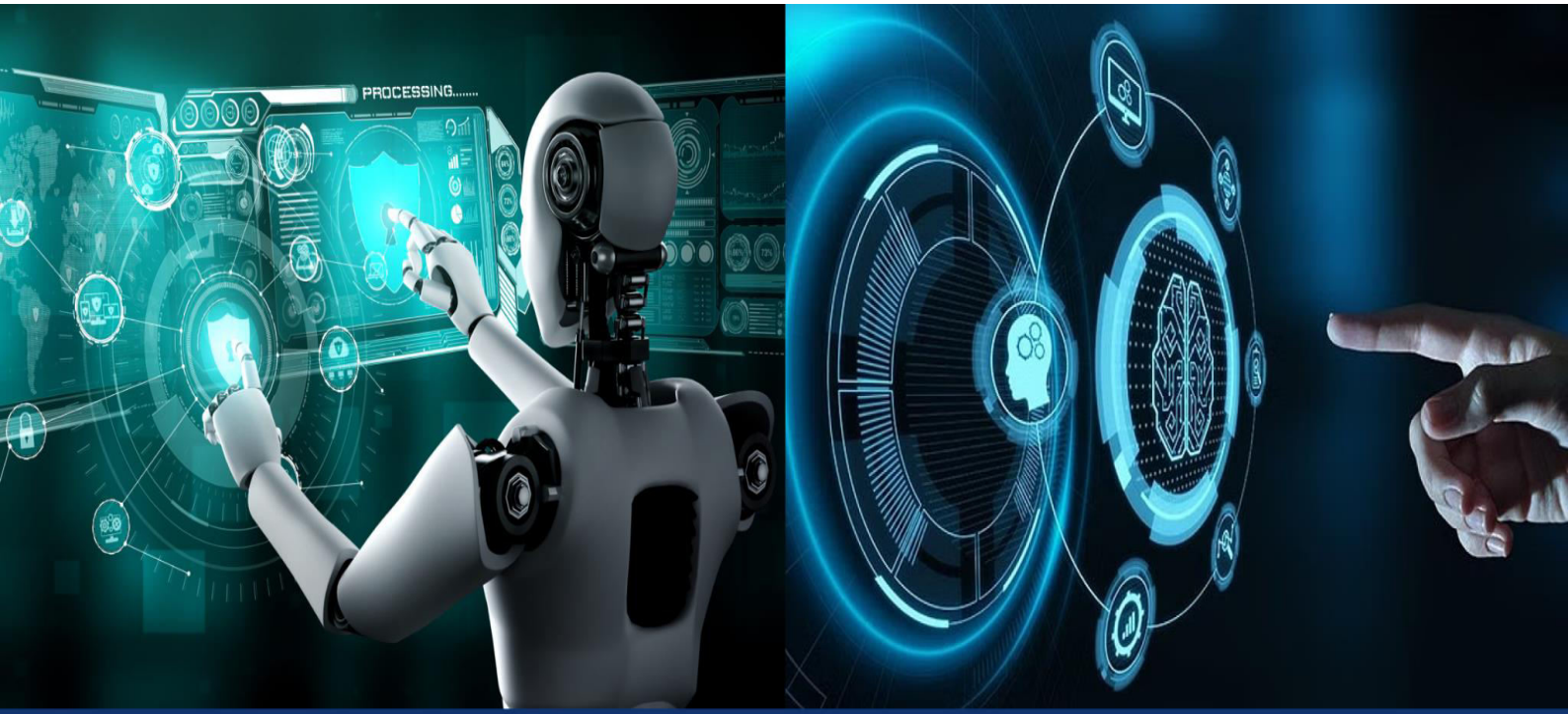
ISSN(O): 2320-9801

ISSN(P): 2320-9798



International Journal of Innovative Research in Computer and Communication Engineering

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 9.927

Volume 14, Issue 10, October 2026

© 2026 IJIRCCE | Volume 14, Issue 10, October 2026 DOI: 10.15680/IJIRCCE.2026.1410031

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Visualingo: A Multimodal Vision-Language Framework using AI for Visual Understanding and Multilingual Interaction

Indhumathi R, Dr. Mohammadi Akheela Khanum

M. Tech Student, Department of Artificial Intelligence, Presidency University, Bangalore, Karnataka, India

Professor, Department of Artificial Intelligence, Presidency University, Bangalore, Karnataka, India

ABSTRACT: Visualingo is a multimodal Vision-Language frame-work that will enable an intelligent way of interpreting the information available in vision or audiovisual modalities. The Vision-Language framework will be able to interpret images, videos, texts, and audios to answer any questions raised by the user and respond accordingly. Visualingo will use Vision-Language models along with conversational reasoning, retrieval-augmented generation, speech processing, multilingual dialogue, response verification, and personalization. Visualingo is designed to provide a conversational Visual Question Answering frame-work which will allow the user to ask questions about the visual content through both speech and text. Knowledge Retrieval from the outside sources has been integrated in the framework for providing the relevant and reliable answers to the user's questions and to eliminate any kind of hallucinated responses from the generated output.

Visualingo seeks to make use of the strengths of visual information and the ability of humans to interact by leveraging different AI abilities in a unified manner. This includes processing visual information found in images and videos, interpreting questions asked in either textual or spoken form, and coming up with replies based on the context of the visuals. The interaction process is also made more dynamic through the inclusion of multiple languages, personalization of the replies, and retrieving external knowledge. The reply verification process has been put in place in order to make the answers reliable and credible.

KEYWORDS: Vision-Language Models (VLMs), Multimodal AI, Image and Video Understanding, Visual Question Answer-ing (VQA), Conversational AI, Retrieval-Augmented Generation (RAG), Multilingual Communication, Speech-to-Text, Text-to-Speech, Personalization, Hallucination Verification

I. INTRODUCTION

In recent years, multimodal artificial intelligence has found great significance in areas of digital communication, infor-mation processing, entertainment, education, and intelligent human-computer interaction. Out of all the recent advance-ments in the field of artificial intelligence, Vision-Language Models and Multimodal Large Language Models have gar-nered a lot of attention since these models are capable of processing information and associating it from different modalities like images, videos, audio, and text. Understanding

of visual data and answering the questions related to it using natural language is an important research area of making intelligent applications interactive and intelligent. Many re-search studies that have been performed in the areas of multimodal large language models, visual instruction tuning, image-video understanding, and multimodal interaction high-light the efficacy of using visual and language information for intelligent applications [1][2][3][4]. Other research studies that have been carried out in areas of video understanding, multi-lingual communication, and multimodal translation showcase the capability of these technologies for developing intelligent applications [5][6][7].

