Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A Modular AI-Driven Multimodal Framework for Real-Time Classroom Interpretation for Deaf Students

Shahir VK

drshahirvk01@gmail.com

AWH Engineering College

Fathima Hisa

AWH Engineering College

Research Article

Keywords: Inclusive Education, Artificial Intelligence, Indian Sign Language, Automatic Speech Recognition, Large Language Models, Multimodal Learning, Assistive Technology, Classroom Interpretation

Posted Date: November 10th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10518658/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

A Modular AI-Driven Multimodal Framework for Real-Time Classroom Interpretation for Deaf Students

Dr. Shahir V. K.[1], Fathima Hisa[2]

[1] *Principal Investigator, AICTE Funded Research Project, AWH Engineering College, Kozhikode, Kerala, India*

[2] *Research Associate, AICTE Funded Research Project, AWH Engineering College, Kozhikode, Kerala, India*

Corresponding author: *Dr. Shahir V.K.*
Email: drshahirvk01@gmail.com

Abstract—Despite significant advances in inclusive education, real-time classroom communication remains a major challenge for students with profound hearing impairment, particularly in science, technology, engineering, and mathematics (STEM) learning environments where continuous verbal interaction and concept-oriented instruction are essential. Existing assistive technologies, including speech-to-text captioning, frequency modulation (FM) systems, and human sign language interpreters, either provide limited contextual understanding, lack scalability, or are unavailable in many educational institutions. This paper presents a modular artificial intelligence (AI)-driven multimodal framework for real-time classroom interpretation designed to enhance accessibility for deaf learners. The proposed framework integrates Automatic Speech Recognition (ASR), educational semantic processing using Large Language Models (LLMs), Indian Sign Language (ISL) gloss generation, intelligent sign retrieval, and avatar-assisted visual communication within a unified architecture. In addition, the framework incorporates a systematic dataset engineering pipeline comprising vocabulary organization, automated metadata generation, annotation management, and gloss-based indexing to facilitate efficient retrieval and future scalability. Unlike conventional systems that primarily focus on isolated speech transcription or sign language translation, the proposed architecture emphasizes modularity, educational context awareness, extensibility, and deployment readiness for real classroom environments. The paper details the overall system architecture, workflow, and implementation methodology while highlighting the integration of multiple AI components into a cohesive assistive learning platform. The proposed framework establishes a scalable technological foundation for developing next-generation intelligent classroom interpretation systems and contributes toward advancing accessible, inclusive, and AI-enabled education for deaf students.

Keywords— Inclusive Education, Artificial Intelligence, Indian Sign Language, Automatic Speech Recognition, Large Language Models, Multimodal Learning, Assistive Technology, Classroom Interpretation.

1.Introduction

The pursuit of equitable and inclusive education has become a global priority, with increasing emphasis on ensuring that learners with disabilities have equal opportunities to participate in mainstream educational environments. Despite significant policy initiatives and technological progress,

communication remains one of the most persistent barriers for students with hearing impairment, particularly those with profound or complete hearing loss. Classroom learning relies heavily on continuous verbal interaction, spontaneous discussions, contextual explanations, and instructor–student engagement, making auditory communication the primary medium for knowledge transfer. Consequently, deaf learners often encounter considerable difficulties in following classroom instruction, resulting in reduced participation, limited conceptual understanding, and disparities in academic achievement.[1], [2], [3], [4]

In India, inclusive education has received renewed attention through the National Education Policy (NEP) 2020, which advocates the adoption of assistive technologies and promotes the standardization and wider use of Indian Sign Language (ISL) to improve educational accessibility. At the same time, the Rights of Persons with Disabilities Act (RPwD Act, 2016) reinforces the need to provide accessible learning environments for persons with disabilities. Although these initiatives establish a strong policy foundation, the practical implementation of accessible classroom communication remains limited. Educational institutions, particularly those offering science, technology, engineering, and mathematics (STEM) programmes, continue to rely predominantly on verbal instruction, where technical terminology, rapid concept transitions, mathematical explanations, and interactive discussions are difficult to interpret without continuous language support.[1], [5], [6], [7]

The challenge is particularly evident for students with profound hearing impairment. While learners with residual hearing may benefit from hearing aids, cochlear implants, or frequency modulation (FM) amplification systems, these technologies primarily enhance audio perception and offer limited benefit to individuals with little or no functional hearing. Human sign language interpreters provide an effective communication medium; however, their availability remains insufficient to satisfy the growing demand across educational institutions. Furthermore, employing dedicated interpreters for every classroom is often financially and logistically impractical, particularly in developing countries. Consequently, many deaf students continue to depend on written notes, lip reading, or delayed instructional support, all of which significantly reduce real-time classroom engagement and independent learning.[1], [2], [8], [9]

Recent advances in artificial intelligence have created new opportunities to address these longstanding accessibility challenges. Deep learning-based Automatic Speech Recognition (ASR) systems have demonstrated substantial improvements in transcription accuracy across multilingual and acoustically challenging environments. Similarly, Large Language Models (LLMs) have shown remarkable capabilities in semantic understanding, contextual reasoning, language simplification, and educational content generation. Parallel developments in computer vision, multimodal learning, and virtual human animation have further enabled intelligent sign language generation and avatar-based communication. Collectively, these technologies provide a promising foundation for developing integrated assistive systems capable of transforming spoken classroom instruction into accessible visual communication.[10], [11], [12], [13], [14]

Despite these technological advancements, existing assistive solutions remain fragmented. Many speech-to-text systems focus exclusively on transcription without considering the linguistic structure of sign languages or the educational context in which communication occurs. Similarly, several sign language translation studies concentrate on isolated word recognition, gesture classification, or offline translation tasks, with limited consideration for continuous classroom discourse, technical vocabulary, or contextual interpretation. Existing avatar-based systems primarily emphasize visual animation, often overlooking the upstream processes required for semantic understanding, gloss generation, educational content simplification, and real-time multimodal integration. As a result, there remains a considerable gap between laboratory-scale research prototypes and deployable classroom interpretation systems capable of supporting inclusive education in authentic learning environments.[15], [16], [17], [18]

Addressing this gap requires more than the integration of individual artificial intelligence models. A practical classroom interpretation system must function as a coordinated multimodal ecosystem capable of acquiring classroom speech, understanding instructional context, simplifying educational content, generating linguistically appropriate Indian Sign Language glosses, retrieving relevant sign

representations, and presenting the interpreted content through accessible visual interfaces with minimal latency. Such a system must also remain modular, scalable, and adaptable to diverse classroom scenarios while supporting future enhancements without extensive architectural modifications.

The research presented in this paper has been undertaken as part of an All India Council for Technical Education (AICTE) Research Promotion Scheme (RPS) – Lab to Market funded research project aimed at developing an intelligent assistive communication framework for inclusive classroom environments. The project, supported with a total funding of INR 34.5 lakhs, focuses on translating recent advances in artificial intelligence into practical educational technologies capable of addressing real-world accessibility challenges. The proposed framework represents the foundational architectural and methodological component of this larger research initiative and establishes the technological basis for the development, implementation, and future deployment of an AI-assisted multimodal classroom interpretation system for deaf learners.[5]

Motivated by these challenges, this paper presents a modular multimodal artificial intelligence framework for real-time classroom interpretation designed to improve communication accessibility for deaf students in inclusive educational environments. Rather than proposing a standalone translation algorithm, the study introduces an end-to-end architectural framework that integrates speech acquisition, robust Automatic Speech Recognition, semantic language understanding, educational content simplification, Indian Sign Language gloss generation, intelligent sign retrieval, and avatar-assisted visual communication within a unified and extensible pipeline. In addition, the framework incorporates a structured dataset engineering methodology, including automated metadata generation, annotation management, vocabulary organization, and gloss-based indexing, thereby establishing a scalable foundation for future development, evaluation, and deployment.

The principal contributions of this work are summarized as follows:

A modular multimodal architecture for real-time classroom interpretation that integrates speech processing, semantic understanding, sign language generation, and visual communication within a unified framework.
A systematic dataset engineering pipeline supporting scalable vocabulary organization, annotation, automated metadata generation, and efficient gloss-based retrieval for Indian Sign Language resources.
A context-aware educational processing workflow that combines Automatic Speech Recognition and Large Language Models to improve semantic representation prior to sign language translation.
A deployment-oriented framework designed to support extensibility, maintainability, and practical implementation in inclusive classroom environments.

The remainder of this paper is organized as follows. Section II reviews recent developments in artificial intelligence, sign language translation, educational assistive technologies, and multimodal learning systems. Section III analyses existing research gaps and motivates the proposed framework. Section IV presents the overall multimodal architecture, while Section V describes the dataset engineering methodology and annotation framework. Section VI discusses prototype implementation and system integration. Finally, Sections VII and VIII present the discussion, future research directions, and concluding remarks.

2. Background and Related Work

Recent advances in artificial intelligence have accelerated the development of intelligent assistive technologies aimed at improving accessibility for individuals with hearing impairment. Significant research efforts have been directed toward Automatic Speech Recognition (ASR), sign language recognition and translation, natural language processing, educational artificial intelligence, virtual sign language avatars, and multimodal human–computer interaction. These technologies have demonstrated considerable potential in addressing specific communication challenges; however, most existing solutions focus on isolated tasks rather than providing a comprehensive framework capable of

supporting continuous classroom communication for deaf learners. As a result, the practical deployment of integrated assistive systems in mainstream educational environments remains limited. [9], [12], [13], [14], [15], [18], [19], [20], [21]

To understand the current state of research and identify existing limitations, this section presents a structured review of recent literature spanning six complementary research domains. The discussion begins with inclusive education and communication challenges experienced by deaf learners, followed by advances in Automatic Speech Recognition for educational applications. Subsequently, research on Indian Sign Language recognition and translation, Large Language Models for educational semantic understanding, sign language avatars, and multimodal artificial intelligence systems for accessible learning is critically examined. Finally, a comparative analysis of representative studies is presented to identify unresolved research challenges and motivate the proposed modular multimodal framework.[2], [8], [10], [12], [13], [14]

2.1 Inclusive Education and Communication Challenges for Deaf Learners

Inclusive education has evolved from a policy-driven objective into a fundamental principle of equitable and sustainable educational development. International frameworks, including the United Nations Convention on the Rights of Persons with Disabilities (UNCRPD) and Sustainable Development Goal 4 (SDG 4), emphasize the provision of quality education that is accessible to all learners regardless of their physical or sensory abilities. In India, the Rights of Persons with Disabilities (RPwD) Act, 2016 and the National Education Policy (NEP) 2020 further reinforce the importance of creating inclusive learning environments through the adoption of assistive technologies, accessible educational resources, and the promotion of Indian Sign Language (ISL) as a recognized means of communication. These initiatives have significantly increased awareness of accessibility requirements within educational institutions; however, translating policy into effective classroom practice remains a considerable challenge.[1], [4], [22], [23], [24], [25], [26]

Among students with disabilities, learners with profound hearing impairment encounter unique barriers that directly influence their academic participation and learning outcomes. Classroom teaching is predominantly based on continuous spoken communication, spontaneous interaction, collaborative discussions, and instructor explanations. Consequently, students who rely primarily on sign language frequently experience difficulties in accessing instructional content in real time, particularly when communication support is unavailable or insufficient. The challenge becomes even more pronounced in higher education, where lectures often involve specialized terminology, abstract concepts, mathematical expressions, technical demonstrations, and rapid transitions between topics. Delays in communication or incomplete interpretation may therefore affect both conceptual understanding and classroom engagement.[2], [8], [9], [21]

