

Multiple-Instance Contrastive Alignment for Pose-Based Sign Spotting

Md. Mosharaf Hossain Apurbo, Md. Asif Ali, Kazi Reyazul Hasan*

Department of Computer Science & Engineering
Bangladesh University of Engineering & Technology, Dhaka, Bangladesh
mosharaf.apurbo57@gmail.com, ijsasif45@gmail.com,
kazireyazulhasan@gmail.com

Abstract

Sign spotting, the task of deciding whether an isolated query sign occurs inside a clip of continuous signing, enables sign language video retrieval without gloss or text annotation. We describe our system for the WSLP@EMNLP-2026 Word/Sign Presence Prediction shared task on Indian Sign Language: a lightweight (2.86M-parameter) pose-based Siamese encoder that embeds both clips in one space, forms an explicit frame-to-frame cosine similarity map, and reduces it with multiple-instance pooling over temporal windows. Training combines binary cross-entropy with a multiple-instance contrastive objective (MIL-NCE). The system reaches 72.69 accuracy and 71.95 macro-F1 on the blind test set, against 61.66 and 58.42 for the published pose baseline on the same data. Ablations identify contrastive alignment as the component that makes the architecture work: the MIL-NCE objective is what places the two clips in a comparable metric space, and training does not leave chance without it. We further document a label-balance property of the released splits that fixes 76.3% of blind test labels without a model, and recommend a change for future editions.

1 Introduction

Most sign language processing targets either *isolated* recognition, which classifies a pre-segmented sign against a fixed vocabulary, or *continuous* recognition and translation, which maps a full clip to gloss or text (Koller et al., 2015; Camgoz et al., 2020). Neither supports the operation an archive search actually needs: deciding whether a particular sign occurs somewhere inside unsegmented signing. That operation is **sign spotting** (Momeni et al., 2020; Jiang et al., 2021b). We study it in its minimal, gloss-free form under the WSLP@EMNLP-2026 Word/Sign Presence Prediction shared task on Indian Sign Language (ISL),

a language whose annotated corpora have emerged only recently (Sridhar et al., 2020; Joshi et al., 2022, 2023). Given an isolated query clip and a continuous sentence clip, the system predicts a binary presence label from 2D pose keypoints alone.

The difficulty is the *citation-to-continuous* gap: a query sign is articulated slowly and deliberately, bounded by rest, whereas the same sign inside a sentence is temporally compressed, spatially reduced and co-articulated with its neighbours (Momeni et al., 2020; Jiang et al., 2021b). The benchmark also closes the usual shortcuts — the positive rate is exactly 0.5 and per-query rates are near chance — so a model that does not genuinely compare its two inputs scores at chance.

Our system places both clips in one metric space with a shared encoder and reads the decision off an explicit frame-to-frame similarity map through multiple-instance pooling. Our contributions are:

1. A 2.86M-parameter Siamese pose encoder with an explicit alignment head, reaching 72.69 accuracy and 71.95 macro-F1 on the blind test set against 61.66 and 58.42 for the published pose baseline (§3, §5).
2. A thirteen-configuration ablation showing that the MIL-NCE objective is the single load-bearing component: without it the model is at chance, and with it every other component we added can be removed, individually within 1.5 accuracy points or all together at no cost (§5).
3. A label-balance property of the released splits that determines 76.3% of blind test labels from the released files alone, with a recommended fix for future editions (§6).

