SignEval 2026 Challenges Results

Ahmed Abul Hasanaath[1]*, Raffaele Mineo[2]*, Hamzah Luqman[1], Sarah Alyami[3], Maad Alowaifeer[1], Amelia Sorrenti[2], Gaia Caligiore[2], Sabina Fontana[2], Egidio Ragonese[2], Giovanni Bellitto[2], Federica Proietto Salanitri[2], Concetto Spampinato[2], Motaz Alfarraj[1], Mufti Mahmud[1], Simone Palazzo[2], Nour Imane Zeghib[1]

[1] King Fahd University of Petroleum & Minerals, [2] University of Catania, [3] Imam Abdulrahman Bin Faisal University

Abstract

This paper summarizes the results of the second Multimodal Sign Language Recognition Challenge, SignEval 2026, organized at CVPR 2026. The challenge featured two tracks: (i) a Continuous Sign Language Recognition (CSLR) task based on the newly curated Isharah dataset, a Saudi Sign Language dataset, and (ii) an 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 evaluates a model's ability to generalize across different signers, a critical property for scalable real-world CSLR systems. The Unseen-Sentences task assesses the model's capability to recognize novel sentence compositions by leveraging learned grammar and semantics. The challenge utilized two leaderboards to showcase submitted methods, with participants setting new benchmarks and achieving state-of-the-art results across both tracks. More information on the challenge, tasks, leaderboards, baselines, and development kits is available at: https://m-slrt.github.io/MSLR2026/.

1. Introduction

Sign language recognition has become an active research area in computer vision and multimodal learning, driven by the need for accessible communication technologies and by the growing availability of benchmark datasets across different sign languages and sensing modalities. Despite recent progress, two challenges remain central: robust continuous sign language recognition in realistic sentence-level settings, and privacy-preserving isolated sign language recognition when visual sensing is unavailable or undesirable. These two directions motivate SignEval 2026, the second challenge organized within the CVPR Multimodal Sign Language Recognition Workshop.

SignEval 2026 is structured into two complementary tracks. Track 1 addresses Continuous Sign Language Recognition (CSLR) using the Isharah dataset [5], a large-scale Saudi Sign Language (SSL) benchmark containing sentence-level samples recorded in realistic environments. To evaluate different forms of generalization, this track is further divided into signer-independent and unseen-sentence tasks. Track 2 focuses on Isolated Sign Language Recognition (ISLR) using MultiMeDaLIS [47], a multimodal Italian Sign Language (ISL) benchmark designed for patient-doctor communication. In contrast to the multimodal isolated track of SignEval 2025 [43], the 2026 edition restricts evaluation to radar-only Range-Time Maps (RTMs), encouraging methods that operate without identity-bearing visual information and are therefore suitable for privacy-sensitive settings such as hospitals and outpatient environments. Together, the two tracks provide a broad view of the current sign language recognition landscape, spanning both sequence-level linguistic modeling and compact privacy-aware sensing.

This paper presents the datasets, task definitions, evaluation settings, leaderboard outcomes, and participating systems of SignEval 2026. For Track 1, we summarize the results of the signer-independent and unseen-sentence CSLR tasks and discuss the dominant modeling strategies adopted by the top teams. For Track 2, we report the outcomes of the radar-only isolated-sign challenge, compare them with contextual baselines and prior MultiMeDaLIS results, and analyze the main methodological trends emerging from the submitted reports.

*These authors contributed equally to this work.
Corresponding authors: R. Mineo (raffaele.mineo@unict.it), Hamzah Luqman (hluqman@kfupm.edu.sa).

2. Literature Review

2.1. Continuous Sign Language Recognition

CSLR is commonly formulated as a sequence-to-sequence learning problem that involves three core components: visual feature extraction, temporal sequence modeling, and alignment or decoding. The feature extraction stage captures spatial appearance and short-term motion patterns, the sequence modeling stage learns long-range temporal dependencies, and the alignment component establishes correspondence between video frames and gloss sequences [19].

For visual representation learning, convolutional neural networks (CNNs) are widely used [10, 11, 28, 72]. Early approaches relied on 2D CNNs for frame-level features [11, 28], while later works explored 3D CNNs to model short-term spatio-temporal cues [10, 72]. More recently, Vision Transformers (ViTs) have been introduced to capture global spatial dependencies [3, 41]. Pose-based representations have also been explored to explicitly model human motion [24, 33]. Despite their effectiveness, these approaches often struggle with fine-grained motion modeling and generalization.

For temporal modeling, most methods employed temporal convolutional networks (TCNs) followed by recurrent neural networks (RNNs) to efficiently capture both short-term and long-term temporal dependencies [31, 34]. More recently, Transformer-based architectures [9, 13, 71] have demonstrated strong performance by capturing global dependencies through self-attention.

A key challenge in CSLR is learning from unsegmented sequences with only gloss-level annotations. While early methods relied on hidden Markov models (HMMs) [36, 69], Connectionist Temporal Classification (CTC) has become the dominant approach due to its ability to enable end-to-end training without frame-level supervision. Building on this, Cui et al. [12] proposed a recurrent CNN framework with CTC-based alignment and iterative refinement, while Cheng et al. [11] introduced a fully convolutional model with a gloss feature enhancement module.

Beyond unimodal approaches, multimodal learning has been explored to improve representation quality. Methods such as TwoStreamSLR and STMC integrate RGB and pose information [10, 70], while other works incorporate attention mechanisms, correlation modeling, and multi-stream designs to enhance spatio-temporal representations [27, 29]. To address data limitations, some studies leverage cross-lingual data or self-supervised pretraining [24, 67]. In contrast, Alyami and Luqman [4] proposed CLIP-SLA, which adapts pretrained vision-language models using parameter-efficient modules to align visual features with gloss embeddings, improving generalization while maintaining efficiency, particularly in low-resource settings.

