

The SignEval 2025 Challenge at the ICCV Multimodal Sign Language Recognition Workshop: Results and Discussion

Hamzah Luqman^{1*}, Raffaele Mineo^{2*}, Murtadha Aljubran³, Ahmed Abul Hasanaath¹, Amelia Sorrenti², Sarah Alyami⁴, Sadam Al-Azani¹, Maad Alowaifeer¹, JiHwan Moon⁵, Vaclav Javorek⁶, Tomas Zelezny⁶, Marek Hruz⁶, Gaia Caligiore⁷, Silvio Giancola⁸, Senya Polikovskiy⁹, Motaz Alfarraj¹, Sabina Fontana², Mufti Mahmud¹, Muhammad Haris Khan³, Kamrul Islam¹⁰, Sevgi Gurbuz¹⁰, Egidio Ragonese², Giovanni Bellitto², Federica Proietto Salantri², Concetto Spampinato², Simone Palazzo²

¹ King Fahd University of Petroleum & Minerals, ² University of Catania, ³ Mohamed Bin Zayed University of Artificial Intelligence, ⁴ Imam Abdulrahman Bin Faisal University, ⁵ Yonsei University, ⁶ University of West Bohemia, ⁷ University of Modena and Reggio Emilia, ⁸ King Abdullah University of Science and Technology, ⁹ Max Planck Institute for Intelligent Systems, ¹⁰ North Carolina State University

Abstract

This paper summarizes the results of the first multimodal sign language recognition challenge, SignEval 2025, organized at ICCV 2025. The challenge featured two tracks: (i) Continuous sign language recognition (CSLR) task based on the newly curated Isharah dataset, a Saudi Sign Language dataset, and (ii) Isolated sign language recognition (ISLR) task using the MultiMeDaLIS dataset, a multimodal Italian Sign Language corpus tailored for doctor-patient communication. Two tasks are defined within the CSLR track: Signer-Independent and Unseen-Sentences. The Signer-Independent task tests the model's ability to generalize across signers, a critical property for scalable real-world CSLR systems. The Unseen-Sentences task evaluates the model's capability to recognize novel sentence compositions by leveraging learned grammar and semantics. The ISLR track utilized MultiMeDaLIS, a multi-modal dataset. The participants of this track were challenged to classify isolated signs using only radar and RGB modalities. The challenge utilized two leaderboards to showcase methods, with participants setting new benchmarks and achieving state-of-the-art results on both tracks. More information on the challenges, tasks, leaderboard, baselines and development kits are available on <https://multimodal-sign-language-recognition.github.io/ICCV-2025/>.

1. Introduction

Sign language recognition (SLR) is a critical field in computer vision and human-computer interaction, aimed at enabling seamless communication between deaf or hard-of-hearing individuals and the hearing community [12]. Unlike spoken languages, sign languages are expressed through complex combinations of manual features, such as hand shapes, orientations, and movements, and non-manual features, such as facial expressions and body posture [37].

SLR can be categorized into word-level or isolated SLR (ISLR) and sentence-level or continuous SLR (CSLR). ISLR focuses on identifying individual signs or isolated sign words [27]. It typically involves the classification of short temporal segments that correspond to specific signs. In contrast, CSLR aims to transcribe full sign language sentences into gloss sequences. This makes CSLR inherently more challenging than ISLR, as it requires modeling long-range temporal dependencies, dealing with coarticulation effects, and understanding compositional linguistic structures [2]. Although significant progress has been achieved at SLR, several challenges remain, such as variability across signers, modality constraints, and the need for scalable systems that generalize beyond training conditions [35].

Most of the proposed models for SLR have focused on RGB-based inputs [1, 19]. Pose representation, which encodes signer body keypoints, offers a compact yet informative abstraction of signing behavior [18, 21]. Representing signs using pose information helps mitigate signer overfitting, which is crucial for developing signer-independent SLR systems. Additionally, it supports signer's privacy preservation, enhances cross-domain robustness, and en-

*These authors contributed equally to this work.

Corresponding authors: H. Luqman (hluqman@kfupm.edu.sa), R. Mineo (raffaele.mineo@unict.it).

ables lightweight computation. Despite these advantages, the use of pose data in SLR remains relatively underexplored compared to other modalities.

Recognizing the need to advance SLR, the *SignEval 2025* challenge introduced two tracks for **CSLR** and **ISLR**. The first track focused on CSLR based on *Isharah* [3] dataset. Two shared tasks have been organized within this track: (i) Signer-Independent, which evaluates the generalization of the CSLR systems across unseen signers, and (ii) Unseen-Sentences, which evaluates the generalization of the CSLR systems to new sentences. Participants were provided with the pose data to develop their models, which prioritize structure, scalability, and signer invariance. The second track of the challenge targeted ISLR using the *MultiMeDaLIS* dataset [6], a multimodal dataset specifically designed for medical and alphabetic gestures in Italian Sign Language (LIS, acronym for *Lingua Italiana dei Segni*). MultiMeDaLIS provides 25,830 isolated sign instances across 126 classes, captured synchronously using three complementary privacy-preserving modalities: 60 GHz radar, RGB video, and skeletal data. This multimodal setup enables the study of ISLR beyond vision-only approaches, fostering research into sensor fusion, modality robustness, and medical communication applications. This paper summarizes the outcomes of this challenge, including both the CSLR and ISLR tracks. We report the datasets, tasks, participant submissions, and best-performing approaches across both tracks, which offer insights into current trends and remaining challenges in SLR.

2. Literature Review

2.1. Continuous Sign Language Recognition

Several approaches have been proposed for CSLR, including convolutional neural networks (CNNs) such as [8], attention-based models and Transformers such as [7, 46], temporal alignment techniques such as [9], weakly supervised learning frameworks such as [9], and multi-modal or cross-modality methods such as [24].