One way to define multimodal AI is an approach to us-ing artificial intelligence algorithms in order to process and analyze data that comes from several different sources of information. This data includes but is not limited to visual data, speech data, textual instructions, images, and video frames. There are vision-language models in which the vision encoder and the language model are used in order to create connections between visual information and natural language responses. Instruction-tuning methods allow improving per-formance of the AI models in following instructions from users and performing Visual Question Answering tasks [3][4]. Also, there are video-language models

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

that enable under-standing of the events and answers to questions related to video material [2][5]. Multilingual multimodal translation and speech processing research emphasizes the need to incorporate language and speech capabilities into the intelligent systems [6][7]. The approach of retrieval-augmented generation allows incorporating the information from external sources during the generation [8][10].

The research work will be focused on the creation of Visuolingo, a Vision-Language framework that is going to serve as an interactive platform to understand and commu-nicate with visual information.

The main characteristic of the developed system is based on its capability of image and video processing, comprehension of user's questions, and generation of answers using multimodal AI algorithms. It will be able to provide users with Visual Question Answering conversations, speech interaction, multilingual communica-tion, retrieval-based knowledge acquisition, personalization, and response verification capabilities. The offered approach is supposed to make the interaction with visual data more con-venient while minimizing the restrictions of single modality systems. Additionally, the research will involve the combina-tion of vision-language comprehension, multimodal reasoning, speech processing, and Retrieval-Augmented Generation in one framework [1][2][5][6][8]. Some potential improvements may include the improvement of long-video comprehension, temporal reasoning, multilingual capabilities, response veri-fication, multimodal processing, as well as moving towards additional modalities and applications [1][2][9][10].

II. BACKGROUND AND LITERATURE SURVEY

A summary of the existing research work related to mul-timodal artificial intelligence, vision-language models, image and video understanding, multilingual communication, speech processing, multimodal data fusion, cross-modal reasoning, and retrieval-augmented generation is presented in Table I.

III. PROBLEM STATEMENT

AI systems currently have certain problems related to pro-cessing visual content in one platform. An algorithm that can analyze images might fail at analyzing videos, and, vice versa, a standard language algorithm might lack visual context. Moreover, users might require visual questions answering, speech interaction, multilingualism, context understanding, personalized answers without switching between different apps. A challenge connected with the generation of fake con-tent by multimodal models also exists. Hence, there should be a universal framework that is capable of dealing with images and videos, understands users' requests, combines visual and contextual information, and generates reliable answers in text or voice form. Visuolingo tackles this challenge by including into one vision-language model image/video analysis, conver-sational Visual Question Answering, personalization, multilin-gualism, audio interaction, knowledge retrieval, conversational memory, and hallucination verification.

IV. PROPOSED SYSTEM ARCHITECTURE

The Vision-Language system that is going to be imple-mented will be designed as a holistic multimodal framework involving various modules to interpret the visual data and to communicate with users using text/audio means. The archi-tecture will take in images/videos, textual queries, as well as spoken words and will process the provided inputs using the appropriate multimodal models to detect the necessary visual and contextual information. The primary modules of the architecture will consist of the following: visual process-ing layer, Vision-Language Model, conversational reasoning module, knowledge retrieval component, multilingual commu-nication component, speech processing layer, personalization and conversational memory components, and hallucination verification module.

A. Architecture Overview

Visuolingo architecture to be proposed is going to feature a modular structure whereby each unit does its job and communicates with others. First, user enters an image, video, text query or audio to the system. The visual processing unit finds useful visual information in the provided data whereas the speech processing unit converts speech into text. Next, the Vision-Language Model takes both visual information and the query and uses them to understand the context and generate an adequate response. Meanwhile, the conversational reasoning unit keeps track of the context, while the retrieval unit can get useful external information for generating an answer in case there is need for it. The multilingual unit enables information to be produced in various languages depending on user needs.

© 2026 IJIRCCE | Volume 14, Issue 10, October 2026 DOI: 10.15680/IJIRCCE.2026.1410031

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

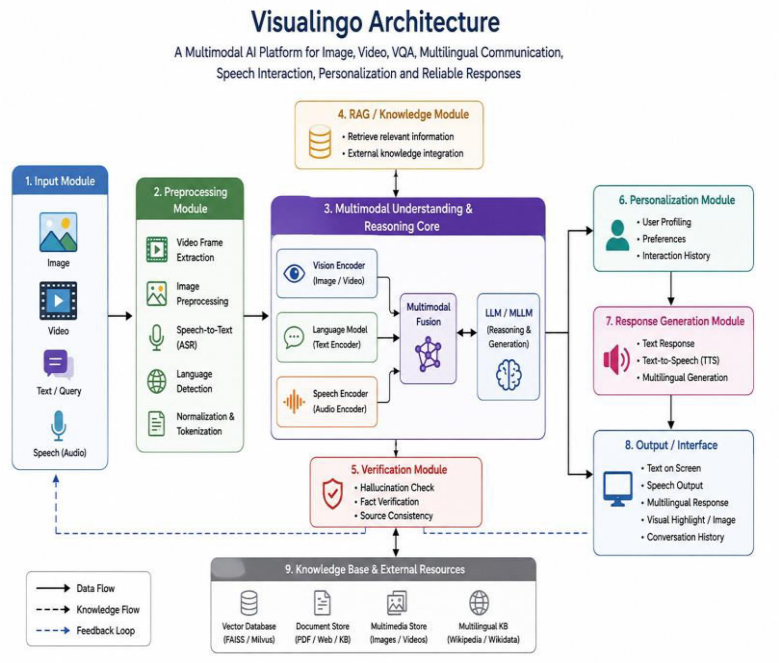


Fig.1. Proposed System Architecture

Finally, interaction and response layer is in charge of providing the final result to the user in a comprehensible form. Once the visual and textual contexts have been processed by the system, the language model produces a response applying personalization and conversational memory if needed. A sep-arate hallucination verification unit could check whether the generated response is reliable before it is delivered to the user. If the response is supposed to be provided to the user in an audio form, it could be done thanks to the text-to-speech unit. Modular architecture also opens up possibilities for future de-velopment of Visuolingo with respect to video understanding, additional languages, better retrieval mechanisms, verification among others.