Recent works have focused on improving spatio-temporal modeling [25, 66, 66, 71]. Wang et al. [66] proposed STNet with enhanced spatial and temporal modules, while Zhu et al. [71] introduced MAM-FSD, which combines motor attention and self-distillation for improved motion modeling. More recently, Hu et al. [25] proposed MTCNet, which models motion dynamics through cross-frame refinement and temporal-channel recalibration, achieving strong performance on benchmark datasets. In addition, Conformer-based models have been explored to combine convolution and self-attention for joint local and global modeling [2].

2.2. Isolated Sign Language Recognition

ISLR assigns a single gloss to a temporally bounded signing clip. Although simpler than continuous sign language recognition, it remains challenging because discriminative information is distributed across handshape, motion trajectory, body pose, and non-manual components.

Progress in ISLR has been largely driven by visual benchmarks such as MS-ASL, WLASL, and AUTSL, which supported CNN-based, 3D convolutional, transformer-based, and pose-driven models for word-level recognition [7, 14, 40, 57, 61]. De Coster et al. showed that combining pose flow, hand crops, and self-attention improves RGB-based recognition [14], while pose-based transformers and other landmark-driven methods demonstrated that structured body, hand, and facial keypoints can provide competitive performance with lower computational cost [7, 40]. Multimodal visual approaches further combined RGB-D, motion, and skeletal cues through multi-branch architectures, confirming the effectiveness of appearance-based sensing when both manual and non-manual components are available [62].

For Italian Sign Language in healthcare-oriented settings, MultiMeDaLIS introduced a multimodal benchmark for patient-doctor communication, with synchronized RGB, RGB-D, lidar, and 60 GHz radar acquisitions over 126 isolated classes, including 100 medical signs and 26 alphabet letters [47]. Subsequent work on the same corpus explored multisource visual learning from RGB, depth, optical flow, and skeleton representations [8], while the SignEval 2025 challenge established MultiMeDaLIS as a reference benchmark for comparative ISLR evaluation [43].

At the same time, camera-based systems remain problematic in privacy-sensitive environments such as hospitals and outpatient settings. This has motivated research on RF and radar sensing, which capture motion without storing identity-bearing imagery and are less affected by illumination variability [16–18, 45]. Early studies showed that RF sensing can support sign classification in non-visual settings [18], while related analyses highlighted the relevance of radar micro-Doppler patterns for sign characterization and corpus development [17, 45]. Later works extended this line

to word-level recognition, trigger-sign detection, and more realistic sequential and interactive settings [16, 37, 38, 54]. In parallel, radar representations have also proved effective in related gesture and motion analysis tasks through Range-Doppler, FMCW, and spectrogram-like encodings [6, 15, 32]. However, much of this literature focuses on ASL, generic gestures, or limited-vocabulary settings, leaving open the question of radar-based recognition for larger and domain-specific sign lexicons.

This gap is particularly relevant for MultiMeDaLIS. Mineo et al. proposed a radar imaging framework based on multi-antenna Range-Doppler Maps and Moving Target Indication for isolated LIS recognition in medical communication scenarios [48]. TRACE extended this direction by aligning radar features with textual embeddings through prompt-conditioned contrastive learning [49]. More recently, SABRE introduced the first benchmark on the frequency-domain radar branch of MultiMeDaLIS using Range-Time Maps (RTMs), together with a dedicated baseline combining convolutional encoding, bidirectional temporal modeling, attention pooling, and metric supervision [51]. These developments motivate the isolated track considered in this work, which focuses on radar-only recognition of isolated LIS signs in a compact, privacy-preserving, and clinically grounded setting [43, 47, 51].

3. Track 1: Continuous Sign Language Recognition Track

3.1. Task Description

The first track in the SignEval 2026 challenge focuses on CSLR, where the objective is to predict gloss sequences (i.e., sentence-level sign representations) from pose-based input signals. Participants are provided with pose features extracted from the Isharah dataset [5] and are expected to design models capable of generalizing to both unseen signers and previously unseen sentence compositions. The *Isharah* dataset contains more than 30,000 RGB video clips representing approximately 2,000 distinct sentences in SSL, performed by 18 different signers. Within this track, two separate evaluation tasks are defined: *signer-independent* and *unseen-sentence* recognition.

Task 01 - Signer-Independent. This task evaluates how well a model generalizes across different individuals. Models are trained using data from a subset of signers and are then tested on completely unseen signers. The training set includes 19,500 samples collected from 12 signers. The development set consists of 1,950 samples from a signer not included in training. The test set contains 7,800 samples from five additional signers who are absent from both training and development splits.

Task 02 - Unseen-Sentences. This task focuses on evaluating the model's ability to recognize sentence structures that were not observed during training. While the model has seen individual glosses in different contexts, it has not encountered the exact sentence combinations present in the evaluation data. The dataset for this task contains 20,900 training samples, along with 550 samples each for development and testing.

3.2. Dataset

Figure 1. **Samples from the Isharah dataset [5].** The pose information of the signers are used in the CSLR track of the SignEval challenge.

Both CSLR tasks are based on the Isharah dataset [5], a large-scale and diverse resource designed to support research in sign language recognition and translation. The dataset comprises 30,000 video clips of SSL sentences performed by 18 fluent signers, including deaf individuals, hard-of-hearing participants, and professional interpreters. It covers around 2,000 unique sentences spanning multiple real-world domains such as healthcare, transportation, education, legal services, and emergency situations. Samples from the dataset are shown in Figure 1.

