

Article

Virtual Trainer for Learning Mexican Sign Language Using Video Similarity Analysis

Felipe de Jesús Rivera-Cervantes ¹, Diana-Margarita Córdova-Esparza ^{1,*}, Juan Terven ²,
Julio-Alejandro Romero-González ¹, Jaime-Rodrigo González-Rodríguez ¹, Mauricio-Arturo Ibarra-Corona ¹
and Pedro-Alfonso Ramírez-Pedraza ^{2,3}

- ¹ Facultad de Informática, Universidad Autónoma de Querétaro, Av. de las Ciencias S/N, Juriquilla, Querétaro 76230, Mexico; frivera22@alumnos.uaq.mx (F.d.J.R.-C.); julio.romero@uaq.mx (J.-A.R.-G.); jgonzalez87@alumnos.uaq.mx (J.-R.G.-R.); mauricio.ibarra@uaq.mx (M.-A.I.-C.)
² Centro de Investigación en Ciencia Aplicada y Tecnología Avanzada-Unidad Querétaro, Instituto Politécnico Nacional, Cerro Blanco 141, Colinas del Cimatario, Querétaro 76090, Mexico; jrtervens@ipn.mx (J.T.); aramirez_ixm@ipn.mx or pedro.ramirez@secihti.mx (P.-A.R.-P.)
³ Secretaría de Ciencia, Humanidades, Tecnología e Innovación SECIHTI, Investigadores por México, Alvaro Obregón, Ciudad de México 03940, Mexico
* Correspondence: diana.cordova@uaq.mx

Abstract

Learning Mexican Sign Language (MSL) benefits from interactive systems that provide immediate feedback without requiring specialized sensors. This work presents a virtual training platform that operates with a conventional RGB camera and applies computer vision techniques to guide learners in real time. A dataset of 335 videos was recorded across 12 lessons with professional interpreters and used as the reference material for practice sessions. From each video, 48 keypoints corresponding to hands and facial landmarks were extracted using MediaPipe, normalized, and compared with user trajectories through Dynamic Time Warping (DTW). A sign is accepted when the DTW distance is below a similarity threshold, allowing users to receive quantitative feedback on performance. Additionally, an experimental baseline using video embeddings generated by the Qwen2.5-VL, VideoMAEv2, and VJEPa2 models and classified via Matching Networks was evaluated for scalability. Results show that the DTW-based module provides accurate and interpretable feedback for guided practice with minimal computational cost, while the embedding-based approach serves as an exploratory baseline for larger-scale classification and semi-automatic labeling. A user study with 33 participants evidenced feasibility and perceived usefulness (all category means significantly above neutral; Cronbach's $\alpha = 0.81$). Overall, the proposed framework offers an accessible, low-cost, and effective solution for inclusive MSL education and represents a promising foundation for future multimodal sign-language learning tools.

Keywords: Mexican Sign Language; practice system; RGB camera; Dynamic Time Warping; video embeddings; Qwen2.5-VL; VideoMAEv2; VJEPa2



Academic Editor: Reza Saatchi

Received: 14 October 2025

Revised: 12 November 2025

Accepted: 19 November 2025

Published: 21 November 2025

Citation: Rivera-Cervantes, F.d.J.; Córdova-Esparza, D.-M.; Terven, J.; Romero-González, J.-A.; González-Rodríguez, J.-R.; Ibarra-Corona, M.-A.; Ramírez-Pedraza, P.-A. Virtual Trainer for Learning Mexican Sign Language Using Video Similarity Analysis. *Technologies* **2025**, *13*, 540. <https://doi.org/10.3390/technologies13120540>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Mexican Sign Language (MSL) is essential for communication, education, and social inclusion within the Deaf community in Mexico. However, its formal teaching remains limited and heterogeneous, partly due to the inherent complexity of coordinating hands, arms, head, torso, and non-manual markers (facial expressions) to convey precise information.

This situation perpetuates communication barriers that hinder access to basic services and rights for people with hearing disabilities.

The use of virtual trainers for sign language instruction represents an innovative solution that helps reduce communication barriers and enhances the teaching–learning process. Several studies have explored this field using neural networks and/or immersive environments. Examples include tutoring and accessibility systems based on deep learning models [1–4]; virtual reality proposals employing 3D avatars and interactive scenarios [5–7]; and sensor-based solutions using devices such as Leap Motion for manual gesture classification [8,9]. In parallel, translation and representation systems have been developed for various sign languages through avatars capable of converting text or speech into animated sequences that reproduce the corresponding gestures [10,11].

Among the most recent works in the literature dedicated to the teaching and learning of sign languages, several studies stand out for their focus on integrating digital tools and interactive environments to improve the acquisition of visual–gestural language. These contributions range from didactic proposals mediated by online training platforms to immersive experiences in virtual and augmented reality environments, all aimed at strengthening comprehension, participation, and educational inclusion of both Deaf and hearing individuals.

The system proposed by [12] was developed to improve the learning and communication of Chinese Sign Language (CSL) through an interactive smartphone-based platform. Recognizing the limitations of traditional learning methods such as books and videos, this tool provides a more dynamic and engaging approach for both Deaf and hearing users. The system suggests the top three possible signs, helping users recall the correct gesture. The study concludes that this interactive approach significantly improves the efficiency of sign language learning and contributes to reducing communication barriers between Deaf and hearing individuals.

In the work of Lara et al. [13], a virtual robotic training platform was designed and implemented to teach Mexican Sign Language (MSL) to both deaf and hearing users. The proposed system, named MSL–Virtual Robotic Trainer (MSL–VRT), was modeled after the human upper limb—comprising the hand, forearm, and wrist—and incorporates 25 degrees of freedom (DOF) to perform finger-spelling corresponding to the 21 static one-handed signs of the MSL alphabet.

Gutiérrez Navarrete [14] proposed a project aimed at facilitating initial communication between Deaf and hearing individuals through a mobile learning application for Colombian Sign Language (LSC). The initiative sought to overcome communication barriers and promote interaction between both groups in everyday contexts. Considering the large population of people with hearing disabilities in Colombia, effective communication was identified as an urgent social need. The proposed application was designed to teach LSC using practical daily-life phrases, fostering empathy, inclusion, and awareness of the Deaf community. The results were promising, supported by non-probabilistic sampling methods. The authors emphasized the importance of continuing research and developing inclusive strategies to foster a more equitable and accessible society, noting that the prototype could be further refined to enhance usability and overall performance.

The proposal presented by [15] addresses the limited availability of sign language learning resources that provide immediate feedback to learners. This study evaluates the effectiveness of an online tool designed to offer real-time video feedback, accessible even on standard laptops. By integrating Google’s MediaPipe technology with a robust logical framework, the prototype aims to deliver accurate and instantaneous responses to support effective American Sign Language (ASL) instruction. Focusing on the ASL alphabet, the research analyzes computer vision-based learning tools for static and dynamic

signs, assessing the potential and current limitations of the system in advancing sign language education.

The article developed by [16] introduces an innovative approach to building an intelligent avatar tutor for teaching Bangla Sign Language (BSL). The proposed system applies image processing and machine learning techniques to recognize and analyze hand gestures from a dedicated dataset, generating corresponding animations that allow the avatar to reproduce numbers and alphabet signs interactively. The study involved pre-processing and collecting gesture data to train the model, resulting in an intuitive and user-friendly interface that enables learners to physically practice signs, improve sensorimotor engagement, and receive immediate feedback through cosine similarity evaluation.

The study by [17] presents a web application based on machine learning designed to make sign language learning more accessible. Representing a step forward in sign language education, the proposed platform differs from traditional approaches by assigning users various words to spell through hand signs. The learners must correctly sign each letter to complete the word and earn points, creating a gamified and engaging experience. The paper details the system's development, main features, and underlying machine learning framework. Implemented with HTML, CSS, JavaScript and Flask, the Web Application uses the user's webcam to capture live video and display predictions of real-time models, enabling interactive practice sessions.

The system proposed by [18] was designed, developed, and evaluated to support sign language learning and communication within Social Virtual Reality (SVR) environments. Building on insights from previous research, the system was implemented as a Virtual Reality Learning Environment (VRLE) that replicates the real-life processes of teaching and assessing sign language learners. The study's findings indicate that, although current technological solutions have proven effective in transferring non-verbal cues (NVCs) and enhancing interaction, several challenges remain. Further research is required to improve the overall quality and effectiveness of sign language communication in virtual environments.

The study conducted by Erofeeva et al. [19] examined how virtual reality (VR) can foster inclusive and immersive learning through community-based sign language (SL) education within the VRChat platform. Using multimodal video analysis of two French SL classes taught by a Deaf instructor and involving learners with diverse sensory, linguistic, and technological profiles, the research explored how VR tools—such as air pens and spatial-directional emojis—facilitated communication and engagement. The findings suggest that these tools helped reduce technological and communicative asymmetries, although their effectiveness depended on participants' interactional competencies.

Coy et al. [20] examined the opportunities and challenges of an approach designed to facilitate real-time communication between Deaf students and hearing teachers who do not use sign language, integrating speech and language technologies, computer vision, machine translation systems, and 3D avatars powered by artificial intelligence (AI).

Similarly, the initiative developed by [21] at the National Taiwan Library promotes digital and interactive strategies to enhance communication between sign language users and the hearing community, reflecting the growing emphasis on inclusivity for people with disabilities. The "Interactive System of Digital Sign Language Picture Books," which combines augmented reality (AR) and artificial intelligence (AI), aims to create more inclusive and engaging learning experiences. The results revealed that ease of use significantly influenced perceived usefulness, attitude toward use, and intention to use, demonstrating that such digital tools can enhance learning engagement and foster more effective social communication.