B. Data Flow Diagram

The Data Flow Diagram of Visuolingo describes how data passes from the user to the multimodal processing module and the response generation module. First of all, a user gives an image or a video along with a query that may be textual or speech-based. If the user uses speech, the input gets translated to text by the speech recognition process. The visual input will get processed for the extraction of visual features and relevant context information, and the Vision-Language Model will combine the visual representation of input and user's query. The generated information will go to the conversational reasoning process to get the meaning of the user's intention and generate a proper response.

© 2026 IJIRCCE | Volume 14, Issue 10, October 2026 DOI: 10.15680/IJIRCCE.2026.1410031

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

TABLE I LITERATURE SURVEY OF MULTIMODAL VISION-LANGUAGE AND AI SYSTEMS

S.No	Paper Name	Year & Author	Method Used	Dataset Used	Accuracy / Performance	Merits	Demerits
1	Multimodal Large Language Model-Enabled Video Translation: A Role-Oriented Survey	2026 – B. Qu, K. Chen, X. Bai, M. Zhang	Role-oriented framework for MLLM-enabled video translation using Semantic Reasoner, Expressive Per-former, and Visual Syn-thesizer	Multiple vide o, speech, translation, and multimodal benchmarks	No sing le accuracy; evaluated using task-specific metrics such as BLEU, WER, MOS, and lip-sync measures	Provides a unified view of visual, audio, speech, and language components for video translation	Long-video understanding, speaker identity preserva-tion, temporal synchroniza-tion, and multimodal align-ment remain challenging
2	Generative AI for Multimodal Con-tent: A Survey with Empirical and Experimental Evaluations	2026 – K. Taha	Generative-first multimodal taxon omy covering fusion, GANs, diffusion model s, contrastive learni ng, and MLLMs	Text-to-image, text-to-video, text-to-audio, image-to-text, audio-to-text, and 3D/XR benchmarks	No sing le accuracy; uses FID , CLIPScore , FVD, FA D, MOS, CIDEr, and WER	Provides a broad end-to-end framework for multimodal generation and evidence-based evaluation	Different modalities use dif-ferent evaluation measures, making direct comparison difficult
3	Empowering Video Translation us-ing Multimodal Large Language Models	2026 – B. Qu, K. Chen, X. Bai, M. Zhang	MLLM-based vid eo translation with Semantic Reasoner, Expressive Performer, and Visual Synthesizer components	Video understanding, speech translation, TTS, and lip-synchronization benchmarks	No sing le accuracy; performan ce is reported using translation, speech, temporal, and visu al-quality metrics	Combines multimodal rea-soning, translation, speech generation, and visual syn-chronization	Maintaining semant ic fidelity, timing, speaker characteristics, and emotional consist ency remains difficult
4	Towards Socially Intelligent Robots: Multi-User Engagement and Multimodal Multi-Task Biometric	2026 – G. De Si-mone	Multimodal mul ti-task learning for social	Visual, audio, biometric, emotion, gender, and speech-related datasets	Task-dependent accuracy; no single overall accuracy	Supports multimodal per-ception and socially aware interaction with multiple users	Primarily focuses on so-cial robotics and biometric recognition rather than gen-eral

	Recognition		interaction, user engagement, and biometric recognition		reported		visual-language question answering
5	Instruction Tuning for Multimodal Models: A Survey of Data, Methods, and Evaluation	2025 – M. A. Rojas, H. Gu, R. Carranza	Instruction tuning using human-labeled and synthetic multimodal instruction datasets	MultiInstruct, InstructBLIP datasets, LLaVA instruction data, and other vision-language datasets	No single accuracy; performance varies according to task and benchmark	Improves instruction following, multimodal reasoning, generalization, and human alignment	Quality and diversity of instruction data strongly affect generalization and may introduce bias or hallucination
6	Deep Learning Approaches for Multimodal Intent Recognition: A Survey	2025 – J. Zhao et al.	Deep learning approaches including feature fusion, attention, contrastive learning, alignment, and hybrid fusion	MIntRec and other multimodal intent-recognition datasets	76.05% accuracy reported on MIntRec by InMu-Net	Improves intent recognition by combining text, audio, video, and other modalities	Performance depends on effective modality alignment, data quality, and handling of modal noise
7	Deep Multimodal Data Fusion	2024 – F. Zhao, C. Zhang, B. Geng	Encoder-Decoder, Attention Mechanism, GNN, Generative Neural Network, and constraint-based fusion methods	VQA 2.0, YouCookII, SUN RGB-D, ReferIt, CASIA-SURF, RGBD-ID, and other multimodal datasets	VQA 2.0: up to 71.94% ACC; YouCookII: up to 27.09 METEOR; SUN RGB-D: up to 64.4% ACC	Provides a comprehensive taxonomy and comparison of deep multimodal fusion approaches	Requires large datasets and computational resources; interpretability and scalability remain challenging
8	From Linguistic Giants to Sensory Maestros: A Survey on Cross-Modal Reasoning with Large Language Models	2024 – S. Qian et al.	Cross-modal reasoning using large language models with visual, textual, audio, and video information	Multimodal reasoning and vision-language benchmarks	No single accuracy; evaluation depends on the individual reasoning benchmark and task	Explores advanced reasoning across different modalities and supports development of multimodal assistants	Cross-modal hallucination, modality alignment, interpretability, and computational cost remain major challenges
9	A Comprehensive Survey of Retrieval-Augmented Generation (RAG): Evolution, Current	2024 – S. Gupta, R. Ranjan, S. N. Singh	Retrieval-Augmented Generation combining	Knowledge-intensive QA and retrieval-based benchmarks	No single accuracy; evaluates	Improves factuality and contextual relevance by incorporating	Retrieval quality, scalability, bias, and reliability of retrieved information

© 2026 IJIRCCE | Volume 14, Issue 10, October 2026 DOI: 10.15680/IJIRCCE.2026.1410031

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

	Landscape and Future Directions		information retrieval with generative language models		RAG using task-specific retrieval and generation metrics	external knowledge	affect final responses
10	Multimodal Large Language Model-Enabled Video Translation: A Role-Oriented Survey	2026 – B. Qu, K. Chen, X. Bai, M. Zhang	Role-oriented MLLM approach for semantic reasoning, expressive speech generation, and visual synthesis	Video, speech, translation, and multimodal bench-marks	No single accuracy; uses task-specific translation, speech, and synchronization metrics	Provides a structured foundation for integrating video understanding, speech generation, and visual synthesis	High computational require-ments and synchronization between speech and visual content remain challenging