All recordings were captured using the front-facing cameras of smartphones in natural environments, resulting in significant variability in background, lighting, and recording conditions, thereby reflecting realistic deployment scenarios. Pose features are extracted using the MediaPipe Holistic model [68], which provides landmarks for the signer's body, hands, and face on a per-frame basis. From these outputs, we select a total of 86 keypoints, including 21 landmarks per hand, 24 upper-body landmarks, and 20 keypoints corresponding to the lip contour.

3.3. Evaluation Metrics

Performance in the CSLR tasks is assessed using the *Word Error Rate (WER)* as the primary evaluation metric. WER

quantifies the discrepancy between the predicted gloss sequence and the ground-truth sequence by considering substitutions (Sub.), deletions (Del.), and insertions (Ins.). It is computed as:

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

3.4. Results and Participating Teams

3.4.1. Baselines

For both CSLR tasks, we developed a pose-based baseline model. For each frame, 42 landmarks corresponding to both hands are used, forming a temporal sequence $P = p_1, p_2, \ldots, p_T$ with dimensionality $T \times 42 \times 2$. This representation is reshaped into $T \times 84$ and passed through a linear embedding layer to project it into a d-dimensional feature space. The architecture is based on a Transformer encoder [64], leveraging stacked self-attention layers to model temporal dependencies. Specifically, the model comprises four Transformer encoder layers augmented with sinusoidal positional encodings to retain temporal order information. No auxiliary losses are introduced between layers, enabling the model to progressively accumulate temporal context through residual connections. Following the Transformer stack, temporal resolution is gradually reduced via a combination of two 1D average pooling layers and TCNs, which together capture longer-range dependencies. Finally, a multi-layer perceptron produces frame-wise gloss predictions.

3.4.2. Results

The results for the CSLR track are presented in Table 1 (Singer-Independent) and Table 2 (Unseen-Sentences). The top-performing teams converged on two dominant paradigms: graph convolutional networks (GCNs) for structured keypoint modeling and attention-based encoders (including Conformer and Transformer variants) for long-range temporal dependencies. Teams such as TEMPO [21], LiftSign [52], VIPL_SLP, El-iitk, Suvajit_PatraL, BioRG [42], and Tawasul [1] built on multi-stream or hybrid architectures that separately process spatial and motion features, followed by sequence modeling via BiLSTMs, GRUs, or Transformer blocks. Many teams also applied cross-stream consistency, temporal pooling, or attention mechanisms to improve generalization.

Task 1 proved to be more manageable, with the top three teams (TEMPO, LiftSign, VIPL_SLP) achieving Test WERs below 6.5%. TEMPO narrowly led the leaderboard with a WER of 5.68%, benefiting from a two-stream ST-GCN backbone augmented with temporal modeling modules and robust regularization. LiftSign and VIPL_SLP followed closely, combining multi-stream processing and ensemble strategies to improve temporal receptive fields and mitigate overfitting.

By contrast, Task 2 highlighted the challenge of generalizing to unseen sentences, where WERs approximately quintupled for most teams. LiftSign and VIPL_SLP maintained the top positions with Test WERs of 27.35% and 27.62%, respectively, illustrating that careful temporal modeling, dual supervision, and ensemble strategies can partially compensate for the lack of exposure to novel sentence structures. Teams such as El-iitk, TEMPO, BioRG, and Tawasul also demonstrated competitive performance, leveraging hybrid attention and cross-stream refinement to extract signer-agnostic features.

Across both tasks, a notable trend is the divergence between development and test WERs, reflecting overfitting risks when models rely excessively on development data. Task 2 particularly exposes extreme over-adaptation, underscoring the necessity of robust regularization, diversified development folds, and mechanisms that emphasize learning sign gestures rather than memorizing sequences. Overall, these results highlight that hybrid architectures combining spatial GCNs with attention-based temporal encoders continue to be the most effective approach for CSLR, especially when complemented by cross-stream consistency and ensemble modeling.

#	Team Name	Test WER (%)
1	TEMPO [21]	5.68
2	LiftSign [52]	5.70
3	VIPL_SLP	6.37
4	El-iitk	7.14
5	Algosalih	8.13
6	Nttruong	8.99
7	Hang	9.41
8	Suvajit_Patra	14.72
9	BioRG [42]	16.35
10	Tawasul [1]	16.62
11	Baseline	20.12
12	SmartALC	36.97
13	Farhankhan	38.14
14	CyberTI	38.49
15	Lili12	42.17
16	Astral_fate	47.82
17	IAsystems	58.44
18	Peneter	64.15
19	TKA	68.4
20	IDEAMCVG	69.16

Table 1. The leaderboard results for Task 1 (Signer-Independent) of the CSLR track.

3.4.3. Participating Teams

TEMPO team [21] proposed a pose-based CSLR framework built upon an enhanced Two-Stream CoSign-2s [10, 33] backbone, leveraging both spatial and temporal dynamics from 2D skeleton sequences. The input consists of 86 keypoints per frame, enriched through feature engineering to include part-normalized coordinates, bidirectional tem-

#	Team Name	Test WER (%)
1	LiftSign [52]	27.35
2	VIPL_SLP	27.62
3	El-iitk	28.28
4	TEMPO [21]	39.73
5	Algosalih	43.17
6	BioRG [42]	50.72
7	TCVG	71.06
8	Farhankhan	71.64
9	KLsys	72.38
10	IAsystems	72.38
11	CyberTI	72.77
12	Hang	73.23
13	TKA	75.63
14	Peneter	78.38
15	Tawasul [1]	78.72
16	Mshamani	79.77
17	AADE	79.92
18	Baseline	80.31

Table 2. The leaderboard results for Task 2 (Unseen-Sentences) of the CSLR track.

poral offsets, and confidence scores, forming a $T \times 86 \times 7$ representation. The model separates static and motion information into two streams processed via ST-GCNs, with part-wise representations fused using complementary masking to encourage signer-invariant features. On top of this backbone, several temporal modeling modules are explored, including the TAPE adapter [20], MSTCN [3], and a hybrid BiLSTM–Transformer sequence model. Additional regularization strategies such as Stochastic Sequence Depth and Cross-Stream Consistency are employed to improve generalization and reduce overfitting. Supervision is applied only on the fused representation through an objective combining CTC and reconstruction losses. Final predictions are obtained via CTC decoding with beam search, followed by a vocabulary-constrained dynamic programming post-processing step to refine sequence outputs.