2 Related Work

Sign spotting. Early systems mined sign instances from broadcast footage using subtitles as

*Corresponding author.

weak supervision — weakly aligned (Buehler et al., 2009) or through subtitle–sign co-occurrence (Pfister et al., 2013) — matching candidates with dynamic time warping (Sakoe and Chiba, 1978). The closest modern line is the Oxford series: Albanie et al. (2020) use mouthing as a weak supervisor, and Momeni et al. (2020) add subtitles and an isolated dictionary, training a shared embedding space with a multiple-instance contrastive loss (MIL-NCE) so a dictionary exemplar lands near the same sign performed inside continuous signing. Varol et al. (2022) scale this dictionary route, Varol et al. (2021) localise signs with attention, and Bull et al. (2021) align subtitles at the sentence level. Jiang et al. (2021b) attack the problem one-shot with a Transformer that reconciles the temporal-scale mismatch between an isolated query and its co-articulated instance. Our system is the “lookup” branch of Momeni et al. (2020), restricted to pose input, a binary presence output and a single GPU. Treating a sentence as a bag of temporal windows that is positive if *any* window matches is a standard MIL formulation (Dietterich et al., 1997; Ilse et al., 2018); the MIL-NCE objective itself follows Miech et al. (2020), an InfoNCE loss (van den Oord et al., 2018) whose positive is a maximum over candidate alignments rather than one fixed pair.

Pose-based sign modelling. Skeleton input is standard for low-resource sign recognition (Jiang et al., 2021a; Hu et al., 2023), and Selvaraj et al. (2022) release pose-based pretrained models covering under-served languages including ISL. We keep the encoder small: no in-domain pretraining corpus is available for this task.

The WSLP shared task. Johnny et al. (2025) describe the system for the 2025 edition. They encode each frame with a 2D CNN over the MediaPipe keypoint grid and pass [CLS] query [SEP] sentence through a single BERT-style encoder (Devlin et al., 2019), classifying from the pooled [CLS] token. We keep their framing — video-to-video, no gloss, pose rather than RGB — but replace the joint encoder over a concatenated pair with a shared encoder and an explicit alignment map. Their ablations found plain BCE to beat contrastive training; ours find the opposite, and §5 argues the architectural difference is what reconciles the two results.

Block	Dim	Normalisation
Body (15 joints)	30	shoulder-centred, width-scaled
Body velocity	30	first difference
Right hand	42	wrist-centred, RMS-scaled
Left hand	42	wrist-centred, RMS-scaled
Geometry	9	wrist→nose, wrist→wrist
Validity mask	4	pose / RH / LH / face
Face (37 points)	74	nose-anchored, RMS-scaled
Total	231	

Table 1: Per-frame feature layout (§3.1).

3 Method

Figure 1 shows the pipeline. A sentence clip of T_s frames and a query clip of T_q frames are converted to normalised pose features, embedded by one shared encoder, and compared through an explicit frame-to-frame cosine similarity map that a multiple-instance head reduces to a presence score.

3.1 Pose Features

We work from 2D keypoints rather than RGB to suppress background, appearance and signer-identity cues (Johnny et al., 2025; Selvaraj et al., 2022). The signer is localised per frame with YOLOv8-pose (Jocher et al., 2023), frames are cropped to a shoulder-centred box, and MediaPipe (Lugaresi et al., 2019) landmarks give 25 body joints, 21 joints per hand and a 37-point reduced face mesh.

From those coordinates we derive the 231-dimensional descriptor in Table 1. Body joints are centred on the shoulder midpoint and scaled by shoulder width; hands are centred on the wrist and scaled by root-mean-square radius, which isolates handshape from camera distance and body size; the face mesh is nose-anchored and scaled the same way. Undetected articulators are forward-then backward-filled, and a validity mask records which components were imputed. Features are derived on the fly, so augmentation acts on raw coordinates before normalisation.

3.2 Shared Pose Encoder

One encoder embeds both clips into a common $d=256$ space, so that cosine similarity between a sentence frame and a query frame is meaningful. Each feature block has its own linear stem, with the left- and right-hand stems sharing weights so a handshape is encoded identically on either hand. A residual grouped temporal convolution adds local context, after which we add a learned positional