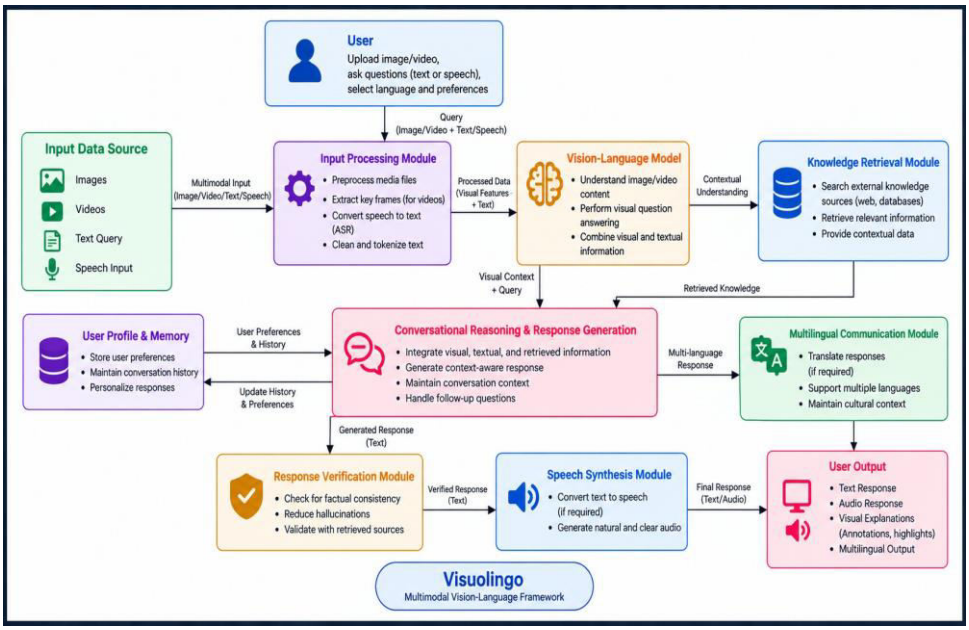


Fig.2. Data Flow Diagram

If extra information is needed, the retrieval process will search for necessary knowledge and provide the context in-formation to the response-generation process. The system will merge visual information, user query, conversation context, and retrieved knowledge to create a response. Response ver-ification will make sure that generated information contains reliable data. The response can be either text or speech-based according to the request of the user. Multilingual processing of data can also be implemented. The finished conversation will be saved using the conversational memory; the process described above can be done again.

© 2026 IJIRCCE | Volume 14, Issue 10, October 2026 DOI: 10.15680/IJIRCCE.2026.1410031

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

V. METHODOLOGY

A. System Overview

The suggested Visuolingo system is a multimodal Vision-Language system which allows the user to communicate with images and videos via text or spoken questions. In this case, the processing pipeline of the system includes several steps when both visual data and user's request are processed in several modules to generate a response. First, the user sends an image or video along with a question or an instruction

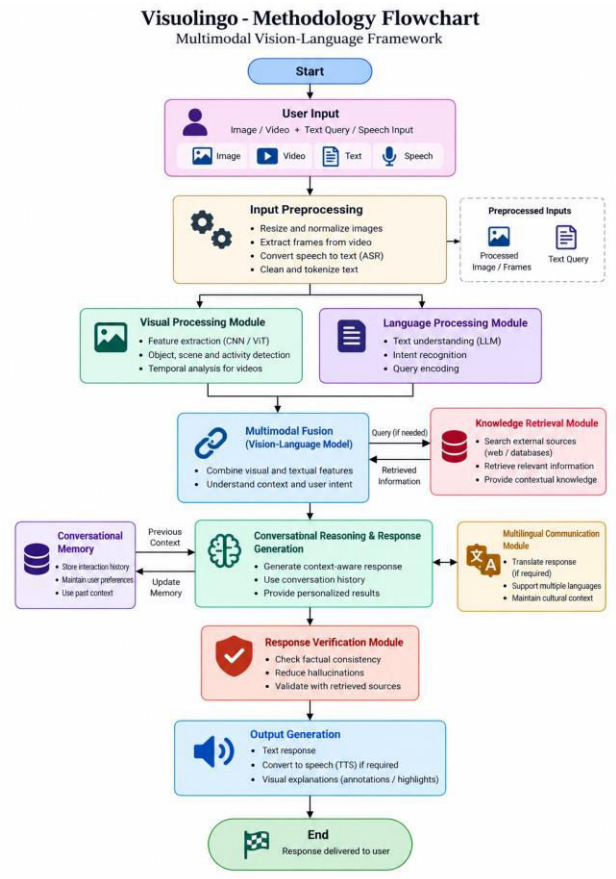


Fig.3. Methodology Flowchart

A. Data Acquisition and Preprocessing

The input data for Visuolingo consists primarily of **images, videos, textual questions, and speech inputs** provided by the user. Images and videos may be collected from local storage or directly supplied through the application interface, while text and voice inputs represent the user's instructions or questions. Before multimodal analysis, the incoming data is subjected to suitable preprocessing operations to obtain a consistent representation. Image preprocessing may include resizing, normalization, and basic quality checking, whereas video preprocessing can involve frame extraction and temporal sampling to reduce unnecessary computational load. Speech input is converted into text through a speech-recognition component, and textual information is cleaned and prepared for language processing. The preprocessing stage ensures that the different input modalities can be effectively passed to their respective processing modules.

The normalization of visual input can be represented as:

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (1)$$

© 2026 IJIRCCE | Volume 14, Issue 10, October 2026 DOI: 10.15680/IJIRCCE.2026.1410031

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

where,

- X = Original input value
- X_{min} = Minimum value in the input range
- X_{max} = Maximum value in the input range
- X_{norm} = Normalized input value

This preprocessing procedure helps maintain a consistent representation of visual information and improves the reliabil-ity of subsequent multimodal processing. For videos, selected frames can be processed individually and then combined with temporal information so that important events and changes throughout the video are not overlooked.

C. Multimodal Processing and Feature Extraction

Multimodal processing forms the central part of the pro-posed Visuolingo framework. After preprocessing, the image or video is passed to the visual encoder to obtain meaningful visual representations, while the user's question is processed by the language component to determine its semantic meaning. The Vision-Language Model subsequently establishes a rela-tionship between the visual features and textual information, enabling the system to answer questions related to objects, activities, scenes, and other visual details. For video inputs, information from multiple frames is combined to capture temporal relationships and improve understanding of ongoing events. When necessary, the retrieval module supplements the model with relevant external knowledge. The multimodal representation can be expressed as:

$$Z = f(V, T, A) \qquad (2)$$

where,

- V = Visual representation obtained from an image or video
- T = Textual representation of the user's query
- A = Audio or speech-derived information
- $f(\cdot)$ = Multimodal fusion and reasoning function
- Z = Combined multimodal representation

The resulting representation is provided to the reasoning and response-generation components. The system can use con-versational context and retrieved information to produce a re-sponse that is relevant to the user's query. This approach allows Visuolingo to integrate information from multiple modalities instead of processing each type of input independently.