LiftSign team [52] built upon the CoSign-2s [33] architecture. The team adopts a multi-stream ST-GCN backbone that processes 86 keypoints partitioned into anatomical regions (body, hands, and mouth). Separate static and motion streams capture spatial configurations and temporal dynamics, respectively, while a fusion stream integrates both to learn higher-level spatiotemporal representations. To enhance temporal modeling, standard pooling operations with a TLP module [26] were used. The extracted features are further modeled using a two-layer BiLSTM to capture long-range dependencies, with dual CTC supervision applied to both convolutional and sequential representations. In addition, complementary regularization is employed through masked temporal views to enforce consistency. A cross-stream knowledge distillation strategy is introduced, where the fusion stream guides the static and motion streams to improve representation quality. During inference, predictions from all three streams are combined using a temperature-scaled ensemble decoding scheme, followed by a CTC beam search to produce the final gloss sequence.

VIPL_SLP proposed an approach built upon the CoSign-1s [33] baseline. Their method focuses on improving temporal modeling through an ensemble strategy composed of multiple models with varying temporal receptive field configurations, achieved by altering the arrangement of 1D convolution and pooling layers. To enhance robustness in the signer-independent setting, they construct diverse training subsets by resampling signers, encouraging the model to generalize across different individuals. Additional temporal augmentation is applied via sequence rescaling. The final predictions are obtained by combining multiple trained models.

El-iitk proposed a two-stream framework that models both spatial and temporal data through a combination of graph-based feature extraction and sequence modeling. The input consists of 86 keypoints grouped into anatomical regions, with an additional motion stream derived from frame-wise joint offsets to capture temporal changes. Each stream is processed through group-wise GCNs to learn localized spatial representations, followed by temporal aggregation using 1D convolutional layers. For long-range dependency modeling, the framework explores Bi-LSTM for sequential context learning and a Transformer encoder with multi-head self-attention and positional encodings. Both short-term and long-term representations are supervised using a dual CTC objective to improve alignment learning. Complementary regularization is applied across streams to encourage diverse yet consistent representations.

Suvajit_PatraL team proposed a parameter-efficient framework that jointly models spatial, temporal, and sequential dependencies. Their approach employs a spatio-temporal attention encoder, in which attention mechanisms replace conventional kernels to dynamically aggregate information across temporal neighborhoods and intra-frame joint relationships. Multiple stacked attention blocks with residual connections and feed-forward layers are used to learn hierarchical spatiotemporal representations. For sequence modeling, the encoded features are processed using bidirectional GRUs to capture temporal dependencies, followed by a BERT layer to incorporate higher-level language context. The final gloss predictions are obtained through a linear projection layer.

BioRG team proposed a CSLRTransformer framework, which augments raw 86-keypoint coordinates with per-joint velocity, forming a spatiotemporal feature tensor that captures both position and motion. Spatial relationships

are modeled through an anatomy-aware GCN that combines predefined skeletal connectivity with a learnable adjacency component. To further emphasize dynamic information, a Cross-Frame Motion Refinement module adaptively enhances motion-sensitive features through gated residual connections. Temporal dependencies are captured using stacked Transformer blocks with RoPE. The encoded features are then temporally compressed via multi-stage pooling and Temporal Channel Adaptive Recalibration modules, which apply dilated convolutions and channel-wise attention to capture multi-scale temporal patterns. Finally, a lightweight MLP head produces gloss logits, and the model is trained end-to-end using the CTC objective.

Tawasul team [1] proposed a multi-stage framework that combines region-wise feature modeling with hybrid sequence learning. The input consists of normalized pose sequences partitioned into four anatomical regions: body, left hand, right hand, and face. Each region is processed independently through dedicated convolutional encoders that extract local temporal features. The multiple streams are then fused using a gated attention mechanism. The fused representation is passed through a stack of Conformer encoder blocks. For training, the model employs a hybrid objective that combines CTC loss with cross-entropy supervision from an autoregressive Transformer decoder.

4. Track 2: Isolated Sign Language Recognition Track

4.1. Task Description

The second track of SignEval 2026 focuses on radar-only ISLR. Given a pre-segmented sample containing a single Italian Sign Language (LIS) gesture, the goal is to predict one label among 126 classes. The track is based on MultiMeDaLIS [47] and targets privacy-aware sign recognition in patient-doctor communication scenarios.

Unlike the SignEval 2025 isolated track, which considered multimodal RGB and radar inputs [43], the 2026 edition is restricted to frequency-domain radar only [50]. Participants are required to use Range-Time Maps (RTMs) derived from a 60 GHz radar sensor, while visual streams and alternative radar representations are excluded from scoring. The task also serves as the challenge counterpart of the SABRE benchmark, which introduced RTM-based radar-only evaluation on the frequency-domain branch of MultiMeDaLIS [51]. Participants submit one predicted label for each evaluation sample, and top teams are asked to provide code and model checkpoints for reproducibility.

4.2. Dataset

The isolated-sign track is based on MultiMeDaLIS, a multimodal LIS dataset designed for patient-doctor communication [47]. It includes 126 isolated classes, 100 medical-domain signs and 26 alphabet letters, for a total of 25,830 instances collected across 205 recording sessions. In each session, all classes are performed by a professional signer under a controlled acquisition protocol.