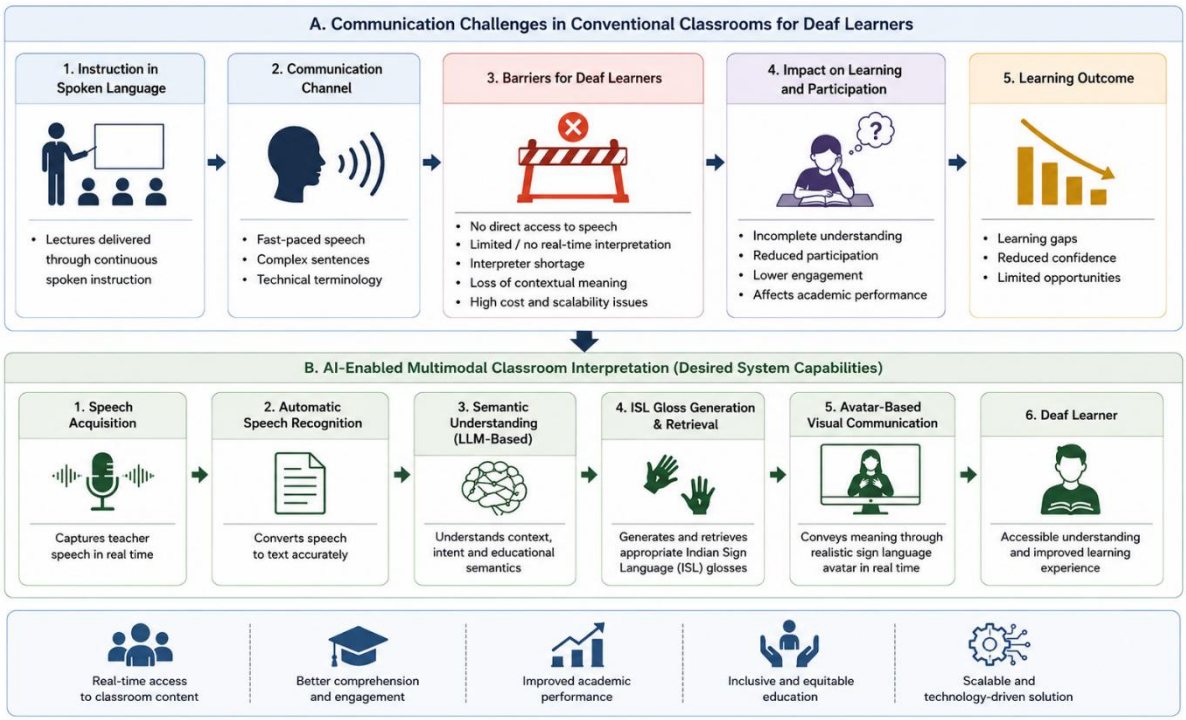


Fig. 1. Communication challenges experienced by deaf learners in conventional classroom environments and the desired capabilities of an AI-enabled multimodal classroom interpretation system. The upper section illustrates the communication barriers that affect classroom participation and learning outcomes, while the lower section presents the key functional components required to support accessible, real-time educational communication.

Several assistive technologies have been introduced to improve educational accessibility for deaf and hard-of-hearing learners. Hearing aids, cochlear implants, and frequency modulation (FM) systems are effective for individuals with residual hearing by amplifying or enhancing auditory signals. However, these technologies provide limited benefit for students with profound hearing loss, for whom auditory information alone is insufficient for effective communication. Closed-captioning and speech-to-text transcription systems have emerged as alternative accessibility solutions by converting spoken language into textual representations. Although these systems improve access to verbal information, they often require learners to divide their attention between reading captions and simultaneously following classroom activities, diagrams, demonstrations, and instructor gestures. This cognitive burden may reduce comprehension, particularly in concept-intensive engineering and science courses where understanding depends not only on words but also on contextual explanations and visual reasoning.[2], [9], [21]

Professional sign language interpreters continue to represent the most effective communication medium for many deaf learners. Their ability to convey linguistic structure, contextual meaning, and non-verbal cues makes them indispensable in many educational settings. Nevertheless, the availability of qualified interpreters remains limited, particularly in developing countries and institutions with constrained resources. Even where interpreters are available, assigning dedicated personnel to every classroom or laboratory session presents significant financial and logistical challenges. Furthermore, classroom communication is inherently dynamic; instructors frequently shift between theoretical explanations, technical terminology, practical demonstrations, and student interactions. Maintaining consistent interpretation quality under such conditions remains a demanding task and may vary depending on interpreter expertise and subject-specific knowledge.[2], [8]

Recent studies have also highlighted that accessibility in education extends beyond communication alone. Effective learning requires the preservation of semantic meaning, instructional intent, and contextual relationships throughout the communication process . Simply converting speech into text or isolated signs does not necessarily ensure meaningful comprehension. Educational content often

contains discipline-specific vocabulary, implicit explanations, references to visual materials, and contextual expressions that require interpretation rather than literal translation. Consequently, assistive technologies intended for educational environments must support semantic understanding in addition to linguistic conversion.[9], [15]

The growing adoption of artificial intelligence has created new opportunities to address these challenges through intelligent and adaptive assistive systems. Advances in Automatic Speech Recognition, natural language processing, multimodal learning, computer vision, and human–computer interaction have enabled increasingly sophisticated communication technologies capable of processing multiple forms of information simultaneously. However, most existing solutions continue to address individual stages of the communication process independently. For example, speech recognition systems focus primarily on transcription, sign language translation systems emphasize gesture recognition or gloss generation, and avatar-based systems concentrate on visual rendering. Comprehensive integration of these technologies into a unified educational framework capable of supporting real-time classroom interpretation remains relatively unexplored.[10], [13], [14], [20]

These observations indicate that achieving genuine educational inclusion requires more than isolated assistive technologies. Future intelligent learning environments should be capable of acquiring classroom speech, understanding instructional context, adapting educational content to learners' communication requirements, generating appropriate sign language representations, and presenting information through accessible multimodal interfaces with minimal latency. Such an integrated approach has the potential to reduce communication barriers, improve classroom participation, and promote independent learning among deaf students in mainstream educational institutions. This need provides the foundation for exploring the technological components reviewed in the subsequent sections, beginning with recent developments in Automatic Speech Recognition for educational applications.

2.2 Automatic Speech Recognition for Educational Accessibility

Automatic Speech Recognition (ASR) has undergone remarkable advancements over the past decade, driven by the rapid evolution of deep learning, self-supervised representation learning, and transformer-based architectures. Modern ASR systems have significantly improved transcription accuracy across multiple languages, accents, and acoustic environments, making speech-to-text technology increasingly applicable to educational, healthcare, and accessibility domains. These developments have transformed ASR from a conventional transcription tool into a fundamental component of intelligent human–computer interaction systems.[10], [27], [28], [29]

Earlier ASR approaches primarily relied on Hidden Markov Models (HMMs) combined with Gaussian Mixture Models (GMMs), which required extensive handcrafted acoustic features and language-specific modeling. Although these methods achieved reasonable performance under controlled conditions, they were highly sensitive to environmental noise, speaker variability, and pronunciation differences. The emergence of deep neural networks substantially improved acoustic modeling by enabling automatic feature extraction from raw speech signals, resulting in improved robustness and recognition accuracy across diverse speaking conditions.[27], [28], [30]

Recent breakthroughs in self-supervised learning have further advanced the capabilities of ASR systems. Models such as wav2vec 2.0, HuBERT, and WavLM learn generalized speech representations from large volumes of unlabeled audio before being fine-tuned for downstream recognition tasks. These approaches have demonstrated strong performance in multilingual and low-resource scenarios while reducing the dependence on manually annotated speech corpora.[27], [28], [29]

The introduction of large-scale transformer-based speech models has further accelerated progress in multilingual speech recognition. OpenAI's Whisper model represents one of the most significant milestones in this direction, having been trained on hundreds of thousands of hours of multilingual and multitask speech data. Whisper demonstrates strong robustness against background noise, varying

speaker accents, spontaneous speech, and domain-specific vocabulary, making it particularly suitable for real-world educational environments where classroom recordings frequently contain overlapping speech, ambient noise, and diverse speaking styles.[10], [31]

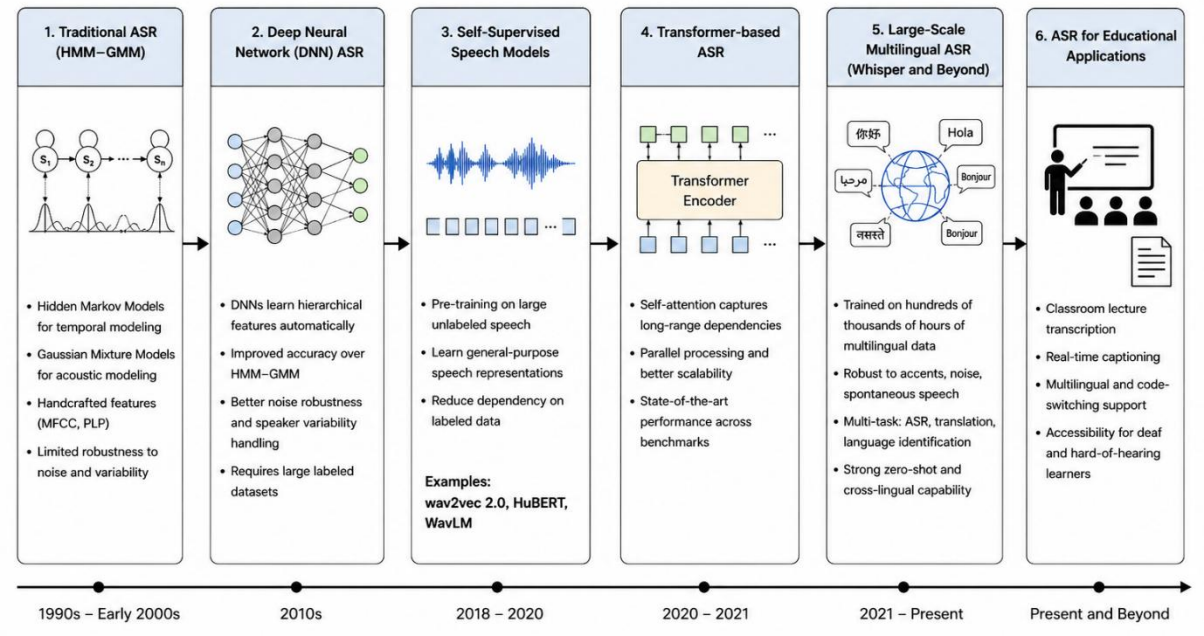


Fig. 2. Evolution of automatic speech recognition technologies from statistical acoustic models to multilingual transformer-based systems for educational applications.