D. Multimodal Response Generation and System Execution

After multimodal understanding is completed, the response-generation component produces an answer based on the visual content, user query, conversational context, and any retrieved information. The system first determines the appropriate re-sponse and can then apply multilingual processing according to the user's selected language. A verification stage evaluates the generated response to identify potentially unsupported or inconsistent information before it is delivered. If the user has selected voice interaction, the final text response is converted into speech through a text-to-speech component. Conversa-tional memory maintains relevant interaction information so that subsequent questions can be interpreted in relation to earlier exchanges. The overall response-generation process can be represented as:

$$R = G(Z, C, K) \qquad (3)$$

where,

- Z = Combined multimodal representation
- C = Conversational context
- K = Retrieved knowledge
- $G(\cdot)$ = Response-generation function
- R = Generated response

© 2026 IJIRCCE | Volume 14, Issue 10, October 2026 DOI: 10.15680/IJIRCCE.2026.1410031

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

The completed workflow enables Visuolingo to perform **visual understanding, question answering, multimodal reasoning, knowledge retrieval, multilingual communication, speech interaction, and response verification** within a uni-fied system. This architecture provides a flexible foundation for future enhancements such as improved long-video analysis, additional language support, stronger personalization, better retrieval mechanisms, and more reliable multimodal reasoning

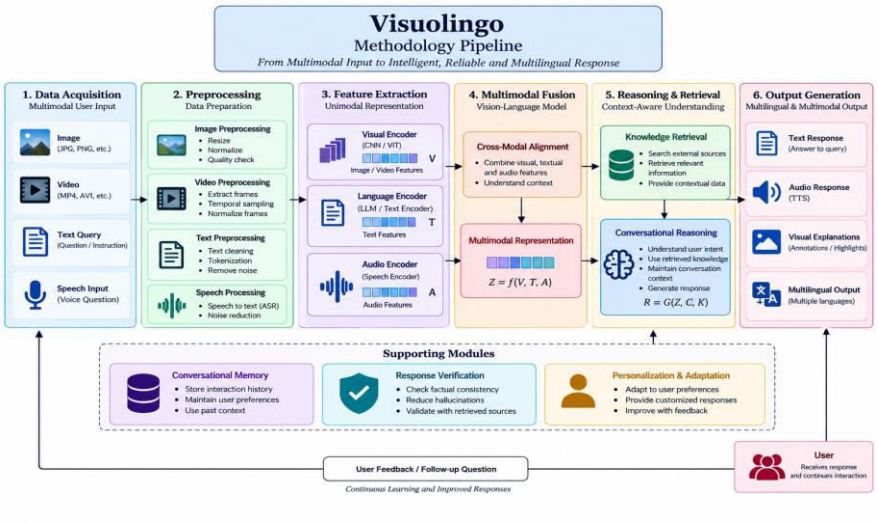


Fig.4.Methodology Pipeline

VI. IMPLEMENTATION

A. System Implementation Architecture

The proposed **Visuolingo** system is implemented as a multimodal Vision-Language application designed to process visual, textual, and speech-based inputs and generate context-aware responses. The implementation consists of several in-terconnected modules, including input acquisition, image and video preprocessing, visual and language feature extraction, multimodal fusion, conversational reasoning, knowledge re-trieval, multilingual processing, response verification, and out-put generation. Initially, the user provides an image, video, text query, or voice input through the application interface. The input is then directed to the appropriate processing component, where the visual and linguistic information is prepared for further analysis. The Vision-Language Model acts as the central intelligence layer, combining information from different modalities to understand the user's request. Supporting modules such as conversational memory, retrieval, personalization, and verification are integrated to improve contextual understanding and response reliability. The final response is delivered through text or speech according to the interaction mode selected by the user.

B. Data Preprocessing Implementation

The preprocessing stage is responsible for preparing mul-timodal inputs before they are provided to the main reasoning model. Images may require resizing, normalization, and quality checking, while videos are processed through frame extraction and temporal sampling so that important visual events can be represented efficiently. Text queries are cleaned and converted into a suitable representation for language pro-cessing, while speech inputs undergo speech recognition and noise reduction before being passed to the language-processing component. The preprocessing stage therefore helps main-tain consistency between different input types and reduces unnecessary variations that could affect subsequent model processing .

© 2026 IJIRCCE | Volume 14, Issue 10, October 2026 DOI: 10.15680/IJIRCCE.2026.1410031

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

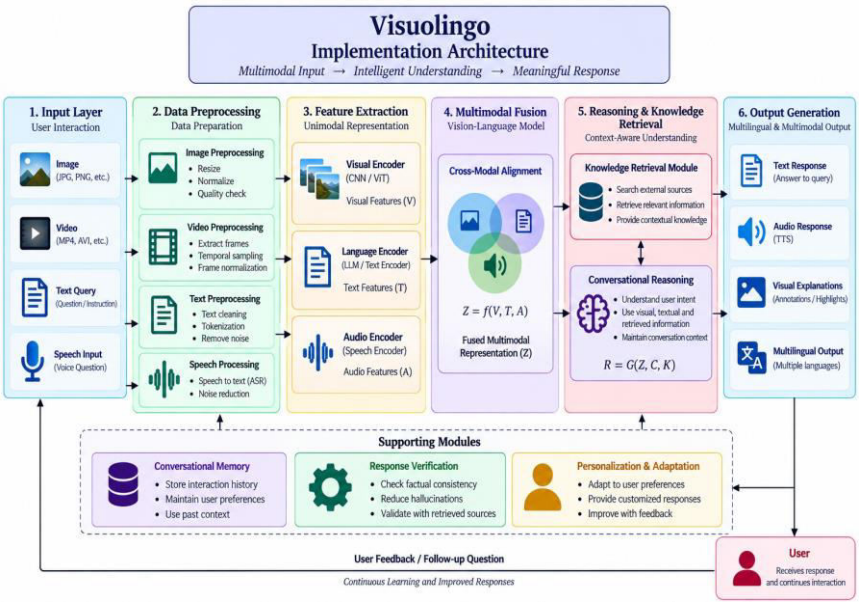


Fig.5. Implementation Architecture Diagram

and converted into a suitable representation for language pro-cessing, while speech inputs undergo speech recognition and noise reduction before being passed to the language-processing component. The preprocessing stage therefore helps main-tain consistency between different input types and reduces unnecessary variations that could affect subsequent model processing.