With the evolution of computer vision and deep learning approaches [22], multimodal models have also gained relevance due to their ability to extract video embeddings and

reason over temporal sequences, allowing new strategies for recognition and comparative analysis [23]. In this context, video embeddings can be combined with few-shot classifiers such as Matching Networks to assign a sample to its most probable class based on similarity metrics [24]. More recently, large-scale multimodal models have made it possible to obtain robust representations from frame sequences, facilitating the exploration of direct classification scenarios in sign language videos [25].

Our study presents two complementary approaches for the analysis of MSL sign videos, applied to a custom data set of 335 recordings produced with professional interpreters: (i) an interactive training system that extracts hand and face key points using MediaPipe and employs Dynamic Time Warping (DTW) to compare the user's performance against a reference; and (ii) an experimental direct video classification method that uses an advanced multimodal model to generate video embeddings and a Matching Network to assign each sample to its most probable class [23–25].

To construct the dataset, twelve MSL lessons were recorded with professional interpreters: one interpreter recorded the alphabet using the left hand, while another recorded the remaining eleven lessons using the right hand. In addition, a second practice dataset was created with a non-expert participant who studied the signs beforehand and recorded them individually. This complementary dataset allowed evaluation of system robustness under learner-like conditions and served as the baseline for DTW experiments.

The first approach, implemented in a lightweight and easily deployable application, focuses on learning through near-real-time quantitative feedback, enabling users to progressively improve their performance. The second approach was explored as a research-oriented alternative for automatic sign recognition without explicit keypoint preprocessing. Although not integrated into the application due to computational requirements, its results show potential for large-scale classification and video pre-labeling tasks.

The main contributions of this work are as follows: (1) an interpretable, low-cost training framework for learning Mexican Sign Language (MSL), which extracts hand and face keypoints from RGB videos and provides real-time quantitative feedback using Dynamic Time Warping (DTW); and (2) an experimental baseline that combines video embeddings with a Matching Network to explore scalability to larger vocabularies and semi-automatic dataset labeling.

The remainder of this article is organized as follows. Section 2 describes the materials and methods, including participant recruitment, digital resource development, and outcome measures. Section 3 presents the quantitative and qualitative results of the proposal. Section 4 discusses these findings in the context of the existing literature, highlighting their strengths, limitations, and implications for practice. Finally, Section 5 summarizes the main conclusions and suggests directions for future research.

2. Materials and Methods

This section describes the methodological framework used to design, implement, and evaluate the proposed system. We first present the construction of two parallel datasets (reference and practice) and the procedure for extracting 2D keypoints from video sequences. We then detail the normalization process, similarity metrics, and temporal alignment using Dynamic Time Warping (DTW), as well as robustness analysis under noise perturbations. Next, we introduce the graphical user interface that supports interactive practice, and the design of a user study conducted to assess usability and learning effectiveness.

2.1. Dataset Construction

We built two parallel datasets of videos under an identical protocol to ensure comparable processing (Figure 1). The reference dataset was recorded with professional interpreters

and contains 335 videos across 12 lessons. The practice dataset was recorded by a non-expert participant who studied the same lessons and reproduced all signs. All sessions were captured with a conventional RGB camera. The videos were manually trimmed frame by frame to preserve only the relevant segment, and the audio was removed.

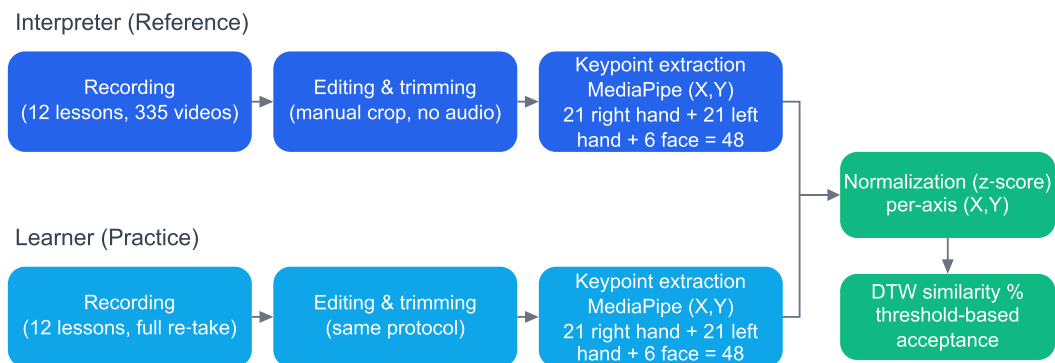


Figure 1. Processing pipeline for both video datasets. Interpreter (reference) and learner (practice) recordings follow the same steps: manual trimming (without audio), extraction of 48 keypoints with MediaPipe, axis-wise normalization (z-score), and DTW alignment to produce a similarity score.

Figure 2 shows four representative signs in the Mexican Sign Language (MSL). Each row corresponds to one sign, and the four images display consecutive frames (left–right) from start to finish.

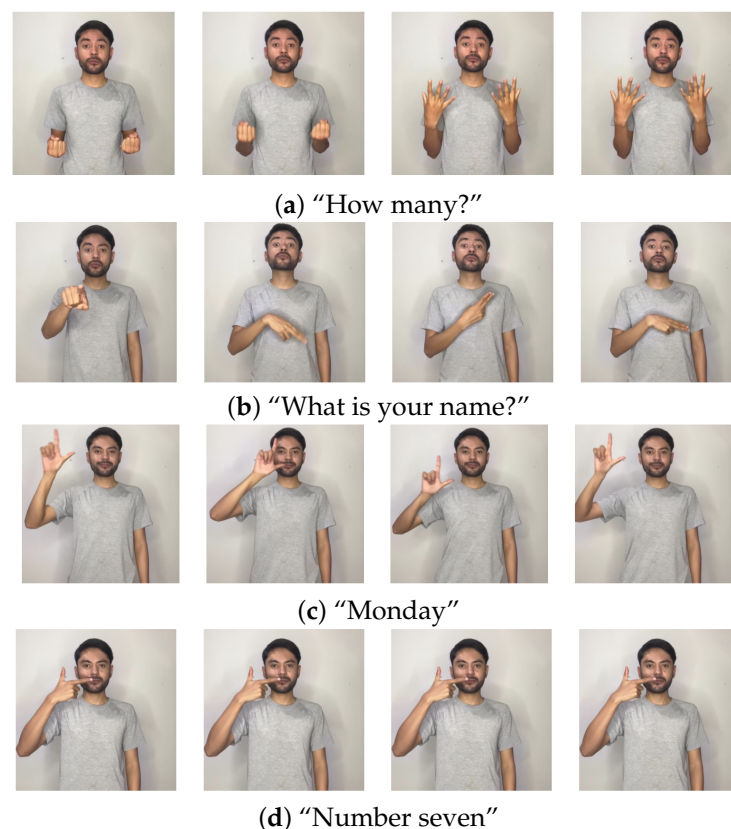


Figure 2. Illustrative examples of four Mexican Sign Language (MSL) signs. Each row shows one complete sign with four consecutive frames (left → right). Images were captured under controlled conditions and are published with informed consent for academic purposes.

2.2. Keypoint Extraction and Data Representation

We used MediaPipe [26] to identify keypoints on the hand and face for each video. The Hands module captures 21 keypoints per hand, while the Face Mesh module identifies

six facial landmarks, namely the right and left eyes, nose tip, mouth, and right and left ears. Consequently, every frame consists of 48 keypoints, each of which has 2D coordinates (x, y) .

Each frame t was vectorized as $\mathbf{p}_t \in \mathbb{R}^{96}$:

$$\mathbf{p}_t = [x_1, y_1, \dots, x_{48}, y_{48}]^\top. \quad (1)$$

Data were saved in CSV format, each file corresponding to a different sign and subject, comprising 96 columns; each row represented a frame. Figure 3 shows the keypoint distribution.

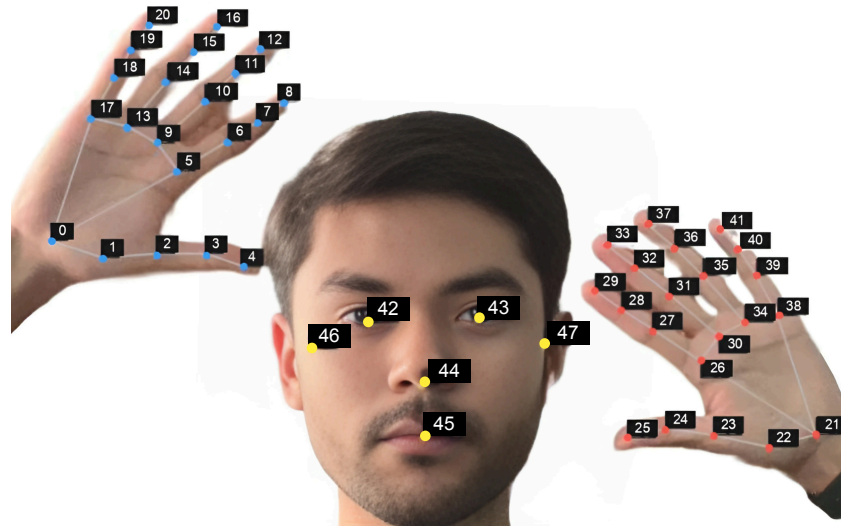


Figure 3. Distribution of the 48 keypoints: 21 on the right hand, 21 on the left hand, and 6 on the face.