Despite these remarkable achievements, current ASR technologies primarily generate textual transcriptions of spoken language. While textual captions improve information accessibility, they do not fully address the communication needs of deaf learners who primarily rely on sign language for conceptual understanding. Educational lectures often contain implicit contextual information, technical terminology, mathematical expressions, and instructor-specific explanations that require semantic interpretation rather than direct word-by-word transcription. Consequently, transcription alone cannot guarantee meaningful learning, particularly in higher education classrooms involving engineering and scientific instruction.[10], [12], [20]

Furthermore, classroom environments present several challenges that continue to influence ASR performance. Variations in microphone placement, reverberation, student interactions, multilingual code-switching, rapid speech delivery, and domain-specific vocabulary may reduce transcription accuracy. These limitations become increasingly significant in real-time educational settings where low latency and high semantic fidelity are essential for effective communication support.[2], [10], [29]

These observations indicate that ASR should be considered as the initial stage of an intelligent classroom interpretation pipeline rather than a complete accessibility solution. Accurate speech transcription must subsequently be complemented by contextual language understanding, semantic refinement, and sign language generation to ensure that educational content remains linguistically meaningful and pedagogically effective for deaf learners. This requirement motivates the integration of advanced language models and multimodal semantic reasoning, which are discussed in the following subsection.

2.3 Sign Language Recognition and Translation

Sign language serves as the primary mode of communication for millions of deaf and hard-of-hearing individuals worldwide. Unlike spoken languages, sign languages are natural visual languages that convey meaning through a combination of hand gestures, facial expressions, body posture, movement,

and spatial orientation. These linguistic components function together to express grammar, emotions, and contextual information, making sign languages structurally distinct from spoken and written languages. Consequently, translating spoken language into sign language requires considerably more than direct word-to-word substitution, demanding an understanding of linguistic structure, semantic relationships, and contextual meaning.[1], [8]

Recent advances in artificial intelligence and computer vision have significantly accelerated research in Sign Language Recognition (SLR) and Sign Language Translation (SLT). Early SLR systems primarily relied on handcrafted visual features extracted from images or videos, followed by conventional machine learning algorithms for gesture classification. Although these approaches demonstrated encouraging results under controlled laboratory conditions, their performance was highly dependent on background uniformity, illumination, camera positioning, and signer-specific variations. Such limitations restricted their applicability in practical real-world environments where visual conditions are considerably more complex.[17], [18]

The introduction of deep learning has transformed sign language processing by enabling automatic feature extraction from large-scale visual datasets. Convolutional Neural Networks (CNNs) have been extensively employed for spatial feature extraction, while Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) networks, and Gated Recurrent Units (GRUs) have improved the modeling of temporal gesture sequences. More recently, transformer-based architectures have demonstrated superior performance by capturing long-range dependencies within continuous sign sequences, significantly improving both recognition accuracy and translation quality. These models have established new benchmarks across several public sign language datasets and have become the dominant paradigm in contemporary sign language research.[14], [18]

While significant progress has been achieved in sign language recognition, Sign Language Translation (SLT) remains substantially more challenging. Recognition systems generally identify individual signs or gesture sequences, whereas translation systems must generate linguistically meaningful sentences that preserve the semantic intent of the original spoken language. Recent studies have therefore focused on gloss-based and gloss-free translation strategies. Gloss-based approaches convert spoken language into an intermediate symbolic representation before generating sign language, whereas gloss-free models attempt to learn direct mappings between spoken or written language and sign sequences. Although gloss-free approaches reduce dependence on manually annotated gloss corpora, they require extensive parallel datasets and remain computationally demanding.[14], [15], [17]

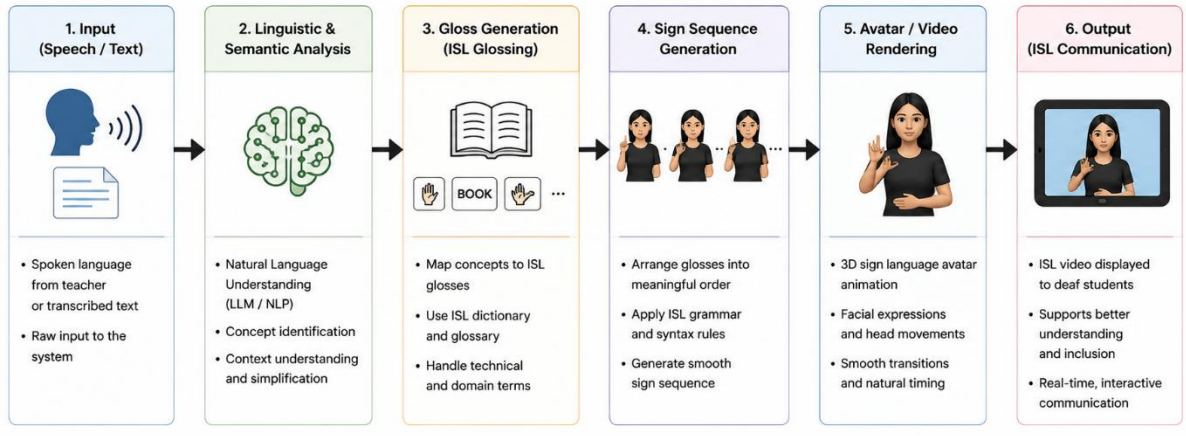


Fig. 3. General pipeline of sign language processing from speech or text input to visual sign language communication.

Indian Sign Language (ISL) presents additional research challenges due to the limited availability of standardized datasets, linguistic resources, and publicly accessible annotated corpora. Compared with American Sign Language (ASL) and British Sign Language (BSL), ISL has received relatively limited attention in artificial intelligence research. Existing datasets often contain isolated words or limited vocabulary collected under controlled experimental settings, making them insufficient for modeling

continuous classroom communication. Furthermore, the diversity of regional signing practices, domain-specific terminology, and limited availability of large-scale annotated educational datasets continue to restrict the development of generalized ISL translation systems.[6], [16]

Recent benchmark initiatives and publicly available datasets have begun addressing these limitations by providing larger collections of annotated ISL videos for recognition and translation tasks. These resources have enabled the development of transformer-based models capable of improving recognition accuracy and facilitating multilingual sign language research. Nevertheless, most existing datasets remain focused on daily conversational vocabulary and lack comprehensive coverage of academic terminology commonly encountered in engineering, science, mathematics, and higher education classrooms. Consequently, many technical concepts continue to require finger spelling or manual interpretation rather than standardized sign representations.[14], [16], [18]

Another important limitation observed in current sign language research is the strong emphasis on isolated algorithmic performance. Most studies evaluate recognition accuracy, translation quality, or gesture generation independently, while comparatively few investigate how these components can be integrated into an end-to-end educational communication system. Practical classroom environments require seamless coordination between speech acquisition, semantic understanding, sign language generation, visual presentation, and user interaction under strict real-time constraints. Addressing these requirements demands a holistic system-level perspective rather than isolated optimization of individual modules.[9], [15]

These observations indicate that effective classroom accessibility cannot rely solely on sign language recognition or translation algorithms. Instead, future educational assistive technologies should integrate sign language processing with contextual language understanding, multimodal semantic reasoning, and adaptive visual communication to provide meaningful instructional support. Such integration becomes particularly important for higher education, where accurate interpretation of technical concepts and domain-specific knowledge is essential for effective learning. The next subsection therefore examines the role of Large Language Models (LLMs) in enabling semantic understanding and context-aware educational communication.

2.4 Large Language Models for Accessible Education

Recent advances in Natural Language Processing (NLP) have significantly transformed the way machines understand, interpret, and generate human language. Early NLP systems primarily relied on handcrafted linguistic rules, statistical language models, and feature-engineering techniques that required extensive domain expertise and large manually annotated datasets. Although these approaches demonstrated satisfactory performance for well-defined tasks such as part-of-speech tagging, named entity recognition, and syntactic parsing, they exhibited limited generalization capabilities and struggled to capture contextual semantics within complex educational discourse.[11], [20], [32]

The introduction of deep learning and transformer-based architectures has fundamentally reshaped language understanding. Unlike recurrent neural networks, transformer models employ self-attention mechanisms that enable efficient modeling of long-range contextual dependencies within text sequences. This architectural advancement has substantially improved performance across a wide range of NLP tasks, including machine translation, summarization, question answering, semantic analysis, and conversational intelligence. Pre-trained transformer models such as BERT, RoBERTa, and T5 further demonstrated that knowledge acquired through large-scale unsupervised pre-training could be effectively transferred to diverse downstream applications with minimal task-specific adaptation.[11]

Building upon these developments, Large Language Models (LLMs) have emerged as a new generation of foundation models capable of performing reasoning, contextual understanding, content generation, and instruction following across multiple domains. Models including GPT-4, Llama 3, and Gemini have demonstrated remarkable capabilities in semantic interpretation, multilingual communication, educational content generation, and complex problem solving. Their ability to understand contextual

relationships rather than relying solely on keyword matching makes them particularly suitable for educational applications, where instructional meaning frequently depends on surrounding explanations, examples, and learner context.[12], [20], [33], [34]

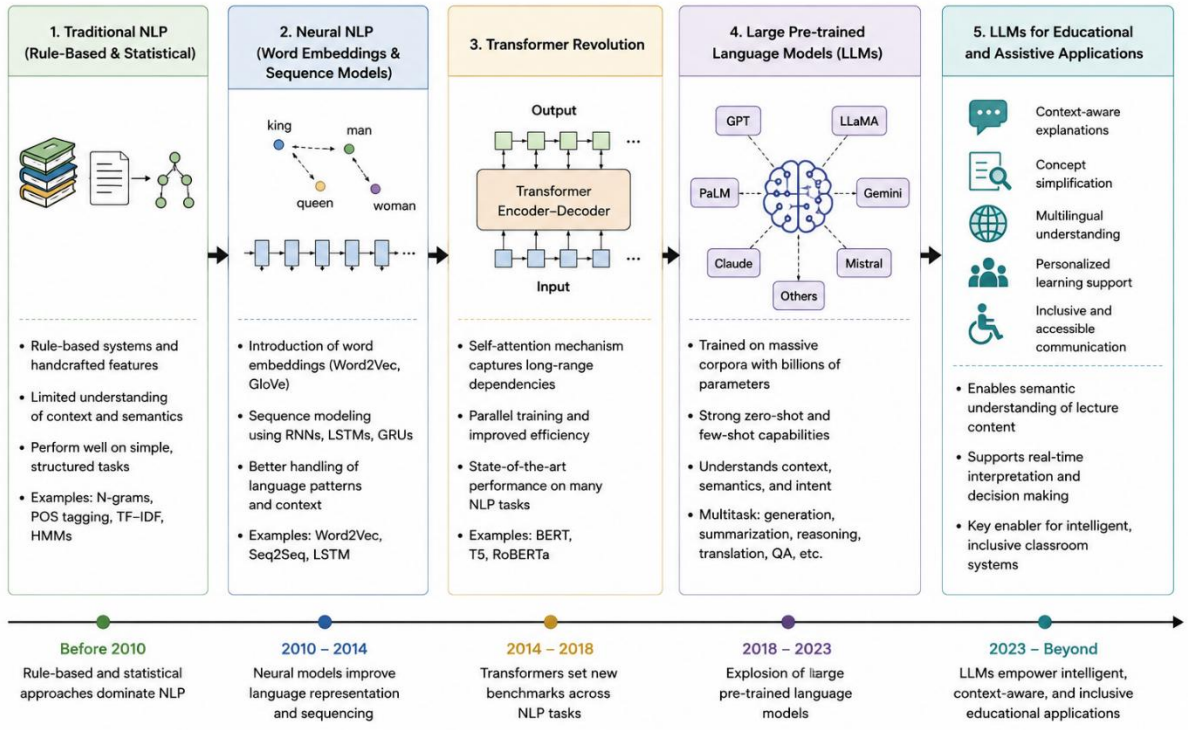


Fig. 4. Evolution of natural language processing from rule-based approaches to transformer architectures and large language models for intelligent educational applications.