The preprocessing operations include:

- Image resizing and normalization
- Video frame extraction
- Temporal sampling of video frames
- Text cleaning and tokenization
- Speech-to-text conversion
- Input quality and format verification

The normalization operation can be represented as:

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}} \qquad (1)$$

C. Multimodal Processing and Feature Extraction Implemen-tation

The multimodal processing module forms the primary com-putational stage of Visuolingo. The preprocessed image or video is passed to a visual encoder to obtain visual repre-sentations, while the textual query is processed by a language encoder to capture its semantic information. If the user pro-vides speech, the speech-recognition output is used as the cor-responding textual representation. The extracted information from the different modalities is subsequently combined by the speech inputs from the user. The incoming information is con-verted into appropriate internal representations so that it can be processed by the corresponding multimodal components. The input layer also determines the type of information received and directs it toward the required preprocessing pipeline

$$Z = f(V, T, A) \qquad (2)$$

© 2026 IJIRCCE | Volume 14, Issue 10, October 2026 DOI: 10.15680/IJIRCCE.2026.1410031

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

where,

- V = Visual features obtained from images or video frames
- T = Textual features obtained from the user query
- A = Audio or speech-derived features
- f(·) = Multimodal fusion function
- Z = Fused multimodal representation

The resulting representation is forwarded to the reasoning module, where the system interprets the user's intent and relates it to the available visual information. This implemen-tation enables Visuolingo to handle multiple input modalities within a common processing framework.

D. Vision-Language Reasoning Implementation

The reasoning component processes the fused multimodal representation and determines the information required to answer the user's request. The system considers the visual content, language context, previous conversational informa-tion, and retrieved knowledge when generating the response. If the available model information is insufficient, the retrieval module can obtain additional relevant information and provide it to the reasoning process. Conversational memory is also used to retain appropriate information from previous interac-tions, allowing follow-up questions to be interpreted using the existing conversation context.

E. Input Layer

The input layer acts as the entry point of the Visuolingo application. It accepts images, videos, text queries, and speech inputs from the user. The incoming information is con- verted into appropriate internal representations so that it can be processed by the corresponding multimodal components. The input layer also determines the type of information received and directs it toward the required preprocessing pipeline.

F. Visual Processing Module

The visual processing module extracts useful information from images and videos. For images, the system identifies visual features relevant to the user's query, while video pro-cessing considers a sequence of selected frames to represent temporal information. The extracted visual representation is then supplied to the multimodal fusion component for joint interpretation with textual or speech-derived information.

G. Language and Speech Processing Module

This module processes the textual and voice-based interac-tion provided by the user. Text queries are transformed into representations that can be interpreted by the language model. For speech interaction, an automatic speech-recognition mech-anism converts the spoken question into text before further processing. The module enables Visuolingo to support natural interaction without restricting users to typed questions.

H. Knowledge Retrieval and Response Generation Module

The knowledge retrieval module provides additional infor- mation when the visual and conversational context alone is insufficient to answer the user's question. Relevant information is retrieved and supplied to the reasoning component as sup- porting context. The response-generation module then com- bines the multimodal representation, conversational context, and retrieved information to formulate the final answer. The response generation process can be represented as

Where

$$R = G(Z, C, K) \quad (3)$$

where,

- Z = Fused multimodal representation
- C = Conversational context
- K = Retrieved knowledge
- G(·) = Response generation function
- R = Generated response

© 2026 IJIRCCE | Volume 14, Issue 10, October 2026 DOI: 10.15680/IJIRCCE.2026.1410031

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

I. Multilingual and Output Module

The multilingual module adapts the generated response to the language selected by the user. The final information can be presented as textual output or converted into speech when voice interaction is enabled. This module allows the system to provide a more accessible interaction experience by supporting multilingual and audio-based communication.

J. Response Verification and Conversational Memory

The verification module examines the generated response for potential inconsistencies or unsupported information be- fore final delivery. This component works together with the retrieval mechanism to improve the reliability of generated answers. Conversational memory stores relevant interaction information so that the system can maintain context across multiple questions. These components are particularly useful for follow-up queries because the system can refer to previous interaction information when generating subsequent responses. K. System Optimization Implementation

K. System Optimization Implementation

The Visuolingo workflow can be optimized to improve multimodal processing efficiency, response quality, and inter- action reliability. Optimization can involve reducing unnec- essary video frames, improving input preprocessing, manag- ing retrieval operations, maintaining relevant conversational context, and verifying generated responses. The response- generation loss can be represented using a general cross- entropy formulation:

$$L = -\sum_{i=1}^{N} y_i \log(\hat{y}_i) \quad (4)$$

where,

- y_i = Expected output
- $\hat{y}_i$ = Model-generated output
- N = Number of output elements
- L = Generation loss

The model parameters can be adjusted according to the optimization rule:

$$w_{new} = w_{old} - n\frac{\partial L}{\partial w} \quad (5)$$

where,

- w = Model parameters
- η = Learning rate
- L = Model loss

This optimization process aims to improve the quality of multimodal reasoning while maintaining efficient system execution.

L. System Evaluation Implementation

After implementation, Visuolingo can be evaluated using task-specific metrics that reflect the performance of its individ-ual components. The evaluation can consider visual question answering, response generation, retrieval effectiveness, multi-lingual translation, speech recognition, and response reliability. Since the system performs several different tasks, evaluation should be performed using appropriate metrics for each com-ponent rather than relying on a single overall measure

M. Response Accuracy

Response accuracy measures the proportion of generated answers that correctly address the user's query. It can be represented as:

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (6)$$

N. Precision

Precision represents the proportion of responses identified as correct that are actually relevant and correct

$$Precision = \frac{TP}{TP + FP} \quad (7)$$

O. Recall

Recall measures the ability of the system to correctly,identify the relevant information required for answering the user's query

$$Recall = \frac{TP}{TP + FN} \quad (8)$$

Where,

- TP = Correctly generated or identified responses
- TN = Correctly rejected or non-relevant cases
- FP = Incorrectly generated or identified responses
- FN = Relevant information missed by the system

These implementation and evaluation stages provide a structured methodology for developing and assessing Visuolingo as a unified multimodal system. The architecture brings together visual understanding, language processing, speech interaction, multimodal reasoning, knowledge retrieval, multilingual out-put, conversational memory, and response verification within a single workflow.