2.3. Normalization and Frame-Wise Similarity Metric

To make interpreter and user trajectories comparable under framing or distance variations, axis-wise normalization (z-score) was applied to each sequence:

$$x' = \frac{x - \mu_x}{\sigma_x}, \quad y' = \frac{y - \mu_y}{\sigma_y}, \quad (2)$$

where μ_x, μ_y and σ_x, σ_y are the mean and standard deviation per axis, computed over the sequence. This centers the point cloud at $(0, 0)$ and standardizes variances, reducing the effects of translation and scale (Figure 4).

Instantaneous similarity between two normalized frames i and j was computed using the Euclidean distance:

$$d(i, j) = \|\mathbf{p}'_i - \mathbf{q}'_j\|_2 = \sqrt{\sum_{k=1}^{96} (p'_{i,k} - q'_{j,k})^2}. \quad (3)$$

Figure 5 illustrates the basic calculation between two homologous points.

2.4. Temporal Alignment with Dynamic Time Warping (DTW)

To compare entire sequences with potential differences in execution speed or minor desynchronizations, Dynamic Time Warping (DTW) was employed. Given the frame-wise distance $d(i, j)$, the accumulated cost is defined recursively as:

$$\text{DTW}[i, j] = d(i, j) + \min\{\text{DTW}[i - 1, j], \text{DTW}[i, j - 1], \text{DTW}[i - 1, j - 1]\}. \quad (4)$$

The final DTW cost acts as a global dissimilarity measure (lower is better). For user feedback, the system also reports a similarity percentage derived from this score (Figure 6).

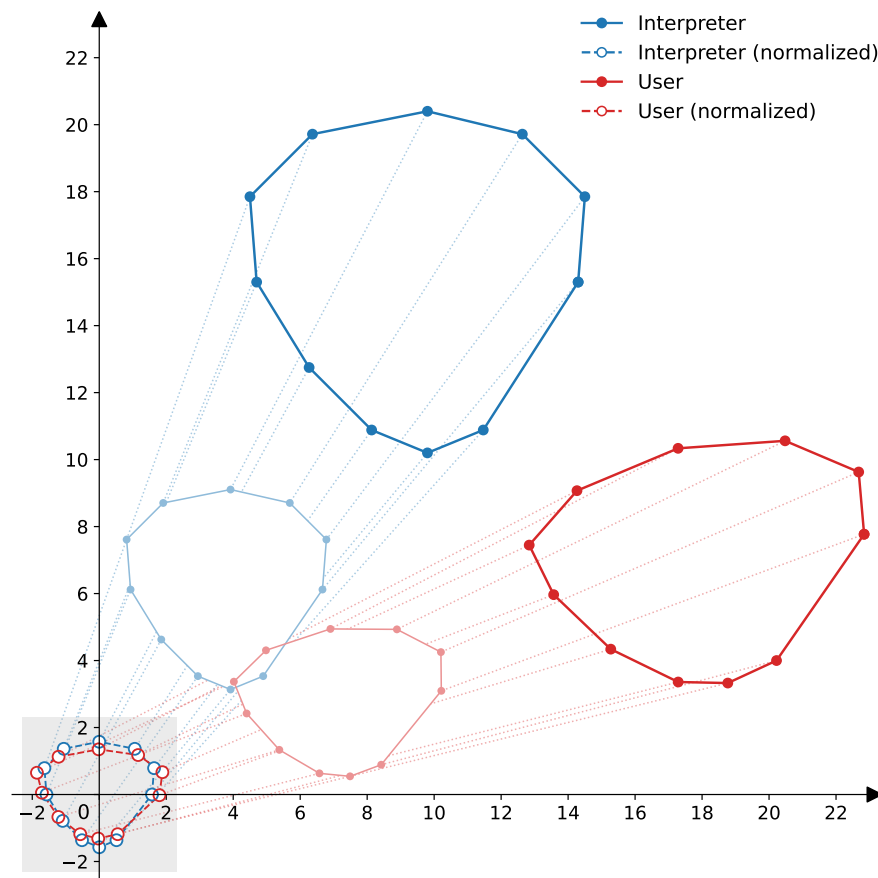


Figure 4. Normalization scheme: original trajectories (blue and red) are transformed into the normalized space (gray box), centered and rescaled for robust comparison.

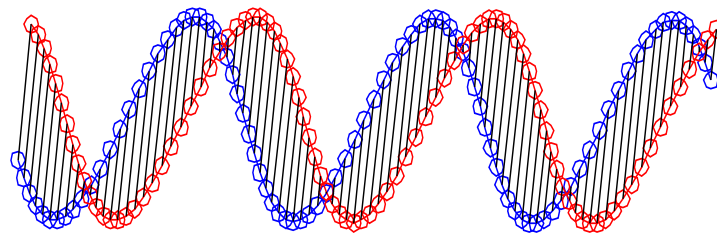


Figure 5. Illustration of the Euclidean distance computation between two time series. The red and blue curves represent two sequences that are compared point-to-point using Euclidean distance. Each black segment links homologous samples at the same time index, showing that Euclidean distance enforces a strict one-to-one temporal alignment even when the sequences differ in phase or speed.

To define a practical acceptance criterion for the execution of the correct sign, we analyzed the distribution of the DTW distances by comparing all lessons ($N = 335$). Because the alphabet lesson was recorded with opposite handedness between interpreter and user, its 27 samples presented systematically higher DTW values (mean = 1419.5, SD = 169.4). Based on this empirical distribution, we adopted a global acceptance threshold of $DTW \leq 600$, which lies between the median and the 75th percentile and is close to the overall mean of the full dataset (≈ 583). This cutoff provides a balanced trade-off between tolerance to natural intra-user variability and rejection of mismatched executions.

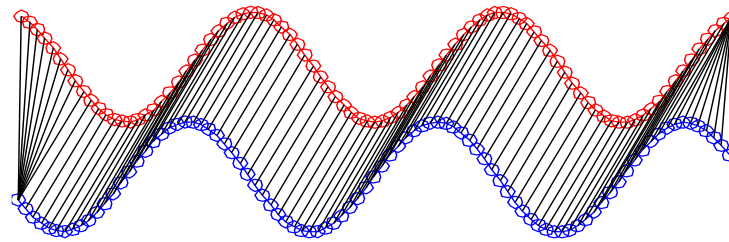


Figure 6. Example of dynamic time warping (DTW) sequence alignment. The red and blue curves denote the two time series being aligned. Unlike Euclidean distance, DTW allows nonlinear temporal alignment between the sequences. The black connecting paths illustrate how DTW warps the time axis to align similar patterns across different time positions, producing an adaptive alignment rather than a fixed point-to-point comparison.

A short sensitivity analysis showed that applying thresholds of 500, 600, and 700 yielded acceptance rates of 55.8%, 69.5%, and 80.5%, respectively, confirming that $DTW \leq 600$ is a robust and interpretable criterion. The alphabet samples were excluded from threshold determination due to the handedness mismatch; alternatively, they could be handled by applying a mirroring step before DTW to remove the systematic bias.

2.5. Noise Robustness

To assess the sensitivity of the system to capture variability and detection errors, perturbed versions of the sequences were generated by adding Gaussian noise with $\sigma \in \{0.5, 1, 2, 3\}$ to the (x, y) coordinates. This allowed evaluation of how DTW distances change under random displacements, simulating variability due to camera or detection artifacts. Results are reported in Section 3.

2.6. Robustness to Lighting and Resolution Changes

To evaluate the robustness of the DTW alignment under realistic capture conditions, we generated user video variants with controlled degradations in lighting and spatial resolution. We defined: Light50 and Light25: moderate and severe brightness reductions, respectively; Res50 and Res25: moderate and severe spatial downscaling; Light50_Res50 and Light25_Res25: moderate and severe combined degradations. The reference condition corresponds to the user pair without any degradation. The extraction of keypoints and the DTW computation were identical to those in the baseline condition.

As an illustrative example of the degradations considered in this subsection, Figure 7 shows the base user video (sign “Number 4”) along with the six lighting and resolution variants, arranged in a single row of seven consecutive panels.



Figure 7. Illustrative example of the lighting and resolution conditions used for robustness analysis. From left to right: (a) original, (b) Light50 (approx. -1 EV), (c) Light25 (approx. -2 EV), (d) Res50 (spatial scale at 50%), (e) Res25 (25%), (f) Light50_Res50 (moderate combined), and (g) Light25_Res25 (severe combined).

2.7. Partial Occlusion Protocol

We simulate occlusions in key regions relevant to Mexican Sign Language (MSL): left hand (L), right hand (R), both hands (B), face (Face) and full frame (All). Three severity levels (10%, 30% and 60%) were tested, as well as two masking patterns (suffixes c/r for

constant/random masks). For each condition, we report the mean and standard deviation of DTW values, as well as the percentage change relative to the reference, as detailed in Section 3.2.

2.8. Graphical User Interface

We built a graphical user interface to evaluate the method with participants. The application is organized into two levels: (i) a main menu with the available lessons and (ii) a practice window where the user views the reference video and performs practice attempts.

In the main menu, lessons are displayed as buttons arranged in rows and columns for an ordered selection (Figure 8).