The integration of LLMs into educational technology has accelerated rapidly in recent years. Current research explores their application in intelligent tutoring systems, automated feedback generation, personalized learning support, question answering, adaptive assessment, and educational content simplification. These systems have shown promising potential for improving learner engagement and providing individualized instructional support across various educational settings. Moreover, multilingual LLMs have expanded opportunities for supporting diverse learner populations by enabling cross-lingual communication and language adaptation.[12], [13], [20]

Despite these remarkable achievements, directly applying general-purpose LLMs to accessibility for deaf learners remains challenging. Classroom lectures frequently contain domain-specific terminology, mathematical expressions, abbreviations, contextual references, and discipline-dependent explanations that require semantic interpretation before they can be effectively communicated through sign language. Furthermore, spoken classroom communication often includes incomplete sentences, spontaneous corrections, code-switching, and instructor-specific teaching styles that conventional language models may not fully interpret without additional contextual information.[12], [15], [34]

Recent studies have therefore investigated the integration of LLMs with multimodal learning systems capable of combining speech, text, visual information, and contextual reasoning. Such multimodal approaches extend beyond conventional text generation by enabling semantic refinement, concept identification, terminology simplification, and context-aware instructional assistance. These capabilities are particularly valuable in inclusive classrooms, where preserving conceptual understanding is considerably more important than performing literal word-by-word translation. Nevertheless, most existing studies investigate individual multimodal tasks independently, while relatively few provide an integrated educational framework capable of supporting continuous real-time classroom communication for deaf learners.[12], [15], [34]

Another practical challenge concerns computational efficiency and deployment. State-of-the-art LLMs generally require substantial computational resources, high-performance GPUs, and large memory capacities, making their deployment difficult within resource-constrained educational institutions. Recent research therefore emphasizes lightweight models, model compression, parameter-efficient fine-tuning, retrieval-augmented generation (RAG), and edge-assisted inference to reduce computational requirements while preserving semantic performance. These developments improve the feasibility of deploying intelligent educational assistants within practical classroom environments.[13], [33], [34]

Overall, the literature demonstrates that Large Language Models possess considerable potential for enhancing educational accessibility through semantic understanding, contextual reasoning, and intelligent language generation. However, current research primarily focuses on isolated applications such as tutoring, question answering, or text generation, with comparatively limited attention devoted to integrating LLMs into complete classroom interpretation systems. These observations highlight the need for a unified multimodal architecture capable of combining speech recognition, semantic understanding, sign language generation, and visual communication to support inclusive learning environments. The following subsection therefore reviews recent advances in digital avatars and visual sign language presentation technologies.

2.5 Digital Avatars for Sign Language Communication

The visual presentation of sign language plays a crucial role in ensuring effective communication for deaf and hard-of-hearing learners. While textual transcription improves accessibility by converting spoken language into readable text, many deaf individuals primarily communicate through sign language, where meaning is conveyed using coordinated hand gestures, facial expressions, body posture, and spatial movement. Consequently, the visual representation of sign language through animated digital avatars has emerged as an active research area aimed at improving accessibility across education, healthcare, public services, and human-computer interaction.[8], [14], [18]

Early sign language animation systems primarily relied on manually designed three-dimensional (3D) avatars with predefined motion sequences representing individual signs. These systems generally utilized motion capture technologies or handcrafted animation libraries to reproduce sign gestures. Although such approaches demonstrated the feasibility of automated sign language visualization, they often produced unnatural movements, limited facial expressions, and discontinuous transitions between consecutive signs. Furthermore, manually constructing large animation libraries required considerable time, domain expertise, and computational resources, limiting their scalability for real-world educational applications.[35], [36], [37]

Recent advances in computer graphics, motion synthesis, and deep learning have substantially improved the realism and expressiveness of sign language avatars. Modern animation techniques employ skeletal motion modeling, inverse kinematics, neural motion generation, and physics-based animation to produce smoother and more natural sign transitions. These approaches enable improved synchronization between hand movements, facial expressions, and upper-body gestures, thereby enhancing the comprehensibility and visual quality of sign language animations. The adoption of deep generative models has further facilitated automatic synthesis of sign sequences from linguistic representations while reducing dependence on manually created animations. [18], [38], [39]

The emergence of artificial intelligence has further expanded the capabilities of digital avatars through multimodal integration. Contemporary systems combine speech recognition, natural language processing, computer vision, and animation technologies to generate dynamic visual representations of spoken content. Instead of displaying isolated signs independently, these systems increasingly attempt to preserve contextual meaning, sentence continuity, and non-verbal linguistic information throughout the communication process. Such multimodal integration represents a significant step toward developing intelligent accessibility systems capable of supporting real-time communication. [12], [13], [14]

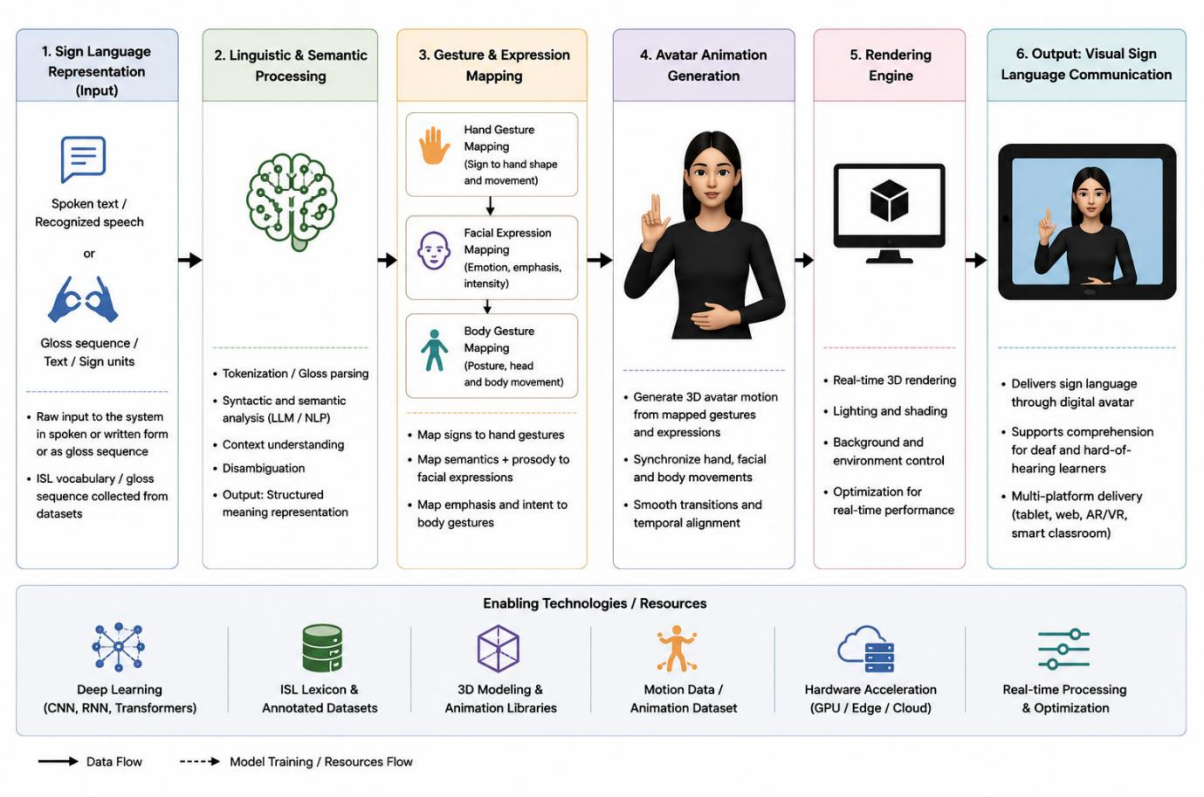


Fig. 5. General workflow of AI-based digital avatar generation for sign language communication.

Despite these technological advancements, several limitations remain unresolved in current avatar-based communication systems. Many existing approaches rely on predefined animation databases that restrict vocabulary expansion and adaptation to new educational domains. Furthermore, the visual quality of generated signs often depends on the availability of standardized sign dictionaries and annotated motion datasets, which remain limited for Indian Sign Language. Existing avatar systems also demonstrate limited capability to represent specialized engineering terminology, mathematical expressions, scientific concepts, and instructor-specific contextual explanations frequently encountered in higher education classrooms. As a result, many educational concepts continue to require manual interpretation or finger spelling, reducing communication efficiency.[6], [35]

Another important challenge concerns real-time deployment. Classroom interpretation requires continuous synchronization between spoken language, semantic understanding, sign generation, and visual rendering with minimal latency. However, many existing avatar systems operate in offline or semi-automatic environments where computational delays, animation rendering time, and network latency reduce their suitability for live classroom interaction. Moreover, user studies have reported that avatar comprehensibility depends not only on animation quality but also on contextual consistency, facial expressions, signing speed, and appropriate representation of linguistic grammar, all of which remain active research challenges.[9], [13], [21]

Recent developments in lightweight rendering engines, neural animation frameworks, and real-time graphics processing have improved the practical deployment of avatar technologies on desktop computers, mobile devices, and web-based platforms. These advances provide opportunities for developing scalable educational accessibility systems capable of supporting live classroom communication while maintaining acceptable computational performance. Nevertheless, most existing studies focus primarily on animation generation or avatar realism, with comparatively limited attention devoted to integrating avatar technologies with intelligent speech understanding and semantic reasoning within a unified educational communication framework.[13], [36], [37]

Overall, digital avatars represent a promising medium for delivering accessible visual communication to deaf learners. However, current research remains largely fragmented across animation generation, gesture recognition, speech processing, and language understanding. The absence of a comprehensive framework capable of integrating these complementary technologies highlights a significant research opportunity for developing intelligent multimodal classroom interpretation systems. This observation motivates the research gap discussed in the following subsection.