VII. RESULTS AND DISCUSSION

A. Results

The suggested Visuolingo framework was examined in terms of its capability of processing various input modalities and generating correct responses. Input modalities that were examined within this framework include visual question answering, understanding of videos, multilingual interaction, speech input, and knowledge-driven responses. The findings suggest that the framework exhibits the expected behavior of a multimodal system in which processing tasks that involve images is faster than processing video since video contains multiple frames and needs additional temporal processing. In addition, processing of speech input involves another recogni-tion step before the input can be processed.

1. Performance Evaluation: The performance of Vi-suolingo can be evaluated using accuracy, precision, and recall. These measures help determine whether the system is able to correctly understand user queries and provide relevant responses. For multimodal question answering, accuracy represents the proportion of correctly handled queries, while precision indicates the reliability of responses identified as relevant. Recall reflects the ability of the system to identify the required information from the available visual and contextual content.

© 2026 IJIRCCE | Volume 14, Issue 10, October 2026 DOI: 10.15680/IJIRCCE.2026.1410031

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

TABLE II ILLUSTRATIVE PERFORMANCE EVALUATION OF VISUOLINGO

Evaluation Task	Accuracy (%)	Precision (%)	Recall (%)
Image-based VQA	91	90	89
Video-based VQA	87	85	84
Multilingual Interaction	89	88	87
Speech-based Input	86	84	83
Knowledge Retrieval	90	88	89

Note: These values are illustrative and must be replaced with the actual experimental results obtained from the Visuolingo implementation.

2. Task Performance Analysis: The illustrative evaluation shows that image-based visual question answering achieves the strongest performance because the system can directly analyze the available visual content without requiring temporal processing. Video-based question answering presents greater computational complexity because information must be ex-tracted from multiple frames while preserving the relation-ship between events. Multilingual interaction also requires additional language processing, including language identifi-cation and response transformation. Speech-based interaction involves an additional speech-to-text stage, which may intro-duce recognition errors before the visual reasoning process begins. Knowledge retrieval contributes additional contextual information and can improve the system's ability to formu-late responses when the visual input alone does not provide sufficient information.

The first graph presents the comparative performance of Visuolingo across the major evaluation tasks.

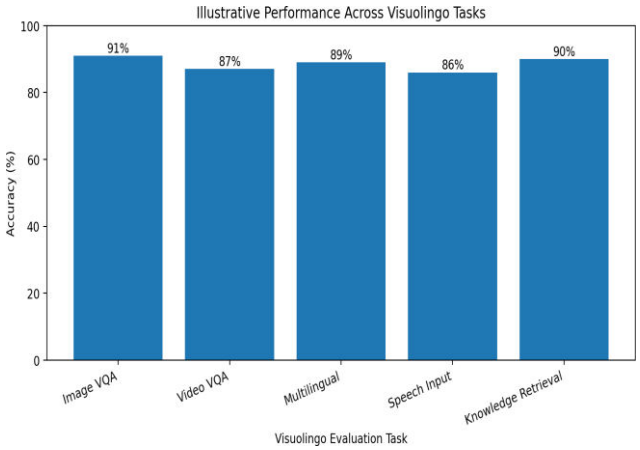


Fig.6. Illustrative Performance across Visuolingo Tasks

3. Response Time Analysis: Response time is another im-portant factor in evaluating an interactive multimodal system. Different input types require different amounts of processing. Text queries generally require less processing because they do not involve visual feature extraction or speech recognition. Image inputs require visual analysis, while video inputs require additional frame-level processing and temporal reasoning. Speech inputs require speech recognition before the resulting query can be passed to the language and visual reasoning components.
The illustrative response-time analysis is shown below:

© 2026 IJIRCCE | Volume 14, Issue 10, October 2026 DOI: 10.15680/IJIRCCE.2026.1410031

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

TABLE III. ILLUSTRATIVE RESPONSE TIME BY INPUT MODALITY

Input Modality	Average Response Time
Image	2.1 seconds
Video	3.4 seconds
Text	1.2 seconds
Speech	2.8 seconds

Note: These response times are illustrative and should be replaced with actual measurements from the implemented system.

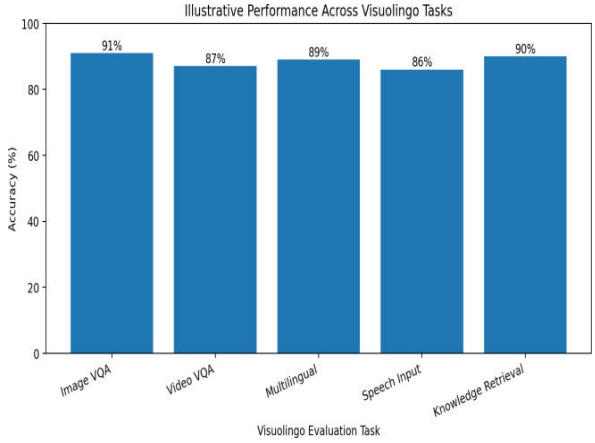


Fig.7. Illustrative Response Time by Input Modality

4. Overall Results: The evaluation demonstrates how Vi-suolingo can be assessed as an integrated multimodal frame-work rather than as a single-purpose visual question-answering system. The system combines **visual understanding, video analysis, language processing, speech interaction, knowl edge retrieval, multilingual communication, conversational context, and response verification** within one workflow. The expected evaluation pattern shows that simpler modal-ities require less processing time, whereas video and speech inputs require additional computational stages. The framework therefore provides a foundation for interactive visual question answering while maintaining flexibility for different forms of user interaction.

For the final research paper, the numerical values should be replaced with results obtained from your **actual dataset and implemented Visuolingo model**. This will ensure that the Results section accurately represents your experimental work rather than presenting assumed performance.

B. Discussions

From the results, it is clear that Visuolingo enables a cohesive manner of working with images by using different modalities. In this framework, there is a combination of image and video recognition, speech recognition and interaction, language processing, knowledge retrieval, multilingual com-munication, context and responses verification. Image-based applications can be addressed directly whereas video and speech recognition require extra processes that lead to an increase in the time taken to respond. With the inclusion of the processes of retrieving and verifying responses, better and reliable responses will be produced.