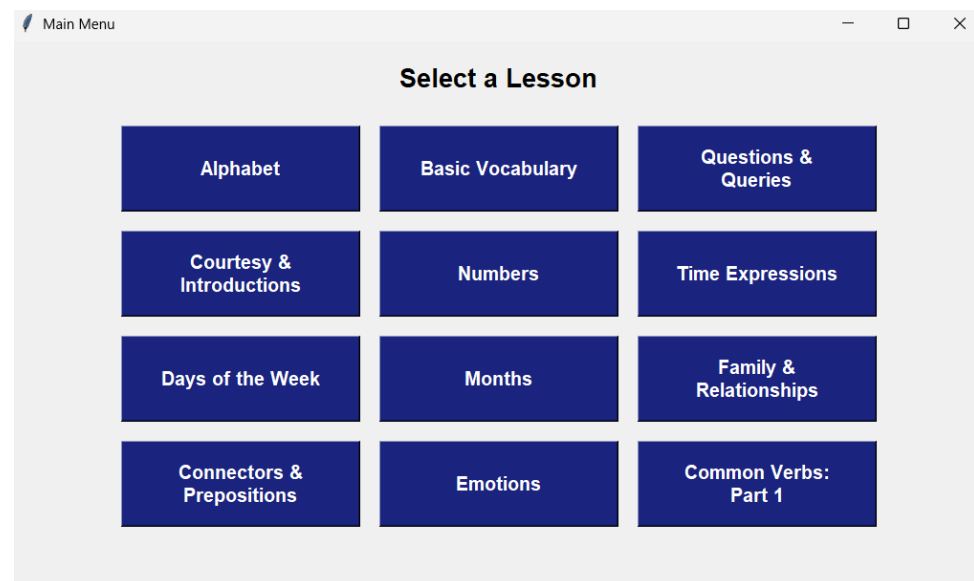


Figure 8. Main menu with available lessons. Each button opens the carousel of videos for the corresponding lesson.

Upon selecting a lesson, the practice window opens as shown in Figure 9. In the upper-left corner, a home button allows returning to the main menu. The central area displays the reference video, above it, two labels appear: the lesson title and the sign name. The bottom strip contains five control buttons: Previous, Play, Pause, Next, and Practice. The Practice button initiates camera capture and evaluation.

As shown in Figure 10a, when the practice button is pressed, a preparation message instructs: “Stand one meter away from the camera so your open palms are visible.” The camera then activates, and the system captures the hands and face keypoints per frame using MediaPipe. The user sequence is normalized and compared with the reference CSV via DTW, providing immediate feedback with similarity percentage as shown in Figure 10b,c.

2.9. User Study Design

A user study was conducted with 33 participants to assess perceived usefulness and user experience. A 5-point Likert questionnaire (1 = strongly disagree, 5 = strongly agree) measured four constructs: ease of use, perceived usefulness, satisfaction, and learning efficiency. The questionnaire’s internal consistency was Cronbach’s $\alpha = 0.81$.



Figure 9. Practice interface. Top: Home button and lesson/sign labels. Center: reference video. Bottom: controls (Previous, Play, Pause, Next, Practice).

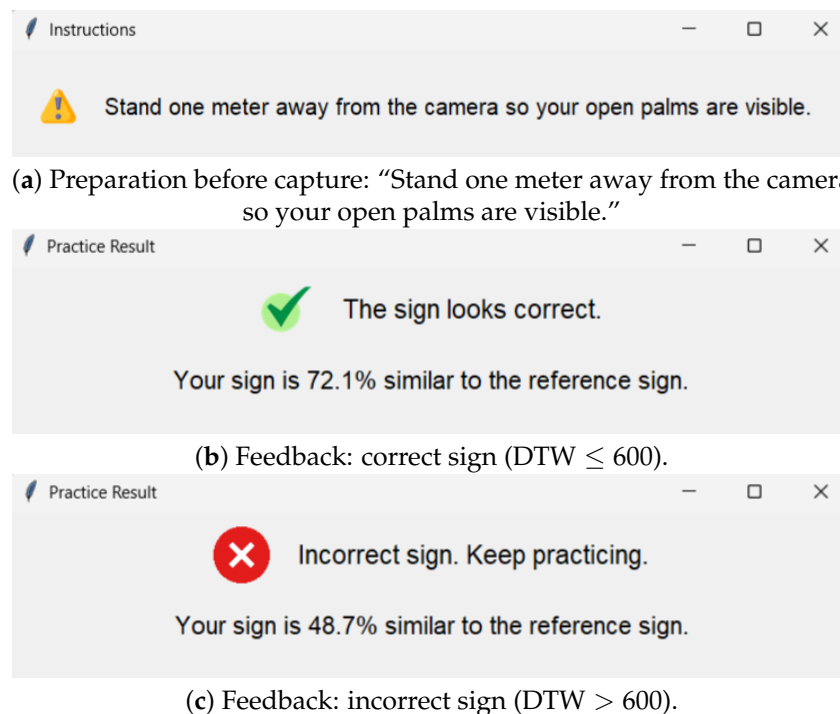


Figure 10. Key practice flow messages shown sequentially: (a) preparation and camera positioning; (b) confirmation of correct execution; (c) notification of incorrect execution.

Each participant: (1) opened the application and selected a lesson; (2) reviewed the reference video and control layout; (3) pressed Practice to capture their attempt after the preparation message; (4) received immediate feedback-based similarity percentage; (5) repeated attempts as needed; (6) completed the questionnaire. During the process, the researcher resolved questions in person or via video call.

3. Results

The experiments with dynamic time warping (DTW) enabled a systematic comparison of user executions against reference performances by professional interpreters. A lower DTW distance indicates greater similarity, whereas larger values reflect discrepancies in shape or timing. This approach made it possible to identify and quantify the differences between movements, adjust for temporal variations, and generate objective metrics for immediate feedback during practice.

In the alphabet lesson, the DTW distances were consistently greater than 1000 due to the difference in the dominant hand (interpreter using the left hand vs. user using the right hand). In contrast, in lessons involving basic vocabulary and short phrases, most distances ranged between 300 and 600, consistent with correct executions.

Table 1 summarizes the mean and standard deviation of DTW distances per lesson.

Table 1. Summary of DTW distances by lesson between the interpreter and users.

Lesson	Mean DTW	Std. Dev.
Alphabet	1419.5	169.4
Basic vocabulary	518.3	255.6
Questions	517.4	148.1
Courtesy and greetings	610.4	278.0
Numbers	390.4	200.1
Adverbs of time	510.6	188.8
Days of the week	271.5	34.3
Months	244.4	50.6
Family and relationships	517.6	217.6
Connectors/prepositions	556.7	216.0
Emotions	483.9	169.7
Common verbs (part 1)	635.3	238.1

3.1. Performance Evaluation of the DTW Method

To quantify the computational requirements of the proposed method based on Dynamic Time Warping (DTW), a single sign from Lesson 10 (Sign_0011) was analyzed. The experiment was conducted on a Windows 11 computer equipped with an Intel Core i7 CPU and 32 GB of RAM. The process included three stages: (1) keypoint extraction from both interpreter and user videos using MediaPipe, (2) data normalization, and (3) DTW distance computation between both temporal trajectories. Execution time was measured in milliseconds (ms), and memory usage in megabytes (MB). RSS_delta represents the net memory change, peak_py indicates the peak memory used by Python 3.10 objects, and peak_rss is the overall process peak (including native video buffers and TensorFlow Lite delegates).

Table 2 summarizes the execution time and memory usage for each stage of the DTW process (keypoint extraction, normalization, and alignment).

The extracted matrices contained 96 normalized coordinates per frame, with dimensions of $A = 198 \times 96$ (0.073 MB) for the interpreter and $B = 147 \times 96$ (0.054 MB) for the user, for a total memory footprint of approximately 0.126 MB. The DTW distance obtained between both sequences was 365.08, indicating strong similarity in temporal alignment and spatial configuration.

Table 2. Performance metrics of each stage for a single sign (Sign_0011).

Stage	Description	Time (ms)	ΔRSS (MB)	Peak Py (MB)	Peak Total (MB)
Keypoint extraction (interpreter)	96 features per frame using MediaPipe	8551.73	+60.04	44.30	291.25
Keypoint extraction (user)	Same process applied to user video	4417.78	+0.70	10.67	232.88
DTW computation (exact)	Temporal–spatial trajectory comparison	285.76	+0.04	0.01	91.94

The results demonstrate that the computational bottleneck lies in the keypoint extraction stage due to video decoding and neural landmark detection, whereas DTW computation itself is lightweight in both time and memory consumption. When starting from precomputed keypoints (CSV files), the inference time drops below 0.3 s, making this approach feasible for real-time feedback on standard CPUs without dedicated GPUs.

Figure 11 presents a representative heatmap of DTW distances between interpreter and user signs. Darker values represent higher similarity, while lighter values indicate discrepancies. As expected, in basic vocabulary lessons most comparisons fell in the 300–600 range, whereas alphabet signs consistently exceeded 1000 due to handedness differences.