2.6 Research Gap and Motivation

The literature reviewed in the preceding subsections demonstrates substantial progress in individual technologies supporting accessibility for deaf and hard-of-hearing learners. Significant advances have been achieved in automatic speech recognition, sign language recognition and translation, large language models, and digital avatar technologies. Collectively, these developments have improved the performance of individual components involved in accessible communication and have expanded the potential of artificial intelligence in inclusive education.[12], [13], [14], [40]

Despite these advancements, existing research remains largely fragmented, with most studies addressing individual components independently rather than as part of a unified educational communication system. Automatic speech recognition systems primarily focus on converting speech into text but generally lack mechanisms for semantic interpretation of educational content. Similarly, sign language recognition and translation research emphasizes gesture recognition accuracy or translation quality while providing limited support for contextual understanding of classroom instruction. Existing large language model applications are predominantly designed for tutoring, question answering, or text generation, whereas digital avatar research concentrates on animation quality and visual realism. Consequently, these technologies are rarely integrated into a comprehensive framework capable of supporting continuous real-time classroom communication.[9], [13], [15], [20]

A further limitation concerns the educational context in which existing systems are evaluated. Most publicly available datasets and benchmark studies focus on conversational language, daily communication, or isolated vocabulary recognition under controlled experimental conditions. Comparatively little attention has been devoted to higher education classrooms, where instructors frequently employ technical terminology, mathematical expressions, scientific concepts, engineering diagrams, abbreviations, and context-dependent explanations. These instructional characteristics require semantic interpretation rather than direct lexical translation, making conventional accessibility solutions inadequate for engineering and technology education.[2], [16], [18]

Current sign language translation systems also encounter practical challenges arising from limited vocabulary coverage and the absence of standardized signs for many technical concepts. In classroom environments, previously unseen engineering terms, programming languages, scientific symbols, and discipline-specific terminology are frequently introduced during instruction. Existing systems typically address such situations through finger spelling or predefined dictionaries, which may interrupt communication flow and increase cognitive load for learners. Moreover, most systems do not dynamically adapt to the instructional context or learning objectives of individual classroom sessions.[6], [15]

Another significant challenge is the absence of semantic reasoning within existing accessibility frameworks. Although modern ASR systems generate accurate textual transcriptions, they generally preserve the original spoken sentence structure without simplifying complex explanations or identifying key educational concepts. Similarly, current avatar systems primarily visualize predefined sign sequences without considering instructional importance, learner comprehension, or contextual relationships among concepts. Consequently, the generated visual communication may not adequately support meaningful learning despite accurate transcription or sign generation.[12], [13], [20]

Furthermore, relatively few studies investigate practical deployment considerations such as computational efficiency, low-latency processing, modular scalability, and integration with existing

classroom infrastructure. Many proposed systems depend on cloud-based processing, high-performance computing resources, or laboratory-controlled environments that may limit their applicability in educational institutions with constrained computational resources. Real-time classroom interpretation therefore requires a balanced architecture capable of integrating multiple artificial intelligence components while maintaining practical deployment feasibility.[20], [33], [34]

Motivated by these observations, this research investigates a modular AI-driven multimodal framework for real-time classroom interpretation that integrates speech acquisition, multilingual automatic speech recognition, semantic language understanding, contextual content refinement, sign language representation, and visual communication within a unified architecture. Rather than treating speech recognition, language understanding, and sign generation as independent processes, the proposed framework adopts a modular design that enables seamless interaction among these complementary technologies. The objective is to facilitate meaningful and context-aware classroom communication for deaf learners while supporting practical deployment in higher education environments.

The proposed work is conducted as part of an AICTE-funded Research Promotion Scheme (RPS) – Lab to Market Project (Funding: ₹34.5 Lakhs), which aims to develop an intelligent assistive technology platform for improving classroom accessibility for deaf students in mainstream higher education. Unlike existing studies that primarily investigate isolated algorithms, the present work emphasizes the design of a scalable, modular, and deployable multimodal framework that serves as the foundation for subsequent implementation, optimization, and real-world validation.[5]

TABLE I

COMPARISON OF EXISTING APPROACHES AND THE PROPOSED MULTIMODAL CLASSROOM INTERPRETATION FRAMEWORK

Feature	Existing Studies	Proposed Framework
Real-time classroom support	Partial	✓
Educational semantic understanding	Limited	✓
LLM integration	Partial	✓
Indian Sign Language support	Limited	✓
Technical vocabulary handling	Limited	✓
Modular architecture	Rare	✓
Real-time multimodal integration	Rare	✓
Higher education focus	Limited	✓

Based on the limitations identified in the existing literature, the next section presents the proposed modular AI-driven multimodal classroom interpretation framework designed to enable real-time, context-aware communication for deaf learners in higher education.

3. Proposed Modular AI-Driven Multimodal Classroom Interpretation Framework

Recent advances in artificial intelligence have significantly improved individual accessibility technologies, including speech recognition, natural language processing, sign language translation, and digital avatar animation. However, as discussed in Section II, existing approaches predominantly address these components as independent research problems, resulting in fragmented solutions that are insufficient for continuous real-time classroom communication. Effective educational accessibility requires seamless interaction among multiple intelligent modules capable of understanding instructional context, generating meaningful sign language representations, and delivering synchronized visual communication with minimal latency.[12], [13], [14], [31]

Motivated by these challenges, this work proposes a modular AI-driven multimodal classroom interpretation framework designed to facilitate real-time communication for deaf students in higher education. Unlike conventional assistive systems that primarily perform speech transcription or isolated sign language translation, the proposed framework integrates multilingual automatic speech

recognition, semantic language understanding, contextual content refinement, sign language representation, expressive avatar animation, and real-time visual rendering into a unified architecture.

The proposed framework has been developed as part of the AICTE Research Promotion Scheme (RPS) – Lab to Market Project, funded by the All India Council for Technical Education (AICTE) under a sanctioned project budget of ₹34.5 lakhs. The primary objective is to develop a deployable intelligent assistive platform capable of improving classroom accessibility for deaf learners in mainstream engineering education while supporting scalability, modularity, and practical implementation.[5]

A modular architecture has been adopted to enable independent development, optimization, and future enhancement of each functional component without affecting the overall system. Such modularity facilitates integration of improved artificial intelligence models, supports hardware scalability, and allows adaptation to different educational domains, spoken languages, and sign language resources. Rather than relying on a single monolithic model, the framework combines specialized AI modules that cooperate to deliver context-aware visual communication in real time.

The overall workflow of the proposed framework is illustrated in Fig. 6. Beginning with classroom speech captured through a wireless microphone, the system performs multilingual speech recognition, semantic analysis using a Large Language Model (LLM), educational concept refinement, sign representation generation, expression mapping, and avatar animation before presenting synchronized visual communication to deaf learners through a tablet or web-based interface. Each functional module is described in detail in the following subsections.

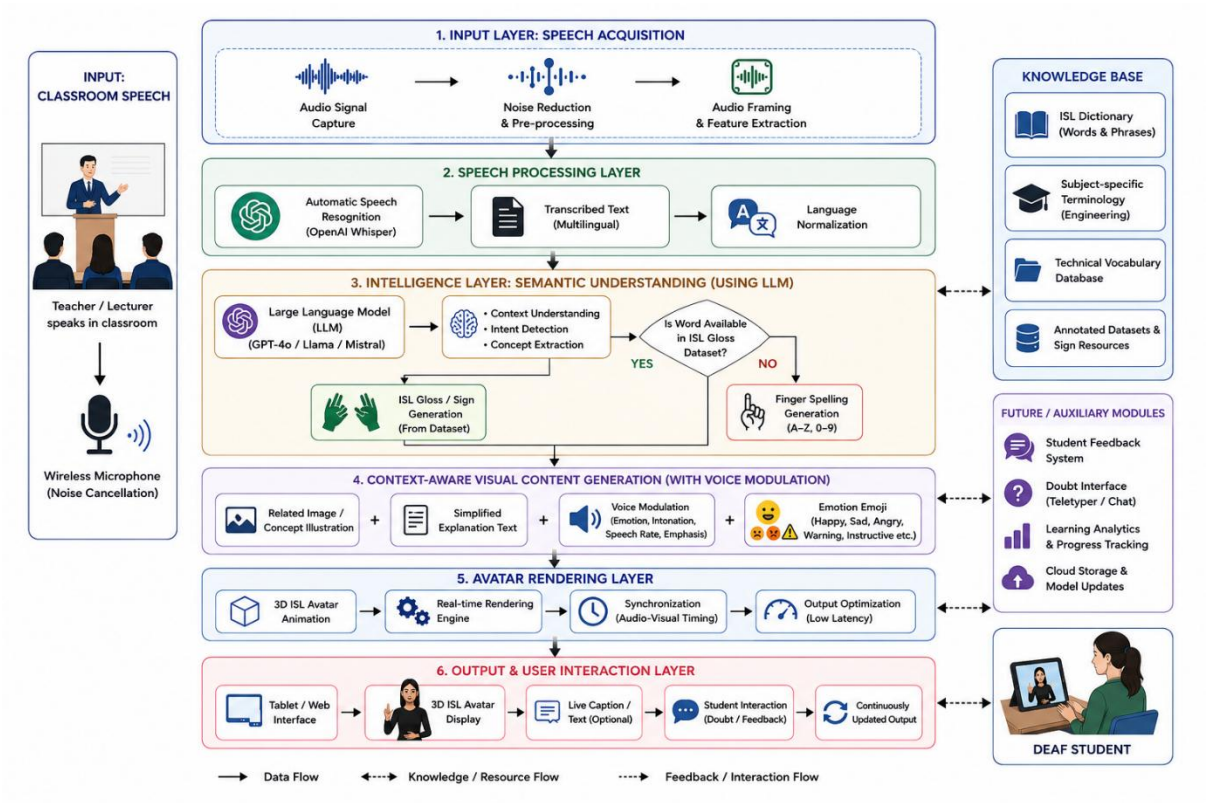


Fig. 6. Overall architecture of the proposed AI-driven multimodal classroom interpretation framework for deaf students.

3.1 Overall System Architecture

The proposed framework adopts a modular pipeline architecture to enable real-time classroom interpretation for deaf learners by integrating automatic speech recognition, semantic understanding, sign language processing, and multimodal visual communication. Unlike conventional speech-to-sign

translation systems that directly map spoken language to predefined sign representations, the proposed framework incorporates a Large Language Model (LLM) to provide contextual reasoning, detect out-of-vocabulary (OOV) technical terms, and generate educational visual support whenever suitable ISL representations are unavailable. The overall architecture of the proposed framework is illustrated in Fig. 6.

The system begins by capturing the instructor's speech using a wireless microphone with noise-cancellation capability. The captured speech signal is transmitted to the processing unit, where multilingual Automatic Speech Recognition (ASR) based on OpenAI Whisper converts the spoken lecture into a textual transcript. Language identification and normalization are subsequently performed to eliminate transcription inconsistencies and produce a clean text representation suitable for downstream processing.[31], [40]

The normalized transcript is then analyzed to determine whether each word or phrase has a corresponding representation in the available Indian Sign Language (ISL) glossary or dataset. For vocabulary that already exists in the ISL lexicon, the corresponding ISL gloss is directly retrieved, enabling efficient generation of sign language without additional semantic processing. This direct retrieval minimizes computational overhead and preserves real-time performance for frequently used vocabulary.[6]

When technical terminology or newly encountered words are absent from the ISL dataset, the framework activates the Large Language Model to perform context-aware semantic reasoning. Rather than functioning as a general translation engine, the LLM is selectively employed to identify the intended meaning of unknown terms, generate appropriate finger spelling sequences, simplify complex concepts where necessary, and retrieve relevant contextual visual information such as images or conceptual illustrations. This selective activation significantly reduces computational complexity while enhancing the system's ability to communicate domain-specific terminology encountered in engineering and higher education classrooms.[12], [33], [34]

The outputs generated from the ISL glossary matching and the LLM-assisted semantic reasoning are integrated to produce a unified multimodal presentation. Along with ISL signs or finger-spelled words, the framework presents contextually relevant images, simplified explanatory text, and visual illustrations to improve conceptual understanding. In addition, voice prosody features—including emotion, intonation, speech rate, and emphasis—are analyzed to generate appropriate visual indicators, such as expressive emojis or instructional cues, enabling deaf learners to perceive important emotional and pedagogical information that is naturally conveyed through spoken communication.

Finally, the integrated multimodal content is rendered through a three-dimensional ISL avatar and delivered to the learner using a tablet or web-based interface. The modular organization of the proposed architecture facilitates independent optimization of individual components, supports future integration of improved AI models and expanded ISL resources, and enables deployment across diverse educational environments without requiring modifications to the overall system architecture.