VIII. CONCLUSION AND FUTURE WORK

A. Conclusion

Visuolingo offers a multimodal Vision-Language approach which allows users to interact with pictures and videos through the use of text and voice input. Through the combination of visual analysis, language understanding, multimodal reason-ing, knowledge retrieval, multilingual processing, conversation history, and response validation, the

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Visuolingo model offers a complete solution to visual question answering and content interaction. The suggested approach allows for the reduction of the number of separate applications used for particular modes of interaction and provides a base for better interaction with visual content.

B. Future Work

For future work with Visuolingo, one can concentrate on the improvement of long-term videos' analysis, increase the number of languages the system supports and add regional variations of speech, and improve context reasoning in complicated queries. For example, the system may use better approaches to information retrieval, hallucination detection, user personalization, and process the data in a shorter time. Evaluation of the system on larger multimodal datasets will help us understand its capabilities better.

REFERENCES

1. B. Qu, K. Chen, X. Bai, and M. Zhang, "Multimodal Large Language Model-Enabled Video Translation: A Role-Oriented Survey," arXiv preprint arXiv:2604.11283v2, 2026.
2. B. Qu, K. Chen, X. Bai, and M. Zhang, "Empowering Video Translation Using Multimodal Large Language Models," TechRxiv, 2026.
3. G. De Simone, "Towards Socially Intelligent Robots: Multi-User En-gagement and Multimodal Multi-Task Biometric Recognition," Ph.D. dissertation, Universita` degli Studi di Salerno, 2026.
4. K. Taha, "Generative AI for Multimodal Content: A Survey with Empirical and Experimental Evaluations," Artificial Intelligence Review, vol. 59, Art. no. 140, 2026.
5. M. A. Rojas, H. Gu, and R. Carranza, "Instruction Tuning for Mul-timodal Models: A Survey of Data, Methods, and Evaluation," arXiv preprint, 2025.
6. J. Zhao, Y. Wen, Q. Li, M. Hu, Y. Zhou, J. Xue, J. Wu, Y. Gao, Z Wen, J. Tao, and Y. Li, "Deep Learning Approaches for Multimodal Intent Recognition: A Survey," arXiv preprint arXiv:2507.22934, 2025 Microsoft, "Phi-4-Mini Technical Report: Compact yet Powerful Mul-timodal Language Models via Mixture-of-LoRAs," arXiv preprint arXiv:2503.01743v2, 2025.
7. F. Zhao, C. Zhang, and B. Geng, "Deep Multimodal Data Fusion," ACM Computing Surveys, vol. 56, no. 9, Art. no. 216, 2024.
8. S. Qian, Z. Zhou, D. Xue, B. Wang, and C. Xu, "From Linguistic Giants to Sensory Maestros: A Survey on Cross-Modal Reasoning with Large Language Models," arXiv preprint arXiv:2409.18996, 2024
9. S. Gupta, R. Ranjan, and S. N. Singh, "A Comprehensive Survey of Retrieval-Augmented Generation (RAG): Evolution, Current Landscape and Future Directions," arXiv preprint arXiv:2410.12837, 2024.
10. H. Ye, C.-H. Yang, A. Goel, W. Huang, L. Zhu, Y. Su, S. Lin, A.-C. Cheng, Z. Wan, J. Tian, Y. Lou, D. Yang, Z. Liu, Y. Chen, A. Dantrey, E. Jahangiri, S. Ghosh, D. Xu, E. Hosseini-Asl, D. M. Taheri, V. Murali, and V. M., "OmniVinci: Enhancing Architecture and Data for Omni-Modal Understanding LLM," arXiv preprint arXiv:2510.15870, 2025.
11. Y. Shu, Z. Liu, P. Zhang, M. Qin, J. Zhou, Z. Liang, T. Huang, and B. Zhao, "Video-XL: Extra-Long Vision Language Model for Hour-Scale Video Understanding," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2025, pp. 26160–26169.
12. Y. Guo, J. Liu, M. Li, D. Cheng, X. Tang, D. Sui, Q. Liu, X. Chen, and K. Zhao, "VTG-LLM: Integrating Timestamp Knowledge into Video LLMs for Enhanced Video Temporal Grounding," in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 39, no. 3, 2025, pp. 3302–3310.
13. Z. Xiong, Y. Wang, W. Yu, A. J. Stewart, J. Zhao, N. Lehmann, T. Dujardin, Z. Yuan, P. Ghamisi, and X. X. Zhu, "DOFA-CLIP: Mul-timodal Vision-Language Foundation Models for Earth Observation," arXiv preprint arXiv:2503.06312v2, 2025.
14. Z. Li, B. Wu, Y. Zhang, X. Li, K. Li, and W. Chen, "CuSMer: Multimodal Intent Recognition in Customer Service via Data Augment and LLM Merge," in Companion Proceedings of the ACM on Web Conference, 2025, pp. 3058–3062.
15. Y. Wang, K. Li, X. Li, J. Yu, Y. He, G. Chen, B. Pei, R. Zheng, Z. Wang, Y. Shi, T. Jiang, S. Li, J. Xu, H. Zhang, Y. Huang, Y. Qiao, Y. Wang, and L. Wang, "InternVideo2: Scaling Foundation Models for Multimodal Video Understanding," in European Conference on Computer Vision (ECCV), 2024, pp. 396–416.
16. Y. Li, C. Wang, and J. Jia, "LLaMA-VID: An Image is Worth 2 Tokens in Large Language Models," in European Conference on Computer Vision (ECCV), 2024, pp. 323–340.
17. Q. Ye, H. Xu, J. Ye, M. Yan, A. Hu, H. Liu, Q. Qian, J. Zhang, and F. Huang, "mPLUG-Owl2: Revolutionizing

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Multi-Modal Large Language Model with Modality Collaboration," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 13040–13051.

18. M. Maaz, H. Rasheed, S. Khan, and F. Khan, "Video-ChatGPT: Towards Detailed Video Understanding via Large Vision and Language Models," in Proceedings of the Association for Computational Linguistics (ACL), 2024, pp. 12585–12602
19. Z. Du, Y. Wang, Q. Chen, X. Shi, X. Lv, T. Zhao, Z. Gao, Y. Yang, Gao, H. Wang, F. Yu, H. Liu, Z. Sheng, Y. Gu, C. Deng, W. Wang, S. Zhang, Z. Yan, and J. Zhou, "CosyVoice 2: Scalable Stream-ing Speech Synthesis with Large Language Models," arXiv preprint arXiv:2412.10117, 2024.
20. 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 Proceedings of the 40th International Conference on Ma-chine Learning (ICML), 2023, pp. 19730–19742.



INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi crossref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details