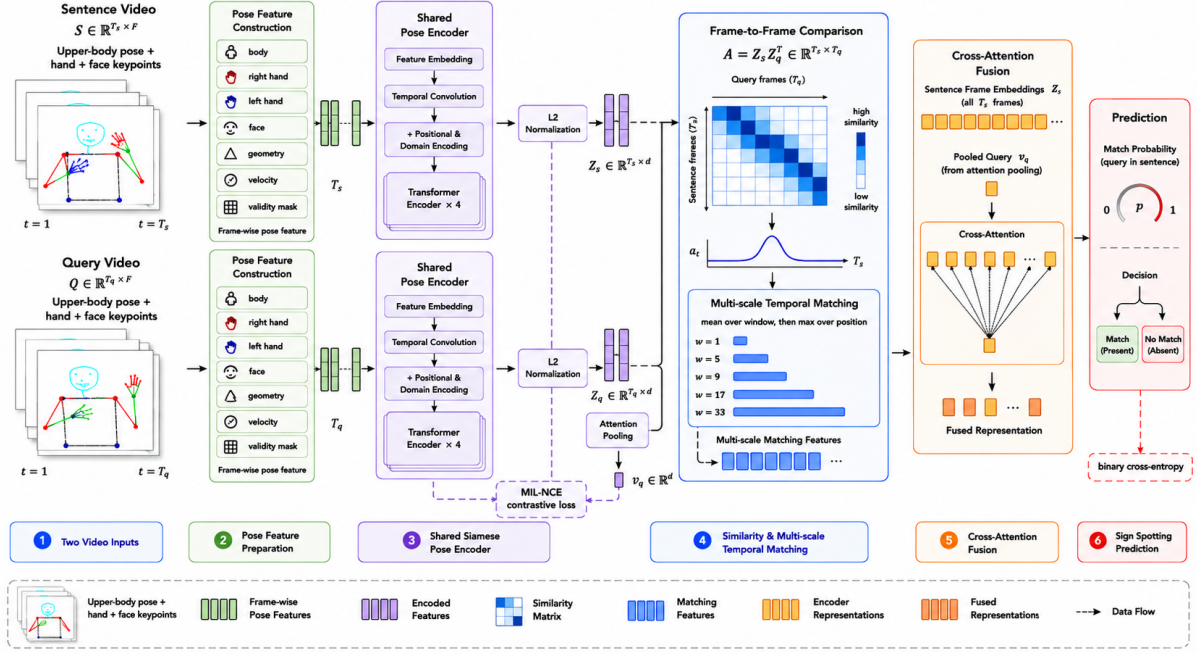


Figure 1: The sign spotting pipeline. One shared (Siamese) encoder maps the continuous sentence clip and the isolated query clip into the same 256-dimensional space, separated only by an additive learned domain embedding. The alignment branch reduces the frame-to-frame cosine similarity map by multiple-instance pooling over temporal windows; the interaction branch cross-attends the pooled query into the sentence frames. Training combines binary cross-entropy with a MIL-NCE contrastive loss applied directly to the embeddings.

embedding and a learned *domain embedding*, $\mathbf{e}_{\text{sent}}$ or $\mathbf{e}_{\text{query}}$. The domain embedding lets the shared trunk absorb systematic framing and tempo differences between citation and co-articulated signing without splitting the geometry into two spaces.

A 4-layer pre-norm Transformer encoder (Vaswani et al., 2017; Xiong et al., 2020) with 8 heads follows. It returns ℓ_2 -normalised per-frame embeddings $\mathbf{Z}^s \in \mathbb{R}^{T_s \times d}$ and $\mathbf{Z}^q \in \mathbb{R}^{T_q \times d}$, together with an attention-pooled query vector $\mathbf{v}_q \in \mathbb{R}^d$.