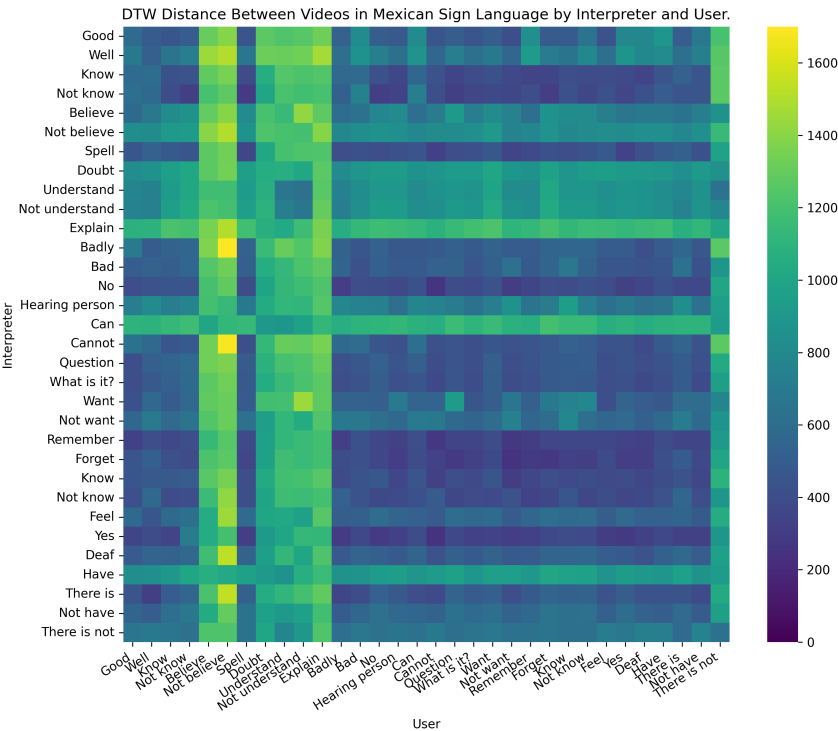


Figure 11. Heatmap of DTW distances between interpreter and user. Darker values indicate higher similarity.

3.2. Robustness Results with DTW

Robustness experiments with Dynamic Time Warping (DTW) quantify how visual conditions affect the similarity between user executions and interpreter references. Lower DTW means higher similarity. We analyze (i) lighting and resolution changes and (ii) partial occlusions.

3.2.1. Effect of Lighting and Resolution

We generate video variants with reduced brightness and spatial resolution. Light50/Light25 decrease brightness by 50% and 25% (about −1 EV and −2 EV). Res50/Res25 down-scale resolution to 50% and 25%. Light50_Res50 and Light25_Res25 apply both factors simultaneously. The baseline condition is the unaltered user–interpreter pair (see Table 3).

Table 3. Lighting/resolution conditions (concise description).

Condition	Type	Brief Description
Baseline	None	Reference without alterations.
Light50	Lighting	Brightness at 50% (−1 EV).
Light25	Lighting	Brightness at 25% (−2 EV).
Res50	Resolution	Spatial resolution at 50%.
Res25	Resolution	Spatial resolution at 25%.
Light50_Res50	Combined	Light 50% + Resolution 50%.
Light25_Res25	Combined	Light 25% + Resolution 25%.

To visualize the overall impact of lighting and resolution, Figure 12 summarizes the average DTW distance by condition (error bars = ± 1 SD).

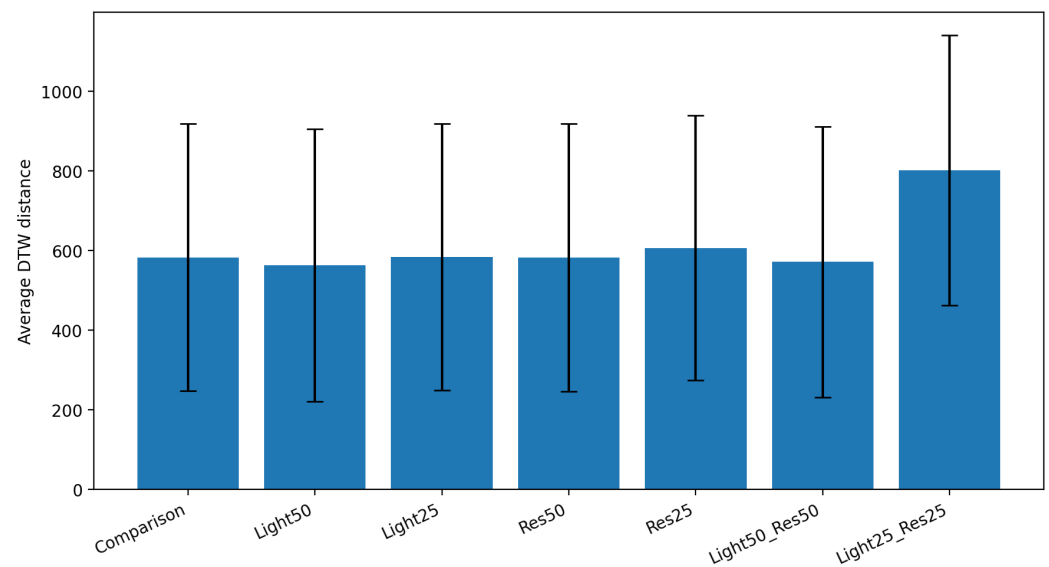


Figure 12. Average DTW distance by condition (± 1 SD). Bars show the mean DTW between interpreter and user under each degradation. Error bars denote ± 1 standard deviation. Lower values indicate higher similarity.

Table 4 reports the mean DTW, standard deviation and percentage change relative to baseline. As shown in the table, moderate degradations (Light50, Res50, Light50_Res50) alter DTW by less than 5%, whereas the severe combined condition (Light25_Res25) yields the largest increase (about +38%), indicating sensitivity to simultaneous low light and low resolution.

Table 4. DTW summary under lighting and resolution degradations.

Condition	Mean DTW	Std. Dev.	% Change vs. Baseline
Baseline	583.0	335.7	0.0%
Light50	563.3	341.5	−3.4%
Light25	583.5	334.9	+0.1%
Res50	582.6	336.1	−0.1%
Res25	606.9	332.8	+4.1%
Light50_Res50	571.5	339.5	−2.0%
Light25_Res25	802.1	339.5	+37.6%

3.2.2. Effect of Partial Occlusions

We occluded key regions at three levels (10%, 30%, 60%) with constant (c) or random (r) masks. Regions: left hand (L), right hand (R), both hands (B), face (Face), and whole frame (All). The notation is occ[Region][Level][Pattern] (see Table 5).

Table 5. Occlusion conditions (concise description).

Type	Region	Brief Description
occL	Left hand	Mask only on left hand.
occR	Right hand	Mask only on right hand.
occB	Both hands	Mask on both hands.
occFace	Face	Mask on facial area.
occAll	Full frame	Mask over the entire frame.
Levels	10%, 30%, 60%	Occluded proportion.
Patterns	c, r	Constant or random mask.

Table 6 summarizes the mean DTW, standard deviation, and relative change for each occlusion region, level, and mask pattern.

Table 6. DTW summary under partial occlusions (by region and severity).

Condition	Mean DTW	Std. Dev.	% Change vs. Baseline
Comp_occL10c	623.4	314.9	0.0%
Comp_occL10r	649.6	329.1	+4.2%
Comp_occL30c	736.7	271.0	+18.2%
Comp_occL30r	752.4	269.6	+20.7%
Comp_occL60c	877.6	268.1	+40.7%
Comp_occL60r	911.8	271.2	+46.2%
Comp_occR10c	635.8	313.6	+2.0%
Comp_occR10r	669.1	320.2	+7.3%
Comp_occR30c	773.5	278.3	+24.1%
Comp_occR30r	789.2	280.4	+26.5%
Comp_occR60c	902.3	284.1	+44.7%
Comp_occR60r	945.8	281.5	+51.7%
Comp_occB10c	713.5	310.9	+14.5%
Comp_occB10r	736.2	317.8	+18.1%
Comp_occB30c	884.2	274.5	+41.8%
Comp_occB30r	901.7	277.0	+44.6%
Comp_occB60c	1067.8	288.6	+71.2%
Comp_occB60r	1095.4	292.9	+75.7%
Comp_occFace10c	652.1	308.2	+4.6%
Comp_occFace10r	689.9	315.3	+10.7%
Comp_occFace30c	790.5	278.4	+26.8%
Comp_occFace30r	812.2	283.1	+29.9%
Comp_occFace60c	961.3	289.5	+54.2%
Comp_occFace60r	992.0	293.6	+59.1%
Comp_occAll10c	756.2	303.7	+21.3%
Comp_occAll10r	783.4	307.9	+25.7%
Comp_occAll30c	1015.5	295.1	+63.0%
Comp_occAll30r	1042.9	298.2	+67.7%
Comp_occAll60c	1325.8	322.4	+112.6%
Comp_occAll60r	1375.0	324.9	+120.6%

DTW increases with occlusion severity. The largest effects occur when both hands or the full frame are masked, while single-hand and 10% masks have smaller impact. This confirms the critical role of hand visibility for temporal alignment.

3.3. Experimental Comparison with Video Embeddings

Beyond DTW, we explored an experimental approach based on direct video classification using advanced multimodal models, aimed at evaluating scalability to larger datasets. We selected three models from the state of the art (SOTA) to test this approach: the Qwen2.5-VL-7B model pre-trained on a dataset constructed through a combination of methods, including cleaning raw web data, synthesizing data, etc. [25]; the VideoMAEv2 model (vit_small_patch16_224) pre-trained on the Kinetics-710 dataset, which presents a dual masking strategy for self-supervised pretraining [27]; and the VJEPa2 model (vjepa2-vitg-fpc64-384), pretrained on internet-scale videos and followed by post-training with a small amount of interaction data [28]. Each selected model is capable of generating video embeddings from frame sequences decomposing each frame into multiple patches represented as N-dimensional vectors, see Figure 13. Each model employs a distinct vision encoder configuration, resulting in feature vectors with different dimensions, as shown Table 7.