Cui et al. [9] presented a weakly supervised framework for continuous sign language recognition, where only ordered gloss labels are available without precise temporal alignment, and training data is limited. Their method employs a recurrent CNN for spatio-temporal feature extraction and sequence modeling. A three-stage optimization strategy was introduced: (1) end-to-end sequence learning using Connectionist Temporal Classification for initial alignment; (2) refinement of the feature extractor using the alignment as stronger supervision; and (3) re-optimization of the sequence model with enhanced features, alongside a weakly supervised detection network for regularization. The approach was evaluated on the RWTH-PHOENIX-Weather multi-signer 2014 dataset without extra supervi-

sion and achieved performance comparable to state-of-the-art methods.

Cheng et al. [8] presented a fully convolutional network to jointly learn spatial and temporal features from weakly annotated video sequences, using only sentence-level supervision. To improve sequence alignment, the authors incorporated a gloss feature enhancement module into the architecture. The method was evaluated on the CSL and RWTH datasets, demonstrating its effectiveness in online recognition settings.

Recently, Ke et al. [24] introduced fine-grained cross-modality consistency loss, aimed at aligning visual and linguistic representations. The FGXM loss promotes consistency between visual context and language understanding, enhancing the integration of multimodal features. The method was evaluated across multiple datasets and model architectures, showing notable gains in both accuracy and data efficiency. To support fairer evaluation, the authors also introduced the unweighted word error rate (uWER), a metric that addresses frequency imbalances between content words and function words in CSLR tasks.

STNet, a spatial-temporal feature-enhanced network, was introduced by [43] to improve sequence modeling in sign language recognition. The authors proposed a spatial resonance module utilizing the optimal transport algorithm to capture global common spatial features between adjacent frames. A frame-wise loss was then introduced to preserve frame-specific characteristics. Additionally, a multi-temporal perception module was designed to enable each frame to attend to a broader temporal context, enhancing multi-scale information interaction. This method was evaluated on three benchmark datasets - PHOENIX14, PHOENIX14-T, and CSL-Daily - with strong performance reported across all.

Zhu et al. [46] proposed a holistic CSLR framework, MAM-FSD, integrating a motor attention mechanism and frame-level self-distillation. The motor attention mechanism focuses on capturing subtle distortions in local motion regions during sign language articulation, enabling the model to extract dynamic visual representations. Complementing this, frame-level self-distillation was employed to enhance feature extraction, improving both inference accuracy and model robustness. The approach was evaluated on three public datasets, achieving state-of-the-art performance.

2.2. Isolated Sign Language Recognition

ISLR focuses on classifying individual signs from pre-segmented video sequences. Early research in this domain frequently employed sensor-based systems, such as instrumented gloves, which provide highly precise kinematic measurements [38]. Despite their accuracy, such devices are inherently intrusive, often hindering the natural execution of signs and failing to capture critical non-manual

features that are integral to sign language communication.

To address these limitations, vision-based approaches have been explored. RGB and depth imaging technologies enable the real-time acquisition of both manual and non-manual signing components [27]. Nonetheless, their deployment in sensitive contexts raises serious privacy concerns and reveals notable shortcomings in terms of robustness, particularly under variable lighting conditions or in dynamic scenes [25, 29].

Consequently, recent studies have explored the use of millimetre-wave radar as a non-invasive alternative for gesture capture. Radar systems, particularly those leveraging micro-Doppler signatures, can encode fine-grained motion dynamics while inherently preserving the visual anonymity of the signer [15, 17, 26]. This modality offers considerable robustness to environmental noise and illumination changes. However, existing radar-based ISLR research predominantly concentrates on limited-vocabulary gesture sets in American Sign Language (ASL), often derived from small-scale datasets [5, 11].

To capture the complexity of both manual and non-manual components, multimodal approaches have been proposed. De Coster et al. [10] employ a transformer-based model that integrates body-pose flow and cropped hand features, achieving competitive results on the AUTSL benchmark. Vahdani et al. [41] fuse RGB-D inputs, optical flow, and skeletal representations within a multi-branch 3D convolutional neural network, attaining 92% accuracy on the ASL-100-RGBD dataset. While these methods demonstrate promising performance, their reliance on raw video data continues to pose significant privacy risks and limits applicability across diverse sign languages.

In this context, radar-enhanced ISLR systems emerge as a promising compromise between recognition accuracy and user confidentiality. Nevertheless, most existing work fails to address the specialised lexical needs of domain-specific applications, such as those found in medical consultations. This observation underlines the motivation for the present study, which leverages high-resolution radar sensing to develop a privacy-preserving ISLR framework tailored to doctor-patient communication scenarios.

3. Track 01: Continuous Sign Language Recognition Track

3.1. Tasks Description

The first track of the SignEval 2025 challenge is a shared task on CSLR, focusing on recognizing gloss sequences (i.e., sign language sentences) from pose-based input streams. Participants are provided with pose modality data derived from the Isharah dataset [3] and are challenged to develop models that can effectively generalize to unseen signers and novel sentence structures. For this task,

a subset of the Isharah dataset was used, consisting of over 14,000 RGB videos covering approximately 1,000 unique sentences in Saudi Sign Language (SSL), performed by 18 different signers.

Two shared tasks have been organized within this track: signer-independent and unseen sentences recognition tasks.

Task 01 - Signer-Independent. In this task, models are trained on a subset of signers and evaluated on entirely unseen signers. This tests the model’s ability to generalize across individuals, a critical property for scalable real-world CSLR systems. The training set for this task consists of 10,000 samples performed by 12 signers. The development set includes 950 samples performed by one signer who does not appear in the training data. The test set contains 3,800 samples performed by five signers who are unseen in both the training and development sets.

Task 02 - Unseen-Sentences. This task evaluates the model’s ability to recognize sign sequences that were not seen during training. Although the model has been exposed to the individual glosses in various contexts during the training, it has never encountered them in the specific sentence structures used for testing. The dataset used for this task consists of 9,900, 550, and 550 samples for the training, development, and test sets, respectively.