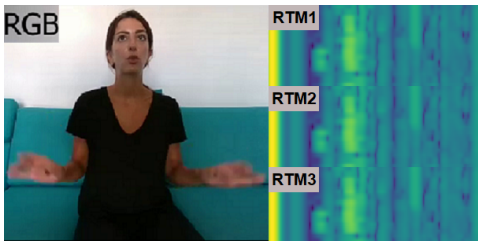

Figure 2. **Example MultiMeDaLIS sample.** Left: RGB frame of an isolated LIS gesture. Right: the corresponding three synchronized RTMs acquired from the radar receive channels. The RGB frame is shown only for illustration, while the challenge uses RTMs exclusively for scoring.

Although MultiMeDaLIS includes multiple sensing modalities, this challenge uses only the frequency-domain radar branch. Each sample is represented by RTMs derived from a 60 GHz mm-wave radar sensor with one transmit antenna and three receive antennas. An example sample is shown in Fig. 2, where an RGB frame is displayed only for illustration together with the three synchronized RTMs. RTMs are the only modality allowed for leaderboard evaluation, while visual streams and alternative radar representations are excluded. Of the original 205 recording sessions, 10 were discarded because at least one class lacked valid data, resulting in 195 usable sessions. These were split at the session level into 117 labeled training sessions, 39 unlabeled development sessions, and 39 unlabeled test sessions.

By restricting the task to RTMs only, the benchmark emphasizes privacy-preserving sign recognition without identity-bearing visual information, while still requiring models to capture fine-grained temporal and range-dependent motion patterns. In this sense, the challenge extends MultiMeDaLIS into a focused testbed for radar-only isolated sign recognition in medically relevant scenarios.

4.3. Evaluation Metrics

The isolated-sign track is evaluated using Top-1 Accuracy, which is also the official ranking metric. The public leaderboard is computed on the development split, while the final ranking is determined on a shared hidden test split.

4.4. Results and Participating Teams

4.4.1. Baselines

The official reference point for the isolated-sign track is provided by the recent SABRE benchmark [51], which introduced RTM-based radar-only evaluation on the frequency-domain branch of MultiMeDaLIS. Within this setting,

standard convolutional backbones, including MobileNetV3 [23], ResNet18 [22], EfficientNet [59], RegNet [53], and InceptionV3 [58], serve as reference RTM baselines, while SABRE represents the strongest published RTM model, reaching 81.6% accuracy [51].

For broader context, Table 3 also reports previously published results on MultiMeDaLIS obtained with RGB, RGB-D, multimodal RGB-radar, and alternative radar representations such as RDM and MTI. These methods are included only as contextual references, since they rely on different sensing modalities or evaluation protocols and are therefore not strictly directly comparable with the 2026 challenge leaderboard.

4.4.2. Results

Table 3 summarizes the outcomes of the SignEval 2026 isolated-sign track together with contextual results from prior literature. AI4Good achieved the best performance, ranking first on both the public and private leaderboards with 92.17% and 93.10% accuracy, respectively. Capybara followed with 91.15% public accuracy and 91.68% private accuracy, while High-Five and Team Alpha formed a second tier around 90% on the private leaderboard.

The results indicate a substantial improvement over the official SABRE baseline. In particular, the top submission improves upon the published RTM reference by 11.5 percentage points on the private leaderboard, while the top four leaderboard entries all remain clearly above the baseline level. Additional teams, such as RTX 5090 and ROLL TIDE RADAR, further confirm that competitive performance can be achieved even under the strict radar-only RTM setting adopted in this edition.

At the same time, the comparison with prior work highlights that RTM-only recognition remains more challenging than multimodal RGB-radar solutions and stronger radar formulations based on RDM and MTI. This suggests that the 2026 track defines a non-trivial benchmark: restrictive enough to emphasize privacy-preserving sensing, yet mature enough to support meaningful methodological progress.

4.4.3. Participating Teams

AI4Good [39] proposed a radar-only framework built around three complementary components, referred to as "Mix, Measure, Merge". The first component studies spectro-temporal data augmentation tailored to RTMs, with more than 300 controlled experiments showing that CutMix and MixUp provide the strongest gains among the tested strategies. The second combines cross-entropy with triplet loss to regularize the embedding space and improve discrimination among physically similar signs. The third relies on a cross-validation-weighted ensemble of 83 models spanning multiple CNN families, including EfficientNet, ConvNeXt, MobileOne, and RegNetY. Their analysis suggests that the main performance gain comes from the training recipe and augmentation design, while architectural diversity further improves robustness at the ensemble stage.

Table 3. **Comparison with previous methods and SignEval 2026 leaderboard results for the isolated-sign track.** For previously published methods, only test accuracy is available, so the public leaderboard column is marked with "-". Earlier results are reported for context only, since they may rely on different sensing modalities or evaluation protocols. All values are reported in %.

Method/Team	Data type	Public	Private/Test
Previous methods			
De Coster et al. [14]	RGB	-	88.4
Caligiore et al. [8]	RGB-D	-	74.6
Vahdani et al. [62]	RGB-D	-	84.1
FusionEnsemble-Net [30]	RGB + 3×RDM	-	99.4
Manjur et al. [46]	RGB + 3×RDM	-	99.7
Uni-Sign [55]	RGB + 3×RDM	-	99.8
Juranek [35]	RGB + 3×RDM	-	99.8
Jhaung et al. [32]	Radar	-	71.9
Debnath et al. [15]	Radar	-	79.3
Arab et al. [6]	Radar	-	81.0
Mineo et al. [48]	3×RDM + 3×MTI	-	93.6
TRACE [49]	3×RDM + 3×MTI	-	99.3
MobileNetV3 [51]	3×RTM	-	53.0
ResNet18 [51]	3×RTM	-	65.2
EfficientNet [51]	3×RTM	-	65.9
RegNet [51]	3×RTM	-	68.4
InceptionV3 [51]	3×RTM	-	69.5
SABRE [51]	3×RTM	-	81.6
SignEval 2026 leaderboard			
AI4Good [39]	3×RTM	**92.17**	**93.10**
Capybara [60]	3×RTM	91.15	91.68
Team Alpha [44]	3×RTM	89.91	89.99
High-Five [65]	3×RTM	89.46	90.09
RTX 5090 [56]	3×RTM	84.88	85.69
ROLL TIDE RADAR	3×RTM	81.38	-
mohboc	3×RTM	80.67	81.05
Khanh Le Nguyen Van	3×RTM	79.32	-
DDS-KGP [63]	3×RTM	72.22	73.69
Haoying	3×RTM	62.33	62.54
Le Ngo Thanh Toan	3×RTM	50.51	48.68