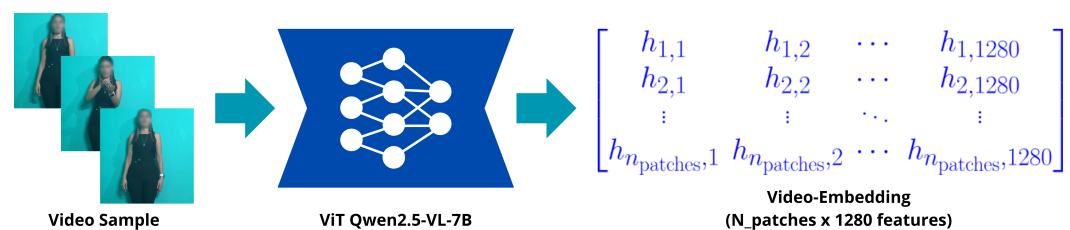


Figure 13. Illustration of the video embedding generation pipeline. Input sign videos are processed by a ViT-based encoder, which decomposes frames into patches and maps them to a matrix of $N_{patches} \times d$ feature vectors that serve as video embeddings for downstream classification.

Table 7. Dimensionality of the video embeddings produced by each foundational model, expressed as the number of features d in the output embedding vector.

Model	Embeddings Generated
Qwen2.5VL	1280
VideoMAEv2	384
VJEPa2	1408

For similarity-based classification, we implemented a Matching Network (MN) [24], comparing query embeddings against a set of supports using cosine distance and soft voting as shown in Figure 14. The support vectors are those that compose the reference videos. Each of these patch embeddings determines the class to which it belongs. The query vectors are those that need to be classified. The process for classifying patch video embeddings involves comparing the cosine distance between each patch of the query vector and each support vector, with the one yielding the smallest distance being set as the prediction for that patch. Due to every video sample having N video-embedding patches, the process is repeated in all the patches that belong to the video, at the same time each prediction is registered to obtain the class voted the most, implementing a soft voting strategic, this class is the one that is set as the preliminar class predicted.

To gain a broader understanding of the advantages and disadvantages of implementing these foundational models as feature extractors, it is necessary to compare their computational resource requirements and processing latency. We compare the amount of memory used by the complete system (feature extractor + Matching Network) and the processing latency of a single sample, see Table 8.

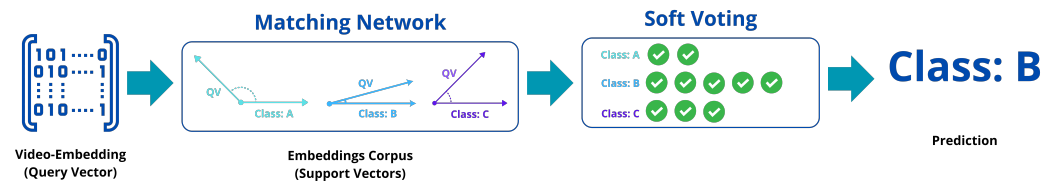


Figure 14. Matching Network-based similarity classification using video embeddings. The query video-embedding vector is compared with a corpus of support embeddings via cosine similarity, producing per-patch class predictions that are aggregated through soft voting to obtain the final predicted class.

Table 8. System requirements. All the experiments were run on a machine with an Intel i7-12700K CPU, an NVIDIA RTX 4080 SUPER (16 GB) GPU, and 32 GB of RAM.

Model	Memory Usage Model	Memory Usage MN	Total Memory	Processing Time (s)
Qwen2.5VL + MN	1205.40 MB	1266.71 MB	2472.11 MB	19.34 s
VideoMAEv2 + MN	84.51 MB	3.03 MB	87.53 MB	3.07 s
VJEPa2 + MN	3946.52 MB	98.114 MB	4044.63 MB	5.33 s

3.4. Embedding Baseline Results

The embedding baseline achieved different results across the foundational models used as feature extractors. In the case of Qwen2.5VL + MN, we achieved a clear class separability between multiple signs with a perfect accuracy score. Figures 15–17 show the classification results using video embedding similarity. This strong performance can be attributed to the Qwen2.5VL’s architecture, which emphasizes video understanding as a core capability. In contrast, VideoMAEv2 and VJEPa2 showed the opposite trend, with poor performance and accuracy (see Table 9). These results can be attributed to the fact that VideoMAEv2 and VJEPa2 are models pre-trained on masked prediction and reconstruction tasks and are commonly finetuned for downstream tasks, making them unsuitable for zero-shot scenarios.

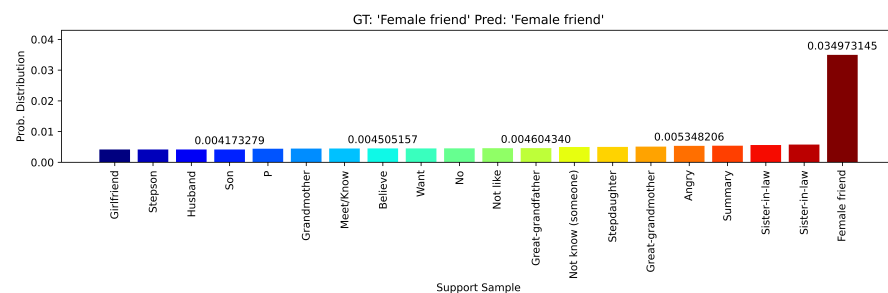


Figure 15. Softmax probability distribution over all support signs for the query sign “Female Friend” using the Qwen2.5VL + MN embedding baseline.

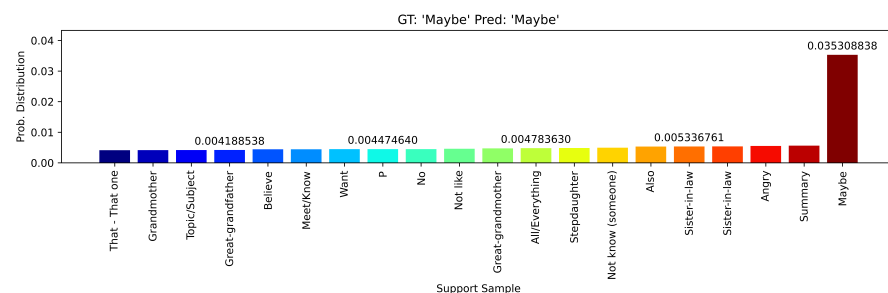


Figure 16. Softmax probability distribution over all support signs for the query sign “Maybe” using the Qwen2.5VL + MN embedding baseline.

Table 9. Zero-shot classification accuracy of the embedding baseline for each foundational model used as feature extractor. Qwen2.5VL + MN achieves perfect accuracy, whereas VideoMAEv2 + MN and VJEPa2 + MN exhibit near-chance performance.

Model	Accuracy
Qwen2.5VL + MN	1.0
VideoMAEv2 + MN	0.003
VJEPa2 + MN	0.003

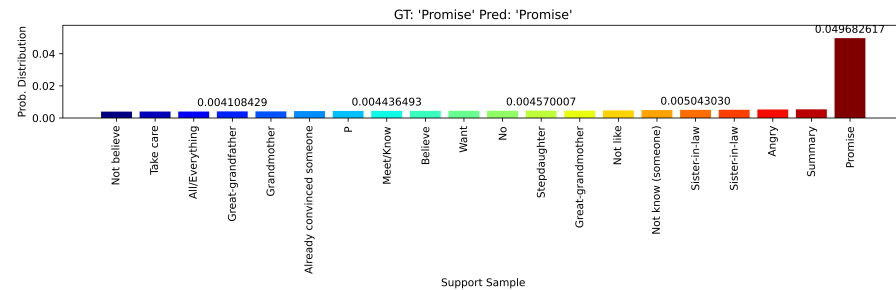


Figure 17. Softmax probability distribution over all support signs for the query sign “Promise” using the Qwen2.5VL + MN embedding baseline.

However, due to the higher computational cost and lower interpretability, this approach was not integrated into the final practice module.

3.5. User Evaluation Results

To assess participant perceptions of the application, a five-point Likert questionnaire (1–5) was administered to 33 participants. The elements were grouped into four categories: ease of use, perceived usefulness, satisfaction, and learning efficiency. An open question was also included for qualitative suggestions.

Figure 18 summarizes the distribution of responses per category (subfigures a–d). The responses were clearly concentrated in the higher range (4–5), suggesting favorable perceptions of the system in all dimensions.

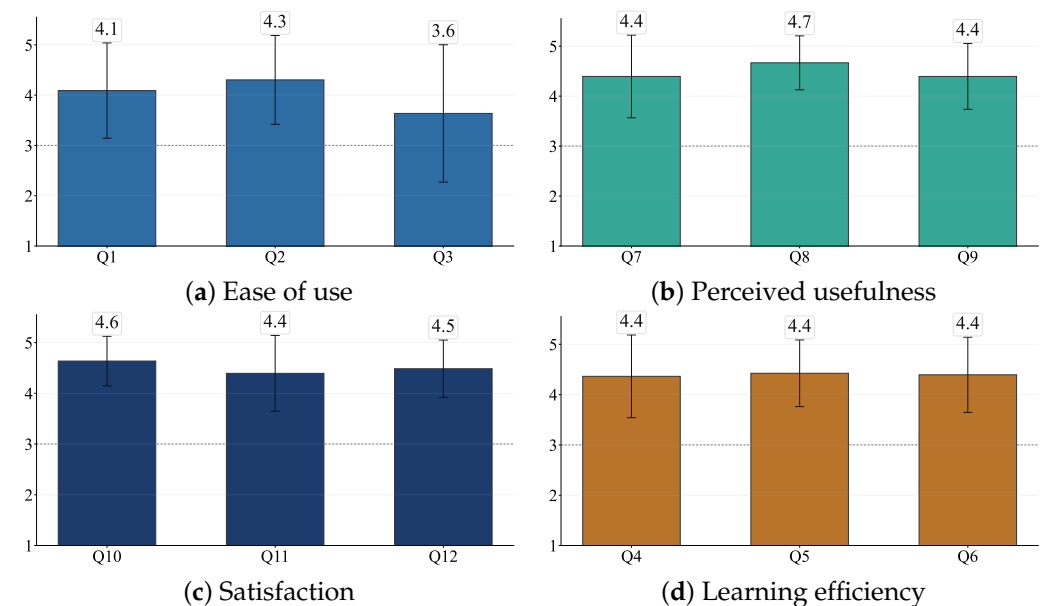


Figure 18. Distribution of responses by questionnaire category (N = 33). The dotted line indicates the neutral reference value (3).