3.2. Dataset

The CSLR tasks leverage the Isharah dataset [3], which is a large-scale, multi-scene resource developed to advance research in CSLR and sign language translation. It contains 30,000 video clips of SSL sentences recorded by 18 proficient SSL users, including deaf individuals, hard-of-hearing participants, and certified interpreters. The videos capture 2,000 unique sentences across domains such as health-care, transportation, education, legal services, and emergency scenarios. These recordings were made by participants using smartphones in natural environments, introducing a wide range of visual variability that mirrors real-world conditions.

For the CSLR shared tasks, we use the Isharah-1000 subset, which consists of 14,000 videos and 1,000 unique sentence types. To extract the pose data from the dataset, the MediaPipe Holistic model [45] is utilized, which estimates full-body, hand, and facial landmarks from each video frame. We select 86 pose keypoints consisting of 21 landmarks from each hand, 24 from the upper body, and 20 keypoints representing the lip contour. The dataset used for the *Signer-Independent* task was split into 10,000, 950, and 3,800 samples for training, development, and testing sets, respectively. Each set contains signers not present in the others. The dataset used for the *Unseen-Sentences* task consists of 9,900, 550, and 550 samples for the training, development, and test sets, respectively.

3.3. Evaluation Metrics

For the CSLR challenge tasks, we employ the *Word Error Rate (WER)* as the primary metric to evaluate system performance. WER measures the proportion of errors in the recognized (hypothesized) gloss sequence relative to the reference (ground-truth) gloss sequence, accounting for substitutions (Sub.), deletions (Del.), and insertions (Ins.) It is defined as:

$$\text{WER} = \frac{\text{Sub.} + \text{Del.} + \text{Ins.}}{\text{ReferenceLength}} \times 100\% \quad (1)$$

3.4. Results and Participating Teams

3.4.1. Baselines

We proposed a baseline for both tasks of the CSLR track using 2D keypoints extracted from hand landmarks. Each frame is represented by 42 keypoints from both hands, resulting in a sequence tensor $P = \{p_1, p_2, \dots, p_T\}$ of shape $T \times 42 \times 2$, which is flattened to $T \times 84$ and processed through a linear projection layer into a d -dimensional hidden space. The proposed baseline adopts an architecture built upon the standard Transformer encoder [42], enhancing temporal modeling through stacked attention layers and temporal pooling. It consists of four Transformer encoder blocks with sinusoidal positional encodings to preserve temporal structure. No intermediate supervision is applied between these blocks, allowing residual accumulation of temporal context. After the Transformer layers, the model performs progressive temporal pooling using two 1D average pooling layers interleaved with temporal convolutional layers (TCNs) to reduce sequence length and capture broader contextual patterns. A final multi-layer perceptron outputs per-frame gloss predictions.

3.4.2. Results

The results of the two tasks in this track are presented in Table 1 (Singer-Independent task) and Table 2 (Unseen-Sentence task). Across both tasks, the top three teams converged on two primary paradigms: Conformer-based encoders and Graph Convolutional Networks (GCNs). Conformer-based encoders interleave convolutional modules for fine-grained and local feature extraction with self-attention layers for capturing long-range dependencies. The GCNs jointly process semantically grouped 3D keypoints (e.g., face, hands, body) in parallel streams before fusing their representations.

Each team conducted a detailed analysis of the Isharah dataset to tailor their architectures and auxiliary modules to the distinct challenges posed by signer variability (Task 1) versus the greater demands of unseen-sentence generalization (Task 2). Almost all solutions adopted a *hybrid attention* strategy using global self-attention to model long-range temporal context and local convolutions or learnable

pooling for adaptive downsampling and refinement of short-range temporal patterns.

Notably, Task 2 proved substantially more challenging, where WERs roughly tripled compared to Task 1 for most teams, underscoring the difficulty of generalizing to novel sentence structures without any prior exposure during training.

As shown in Tables 1 and 2, a clear pattern of overfitting emerges when comparing the WERs of the development and test sets. Some teams that achieve very low WER on the development set often incur substantially higher Test WER, indicating models have memorized idiosyncrasies of the held-out development data rather than learning robust generalizations. For example, in Task 1, the VIPL.SLP team achieved a development WER of 25.16% and a test WER of 7.45%, which indicates a surprisingly better performance on unseen test data. This can be attributed to the use of careful regularization and data augmentation strategies that effectively mitigated overfitting. By contrast, the CV Group (UWB) exhibits a development WER of 48.42% whereas their test WER is 33.82%. This inverted gap points to potential data leakage or overly aggressive model pruning tuned on the development set.

In Task 2, the overfitting gap widens considerably, where the CV Group (UWB) reported a development WER of just 2.77%, yet suffers an 80.27% WER on Test, a more than 28× increase that typifies extreme over-adaptation to development sentences. Similarly, the development WER of the Decoder team is 9.44%, whereas their test WER jumped significantly to 107.27%. The increase confirms that heavily parameterized models can “memorize” development transcripts but fail to generalize to novel sentence structures without stronger regularization or cross-validation rigor.

Overall, the discrepancies between development and test WERs underscore the need for robust regularization and more diverse development folds. Moreover, CSLR models should focus on learning the sign gestures rather than memorizing signer-specific patterns, in order to enhance robustness to signer variation.

3.4.3. Participating Teams

VIPL.SLP [44] proposed a skeleton-based CSLR framework using a unified “CNN/GCN-Conv1D-BiLSTM” pipeline. Based on CoSign [21], keypoints are grouped (e.g., hands, face, body) and processed by group-specific GCNs. Alternatively, they encoded keypoints as Gaussian heatmaps, processed via CNNs. Frame-level features are aggregated with 1D convolutions, followed by a two-layer BiLSTM for temporal modeling. Predictions from both local (frontend) and global (backend) features are supervised with CTC loss. Motion cues from joint-wise offsets and regularization enhance the fusion of multi-stream signals. Models were trained on 86 keypoints. Their single-stream

#	Team Name	Dev WER (%)	Test WER (%)
1	VIPL_SLP [44]	25.16	07.45
2	Kronus [39]	09.90	09.65
3	Decoder [13]	05.62	12.01
4	CPAMI [16]	07.31	13.07
5	Kumass2020	06.94	16.67
6	RW [22]	15.37	20.45
7	StarAtNyte*	–	31.11
8	CV Group (UWB)	48.42	33.82
9	Baseline	25.10	38.50

Table 1. Results for Task 1 (Signer-Independent) in SignEval 2025 Challenge Track 1, showing both development- and test-set WERs. * indicates the team did not submit a report and is not listed among participating teams.