Capybara [60] proposed a Range-Aware Transformer framework with kinematic feature augmentation. Starting from the three RTM channels, their preprocessing pipeline derives additional motion cues such as velocity, acceleration, and range gradients, producing a richer multi-channel representation intended to reduce sensitivity to inter-subject variation and signal-intensity fluctuations. The resulting architecture combines convolutional feature extraction, range-aware pooling, and Transformer-based temporal modeling, while training is strengthened through physics-aware perturbations, CutMix-style regularization, and a two-stage optimization strategy followed by soft-voting ensemble inference. Their report also highlights the importance of preserving early range information and scaling the backbone gradually rather than relying on ag-

gressive early pooling.

High-Five [65] proposed PRISM, a privacy-preserving radar-based recognition framework formulated as a dual-perspective ensemble. A first stream uses a ConvNeXt-Large backbone on the stacked three-channel RTM tensor to capture global macro-level motion patterns, whereas a second multi-branch ResNet stream processes each antenna separately before adaptive fusion, so as to preserve antenna-specific cues and mitigate viewpoint bias. The two streams are combined through weighted soft-voting over stratified 5-fold models, with the global branch assigned a higher weight and the view-specific branch used to disambiguate more difficult cases. Their pipeline also includes radar-specific preprocessing, augmentation, and test-time augmentation, yielding a robust ensemble-oriented solution.

Team Alpha [44] proposed MSBAD, a Multi-Scale Born-Again Distillation framework centered on multi-scale feature extraction and diversity-driven ensembling. Their method augments ConvNeXtV2-Tiny backbones with a backbone-agnostic head that aggregates features from multiple stages through GeM pooling and learned scale attention, capturing both coarse temporal dynamics and fine spatial details. To improve both single-model quality and ensemble diversity, they combine progressive resolution scaling with born-again distillation and cross-architecture knowledge distillation involving ConvNeXtV2, CAFormer-S18, and EfficientNetV2-S models. At inference time, they use a six-model geometric-mean ensemble with multi-view test-time augmentation and an additional consensus filtering step, showing that diversity across architectures and resolutions is more beneficial than simply increasing ensemble size.

RTX 5090 [56] proposed a dual-stream architecture referred to as CAST/RadarCVD-Net, explicitly motivated by the physical properties of RTM signals. Their method first converts RTMs into Cadence Velocity Diagrams (CVDs) through dB-to-linear inversion followed by temporal FFT, aiming to recover motion periodicity without introducing artefacts from operating directly in the log domain. In parallel, the original RTMs are processed as magnitude-based structural representations. The two streams are encoded through separate ConvNeXt-Tiny and EfficientNetV2-S backbones, while a Cross-Antenna Spatial Attention module models the geometric relation among the three radar receivers. Finally, an asymmetric cross-attention fusion module integrates the RTM and CVD branches, yielding a model designed to combine range structure and velocity evidence.

DDS-KGP [63] proposed a lightweight three-stream convolutional baseline operating directly on RTM inputs. Each radar channel is independently normalized and temporally resampled to a fixed length, then encoded through compact convolutional blocks designed to capture temporal-spectral micro-motion patterns with minimal preprocessing overhead. The stream-level features are concatenated and passed to a multilayer perceptron for final classification. Although simpler than the top-ranked ensemble-based solutions, this approach provides a useful reference point for understanding how far a compact convolutional design can go in the RTM-only setting, and shows that competitive radar-only recognition does not necessarily require highly complex fusion mechanisms.

5. Conclusion

This paper presented the SignEval 2026 challenge at the CVPR Multimodal Sign Language Recognition Workshop and summarized the outcomes of its two tracks. Track 1 addressed CSLR on Isharah through signer-independent and unseen-sentence settings, while Track 2 focused on radar-only ISLR on MultiMeDaLIS using RTM representations. The results show clear progress in both settings. In the CSLR track, the strongest systems combined structured pose modeling with attention-based temporal encoders. In the radar-only ISLR track, several teams substantially improved over published RTM baselines, showing that privacy-preserving radar sensing can support strong isolated sign recognition without visual input, while still remaining more challenging than multimodal or richer radar formulations. Overall, SignEval 2026 provides a common benchmark for evaluating sign language recognition across complementary tasks and sensing conditions, and we expect it to support future work on robust CSLR, radar-native modeling, and privacy-aware assistive systems.