3.3 Multi-Scale Multiple-Instance Alignment

We form the frame-to-frame cosine similarity map

$$\mathbf{A} = \mathbf{Z}^s \mathbf{Z}^{q\top} \in \mathbb{R}^{T_s \times T_q}, \quad (1)$$

with query padding masked out. Each sentence frame t is scored against its best-matching part of the query by a smooth maximum,

$$a_t = \sum_j \text{softmax}_j(\mathbf{A}_{t,:})_j A_{tj}, \quad (2)$$

giving an activation trace $\mathbf{a} \in \mathbb{R}^{T_s}$ over sentence frames. Because a sign occupies an unknown number of frames, we pool $\mathbf{a}$ at several temporal scales

$w \in \{1, 5, 9, 17, 33\}$ under a multiple-instance reading, in which the sentence is positive if *some* window matches (Momeni et al., 2020; Jiang et al., 2021b):

$$s_w = \max_t \frac{1}{|W_t^w|} \sum_{u \in W_t^w} a_u, \quad (3)$$

where W_t^w is the set of valid frames in the width- w window centred at t . The five scale scores, a top- k mean of $\mathbf{a}$, its mean and the global maximum of $\mathbf{A}$ form an 8-dimensional alignment descriptor.

In parallel, an interaction branch cross-attends $\mathbf{v}_q$ into the full sentence frame sequence to obtain a context vector $\mathbf{c}$, and forms the comparison vector $[\mathbf{c}; \mathbf{v}_q; \mathbf{c} \odot \mathbf{v}_q; \mathbf{c} - \mathbf{v}_q]$. The concatenated alignment and interaction features are passed to a 2-layer MLP that outputs $\hat{y}$, the probability that the query sign occurs in the sentence.

3.4 Training Objective

$$\mathcal{L} = \mathcal{L}_{\text{BCE}} + 0.3 \mathcal{L}_{\text{NCE}}. \quad (4)$$

The first term is binary cross-entropy on the logit with label smoothing. The second structures the embedding space directly rather than only through the classifier head, following the MIL-NCE objective (Miech et al., 2020; Momeni et al., 2020). For a

batch of B pairs define the sentence-to-query score

$$\sigma_{ij} = \max_t \langle \mathbf{z}_i^s[t], v_{qj} \rangle, \quad (5)$$

which is the multiple-instance step: the alignment position t is a free variable rather than a supervised one. Over the subset $\mathcal{P}$ of rows whose label is positive we take a symmetric InfoNCE loss with temperature $\tau = 0.07$,

$$\mathcal{L}_{\text{NCE}} = \frac{1}{2} \left[\text{CE}_{\mathcal{P}}(\sigma/\tau) + \text{CE}_{\mathcal{P}}(\sigma^\top/\tau) \right], \quad (6)$$

where the target of row i is index i itself. Off-diagonal entries that pair a sentence with a duplicate query, or with a query labelled present for that sentence in training, are masked out so valid matches are never pushed apart.

4 Experimental Setup

4.1 Data

The shared task releases 25,432 labelled training pairs over 7,857 sentence clips and 1,410 query clips, exactly balanced between present and absent, plus 1,413 validation and 1,413 test pairs whose labels are withheld. Five training pairs are dropped because two clips failed pose extraction, leaving 25,427.

Two properties of the released splits shaped our protocol. First, the blind splits reuse training clips in new combinations: 95.97% of test sentences and 99.72% of test queries appear in training, but only 2 of 1,413 exact test *pairs* do. Evaluation therefore measures generalisation to new pairings of seen clips, which we match with a **pair-level** cross-validation split, reporting a sentence-grouped split separately as a bound. Second, the negatives are drawn at random rather than mined: a negative query is positive elsewhere in the same source video for only 1.63% of negative pairs, *below* the 3.30% rate from permuting queries at random. Uniformly sampled extra negatives are therefore distributionally close to the released ones.

The leaderboard reports accuracy with macro-averaged precision, recall and F1; we use macro-F1 ($F1_M$) as the headline score throughout.