#	Team Name	Dev WER (%)	Test WER (%)
1	VIPL_SLP [44]	36.90	28.20
2	CPAMI [16]	55.08	47.78
3	Kronus [40]	–	55.28
4	Kumass2020	62.87	57.68
5	CV Group (UWB)	2.77	80.27
6	Baseline	87.60	81.20
7	Decoder [13]	9.44	107.27
8	Wonjune.kim*	108.63	146.00
9	RW [22]	140.95	283.56

Table 2. Results for Task 2 (Unseen Sentence) in MSLR Challenge Track 1, showing both development- and test-set WERs. * indicates the team did not submit a report and is not listed among participating teams.

model consisted of three-layer group-wise GCNs as feature extractors, followed by 1D convolutions and a BiLSTM with 1024 hidden units for local and global temporal modeling, respectively.

Kronus presented two complementary pose-based architectures, both leveraging 2D pose keypoints and anatomically dividing the input into four regions: right hand, left hand, face/lips, and upper body/torso. Each region is processed independently using a GCN-based stream, followed by either temporal compression or an MLP, and their features are fused for downstream decoding. The first model [39], targeting Signer-Independent CSLR, uses a Transformer encoder and a permutation-trained autoregressive decoder, with auxiliary gloss decoders supervising each region. The second model [40], designed for generalization to unseen sentence compositions, employs a two-stage temporal convolutional backbone, with a shared classifier producing gloss predictions at multiple temporal resolutions, all trained using CTC losses. Both models incorporate aux-

iliary supervision on region streams to ensure meaningful contributions from all parts, enhancing robustness to signer variability and sentence complexity.

Decoder [13] proposes a framework designed for signer-independent recognition, leveraging anatomically partitioned 2D pose sequences. The input skeleton is divided into four regions: right hand, left hand, lips, and torso, each encoded independently using region-specific GCNs followed by temporal compression via 1D convolutions. These part-wise features are concatenated and processed through a Transformer encoder. Gloss sequences are generated using a permutation-trained autoregressive Transformer decoder that models flexible gloss dependencies beyond monotonic left-to-right order. Auxiliary gloss decoders supervise each region stream to enforce semantic alignment and improve robustness, particularly in underrepresented regions. This architecture balances detailed local motion modeling with global temporal context to handle the complexities of continuous sign language.

CPAMI [16] addressed the challenges of signer variability and syntactic generalization by developing two specialized pose-based architectures, each tailored to the task setting. For the Signer-Independent task, they employed a Conformer-based model that combines convolutional and self-attention mechanisms to capture both local and global temporal patterns. The architecture begins with a shallow temporal encoder using 1D convolutions, followed by stacked Conformer blocks that integrate multi-head self-attention, depthwise convolutions, and feed-forward layers. This hybrid modeling approach yields signer-agnostic representations suitable for CTC-based gloss prediction. For the US track, they proposed a Multi-Scale Fusion Transformer architecture featuring a joint attention mechanism and a novel dual-path temporal encoder. This encoder processes input sequences through both fine-grained and down-sampled branches, fusing them into a rich multi-scale representation that is then passed through a Transformer encoder and an MLP-based classifier head. Both systems use CTC loss for end-to-end training without frame-level alignment.

Kumass2020 proposed a Temporal Pose-aware CSLR (TemPo) model to learn structured spatiotemporal representations from 3D pose keypoints. The input to TemPo is a tensor $X \in \mathbb{R}^{T \times D}$, where T is the number of frames and D is the pose vector dimensionality. The architecture includes a group-wise embedding that processes semantic groups of the human pose (e.g., face, hands, body) via dedicated MLPs, producing disentangled feature representations. These embeddings are concatenated and passed through a hybrid backbone combining Transformer encoder layers for capturing global context and TCNs with a channel-wise learnable pooling layer for local feature refinement and adaptive temporal downsampling. To improve robustness and reduce overfitting, the team employed a Se-

semantic Consistency Regularization (SCR) strategy involving structured masking of semantic keypoint groups during training, encouraging consistent predictions through a symmetric KL divergence loss alongside standard CTC losses. During inference, the full pose sequence is used without masking, and CTC decoding produces the final gloss sequence.

RW team [22] proposed a decoder-only autoregressive architecture for continuous sign language recognition, treating the task as a conditional gloss generation problem. Their model, named AutoSign, adapts the DTrOCR framework [14] to generate Arabic gloss sequences from 2D skeleton data extracted using pose estimation methods. The input pose sequence is grouped into four semantic parts (left/right hand, face, body), then processed with part-aware augmentations before being compressed by a two-layer 1D-CNN. This compressor reduces the temporal resolution and increases the feature dimensionality from 134 to 512. Temporal order is preserved via learnable positional embeddings. The processed pose features are projected and concatenated with gloss token embeddings, and passed through a decoder-only transformer based on AraGPT2 [4], which uses causal and cross-attention to autoregressively predict gloss tokens. The system is trained using cross-entropy loss over non-padded gloss tokens.

CV Group (UWB) focused primarily on preprocessing efforts. Their approach involved normalising signer-specific keypoints, pruning inconsistent annotations, and generating gloss-level labels suitable for training. Two primary approaches were evaluated: a customised T5-based model adapted from prior research [34], and Uni-Sign [28], a recently proposed pre-trained architecture designed for Sign Language Recognition. Both methods incorporated Sign Space normalization and augmentation strategies described in [34].