Acknowledgement

The first track of the SignEval 2026 challenge is sponsored by 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] Sadam Al-Azani, Safwan Nabeel, Qasim Al Mahfood, and Mohanad Mohamed. Twasel at SignEval 2026: Adaptive multi-stream pose fusion for continuous sign language recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2026. 4, 5, 6

[2] Neena Aloysius, Geetha M, and Prema Nedungadi. Continuous sign language recognition with adapted conformer via unsupervised pretraining, 2024. 2

[3] Sarah Alyami and Hamzah Luqman. Swin-mstp: Swin transformer with multi-scale temporal perception for continuous sign language recognition. *Neurocomputing*, 617:129015, 2025. 2, 5

[4] Sarah Alyami and Hamzah Luqman. Clip-sla: Parameter-efficient clip adaptation for continuous sign language recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 4137–4147, 2025. 2

[5] Sarah Alyami, Hamzah Luqman, Sadam Al-Azani, Maad Alowaifeer, Yazeed Alharbi, and Yaser Alonaizan. Isharah: A large-scale multi-scene dataset for continuous sign language recognition. *arXiv preprint arXiv:2506.03615*, 2025. 1, 3

[6] 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, 7

[7] Matyáš Boháček and Marek Hrúz. Sign pose-based transformer for word-level sign language recognition. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 182–191, 2022. 2

[8] 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, 2024. 2, 7

[9] 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

[10] Yutong Chen, Ronglai Zuo, Fangyun Wei, Yu Wu, Shujie Liu, and Brian Mak. Two-stream network for sign language recognition and translation. *Advances in Neural Information Processing Systems*, 35:17043–17056, 2022. 2, 4

[11] 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

[12] Runpeng 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

[13] R. Cui et al. Spatial-temporal transformer for continuous sign language recognition. *IEEE Transactions on Multimedia*, 2023. 2

[14] 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. 2, 7

[15] 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, 7

[16] Kenneth Dehaan, Emre Kurtoglu, Sabyasachi Biswas, Caroline Kobek Pezzarossi, Darrin Griffin, Chris Crawford, Ali Gurbuz, Evie Malaia, Abraham Glasser, Raja Kushalnagar, et al. Rf-chesssign: Radar-enabled human-computer interaction in a real-time sign language-controlled game. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4941–4951, 2025. 2, 3

[17] Sevgi Z Gurbuz, Ali C Gurbuz, Evie A Malaia, Darrin J Griffin, Chris Crawford, M Mahbubur Rahman, Ridvan Aksu, Emre Kurtoglu, Robiulhossain Mdrafi, 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. 2

[18] Sevgi Z Gurbuz, Ali Cafer Gurbuz, Evie A Malaia, Darrin J Griffin, Chris S Crawford, Mohammad Mahbubur Rahman, Emre Kurtoglu, Ridvan Aksu, Trevor Macks, and Robiulhossain Mdrafi. American sign language recognition using rf sensing. *IEEE Sensors Journal*, 21(3):3763–3775, 2020. 2

[19] Aiming Hao, Yuecong Min, and Xilin Chen. Self-mutual distillation learning for continuous sign language recognition. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11283–11292, 2021. 2

[20] Ahmed Abul Hasanaath and Hamzah Luqman. Ustm: Unified spatial and temporal modeling for continuous sign language recognition. *arXiv preprint arXiv:2512.13415*, 2025. 5

[21] Ahmed Hassan and Nadine Alsayad. TEMPO at SignEval 2026: Signer-independent temporal modeling and vocabulary-constrained post-processing for arabic CSLR. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2026. 4, 5

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 7

[23] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1314–1324, 2019. 7

[24] 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. 2

[25] Hongguan Hu, Jianjun Peng, Zhidong Xiao, Li Guo, Yi Hu, and Di Wu. Motion-temporal calibration network for con-

tinuous sign language recognition. *Complex & Intelligent Systems*, 12(1):35, 2026. 2

[26] Lianyu Hu, Liqing Gao, Zekang Liu, and Wei Feng. Temporal lift pooling for continuous sign language recognition. In *European conference on computer vision*, pages 511–527. Springer, 2022. 5

[27] 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. 2

[28] X. Hu et al. Multilingual continuous sign language recognition with enhanced spatio-temporal modeling. *Expert Systems with Applications*, 235:109903, 2024. 2

[29] Y. Hu et al. Self-attention based continuous sign language recognition. *Neurocomputing*, 2023. 2

[30] Md Milon Islam, Md Rezwanul Haque, SM Raju, and Fakhri Karray. Fusionensemble-net: An attention-based ensemble of spatiotemporal networks for multimodal sign language recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4924–4930, 2025. 7

[31] H. Jang et al. Self-supervised learning for continuous sign language recognition. *Pattern Recognition*, 2023. 2

[32] Yu-Chiao Jhaung, Yu-Ming Lin, Chiao Zha, Jenq-Shiou Leu, and Mario Köppen. Implementing a hand gesture recognition system based on range-doppler map. *Sensors*, 22(11):4260, 2022. 3, 7

[33] 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. 2, 4, 5

[34] X. Jiao et al. Learning temporal dynamics with 1d convolutions for continuous sign language recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 2

[35] Jakub F Juranek. Multimodal italian sign language recognition with radar-video late fusion on the multimedalis dataset. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5020–5026, 2025. 7

[36] W W Kong and Surendra Ranganath. Towards subject independent continuous sign language recognition: A segment and merge approach. In *Pattern Recognition*, pages 1294–1308. Elsevier, 2014. 2

[37] Emre Kurtoğlu, Ali C Gurbuz, Evie A Malaia, Darrin Griffin, Chris Crawford, and Sevgi Z Gurbuz. Asl trigger recognition in mixed activity/signing sequences for rf sensor-based user interfaces. *IEEE Transactions on Human-Machine Systems*, 52(4):699–712, 2021. 3

[38] Emre Kurtoğlu, Kenneth DeHaan, Caroline Kobek Pezzarossi, Darrin J Griffin, Chris Crawford, and Sevgi Z Gurbuz. Interactive learning of natural sign language with radar. *IET Radar, Sonar & Navigation*, 18(8):1203–1216, 2024. 3

[39] Cristian Lazo Quispe and Gissella Bejarano. Ai4good at signeval 2026: Mix, measure, merge with spectro-temporal augmentation and multi-model fusion for isolated radar-based italian sign language recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2026. 7