3.2 Speech Acquisition and Automatic Speech Recognition Module

The first stage of the proposed framework is responsible for capturing classroom speech and converting it into an accurate textual representation suitable for subsequent semantic processing. Since the effectiveness of all downstream modules depends on the quality of speech transcription, robust speech acquisition and automatic speech recognition are critical components of the overall system.

Classroom lectures are captured using a wireless microphone equipped with noise-cancellation capability to minimize the influence of environmental noise, student conversations, reverberation, and other acoustic disturbances commonly encountered in educational environments. The wireless microphone allows instructors to move freely during lectures while maintaining consistent speech

quality. The received audio signal is forwarded to the processing unit, where it undergoes preprocessing to remove background noise and normalize the speech signal before transcription.

Following preprocessing, the proposed framework employs the OpenAI Whisper automatic speech recognition model to generate multilingual text transcriptions from the instructor's speech. Whisper was selected because of its strong robustness to diverse accents, spontaneous speech, background noise, and multilingual classroom environments. Unlike conventional ASR systems that require language-specific acoustic models, Whisper utilizes a transformer-based encoder-decoder architecture trained on large-scale multilingual speech corpora, enabling reliable transcription across different speakers and teaching styles. These characteristics make Whisper particularly suitable for higher education classrooms, where instructors frequently switch between English, regional languages, technical terminology, and mathematical expressions.[27], [28], [29], [31]

The output generated by the ASR module consists of a continuous lecture transcript together with punctuation and sentence segmentation. To improve the quality of downstream processing, the transcript undergoes language identification and normalization, including correction of transcription inconsistencies, removal of redundant symbols, and standardization of technical expressions. This normalization stage ensures that the subsequent semantic processing module receives a clean and structured textual representation, thereby improving overall system reliability.

Unlike traditional classroom captioning systems that terminate after speech-to-text conversion, the proposed framework treats the ASR output as an intermediate representation rather than the final output. The normalized transcript is subsequently forwarded to the semantic reasoning module, where each word or phrase is analyzed to determine whether a corresponding Indian Sign Language (ISL) representation exists in the available glossary. This design enables efficient utilization of existing ISL vocabulary while selectively invoking advanced language understanding only when necessary.

To support real-time classroom interaction, the speech acquisition and ASR module is designed with low-latency operation as a primary objective. Continuous audio streaming and incremental transcription reduce communication delays between the instructor and deaf learners, enabling the subsequent modules to process speech in near real time. This modular design also permits future replacement of the ASR engine with more advanced speech recognition models without affecting the remaining components of the framework.

Once the lecture transcript has been generated and normalized, the framework performs context-aware semantic analysis to determine whether each recognized word or phrase is available in the ISL glossary or requires LLM-assisted processing. The semantic reasoning module is described in the following subsection.

3.3 Context-Aware Semantic Processing and LLM-Assisted Finger Spelling

After automatic speech recognition generates the normalized lecture transcript, the proposed framework performs semantic analysis to determine the most appropriate method for conveying each recognized word or phrase to deaf learners. Unlike conventional speech-to-sign translation systems that uniformly process all recognized text through a language model, the proposed framework adopts a selective semantic reasoning strategy. The objective is to maximize the utilization of the existing Indian Sign Language (ISL) glossary while invoking the Large Language Model (LLM) only when semantic assistance is required.

Initially, each recognized word or phrase is compared against the available ISL glossary and annotated sign language dataset. If a corresponding ISL gloss exists, the framework directly retrieves the associated sign representation without performing additional language model processing. This direct glossary matching minimizes computational overhead, reduces inference latency, and ensures that

standardized ISL signs are presented whenever available. Consequently, frequently used educational vocabulary can be translated efficiently while preserving real-time classroom interaction.[6]

When a recognized word is not present in the ISL glossary, it is classified as an Out-of-Vocabulary (OOV) term. Such words commonly include engineering terminology, scientific concepts, mathematical expressions, abbreviations, newly introduced technical vocabulary, and proper nouns that are not represented in existing ISL datasets. In these situations, the framework activates the Large Language Model to perform context-aware semantic reasoning rather than direct translation. The LLM first analyzes the surrounding sentence to determine the intended meaning of the unknown word within the instructional context. This contextual understanding is essential because the same word may represent different concepts depending on the subject being taught.[12], [33]

Following semantic interpretation, the LLM generates an appropriate finger-spelling sequence for the OOV term. Instead of attempting to create an artificial or potentially inaccurate sign, the system represents the unknown word using standard alphabet-based finger spelling, allowing the learner to identify the exact technical term being discussed. This approach preserves the original terminology while maintaining consistency with established sign language communication practices.[14], [15]

In addition to finger spelling, the LLM generates supporting educational content to improve conceptual understanding. Based on the interpreted context, the framework retrieves or generates a relevant image, diagram, or conceptual illustration together with a simplified textual explanation of the identified concept. These multimodal learning resources assist students in understanding unfamiliar technical terminology that may not yet have standardized ISL representations. For example, if the instructor introduces the term "Photosynthesis" and no corresponding ISL sign exists in the dataset, the system displays the finger-spelled word alongside an illustrative image and a concise educational explanation. Similarly, engineering concepts such as "Finite Element Analysis," "Backpropagation," or "Fourier Transform" can be presented using finger spelling supported by contextually relevant visual content.[12], [20]

Unlike many existing educational applications that employ LLMs as general-purpose translation engines, the proposed framework uses the language model selectively as an intelligent semantic support module. By activating the LLM only for OOV terms, the proposed architecture significantly reduces computational requirements while preserving high translation quality for vocabulary already available in the ISL glossary. This selective processing strategy represents one of the principal innovations of the proposed framework and enables practical deployment in real-time classroom environments.

3.4 Multimodal Content Generation and Avatar-Based Presentation

The final stage of the proposed framework transforms the processed linguistic and semantic information into an integrated multimodal learning interface designed specifically for deaf learners. Unlike conventional speech-to-sign translation systems that generate only sign language animations, the proposed framework combines Indian Sign Language (ISL) gestures, finger spelling, contextual visual content, and simplified educational text to improve conceptual understanding during live classroom instruction.

For vocabulary available in the ISL glossary, the corresponding ISL gloss is converted into sign language animation through the avatar rendering module. The generated signs are displayed sequentially while preserving the grammatical order defined by the selected ISL representation. This direct gloss-to-avatar mapping enables efficient visualization of frequently used vocabulary without requiring additional semantic processing, thereby reducing computational latency during real-time classroom interaction.[14]

When the semantic reasoning module identifies an out-of-vocabulary (OOV) term, the generated finger-spelling sequence is transmitted to the avatar controller instead of an unavailable sign gesture. The avatar reproduces the corresponding alphabetic hand configurations in sequence, allowing students to

identify the exact technical term introduced by the instructor. This strategy preserves the original terminology while avoiding incorrect or artificially generated sign representations that could alter the intended meaning.[15]

To enhance comprehension of unfamiliar concepts, the framework simultaneously presents context-aware educational resources generated during the semantic processing stage. These resources may include relevant images, conceptual diagrams, and simplified textual explanations associated with the identified technical term. The multimodal presentation enables learners to associate the finger-spelled word with its visual representation and educational context, thereby improving concept retention and reducing cognitive effort during classroom learning.[12], [13]

In addition to semantic information, the proposed framework incorporates voice prosody analysis to preserve instructional cues that are often lost during text-based translation. Characteristics such as emotional tone, emphasis, warning, encouragement, questioning, or instructional stress can be represented using simple visual indicators or expressive emojis displayed alongside the educational content. Rather than attempting to reproduce complex emotional facial expressions through the avatar, this lightweight representation provides learners with additional contextual information regarding the instructor's communication style while maintaining computational efficiency. This functionality represents an auxiliary educational support feature and can be further extended in future work.

The integrated multimodal output is delivered through a tablet or web-based application positioned in front of the learner. The interface synchronizes avatar animation, finger spelling, contextual images, simplified text, and instructional indicators within a single display, allowing students to follow classroom instruction without repeatedly shifting attention between multiple assistive tools. The modular presentation strategy also enables future integration of additional accessibility features, such as interactive concept exploration, personalized learning assistance, and adaptive educational feedback.

Overall, the proposed output generation module extends beyond conventional sign language visualization by combining linguistic accuracy with contextual educational support. This multimodal presentation strategy improves the accessibility of higher education lectures, particularly when technical terminology or newly introduced concepts are not represented in existing ISL resources.

4. Implementation Strategy

The proposed multimodal classroom interpretation framework is designed as a modular and scalable assistive system that can be deployed in higher education classrooms with minimal modification to existing teaching practices. The implementation strategy focuses on integrating commercially available hardware with state-of-the-art artificial intelligence models to achieve real-time classroom interpretation for deaf learners. The modular architecture enables independent development, testing, and future enhancement of individual components while maintaining seamless interaction among all system modules.

The implementation consists of four major functional stages: speech acquisition, multilingual speech recognition, context-aware semantic processing, and multimodal educational content generation. Each module communicates through well-defined interfaces, allowing future replacement or improvement of individual models without affecting the remaining components of the framework.

The proposed implementation has been planned using open-source artificial intelligence frameworks wherever possible to minimize deployment cost and improve scalability. Computationally intensive modules such as speech recognition and semantic processing can be executed either on a local workstation equipped with GPU acceleration or on a cloud-based server depending on the available computational resources. The final multimodal output is delivered to learners through a tablet or web-based interface positioned within the classroom.[12], [33], [40]

The following subsections describe the planned hardware infrastructure, software components, operational workflow, and expected system outputs.

4.1 Hardware Components

The proposed framework requires only a limited set of hardware components, making it suitable for deployment in educational institutions without significant infrastructure modification. A wireless microphone equipped with noise-cancellation capability is used to capture the instructor's speech during classroom lectures. The captured audio is transmitted to a processing workstation responsible for executing the artificial intelligence modules.

The processing workstation performs speech recognition, semantic reasoning, glossary matching, and multimodal content generation. Depending on computational requirements, the workstation may utilize GPU acceleration to improve inference speed during real-time classroom operation. The generated output is delivered to a tablet or web-enabled display accessible to deaf learners. This display presents the synchronized ISL avatar, finger spelling, contextual images, and simplified explanatory text.[6], [40]

The modular hardware configuration also permits future integration of additional devices such as smart boards, classroom cameras, wearable devices, or edge computing platforms without modifying the fundamental architecture of the proposed framework.

4.2 Software Components

The proposed framework integrates several artificial intelligence technologies to perform real-time classroom interpretation. OpenAI Whisper is employed for multilingual automatic speech recognition due to its robustness in noisy classroom environments and its ability to accurately transcribe multilingual speech.[40]

Semantic processing is performed using a Large Language Model that analyzes the lecture transcript, identifies out-of-vocabulary technical terminology, generates finger-spelling sequences, and retrieves contextually relevant educational resources. The framework also maintains an ISL glossary and sign language dataset that enables direct retrieval of standardized ISL representations for vocabulary already available within the dataset.[12], [33], [34]

The avatar rendering module converts ISL gloss sequences and finger-spelling commands into animated sign language using a three-dimensional avatar. Contextual images and simplified educational explanations are simultaneously presented through the learner interface to improve conceptual understanding.[13], [14]

The modular software architecture allows each component to be independently updated as improved datasets, speech recognition models, or language models become available.

4.3 Operational Workflow

During classroom instruction, the instructor's speech is continuously captured through the wireless microphone and transmitted to the speech recognition module. The generated transcript is normalized and analyzed to determine whether each recognized term exists within the available ISL glossary.