4.2 Protocol and Implementation

The leaderboard system is a 5-fold stratified pair-level cross-validation ensemble with horizontal-mirror test-time augmentation, using one decision threshold tuned on the pooled out-of-fold predictions and applied unchanged to the blind rows. It

System	Acc.	F1 _M	P _M	R _M
2D CNN + joint enc. [†]	61.66	58.42	67.16	61.74
Ours	72.69	71.95	75.46	72.74

Table 2: Blind test scores (%), macro-averaged. [†]Johnny et al. (2025), the 2025 edition of this shared task; see the caveat in the text.

also trains on one extra uniformly sampled negative per released pair, redrawn each epoch and rejected if the pair appears in any released file, and applies a gated prior that transfers the label of a near-duplicate query when one exists for the same sentence; the prior fires on 0.2% of rows and we report it for completeness rather than as a contribution. Ablations (§5) use a cheaper protocol: one 80:20 stratified pair-level split, one run per configuration, no ensembling and no synthetic negatives, so the latter is measured in its own row rather than hidden underneath every row.

The model has 2.86M parameters and is trained for 30 epochs with AdamW (Loshchilov and Hutter, 2019) at learning rate 3×10^{-4} , weight decay 0.02, batch size 32, cosine decay after a short warmup, and an exponential moving average of the weights for evaluation. Sequences are capped at 288 sentence and 128 query frames. Augmentation — mirroring, rotation, scaling, translation, jitter, temporal speed perturbation and articulator dropout — is applied to raw coordinates, with a single mirror decision per *pair* so both clips stay in the same reference frame. The decision threshold maximises positive-class F1 subject to the predicted positive rate staying within [0.35, 0.65], which prevents a weak model from collapsing onto the all-ones solution. One run takes about 30 minutes on a single NVIDIA T4.

5 Results

5.1 Blind Test Result

Table 2 gives our official blind-test score: 72.69 accuracy and 71.95 macro-F1, against 61.66 and 58.42 for Johnny et al. (2025). The two editions release identical split statistics, so the comparison is on matched data, but we could not verify that the test labels are identical across editions and we did not rerun their system; we therefore read Table 2 as a strong indication rather than a controlled comparison. Out-of-fold on the training pairs the ensemble scores 72.37 accuracy and 72.07 macro-F1, within half a point of the blind result, so the

Configuration	Acc.	F1 _M	P _M	R _M
Full model	72.63	72.28	73.85	72.63
– MIL-NCE loss	50.92	49.84	51.01	50.92
– multi-scale MIL	73.12	72.64	74.87	73.12
– cross-attention	72.77	72.57	73.44	72.77
– domain token	73.18	72.83	74.43	73.18
– face keypoints	73.54	73.22	74.70	73.54
– velocity	72.47	71.86	74.60	72.47
– geometry	74.09	73.85	74.97	74.09
– vel. and geom.	72.79	72.34	74.38	72.79
– MIL, domain, face	73.32	72.84	75.09	73.32
– all but MIL-NCE	72.63	72.30	73.75	72.63
– mirror TTA	72.04	71.68	73.23	72.04
+ synth. negatives	72.28	71.95	73.38	72.28
Sentence-level split	69.96	69.40	71.39	69.89

Table 3: Ablations on one 80:20 pair-level split (% macro-averaged), one run each. “– all but MIL-NCE” removes the face, velocity and geometry blocks, the domain token, the cross-attention branch and the extra MIL scales together. Only removing MIL-NCE (bold) breaks the model. The last row uses a different validation partition and is a generalisation bound, not a comparable ablation.

cross-validation protocol is a reliable proxy here.

5.2 Ablations

MIL-NCE is the whole system. Removing the contrastive term (§3.4) collapses the model to chance, from 72.63 to 50.92 accuracy, and its ranking quality falls with it (AUC 0.83 → 0.51). Nothing else in Table 3 comes close to that effect. The interpretation is that the classifier head reads a cosine similarity map, and BCE alone gives no gradient that makes that map metric; the contrastive term is what puts the two clips in a comparable space in the first place. This also resolves the apparent disagreement with Johnny et al. (2025), who found contrastive supervision unhelpful: in a concatenated single-trunk encoder the decision never passes through an explicit similarity, so a contrastive term is an unrelated auxiliary objective rather than a prerequisite.