[40] 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. 2

[41] Y. Li et al. Multi-scale vision transformer for continuous sign language recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[42] Rosario Licciardello, Georgia Fargetta, Alessandro Ortis, and Sebastiano Battiato. CSLRTransformer: A pose-only system for continuous sign language recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2026. 4, 5

[43] Hamzah Luqman, Raffaele Mineo, Murtadha Aljubran, Ahmed Abul Hasanaath, Amelia Sorrenti, Sarah Alyami, Sadam Al-Azani, Maad Alowaifeer, Jihwan Moon, Václav Javorek, et al. The signeval 2025 challenge at the iccv multimodal sign language recognition workshop: Results and discussion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5027–5036, 2025. 1, 2, 3, 6

[44] Arkadip Maitra, Suvajit Patra, and Soumitra Samanta. Msbad at signeval 2026: Multi-scale born-again distillation ensemble for radar-based sign language recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2026. 7, 8

[45] Evie Malaia, Joshua Borneman, and Sevgi Gurbuz. Capturing motion: Using radar to build better sign language corpora. In *Proceedings of the LREC-COLING 2024 11th Workshop on the Representation and Processing of Sign Languages: Evaluation of Sign Language Resources*, pages 213–218, 2024. 2

[46] Sultan Mohammad Manjur, Sabyasachi Biswas, and Ali C Gurbuz. A multimodal video and radar fusion framework for high-accuracy isolated sign language recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5002–5011, 2025. 7

[47] 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. 1, 2, 3, 6

[48] 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*, pages 533–543. Springer, 2025. 3, 7

[49] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4894–4902, 2025. 3, 7

[50] Raffaele Mineo, Amelia Sorrenti, Gaia Caligiore, Senya Polikovsky, Sabina Fontana, Egidio Ragonese, Giovanni Bellitto, Federica Proietto Salanitri, Concetto Spampinato, and Simone Palazzo. 2st multimodal italian sign language recognition challenge. https://kaggle.com/competitions/cvpr-mslr-2026-track-2, 2026. Kaggle. 6

[51] Raffaele Mineo, Amelia Sorrenti, Gaia Caligiore, Federica Proietto Salanitri, Giovanni Bellitto, Senya Polikovsky, Sabina Fontana, Egidio Ragonese, Concetto Spampinato, and Simone Palazzo. A benchmark for radar-based italian sign language recognition using frequency-domain range-time maps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2026. 3, 6, 7

[52] Nguyen Nam and Hiroki Takahashi. LiftSign at SignEval 2026: A distilled multi-stream ensemble with temporal lift pooling for skeleton-based sign language recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2026. 4, 5

[53] Ilija Radosavovic, Raj Prateek Kosaraju, Ross Girshick, Kaiming He, and Piotr Dollár. Designing network design spaces. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10428–10436, 2020. 7

[54] M Mahbubur Rahman, Emre Kurtoglu, Robiulhossain Mdrafi, Ali C Gurbuz, Evie Malaia, Chris Crawford, Darrin Griffin, and Sevgi Z Gurbuz. Word-level asl recognition and trigger sign detection with rf sensors. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8233–8237. IEEE, 2021. 3

[55] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4987–4994, 2025. 7

[56] Md. Shakhoyat Rahman Shujon, Sheikh Md. Galib Mahim, Md. Milon Islam, Md Rezwanul Haque, Fakhri Karray, Md Rabiul Islam, and Hamdi Altaheri. Cast at signeval 2026: Channel-aware spatial transfer of pretrained vision backbones via pseudo-image radar representations for isolated sign language recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2026. 7, 8

[57] Ozge Mercanoglu Sincan and Hacer Yalim Keles. Autsl: A large scale multi-modal turkish sign language dataset and baseline methods. *IEEE access*, 8:181340–181355, 2020. 2

[58] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016. 7

[59] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019. 7

[60] Duc-Thien Tran, Le-Vu Nguyen-Dinh, Hoang-Nam Trinh, Thanh-Toan Le-Ngo, and Tinh-Anh Nguyen-Nhu. Capy at signeval 2026: Range-aware transformer with kinematic feature augmentation for robust radar-based sign language recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2026. 7

[61] Hamid Reza Vaezi Joze and Oscar Koller. Ms-asl: A large-scale data set and benchmark for understanding american sign language. *arXiv e-prints*, pages arXiv–1812, 2018. 2

[62] 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. 2, 7

[63] Abhishek Bharadwaj Varanasi, Manjira Sinha, and Tirthankar Dasgupta. Rtm-cnn at signeval 2026: A lightweight convolutional baseline for rtm sequence classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2026. 7, 8

[64] 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

[65] Huayu Wang, Zhiwei Tan, Jia-Xian Jian, Simon Zou, Shiqi Huang, Pau-Choo Chung, and Jenq-Neng Hwang. Prism at signeval 2026: Privacy-preserving radar-based italian sign language recognition via convnext and ensemble learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2026. 7, 8

[66] Zhen Wang, Dongyuan Li, Renhe Jiang, and Manabu Okumura. Continuous sign language recognition with multi-scale spatial-temporal feature enhancement. *IEEE Access*, 2025. 2

[67] Z. Wei et al. Improving continuous sign language recognition with cross-lingual transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2023. 2

[68] 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

[69] Jihai Zhang, Wengang Zhou, and Houqiang Li. A threshold-based HMM-DTW approach for continuous sign language recognition. *ACM International Conference Proceeding Series*, pages 237–240, 2014. 2

[70] H. Zhou et al. Spatio-temporal multi-cue learning for continuous sign language recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021. 2

[71] 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

[72] Chong Zuo et al. C2slr: Consistency-enhanced continuous sign language recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 2