4. Track 02: Isolated Sign Language Recognition

4.1. Task Description

The ISLR Challenge seeks to advance multimodal sign-language and gesture recognition by providing a comprehensive dataset that pairs RGB video with 60 GHz mm-wave radar Doppler (RDM) measurements. Its structured yet flexible competition format invites teams to devise and evaluate novel methods that harness the complementary strengths of these modalities and push the boundaries of multimodal fusion. By leveraging the newly released MultiMeDaLIS dataset [31], the challenge encourages systems able to capture subtle hand shapes, intricate motion dynamics, and fine-grained temporal cues, setting a new benchmark for multimodal understanding.

4.2. Dataset

MultiMeDaLIS [31] is an Italian Sign Language (LIS) dataset tailored to patient–doctor communication. It comprises 126 lexical items - 100 healthcare-specific signs and the 26 letters of the Italian one-hand manual alphabet - each performed 205 times by a proficient signer, yielding 25,830 annotated instances. Data were recorded synchronously with four privacy-preserving, non-invasive sensors: a three-antenna 60 GHz radar array (13 fps), 720p 30 fps RGB-D video with facial-landmark tracking (Intel RealSense D455), dedicated facial-expression streams from a Kinect v1, and 1080p 25 fps depth point clouds from a Zed 2. This multimodal capture spans both manual articulators and subtle non-manual cues (e.g., facial movements, torso shifts, mouthing), yielding a dataset larger and more suitable for supervised learning than previously available radar-based or RGB-D-based sign-language corpora. The dataset is partitioned into 12,600 labelled training samples across 126 gesture-specific directories, and 4,410 validation samples plus 7,560 test samples stored in class-agnostic directories with concealed labels. Each sample directory contains one frontal RGB video and three temporally aligned range-Doppler clips, preserving multimodal synchrony while masking class information in the file paths.

4.3. Evaluation Metrics

For the ISLR track, we evaluate the performance of the recognition models using the standard classification metrics: *accuracy*, *sensitivity* (recall), *specificity*, and the F_1 -*score*. All these metrics are derived from the confusion matrix, which comprises true positives, true negatives, false positives, and false negatives.

4.4. Results and Participating Teams

4.4.1. Baselines

To contextualize the results achieved by participants in the challenge, we define a set of reference baselines that include both standard deep learning architectures and recent domain-specific approaches relevant to Italian Sign Language (LIS) recognition in medical contexts.

As a core benchmark, we adopt the ResNet(2+1)d architecture as in Caligiore et al. [6], which extends standard 2D convolutional networks by factorizing 3D spatiotemporal filters into separate spatial and temporal components. This architecture is well-suited to capture motion dynamics in video-based sign language recognition while maintaining a balanced computational cost. We evaluate ResNet(2+1)d under three input modalities: RGB only, RGB combined with depth (RGB-D), and RGB combined with Range-Doppler Maps (RGB+RDM), the latter incorporating radar-derived motion information into the visual stream.

Moreover, Mineo et al. have recently introduced two radar-only architectures for clinical LIS: one leveraging

	Data type	Acc.	Sens.	Spec.	F1-Sc.
R(2+1)d [6]	RGB	71.8	64.5	79.1	58.2
R(2+1)d [6]	RGB-D	74.6	66.4	82.8	59.9
Mineo et al. [32]	RDM+MTI	93.6	87.9	99.3	87.7
TRACE [33]	RDM+MTI	99.3	98.8	99.9	98.8
R(2+1)d [32]	RGB+RDM	79.4	70.2	88.6	75.5
ROLL TIDE RADAR [36]	RGB+RDM	99.8	99.9	100.0	99.9
Former Human [23]	RGB+RDM	99.8	99.8	100.0	99.8
IMPRESS NCSU [30]	RGB+RDM	99.7	99.7	100.0	99.7
CPAMI (UW) [20]	RGB+RDM	99.4	99.4	100.0	99.4
ML	RGB+RDM	99.3	99.3	100.0	99.4
SabyasachiBiswas147	RGB+RDM	97.1	97.1	100.0	97.1
Matteo Cacioppo	RGB+RDM	78.1	78.1	99.8	78.9

Table 3. Baseline and Leaderboard for the SignEval Challenge Track 2.

a residual autoencoder on Range-Doppler Maps with a subsequent Transformer for classification [32], and another (TRACE) combining a ResNetMC backbone with CLIP-based text alignment to directly associate radar features with natural-language labels, thereby improving semantic coverage of medical sign vocabularies [33].

4.4.2. Results

Table 3 reports performance on the ISLR challenge track. The R(2+1)d model from Caligiore et al. [6] serves as a visual baseline, achieving 71.8% accuracy on RGB alone and 74.6% when incorporating depth. The two radar-only approaches by Mineo et al. [32, 33] demonstrate the power of RDM+MTI inputs, both reaching 93.6% accuracy, which far exceeds the RGB-only baseline and confirms the efficacy of radar for privacy-preserving ISLR.

Among the challenge participants, fusion of RGB+RDM modalities yields near-perfect scores: *ROLL TIDE RADAR* and *Former Human* both achieve 99.8% accuracy, while *IMPRESS NCSU* and *CPAMI (UW)* closely follow at 99.7% and 99.4%, respectively. These results underline the advantage of multimodal integration—particularly late fusion strategies—in leveraging complementary visual and motion cues. Notably, leaderboard entries such as *ML* and *SabyasachiBiswas147* maintain similarly high accuracy (99.3% and 97.1%), whereas *Matteo Cacioppo* submission reaches 78.1%, suggesting that model architecture and fusion design critically impact final performance. Overall, the top-ranked teams substantially outperform both the vision-only and radar-only baselines, confirming that joint RGB–radar systems offer the most robust solution for isolated sign recognition in clinical contexts.