If a matching ISL representation is available, the corresponding gloss is directly retrieved and forwarded to the avatar generation module. If no suitable ISL representation exists, the Large Language Model performs context-aware semantic analysis, generates an appropriate finger-spelling sequence, and retrieves supporting educational resources including relevant images and simplified textual explanations.[12]

The multimodal output is synchronized and presented through the learner interface in real time, enabling deaf students to simultaneously observe ISL signs, finger-spelled technical terms, and contextual visual aids while following classroom instruction.

4.4 Expected System Outputs

The proposed framework is expected to generate multiple synchronized outputs that collectively improve classroom accessibility. These outputs include:

Accurate multilingual speech transcription.
ISL avatar animation for vocabulary available in the ISL glossary.
Finger spelling for out-of-vocabulary technical terminology.
Context-aware educational images and conceptual illustrations.
Simplified explanatory text generated through semantic reasoning.
Visual indicators representing important instructional cues derived from voice prosody, where applicable.

By combining these complementary outputs within a single interface, the proposed framework aims to reduce communication barriers encountered by deaf learners while improving conceptual understanding of technical subjects in higher education.

5. Design Analysis and Expected Performance

This section evaluates the proposed framework from a system-design perspective by discussing its expected operational characteristics, advantages over existing approaches, computational behaviour, and anticipated impact on classroom accessibility. Although the complete prototype is currently under development as part of the AICTE Research Promotion Scheme (RPS) – Lab to Market Project, the modular architecture enables qualitative assessment of the proposed system based on the functional contribution of each module.[13]

5.1 Comparative Analysis

To evaluate the proposed framework from a functional perspective, a qualitative comparison was conducted with conventional speech-to-text systems and existing speech-to-sign translation approaches. The comparison focuses on key characteristics that directly influence classroom accessibility for deaf learners, including real-time interpretation, utilization of Indian Sign Language (ISL) resources, handling of out-of-vocabulary (OOV) technical terms, contextual semantic understanding, and support for multimodal educational communication. The objective of this analysis is to identify the functional advantages introduced by the proposed framework rather than to present experimental performance metrics.[15], [18], [20]

TABLE II

QUALITATIVE COMPARISON OF THE PROPOSED FRAMEWORK WITH EXISTING APPROACHES

Feature	Speech-to-Text Systems	Speech-to-Sign Systems	Proposed Framework
Real-time speech transcription	✓	✓	✓
Direct ISL glossary matching	✗	✓	✓
Context-aware semantic reasoning	✗	Limited	✓
Automatic OOV word detection	✗	✗	✓
LLM-assisted finger spelling	✗	✗	✓
Context-aware visual support	✗	✗	✓
Simplified educational explanation	✗	✗	✓
Support for technical terminology	Limited	Limited	✓
Modular system architecture	Limited	Limited	✓
Scalability for future enhancement	Limited	Limited	✓

As shown in Table II, conventional speech-to-text systems are primarily designed to convert spoken language into textual transcripts and therefore do not provide sign language interpretation or semantic assistance for deaf learners. Although these systems enable caption generation, they cannot effectively communicate technical concepts that require visual or linguistic representation beyond plain text.[2], [40]

Existing speech-to-sign translation systems improve accessibility by converting speech into sign language representations. However, their performance largely depends on the availability of predefined sign vocabularies and annotated sign language datasets. When previously unseen technical terms, scientific concepts, abbreviations, or subject-specific vocabulary are encountered, many existing systems either fail to generate appropriate signs or rely on manual intervention. Furthermore, most available systems perform direct translation without providing additional contextual support to facilitate conceptual understanding.[14], [15], [41]

The proposed framework addresses these limitations through a modular processing strategy that combines direct ISL glossary matching with selective LLM-assisted semantic reasoning. Rather than invoking the language model for every recognized word, the framework first searches the ISL glossary for an existing sign representation. Only when an out-of-vocabulary term is detected is the LLM activated to determine the contextual meaning of the word, generate an appropriate finger-spelling sequence, and retrieve supporting educational content such as relevant images and simplified textual explanations. This selective processing strategy preserves computational efficiency while extending the system's capability to communicate specialized terminology commonly encountered in engineering and higher education classrooms.

Overall, the comparative analysis demonstrates that the proposed framework extends beyond conventional speech transcription and sign language translation by integrating contextual semantic understanding with multimodal educational support. These capabilities enable a more comprehensive accessibility solution while maintaining compatibility with existing ISL resources and providing a scalable foundation for future enhancements.

5.2 Illustrative Processing Scenarios

To demonstrate the operational behavior of the proposed framework, representative classroom scenarios are presented in this subsection. These examples illustrate how the framework dynamically selects different processing paths depending on whether a recognized term is available in the Indian Sign Language (ISL) glossary. Rather than processing every recognized word through the Large Language Model (LLM), the proposed architecture activates semantic reasoning only for out-of-vocabulary (OOV) terms. This selective processing strategy reduces computational overhead while preserving the ability to communicate technical concepts that are not represented in the existing ISL dataset.[12]

TABLE III

ILLUSTRATIVE PROCESSING SCENARIOS OF THE PROPOSED FRAMEWORK

Spoken Term	Available in ISL Dataset	Processing Path	Generated Output
Water	Yes	Direct ISL Gloss Retrieval	ISL Avatar displays corresponding sign
Voltage	Yes	Direct ISL Gloss Retrieval	ISL Avatar displays corresponding sign
Photosynthesis	No	LLM → Finger Spelling → Context Support	Finger spelling + Related image + Simplified explanation
Transformer	No	LLM → Finger Spelling → Context Support	Finger spelling + Contextual illustration
Backpropagation	No	LLM → Finger Spelling → Context Support	Finger spelling + Educational explanation
Fourier Transform	No	LLM → Finger Spelling → Context Support	Finger spelling + Concept visualization

As illustrated in Table III, the proposed framework first determines whether the recognized vocabulary exists in the available ISL glossary. For commonly used words such as Water and Voltage, the corresponding ISL sign is directly retrieved and rendered through the avatar without invoking the LLM. This direct retrieval mechanism minimizes processing latency and enables efficient real-time interpretation for vocabulary already represented within the ISL dataset.

For technical or specialized terms that are absent from the available ISL glossary, the framework automatically switches to the semantic reasoning pathway. The LLM analyzes the instructional context of the recognized word, generates an appropriate finger-spelling sequence, and retrieves supporting educational resources such as relevant images, conceptual illustrations, or simplified textual explanations. This multimodal presentation enables learners to identify unfamiliar terminology while simultaneously understanding its educational meaning.[12], [20]

Unlike conventional speech-to-sign systems that either omit unknown vocabulary or require manual intervention, the proposed framework provides an automated mechanism for handling OOV technical terms without disrupting the continuity of classroom communication. This selective activation of the LLM represents one of the principal innovations of the proposed framework, allowing standardized ISL signs to be preserved whenever available while extending accessibility to previously unsupported academic vocabulary.

The illustrative scenarios demonstrate that the proposed architecture is capable of supporting both everyday classroom communication and domain-specific higher education content within a unified multimodal framework. This design significantly enhances the practical applicability of the system in engineering education, where technical terminology frequently exceeds the vocabulary currently available in publicly accessible ISL resources.

5.3 Advantages and Practical Implications of the Proposed Framework

The proposed framework introduces several functional and practical improvements over conventional classroom accessibility systems by combining modular artificial intelligence components within a unified interpretation pipeline. Unlike existing speech-to-text or speech-to-sign solutions that primarily focus on isolated translation tasks, the proposed architecture integrates multilingual speech recognition, direct ISL glossary matching, selective LLM-assisted semantic reasoning, finger spelling, contextual educational support, and avatar-based communication into a single end-to-end framework. This integration enables a more comprehensive and scalable solution for improving accessibility in higher education classrooms.

One of the principal advantages of the proposed framework is the selective utilization of the Large Language Model (LLM). Instead of processing every recognized word through the language model, the framework first searches the available Indian Sign Language (ISL) glossary for a corresponding sign representation. Only when an out-of-vocabulary (OOV) term is detected does the system invoke the LLM to perform semantic reasoning, generate finger spelling, and provide contextual educational support. This selective activation significantly reduces computational overhead while preserving the benefits of advanced language understanding for specialized technical vocabulary.

Another important contribution of the proposed framework is its ability to support engineering and higher education terminology. Existing ISL datasets are primarily composed of everyday vocabulary and therefore provide limited coverage of domain-specific concepts encountered in science, engineering, and technology courses. By combining finger spelling with context-aware images and simplified explanations, the proposed framework enables learners to understand unfamiliar technical concepts without requiring newly defined ISL signs for every specialized term. This capability makes the framework particularly suitable for multidisciplinary higher education environments.[16]

TABLE IV

ADVANTAGES OF THE PROPOSED FRAMEWORK

Component	Existing Systems	Proposed Framework
ISL Translation	Fixed vocabulary	ISL glossary + LLM-assisted finger spelling
Unknown Technical Terms	Not handled / Manual intervention	Automatic OOV detection and finger spelling
Semantic Understanding	Limited	Context-aware reasoning using LLM
Educational Support	Sign language only	Images, simplified text, contextual explanation
System Design	Monolithic	Modular and scalable
Deployment	Specialized setups	Classroom-friendly, modular implementation
Future Expansion	Limited	Easy integration of improved AI models and larger ISL datasets

The modular architecture further improves the scalability and maintainability of the proposed system. Each functional module—including speech acquisition, automatic speech recognition, semantic processing, ISL glossary matching, finger spelling generation, contextual content generation, and avatar rendering—operates independently through clearly defined interfaces. Consequently, future improvements to any individual component, such as upgraded speech recognition models, expanded ISL datasets, or more advanced language models, can be incorporated without requiring major modifications to the remaining system.

The proposed framework also enhances the learning experience by providing multimodal educational support rather than relying solely on sign language translation. In addition to ISL signs and finger spelling, the learner receives relevant images, conceptual illustrations, simplified explanatory text, and optional instructional indicators derived from the instructor's voice characteristics. Presenting multiple complementary forms of information enables students to associate unfamiliar technical terminology with meaningful educational context, thereby supporting improved comprehension and knowledge retention.

From a deployment perspective, the proposed framework has been designed using commercially available hardware and predominantly open-source software components. The use of a wireless microphone, multilingual speech recognition, local or cloud-based processing, and tablet or web-based visualization enables practical deployment in existing classrooms with minimal infrastructure modification. This cost-effective implementation strategy supports the broader objective of improving accessibility within mainstream educational institutions.

Overall, the proposed framework extends beyond conventional assistive technologies by integrating linguistic accessibility with contextual educational assistance. Through its modular design, selective semantic reasoning, and multimodal communication strategy, the framework provides a practical foundation for future intelligent classroom accessibility systems capable of supporting deaf learners in higher education.

6. Conclusion and Future Work

This paper presented a modular AI-driven multimodal classroom interpretation framework designed to improve educational accessibility for deaf students in higher education. The proposed architecture integrates multilingual automatic speech recognition, Indian Sign Language (ISL) glossary matching, selective Large Language Model (LLM)-assisted semantic reasoning, automatic finger spelling for out-of-vocabulary (OOV) terms, contextual educational support, and avatar-based visual communication within a unified real-time framework. Unlike conventional speech-to-sign systems that rely solely on predefined sign vocabularies, the proposed framework introduces a selective processing strategy in which the LLM is activated only when suitable ISL representations are unavailable. This approach preserves computational efficiency while extending the system's capability to communicate technical and domain-specific terminology frequently encountered in engineering education.