Table 10 integrates, for each item, the mean, standard deviation, p -value of a one-sample t -test vs. 3, Cohen's d effect size, and 95% confidence intervals. In all categories, the mean was significantly higher than 3 ($p < 0.05$).

Table 10. Integrated results per item (N = 33): mean, standard deviation, p -value (one-sample t -test vs. 3), Cohen's d , and 95% confidence intervals.

Category/Item	Mean	Std. Dev.	p -Value	d	95% CI Low	95% CI High
<i>Ease of use</i>						
Menu usability	4.09	0.95	1.86×10^{-7}	1.15	3.75	4.43
Clear instructions	4.30	0.88	1.10×10^{-9}	1.48	3.99	4.61
Functional buttons	3.64	1.37	0.0116	0.47	3.15	4.13
<i>Perceived usefulness</i>						
Clarity of videos	4.36	0.82	7.33×10^{-11}	1.66	4.07	4.65
DTW comparison useful	4.42	0.66	1.04×10^{-13}	2.15	4.19	4.65
App design useful	4.39	0.75	4.10×10^{-12}	1.85	4.12	4.66
<i>Satisfaction</i>						
Overall satisfaction	4.39	0.83	4.94×10^{-11}	1.67	4.10	4.68
Recommendation to others	4.67	0.54	4.14×10^{-18}	3.09	4.48	4.86
Motivation to practice	4.39	0.66	1.55×10^{-13}	2.11	4.16	4.62
<i>Learning efficiency</i>						
Lesson organization	4.64	0.49	3.74×10^{-19}	3.35	4.47	4.81
Feedback (DTW) useful	4.39	0.75	4.10×10^{-12}	1.85	4.12	4.66
Learned new signs	4.48	0.57	4.31×10^{-16}	2.60	4.28	4.68

We applied the Holm–Bonferroni correction for multiple comparisons; significance remained after adjustment.

Qualitative Results

The open question (Q13) highlighted recurring improvement opportunities: (i) visual optimization of the interface, (ii) improved robustness under poor lighting or low-quality cameras, and (iii) more detailed feedback to guide practice. However, several participants expressed their full satisfaction with the application.

Taken together, the quantitative evidence: means significantly greater than 3, medium to large effect sizes and confidence intervals above neutral, and the psychometric evidence (reliability, correlations and PCA) indicate that the application is perceived as useful, clear, and motivating to learn the Mexican Sign Language.

4. Discussion

The Dynamic Time Warping (DTW)-based approach proved effective for guided practice of Mexican Sign Language (MSL), as it provides immediate and quantitative feedback using only a conventional RGB camera. The similarity threshold allowed for consistent differentiation between correct and incorrect executions, particularly in lessons involving basic vocabulary and short phrases. These findings indicate that motion–trajectory alignment can serve as a reliable metric for sign fidelity in low-cost configurations, making it suitable for educational applications without requiring specialized sensors or wearable devices.

These examples (Figure 2) visually illustrate the temporal evolution of the recorded signs, reinforcing the interpretability and reproducibility of the dataset for subsequent DTW analysis.

In addition to DTW, several alternative methods could be considered for aligning or comparing temporal motion sequences between the interpreter and the user. Examples

include frame-wise Euclidean or cosine distance, cross-correlation measures, and probabilistic or parametric models such as Dynamic Movement Primitives (DMPs) and the soft-DTW variant. We selected classical DTW for this work because it provides an explicit alignment path, is robust to differences in sequence length and execution speed, and produces interpretable distance matrices that can be directly visualized as feedback for learners. In contrast, methods such as soft-DTW or embedding-based similarity scores are more suitable for large-scale automatic recognition but less transparent for educational feedback. Therefore, DTW was favored for its balance between simplicity, computational efficiency, and pedagogical interpretability.

A complementary baseline using video embeddings and Matching Networks was also explored to evaluate scalability toward larger vocabularies. Although this approach achieved clear class separability and high accuracy, it demands greater computational resources and offers limited interpretability at the frame level. In contrast, DTW offers lightweight computation and transparent alignment visualization, which are advantageous for real-time feedback and educational use. Therefore, both strategies are considered complementary: DTW is ideal for interactive practice and formative evaluation, while the embedding-based model has potential for large-scale recognition and automatic dataset annotation.

The computational analysis (Table 2) demonstrated that keypoint extraction dominates resource consumption, while DTW computation remains lightweight. This supports the method's suitability for real-time educational feedback.

The main strength of the proposed system lies in its accessibility and ease of deployment. It can be used on standard laptops or tablets and provides an intuitive interface for learners. However, its performance depends on accurate keypoint detection; tracking errors or occlusions can increase dissimilarity scores, especially under poor lighting or low-resolution conditions. The visual examples in Figure 7 help interpret how lighting and resolution degradations visually impact keypoint detection and, consequently, DTW alignment stability.

Quantitatively, moderate lighting or resolution degradations altered DTW distances by less than 5%, indicating robustness for standard webcams. However, severe combined conditions (Light25_Res25) and occlusions affecting both hands increased DTW by over 35–70%, which could lead to false negatives in practical use. These results underscore the need for adaptive thresholds and more robust keypoint tracking under non-controlled environments. In addition, the data set was limited to 12 lessons and a small group of interpreters, which may constrain generalization to broader contexts of MSL use.

The results of the user study further support the educational value of the system. Participants reported high satisfaction, perceived usefulness, and learning motivation. The psychometric validation of the questionnaire (Cronbach's $\alpha = 0.81$ and all means significantly above the neutral point after Holm–Bonferroni correction) supports the reliability and internal consistency of the participants' responses, confirming the robustness of subjective evaluations. These outcomes are consistent with previous research showing that interactive and feedback-oriented learning environments enhance engagement and retention in sign language education. The DTW-based feedback mechanism not only improved self-assessment but also encouraged repeated practice, demonstrating its potential as a tool for inclusive and autonomous learning.

5. Conclusions

This study presented a virtual training system for learning Mexican Sign Language (MSL) using a computer vision approach based on Dynamic Time Warping (DTW) and, experimentally, video embeddings with Matching Networks. The DTW-based module allowed learners to practice signs with real-time feedback through a conventional RGB

camera, providing an interpretable and efficient similarity metric that supports immediate correction. The embedding-based baseline, in turn, achieved high classification accuracy, demonstrating potential for large-scale vocabulary expansion and automatic dataset labeling. Together, these complementary methods highlight the value of combining interpretability and scalability for inclusive sign language education.

Limitations and Future Work

The main limitations of this study include the relatively small dataset (12 lessons with few interpreters), which limits model generalization, and the DTW metric's dependence on accurate keypoint detection, making it sensitive to occlusions and lighting variations. The evaluation also focused mainly on short, isolated signs rather than continuous signing, and the participant group lacked broad diversity in age, hearing condition, or cultural background. Furthermore, while accurate, the embedding-based approach requires substantial computational resources, hindering real-time deployment on low-power devices.

Building directly upon the robustness findings, we will (i) adopt adaptive DTW thresholds conditioned on capture quality (lighting and resolution) and (ii) improve keypoint extraction under occlusions, especially when both hands are partially hidden. Concretely, we will calibrate per-lesson and per-condition thresholds using validation splits, integrate lightweight exposure and contrast normalization for low-light scenes, and add occlusion-aware hand tracking and user prompts to preserve alignment stability in non-controlled environments.

Future work will expand the dataset with additional interpreters and vocabulary, including participants from diverse demographic and linguistic backgrounds, to improve representativeness and generalization between populations. We also plan to evaluate the transferability of the model between different user groups, enhance robustness under various recording conditions, and explore adaptive similarity metrics that personalize feedback. Lightweight multimodal models can merge the interpretability of DTW with the scalability of embeddings, enabling hybrid systems suitable for both teaching and large-scale MSL analysis.

Future research could also explore end-to-end recurrent architectures such as LSTM- or GRU-based models to learn temporal dependencies directly from raw video sequences, provided that larger and more balanced datasets become available. This dual analysis highlights a complementary pathway: DTW provides interpretable, low-latency feedback for learners, while embedding-based classification enables scalability for automatic dataset expansion. Integrating both could yield hybrid systems capable of delivering personalized, explainable, and large-scale sign language instruction. Finally, robustness experiments under controlled degradations (lighting, resolution, and occlusions) showed that DTW similarity remains stable under moderate conditions, with significant degradation only when both hands or the full frame were occluded (Table 6). These findings confirm the reliability of the method for typical classroom and home environments.