4.4.3. Participating Teams

ROLL TIDE RADAR [36] combines a state-of-the-art video-based ISLR model with a novel radar envelope pipeline. For the RGB branch, they adopt Uni-Sign - a dual-

encoder architecture that fuses a vision encoder and a pose encoder via a prior-guided fusion module before querying mT5-Base for classification - fine-tuned on MultiMeDaLIS with an Italian-language head. For the radar branch, they extract kinematic envelopes (upper and lower) from single-antenna RDMs to condense the 3D data cube into paired 2D heatmaps, which are then resized and fed into pretrained image classifiers (ConvNeXtV2, Swin Transformer V2, DeiT, and VGG16) in order to form RDM-Env. The Uni-Sign branch uses the publicly released WLASL100 weights and is fine-tuned for 20 epochs with AdamW (learning rate 3×10^{-4}), a batch size of 8, and a cross-entropy loss on pose-processed clips from RTMPose. The radar envelope models are trained by first computing energy-threshold envelopes (95% upper, 5% lower) per frame, stacking them into $2 \times T \times R$ matrices, and generating RGB-style heatmaps. Each candidate classifier is then fine-tuned on these 2D inputs for 20 epochs using AdamW at 3×10^{-4} , batch size 8, and standard ImageNet normalization. Performance is monitored via validation accuracy, with the best model selected for late fusion.

Former Human [23] implements a two-branch architecture. The RGB branch is a 3D-ResNet-18 model pretrained on Kinetics-400 and fine-tuned to the 126-class vocabulary by replacing its final fully-connected layer with a 126-unit linear head. The Radar branch employs a diffusion-based U-Net trained to reverse Gaussian noise over 400 timesteps on RDMs. Latent features are extracted from the fourth down block and fed to one of three classifier heads (linear, GRU, or partially fine-tuned U-Net + spatial-temporal head) to produce 126 logits. The RGB branch was trained for 15 epochs on a single NVIDIA P100 GPU using AdamW with a learning rate of 1×10^{-4} and default weight decay. They processed 16-frame RGB clips, resized to 224x224 and normalized with ImageNet statistics, in batches of eight. Cross-entropy loss was optimized while logging metrics to TensorBoard and a CSV file at each iteration; they saved both the final and best-performing model weights. For the radar branch, three strategies were explored: first, a linear head trained with AdamW (1×10^{-3}) on frozen diffusion features; second, a Bi-GRU head trained with AdamW (1×10^{-4} , weight decay 1×10^{-5}) and label smoothing (0.1), again with the backbone frozen; and third, a fine-tuned setup in which early U-Net blocks remained frozen while deeper layers and a new spatial-temporal head were updated with differential learning rates (5e-5 for U-Net, 1e-3 for spatial extractor, 2e-3 for the GRU+MLP), including a 5-epoch warm-up and cosine learning-rate schedule. Finally, an ensemble late-fusion (0.1 radar : 0.9 RGB) combined one-hot logits from both branches.

IMPRESS NCSU [30] approach begins with the Frequency-Modulated Continuous Wave radar signals provided as RDM sequences. After describing each chirp

$f_c(t) = f_0 + \frac{B}{T_c}t$ and mixing to obtain the IF signal $s_{IF}(t)$, they compute 2D FFTs across fast-time and slow-time to form RDMs. Rather than applying STFT to raw returns, they collapse the RDM along the range axis—summing bins $r_{\text{start}} = 50$ to $r_{\text{end}} = 200$ —to derive equivalent micro-Doppler images $S_{\mu\text{-D}}(\omega, t)$. This yields a compact Doppler-time spectrogram per frame, which is then normalized and resized into 96×96 grayscale inputs. Classification is performed by a three-branch CNN: each branch processes one antenna’s micro-Doppler image through two 64-filter and two 128-filter convolutional blocks (3×3 kernels, Batch-Norm, ReLU, max-pool), then relational features are concatenated, passed through a 512-unit fully-connected layer with batch-norm, ReLU, and 0.5 dropout, and finally softmaxed over 126 classes. All models were trained using the Adam optimizer with an initial learning rate of 10^{-3} and standard cross-entropy loss. Inputs were batched at 64 samples per iteration, and training ran for 30 epochs with learning-rate reduction on plateau and early stopping based on validation accuracy. To ensure efficient use of the provided RDM data, the range-collapsed $\mu\text{-D}$ images were pre-computed and cached, reducing on-the-fly FFT overhead. Model checkpoints were saved at each validation improvement, and the best weights were retained for final evaluation.

CPAMI (UW) [20] proposes FusionEnsemble-Net, a hierarchical ensemble that ingests both RGB video and RDM data through four parallel spatiotemporal backbones (3D ResNet-18, MC3-18, R(2+1)D-18, Swin-B). Each backbone extracts modality-specific features, which are then temporally modeled via LSTM or transformer encoders. An attention fusion module dynamically weights and combines the RGB and radar representations for each backbone, producing four fused feature vectors. Finally, each vector is classified by its own linear head, and the per-head probability outputs are averaged to yield the final sign prediction. The entire network was implemented in PyTorch and trained end-to-end on two NVIDIA A6000 GPUs. All spatiotemporal backbones were initialized with Kinetics-400 (for 3D CNNs) or ImageNet (for Swin-B) pre-trained weights. The RGB clips and the synchronized three-antenna RDM sequences were resized to 224×224, normalized, and batched at size 4. They optimized the cross-entropy loss using AdamW with a cosine annealing learning-rate schedule, running for 30 epochs. Temporal modeling layers comprised two-layer LSTMs (hidden size 512) or transformer encoders (8 heads). During inference, predictions from all four classification heads were averaged, resulting in a robust ensemble output.

5. Conclusion

This paper presents a summary of SignEval 2025, the first multimodal SLR challenge, held as part of the MSLR work-

shop at ICCV 2025. The challenge comprised two distinct tracks: CSLR, utilizing pose-based input from the Isharah dataset, and ISLR, based on the multimodal MultiMeDaLIS dataset. The CSLR track included two shared tasks, focusing on signer-independent recognition and unseen-sentences generalization. We provide an overview of the tasks, datasets, and evaluation protocols, which highlight the diversity and complexity of the problems addressed. Multiple teams participated in the challenge, submitting a wide range of innovative solutions that employed deep learning architectures, attention-based fusion strategies, and modality-specific adaptations. The benchmark results and methodologies discussed in this paper offer valuable insights into the current state of SLR research. They underline the potential of pose-based representations for developing scalable CSLR systems and the effectiveness of multimodal fusion for achieving robust ISLR performance. We hope this benchmark will foster future research on inclusive, privacy-aware, and multimodal sign language technologies. Future editions of the MSLR challenge aim to broaden the scope to include continuous multimodal recognition, signer adaptation, and multilingual sign language understanding.