The proposed framework further enhances classroom accessibility by combining multiple communication modalities, including ISL signs, finger spelling, contextual images, conceptual

illustrations, and simplified educational text within a single integrated learning interface. By leveraging existing ISL resources whenever possible and intelligently supporting unknown vocabulary through semantic reasoning, the framework addresses one of the major limitations of current sign language interpretation systems. The modular system architecture also facilitates scalability, allowing future improvements in speech recognition, language models, ISL datasets, and avatar technologies to be incorporated without redesigning the complete framework.

The development of this framework is being carried out as part of the AICTE Research Promotion Scheme (RPS) – Lab to Market Project, with the objective of developing a practical assistive technology that can be deployed in mainstream classrooms to support inclusive engineering education. The proposed design establishes a foundation for future intelligent classroom accessibility systems capable of bridging communication gaps between instructors and deaf learners while preserving both linguistic accuracy and conceptual understanding.

Future work will focus on the implementation and experimental validation of the proposed framework in real classroom environments. Planned developments include expanding the ISL glossary with domain-specific educational vocabulary, improving multilingual support, optimizing real-time system latency, and evaluating the framework with deaf learners and educators through user-centered studies. Additional enhancements such as adaptive personalization, emotion-aware avatar expressions, interactive student feedback mechanisms, and continuous learning from classroom interactions will also be investigated to further improve accessibility and educational effectiveness.

ACKNOWLEDGMENT

The authors gratefully acknowledge the financial support provided by the All India Council for Technical Education (AICTE) under the Research Promotion Scheme (RPS) – Lab to Market initiative for the development of this research. This work forms part of an AICTE-funded project with a sanctioned budget of ₹34.5 lakhs, aimed at developing an AI-driven multimodal classroom interpretation framework to improve educational accessibility for deaf learners in higher education.

The authors also express their sincere gratitude to AWH Engineering College, Kozhikode, Kerala, for providing institutional support, research facilities, and encouragement throughout the course of this work. Special appreciation is extended to all faculty members, domain experts, and members of the project team whose valuable discussions and suggestions contributed to the development of the proposed framework.

REFERENCES

[1] United Nations, “Convention on the Rights of Persons with Disabilities (CRPD) | United Nations Enable,” United Nations. Accessed: Jul. 19, 2026. [Online]. Available: https://www.un.org/development/desa/disabilities/convention-on-the-rights-of-%20persons-with-disabilities.html?utm_source=chatgpt.com

[2] “World report on hearing.” Accessed: Jul. 19, 2026. [Online]. Available: https://www.who.int/publications/i/item/9789240020481

[3] “Global education monitoring report 2023, Southeast Asia: technology in.” Accessed: Jul. 19, 2026. [Online]. Available: https://www.unesco.org/en/articles/global-education-monitoring-report-2023-southeast-asia-technology-education-tool-whose-terms-summary

[4] “The Salamanca Statement and Framework for Action on Special Needs Education - UNESCO Digital Library.” Accessed: Jul. 20, 2026. [Online]. Available: https://unesdoc.unesco.org/ark:/48223/pf0000098427

[5] “Government of India, All India Council for Technical Education |.” Accessed: Jul. 21, 2026. [Online]. Available: https://www.aicte.gov.in/

[6] “Homepage | Indian Sign Language Research and Training Center ,Government of India | India.” Accessed: Jul. 20, 2026. [Online]. Available: https://islrtc.nic.in/

[7] C. I. of E. Technology, “BASIC INDIAN SIGN LANGUAGE (ISL) COURSE 2026 |.” Accessed: Jul. 21, 2026. [Online]. Available: https://ciet.ncert.gov.in/activity/ciet.ncert.gov.in/activity/isl2026?lang=en
[8] G. Tang, “Sign language and inclusive deaf education: An Asian perspective,” *Deafness & Education International*, vol. 26, no. 1, pp. 1–5, Jan. 2024, doi: 10.1080/14643154.2024.2302702.
[9] A. Coy, P. S. Mohammed, and P. Skerrit, “Inclusive Deaf Education Enabled by Artificial Intelligence: The Path to a Solution,” *Int J Artif Intell Educ*, vol. 35, no. 1, pp. 96–134, Mar. 2025, doi: 10.1007/s40593-024-00419-9.
[10] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, “Robust Speech Recognition via Large-Scale Weak Supervision,” Dec. 06, 2022, *arXiv*: arXiv:2212.04356. doi: 10.48550/arXiv.2212.04356.
[11] A. Vaswani *et al.*, “Attention Is All You Need,” Aug. 02, 2023, *arXiv*: arXiv:1706.03762. doi: 10.48550/arXiv.1706.03762.
[12] OpenAI *et al.*, “GPT-4 Technical Report,” Mar. 04, 2024, *arXiv*: arXiv:2303.08774. doi: 10.48550/arXiv.2303.08774.
[13] S. Yin *et al.*, “A Survey on Multimodal Large Language Models,” *National Science Review*, vol. 11, no. 12, p. nwae403, Nov. 2024, doi: 10.1093/nsr/nwae403.
[14] N. C. Camgoz, O. Koller, S. Hadfield, and R. Bowden, “Sign Language Transformers: Joint End-to-end Sign Language Recognition and Translation,” Mar. 30, 2020, *arXiv*: arXiv:2003.13830. doi: 10.48550/arXiv.2003.13830.
[15] A. Núñez-Marcos, O. Perez-de-Viñaspre, and G. Labaka, “A survey on Sign Language machine translation,” *Expert Systems with Applications*, vol. 213, p. 118993, Mar. 2023, doi: 10.1016/j.eswa.2022.118993.
[16] R. Damdoo and P. Kumar, “An integrative survey on Indian sign language recognition and translation,” *IET Image Processing*, vol. 19, no. 1, p. e70000, 2025, doi: 10.1049/ipr2.70000.
[17] O. Koller, “Quantitative Survey of the State of the Art in Sign Language Recognition,” Aug. 29, 2020, *arXiv*: arXiv:2008.09918. doi: 10.48550/arXiv.2008.09918.
[18] M. Kalinowski and B. Kostek, “Machine learning in sign language: A comprehensive analysis and trend survey,” *Computer Science Review*, vol. 60, p. 100895, May 2026, doi: 10.1016/j.cosrev.2026.100895.
[19] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, “Robust Speech Recognition via Large-Scale Weak Supervision,” arXiv.org. Accessed: Jul. 21, 2026. [Online]. Available: https://arxiv.org/abs/2212.04356v1
[20] W. X. Zhao *et al.*, “A Survey of Large Language Models,” Mar. 18, 2026, *arXiv*: arXiv:2303.18223. doi: 10.48550/arXiv.2303.18223.
[21] S. K, S. V, and S. K. D, “AI Powered Learning Platform for Deaf and Mute People Using Sign Language,” in *2026 Contemporary Computing Innovations Conference (CCIC)*, Feb. 2026, pp. 1–6. doi: 10.1109/CCIC68129.2026.11485962.
[22] “Rights of Persons with Disabilities Act, 2016.,” *An Act to give effect to the United Nations Convention on the Rights of Persons with Disabilities and for matters connected therewith or incidental thereto.*, Dec. 2016, Accessed: Jul. 21, 2026. [Online]. Available: http://indiacode.nic.in/handle/123456789/2155
[23] “ISL (Indian Sign Language Portal) || FDMSE || RKMVERI || COIMBATORE CAMPUS – Dictionary on Indian Sign Language (ISL) signs by FDMSE, Coimbatore.” Accessed: Jul. 21, 2026. [Online]. Available: https://indiansignlanguage.org/
[24] S. Patra, “ISL Project.” Accessed: Jul. 21, 2026. [Online]. Available: https://cs.rkmvu.ac.in/~isl/
[25] A. Joshi *et al.*, “iSign: A Benchmark for Indian Sign Language Processing,” in *Findings of the Association for Computational Linguistics: ACL 2024*, L.-W. Ku, A. Martins, and V. Srikumar, Eds., Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 10827–10844. doi: 10.18653/v1/2024.findings-acl.643.
[26] “nep2020.pdf.” Accessed: Jul. 21, 2026. [Online]. Available: https://aicte.gov.in/sites/default/files/nep2020.pdf
[27] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, “wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations,” Oct. 22, 2020, *arXiv*: arXiv:2006.11477. doi: 10.48550/arXiv.2006.11477.

[28] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units," Jun. 14, 2021, *arXiv*: arXiv:2106.07447. doi: 10.48550/arXiv.2106.07447.

[29] S. Chen *et al.*, "WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing," *IEEE J. Sel. Top. Signal Process.*, vol. 16, no. 6, pp. 1505–1518, Oct. 2022, doi: 10.1109/JSTSP.2022.3188113.

[30] D. Amodei *et al.*, "Deep Speech 2 : End-to-End Speech Recognition in English and Mandarin," in *Proceedings of The 33rd International Conference on Machine Learning*, PMLR, Jun. 2016, pp. 173–182. Accessed: Jul. 21, 2026. [Online]. Available: https://proceedings.mlr.press/v48/amodei16.html

[31] *openai/whisper*. (Jul. 20, 2026). Python. OpenAI. Accessed: Jul. 20, 2026. [Online]. Available: https://github.com/openai/whisper

[32] "Deep Learning." Accessed: Jul. 21, 2026. [Online]. Available: https://www.deeplearningbook.org/

[33] A. Grattafiori *et al.*, "The Llama 3 Herd of Models," Nov. 23, 2024, *arXiv*: arXiv:2407.21783. doi: 10.48550/arXiv.2407.21783.

[34] G. Team *et al.*, "Gemini: A Family of Highly Capable Multimodal Models," May 09, 2025, *arXiv*: arXiv:2312.11805. doi: 10.48550/arXiv.2312.11805.

[35] B. A. Al Abdullah, G. A. Amoudi, and H. S. Alghamdi, "Advancements in Sign Language Recognition: A Comprehensive Review and Future Prospects," *IEEE Access*, vol. 12, pp. 128871–128895, 2024, doi: 10.1109/ACCESS.2024.3457692.

[36] B. Foundation, "Blender - The Free and Open Source 3D Creation Software — blender.org," Blender. Accessed: Jul. 21, 2026. [Online]. Available: https://www.blender.org/

[37] "Unity: Develop, Deploy, and Grow | The World's Leading Game Engine," Unity. Accessed: Jul. 21, 2026. [Online]. Available: https://unity.com

[38] "GitHub - CMU-Perceptual-Computing-Lab/openpose: OpenPose: Real-time multi-person keypoint detection library for body, face, hands, and foot estimation · GitHub." Accessed: Jul. 21, 2026. [Online]. Available: https://github.com/CMU-Perceptual-Computing-Lab/openpose

[39] "MediaPipe Solutions guide | Google AI Edge," Google for Developers. Accessed: Jul. 21, 2026. [Online]. Available: https://developers.google.com/edge/mediapipe/solutions/guide

[40] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust Speech Recognition via Large-Scale Weak Supervision".

[41] R. Damdoo and P. Kumar, "An integrative survey on Indian sign language recognition and translation," *IET Image Processing*, vol. 19, no. 1, p. e70000, Jan. 2025, doi: 10.1049/ipr2.70000.