Author Contributions: Conceptualization, D.-M.C.-E.; Methodology, F.d.J.R.-C., D.-M.C.-E., J.T. and J.-R.G.-R.; Software, F.d.J.R.-C.; Validation, F.d.J.R.-C., D.-M.C.-E., J.T. and J.-R.G.-R.; Formal analysis, F.d.J.R.-C., D.-M.C.-E., J.T., J.-A.R.-G. and M.-A.I.-C.; Investigation, F.d.J.R.-C., D.-M.C.-E., J.T., J.-R.G.-R., J.-A.R.-G. and M.-A.I.-C.; Resources, D.-M.C.-E., J.T., J.-R.G.-R., M.-A.I.-C. and P.-A.R.-P.; Data curation, F.d.J.R.-C.; Writing—original draft preparation, F.d.J.R.-C., D.-M.C.-E., J.T., J.-A.R.-G. and M.-A.I.-C.; Writing—review and editing, F.d.J.R.-C., D.-M.C.-E., J.T., J.-A.R.-G., M.-A.I.-C. and P.-A.R.-P.; Visualization, F.d.J.R.-C., D.-M.C.-E., J.T., J.-R.G.-R., J.-A.R.-G., M.-A.I.-C. and P.-A.R.-P.; Supervision, D.-M.C.-E. and J.T.; Project administration, D.-M.C.-E. and J.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: This study was conducted in accordance with the ethical guidelines of the Universidad Autónoma de Querétaro. Ethical approval was obtained from the Faculty Board (Facultad de Informática) of UAQ under protocol FI/CA-470/116, dated 12 March 2024. All participants provided informed consent prior to their participation.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

Acknowledgments: We thank the Autonomous University of Querétaro for its institutional support, and the Secretariat of Science, Humanities, Technology, and Innovation (SECIHTI) for providing a student scholarship that supported the master's studies. We also acknowledge the use of generative AI tools, specifically Grammarly Assistant (Version 1.0, Grammarly Inc., San Francisco, CA, USA) to improve grammar, clarity, and readability, and GPT-5 (OpenAI, San Francisco, CA, USA) to assist with wording and proofreading of the manuscript.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Arra, A.; Mangampo, H.; Presto, M.L.; Herradura, T. Senyas: A 3D Animated Filipino Sign Language Interpreter Using Speech Recognition. In Proceedings of the Philippine Computing Science Congress (PCSC 2022), Iligan, Philippines, 5–6 May 2022.
2. Adanigbo, O.; Oyewole, T. Development of a Sign Language E-Tutor using Convolutional Neural Network. *FUOYE J. Eng. Technol.* **2023**, *8*, 192–196. [\[CrossRef\]](#)
3. Alam, M.S.; Lamberton, J.; Wang, J.; Leannah, C.; Miller, S.; Palagano, J.; de Bastion, M.; Smith, H.L.; Malzkuhn, M.; Quandt, L.C. ASL champ!: A virtual reality game with deep-learning driven sign recognition. *Comput. Educ. X Real.* **2024**, *4*, 100059. [\[CrossRef\]](#)
4. Shaw, A.; Wünsche, B.C.; Mariono, K.; Ranveer, A.; Xiao, M.; Hajika, R.; Liu, Y. JengASL: A Gamified Approach to Sign Language Learning in VR. *J. WSCG* **2023**, *31*, 34–42. [\[CrossRef\]](#)
5. El Ghouli, O.; Othman, A. Virtual reality for educating Sign Language using signing avatar: The future of creative learning for deaf students. In Proceedings of the 2022 IEEE Global Engineering Education Conference (EDUCON), Tunis, Tunisia, 28–31 March 2022; pp. 1269–1274.
6. Escudeiro, P.; Escudeiro, N.; Norberto, M.; Lopes, J. Virtualsign translator as a base for a serious game. In Proceedings of the 3rd International Conference on Technological Ecosystems for Enhancing Multiculturality, Porto, Portugal, 7–9 October 2015; pp. 251–255.
7. Lima, T.; Rocha, M.S.; Santos, T.A.; Benetti, A.; Soares, E.; de Oliveira, H.S. Innovation in learning—the use of avatar for sign language. In *Human-Computer Interaction. Applications and Services, Proceedings of the 15th International Conference, HCI International 2013, Las Vegas, NV, USA, 21–26 July 2013*; Proceedings, Part II 15; Springer: Berlin/Heidelberg, Germany, 2013; pp. 428–433.
8. Feicán, C.; Cabrera, J.; Arévalo, J.; Ayala, E.; Guerrero, F.; Pinos, E. Sign language trainer using leap motion. In *Advances in Human Factors in Training, Education, and Learning Sciences, Proceedings of the AHFE 2017 International Conference on Human Factors in Training, Education, and Learning Sciences, Los Angeles, CA, USA, 17–21 July 2017*; Springer: Cham, Switzerland, 2018; pp. 253–264.
9. Schioppo, J.; Meyer, Z.; Fabiano, D.; Canavan, S. Sign language recognition: Learning american sign language in a virtual environment. In Proceedings of the Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems, Glasgow, UK, 4–9 May 2019; pp. 1–6.
10. Ghanem, S.; Albidewi, I. An Avatar Based Natural Arabic Sign Language Generation System for Deaf People. *Comput. Eng. Intell. Syst.* **2013**, *4*, 1–7.
11. Chakladar, D.D.; Kumar, P.; Mandal, S.; Roy, P.P.; Iwamura, M.; Kim, B.G. 3d avatar approach for continuous sign movement using speech/text. *Appl. Sci.* **2021**, *11*, 3439. [\[CrossRef\]](#)
12. Zhang, Y.; Min, Y.; Chen, X. Teaching Chinese Sign Language with a Smartphone. *Virtual Real. Intell. Hardw.* **2021**, *3*, 248–260. [\[CrossRef\]](#)
13. Lara Ortiz, D.; Fuentes-Aguilar, R.; Chairez, I. Assisted Sign Language Tutoring System Using a Bioinspired Virtual Controlled Robotic Arm. *SSRN Electron. J.* **2023**. [\[CrossRef\]](#)
14. Gutiérrez Navarrete, C.A. Prototipo de Aplicación para la Enseñanza del Lenguaje de señAs Colombianas de uso Diario, a Personas Oyentes que Conviven con Sordos. Master's Thesis, Universidad Nacional Abierta y a Distancia (UNAD), Bogotá, Colombia, 2025.

15. Daniel, C.A.; Liam, N.M.; Habchi, Y.; Basnet, S. Enhancing Language Learning with Real-Time Sign Language Recognition and Feedback. In Proceedings of the 2024 IEEE MIT Undergraduate Research Technology Conference (URTC), Cambridge, MA, USA, 11–13 October 2024; pp. 1–5. [\[CrossRef\]](#)
16. Islam, A.; Rahman, S.A.S.; Bhowmick, P. A Smart Avatar Tutor for Mimicking Characters of Bangla Sign Language. Ph.D. Thesis, Brac University, Dhaka, Bangladesh, 2024.
17. Lawana, S.P. Development of a Mobile Application for Interactive Sign Language Learning Using AI-Powered Gesture Recognition and Gamification. *Idea Future Res.* **2024**, *2*, 72–80.
18. Kasapakis, V.; Dzardanova, E.; Vosinakis, S.; Agelada, A. Sign language in immersive virtual reality: Design, development, and evaluation of a virtual reality learning environment prototype. *Interact. Learn. Environ.* **2024**, *32*, 6657–6671. [\[CrossRef\]](#)
19. Erofeeva, M.; Klowait, N.; Belov, M.; Soulié, Y. Leveraging VR Tools for Inclusive Education: Implications from Sign Language Learning in VRChat. In *International Conference on Immersive Learning*; Springer: Cham, Switzerland, 2025; pp. 147–164.
20. Coy, A.; Mohammed, P.S.; Skerit, P. Inclusive deaf education enabled by artificial intelligence: The path to a solution. *Int. J. Artif. Intell. Educ.* **2025**, *35*, 96–134. [\[CrossRef\]](#)
21. Hwang, K.F.; Wang, Y.H.; Tsao, T.Y.; Chou, S.H.; Chueh, S.Y.; Wu, C.C.; Ho, K.Y. Adopting Artificial intelligence and artificial reality in an interactive sign language learning system: Acceptance of interactive technology. *Eng. Proc.* **2025**, *89*, 14.
22. Ma, R.; Li, S.; Zhang, B.; Fang, L.; Li, Z. Flexible and generalized real photograph denoising exploiting dual meta attention. *IEEE Trans. Cybern.* **2022**, *53*, 6395–6407. [\[CrossRef\]](#) [\[PubMed\]](#)
23. Suzuki, M.; Matsuo, Y. A survey of multimodal deep generative models. *Adv. Robot.* **2022**, *36*, 261–278. [\[CrossRef\]](#)
24. Vinyals, O.; Blundell, C.; Lillicrap, T.; Kavukcuoglu, K.; Wierstra, D. Matching networks for one shot learning. In Proceedings of the 30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain, 5–10 December 2016; Volume 29.
25. Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. Qwen2. 5-vl technical report. *arXiv* **2025**, arXiv:2502.13923.
26. Lugaresi, C.; Tang, J.; Nash, H.; McClanahan, C.; Uboweja, E.; Hays, M.; Zhang, F.; Chang, C.L.; Yong, M.G.; Lee, J.; et al. Mediapipe: A framework for building perception pipelines. *arXiv* **2019**, arXiv:1906.08172. [\[CrossRef\]](#)
27. Wang, L.; Huang, B.; Zhao, Z.; Tong, Z.; He, Y.; Wang, Y.; Wang, Y.; Qiao, Y. Videomae v2: Scaling video masked autoencoders with dual masking. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 14549–14560.
28. Assran, M.; Bardes, A.; Fan, D.; Garrido, Q.; Howes, R.; Muckley, M.; Rizvi, A.; Roberts, C.; Sinha, K.; Zholus, A.; et al. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *arXiv* **2025**, arXiv:2506.09985.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.