Acknowledgement

The first track of the SignEval challenge is funded by the SharedTech. This work was also supported by the SDAIA-KFUPM Joint Research Center for Artificial Intelligence at King Fahd University of Petroleum and Minerals. Raffaele Mineo and Amelia Sorrenti are PhD students enrolled in the National PhD in Artificial Intelligence, XXXVII and XXXVIII cycles, respectively, course on Health and life sciences, organized by University Campus Bio-Medico of Rome.

References

- [1] Sarah Alyami and Hamzah Luqman. Swin-mstp: Swin transformer with multi-scale temporal perception for continuous sign language recognition. *Neurocomputing*, 617:129015, 2025. 1
- [2] Sarah Alyami, Hamzah Luqman, and Mohammad Ham-moudeh. Reviewing 25 years of continuous sign language recognition research: Advances, challenges, and prospects. *Information Processing & Management*, 61(5): 103774, 2024. 1
- [3] Sarah Alyami, Hamzah Luqman, Sadam Al-Azani, Maad Alowaiifeer, Yazeed Alharbi, and Yaser Alonazian. Isharah: A large-scale multi-scene dataset for continuous sign language recognition. *arXiv preprint arXiv:2506.03615*, 2025. 2, 3
- [4] Wissam Antoun, Fady Baly, and Hazem Hajj. Arabert: Transformer-based model for arabic language understanding. *arXiv preprint arXiv:2003.00104*, 2020. 6

- [5] Homa Arab, Iman Ghaffari, Lydia Chioukh, Serioja Ovidiu Tatu, and Steven Dufour. A convolutional neural network for human motion recognition and classification using a millimeter-wave doppler radar. *IEEE Sensors Journal*, 22(5):4494–4502, 2022. 3
- [6] Gaia Caligiore, Raffaele Mineo, Concetto Spampinato, Egidio Ragonese, Simone Palazzo, and Sabina Fontana. Multisource approaches to italian sign language (LIS) recognition: Insights from the multimedalis dataset. In *Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024)*, pages 132–140, Pisa, Italy, 2024. CEUR Workshop Proceedings. 2, 6, 7
- [7] Necati Cihan Camgoz, Oscar Koller, Simon Hadfield, and Richard Bowden. Sign language transformers: Joint end-to-end sign language recognition and translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10023–10033, 2020. 2
- [8] Ka Leong Cheng, Zhaoyang Yang, Qifeng Chen, and Yu-Wing Tai. Fully convolutional networks for continuous sign language recognition. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIV 16*, pages 697–714. Springer, 2020. 2
- [9] Rungpeng Cui, Hu Liu, and Changshui Zhang. Recurrent convolutional neural networks for continuous sign language recognition by staged optimization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7361–7369, 2017. 2
- [10] Mathieu De Coster, Mieke Van Herreweghe, and Joni Dambre. Isolated sign recognition from rgb video using pose flow and self-attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3441–3450, 2021. 3
- [11] Bidya Debnath, Iffat Ara Ebu, Sabyasachi Biswas, Ali C Gurbuz, and John E Ball. Fmcw radar range profile and micro-doppler signature fusion for improved traffic signaling motion classification. In *2024 IEEE Radar Conference (RadarConf24)*, pages 1–6. IEEE, 2024. 3
- [12] El-Sayed M El-Alfy and Hamzah Luqman. A comprehensive survey and taxonomy of sign language research. *Engineering Applications of Artificial Intelligence*, 114:105198, 2022. 1
- [13] Fatimah Mohamed Emad Elden. Cslrconformer: A data-centric conformer approach for continuous arabic sign language recognition on the isharah dataset. *arXiv preprint arXiv:2508.01791*, 2025. 5
- [14] Masato Fujitake. Dtrocr: Decoder-only transformer for optical character recognition. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 8025–8035, 2024. 6
- [15] Sevgi Z Gurbuz, Ali C Gurbuz, Evie A Malaia, Darrin J Griffin, Chris Crawford, M Mahbubur Rahman, Ridvan Aksu, Emre Kurtoglu, Robiulhossain Mdrafai, Ajaymehul Anbuselvam, et al. A linguistic perspective on radar micro-doppler analysis of american sign language. In *2020 IEEE international radar conference (RADAR)*, pages 232–237. IEEE, 2020. 3
- [16] Md Rezwanul Haque, Md. Milon Islam, S M Taslim Uddin Raju, and Fakhri Karray. A signer-invariant conformer and multi-scale fusion transformer for continuous sign language recognition. In *2025 IEEE/CVF International Conference on Computer Vision Workshops*, 2025. 5
- [17] Marlene Hilzensauer and Klaudia Krammer. A multilingual dictionary for sign languages:” spreadthesign”. In *ICERI2015 Proceedings*, pages 7826–7834. IATED, 2015. 3
- [18] Hezhen Hu, Weichao Zhao, Wengang Zhou, and Houqiang Li. Signbert+: Hand-model-aware self-supervised pre-training for sign language understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):11221–11239, 2023. 1
- [19] Lianyu Hu, Liqing Gao, Zekang Liu, and Wei Feng. Continuous sign language recognition with correlation network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2529–2539, 2023. 1
- [20] Md. Milon Islam, Md Rezwanul Haque, S M Taslim Uddin Raju, and Fakhri Karray. Fusionensemble-net: An attention-based ensemble of spatiotemporal networks for multimodal sign language recognition. In *2025 IEEE/CVF International Conference on Computer Vision Workshops*, 2025. 7, 8
- [21] Peiqi Jiao, Yuecong Min, Yanan Li, Xiaotao Wang, Lei Lei, and Xilin Chen. Cosign: Exploring co-occurrence signals in skeleton-based continuous sign language recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 20676–20686, 2023. 1, 4
- [22] Samuel Johnny, Blessed Guda, Andrew Stephen, and Assane Gueye. Autosign: Direct pose-to-text translation for continuous sign language recognition. In *2025 IEEE/CVF International Conference on Computer Vision Workshops*, 2025. 5, 6
- [23] Jakub Juranek. Multimodal italian sign language recognition with radar–video late fusion on the multimedalis dataset. In *2025 IEEE/CVF International Conference on Computer Vision Workshops*, 2025. 7
- [24] Zhenghao Ke, Sheng Liu, and Yuan Feng. Fine-grained cross-modality consistency mining for continuous sign language recognition. *Pattern Recognition Letters*, 191:23–30, 2025. 2
- [25] Hovannes Kulhandjian, Prakshi Sharma, Michel Kulhandjian, and Claude D’Amours. Sign language gesture recognition using doppler radar and deep learning. In *2019 IEEE globecom workshops (GC Wkshps)*, pages 1–6. IEEE, 2019. 3
- [26] Beichen Li, Jingyu Yang, Yang Yang, Chen Li, and Yutong Zhang. Sign language/gesture recognition based on cumulative distribution density features using uwb radar. *IEEE transactions on instrumentation and measurement*, 70:1–13, 2021. 3
- [27] Dongxu Li, Cristian Rodriguez, Xin Yu, and Hongdong Li. Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1459–1469, 2020. 1, 3
- [28] Zecheng Li, Wengang Zhou, Weichao Zhao, Kepeng Wu, Hezhen Hu, and Houqiang Li. Uni-sign: Toward unified sign language understanding at scale. *arXiv preprint arXiv:2501.15187*, 2025. 6

- [29] Yilong Lu and Yue Lang. Sign language recognition with cw radar and machine learning. In *2020 21st International Radar Symposium (IRS)*, pages 31–34. IEEE, 2020. 3
- [30] Sultan Mohammad Manjur, Sabyasachi Biswas, and Ali C. Gurbuz. A multimodal video and radar fusion framework for high-accuracy isolated sign language recognition. In *2025 IEEE/CVF International Conference on Computer Vision Workshops*, 2025. 7
- [31] Raffaele Mineo, Gaia Caligiore, Concetto Spampinato, Sabina Fontana, Simone Palazzo, and Egidio Ragonese. Sign language recognition for patient-doctor communication: A multimedia/multimodal dataset. In *2024 IEEE 8th Forum on Research and Technologies for Society and Industry Innovation (RTSI)*, pages 202–207. IEEE, 2024. 6
- [32] Raffaele Mineo, Gaia Caligiore, Federica Proietto Salanitri, Isaak Kavasidis, Senya Polikovsky, Sabina Fontana, Egidio Ragonese, Concetto Spampinato, and Simone Palazzo. Radar-based imaging for sign language recognition in medical communication. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2025. 7
- [33] Raffaele Mineo, Amelia Sorrenti, Gaia Caligiore, Federica Proietto Salanitri, Giovanni Bellitto, Senya Polikovsky, Sabina Fontana, Egidio Ragonese, Concetto Spampinato, and Simone Palazzo. Text-aligned radar-based sign language recognition for healthcare communication. In *2025 IEEE/CVF International Conference on Computer Vision Workshops*, 2025. 7
- [34] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020. 6
- [35] Razieh Rastgoo, Kourosh Kiani, and Sergio Escalera. Sign language recognition: A deep survey. *Expert Systems with Applications*, 164:113794, 2021. 1
- [36] Dmitriy Sazonov, Kamrul Islam, Evie Malaia, and Sevgi Gurbuz. Modality-specific benchmarks and radar range-doppler envelope classification for multimodal isolated sign language recognition. In *2025 IEEE/CVF International Conference on Computer Vision Workshops*, 2025. 7
- [37] Ala addin I Sidig, Hamzah Luqman, and Sabri A Mahmoud. Arabic sign language recognition using optical flow-based features and hmm. In *Recent Trends in Information and Communication Technology: Proceedings of the 2nd International Conference of Reliable Information and Communication Technology (IRICT 2017)*, pages 297–305. Springer, 2018. 1
- [38] William C Stokoe Jr. Sign language structure: An outline of the visual communication systems of the american deaf. *Journal of deaf studies and deaf education*, 10(1):3–37, 2005. 2
- [39] Sieu Tran. Region-aware pose modeling and permutation decoding for signer independent sign language recognition. In *2025 IEEE/CVF International Conference on Computer Vision Workshops*, 2025. 5
- [40] Sieu Tran. Generalizable sign language recognition via local temporal convolutions and region-aware pose encoding. In *2025 IEEE/CVF International Conference on Computer Vision Workshops*, 2025. 5
- [41] Elahe Vahdani, Longlong Jing, Matt Huenerfauth, and Yingli Tian. Multi-modal multi-channel american sign language recognition. *International Journal of Artificial Intelligence and Robotics Research*, 1(01):2450001, 2024. 3
- [42] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 4
- [43] Zhen Wang, Dongyuan Li, Renhe Jiang, and Manabu Okumura. Continuous sign language recognition with multi-scale spatial-temporal feature enhancement. *IEEE Access*, 2025. 2
- [44] Yifan Yang, Peiqi Jiao, Zixi Nan, and Xilin Chen. A closer look at skeleton-based continuous sign language recognition. In *2025 IEEE/CVF International Conference on Computer Vision Workshops*, 2025. 4, 5
- [45] Fan Zhang, Valentin Bazarevsky, Andrey Vakunov, Andrei Tkachenka, George Sung, Chuo-Ling Chang, and Matthias Grundmann. Mediapipe hands: On-device real-time hand tracking. *arXiv preprint arXiv:2006.10214*, 2020. 3
- [46] Qidan Zhu, Jing Li, Fei Yuan, and Quan Gan. Continuous sign language recognition based on motor attention mechanism and frame-level self-distillation. *Machine Vision and Applications*, 36(1):1–12, 2025. 2