Everything else is within noise. The ten remaining removals span 72.04 to 74.09 accuracy, a 2.05-point band around a 72.63 baseline, and seven of them *improve* on it. With one seed and one partition per row we cannot separate these from run-to-run variation, so we claim no effect for any of them individually. The strongest statement the table does support is the – all but MIL-NCE row: removing the face block, velocity, geometry, the

domain token, the cross-attention branch and multi-scale pooling all at once — a 157-dimensional feature vector and 2.43M parameters, 15% fewer — matches the full model’s accuracy to two decimal places. The components layered on top of the contrastive objective are not earning their place.

Synthetic negatives. Doubling the training set with uniformly sampled negatives costs 0.35 accuracy points and doubles training time. It is in the leaderboard system because that system was trained before this measurement existed; it should be dropped.

Generalisation bound. Under a sentence-grouped split, where no validation sentence is seen in training, the model reaches 69.96 accuracy and 69.40 macro-F1, i.e. 2.7 and 2.9 points below the pair-level run. This row is not comparable to the rest of the table because its validation partition differs; it bounds how much of the headline score depends on the blind splits reusing training clips (§4.1).

6 A Label-Balance Artefact in the Splits

The released files carry a property with a larger effect on leaderboard scores than any modelling choice in this paper: *for every sentence, once the training, validation and test rows are pooled, positives equal negatives*.

Fix a sentence with c labelled training rows of which p are positive, and k blind rows. The property fixes total positives at $(c + k)/2$, so the blind rows must contain $\text{need} = (c + k)/2 - p$ positives; when $\text{need} \in \{0, k\}$ every blind row for that sentence is determined. It holds on all 7,910 sentences, and hiding 1,413 random labelled rows and re-deriving them recovers all of them exactly. Applying it pins 76.3% of test rows and 75.6% of validation rows, so answering the pinned rows and guessing the rest has expected accuracy 0.8815 with no model at all.

Exploiting this would break no rule, but an accuracy produced by counting rows measures nothing about sign language; the failure mode is a familiar one (Poliak et al., 2018; Geirhos et al., 2020). We did not use the property in any submission, and disclosed it to the organisers before submitting this paper. For future editions we recommend assigning blind rows so per-sentence label counts do not sum to a fixed ratio.

7 Conclusion

We presented a 2.86M-parameter pose-based Siamese system for gloss-free sign spotting in Indian Sign Language, which reaches 72.69 accuracy and 71.95 macro-F1 on the WSLP@EMNLP-2026 blind test set. Placing both clips in one embedding space and reading the decision off an explicit frame-to-frame similarity map is what makes the architecture work, and the ablations locate the reason precisely: the MIL-NCE contrastive objective is the only component whose removal matters, while every feature block and architectural addition we layered on top is within run-to-run noise of a 2.43M-parameter model that keeps the contrastive loss and drops the rest. For this task, effort is better spent on the objective that shapes the shared space than on enriching the encoder that fills it.

Limitations

Every ablation row is one run on one seed and one partition. The MIL-NCE effect is far too large to be noise; no other row is, and we draw no conclusion from any of them individually. In that protocol the retained epoch and the decision threshold are both chosen on the partition that is then scored, so those absolute values are optimistic, though the bias applies equally to all rows and the blind test numbers are unaffected. The decoder is also miscalibrated: it predicts positive on 66.4% of test rows where the true rate is near 50%, trading positive precision for recall. Finally, all results come from one ISL corpus, and we could not obtain further blind measurements because the evaluation portal closed before this paper was written, so the single submitted test score is the only blind evidence we can offer.

References

- Samuel Albanie, Gül Varol, Liliane Momeni, Triantafyllos Afouras, Joon Son Chung, Neil Fox, and Andrew Zisserman. 2020. [BSL-1K: Scaling up co-articulated sign language recognition using mouthing cues](#). In *Proceedings of the European Conference on Computer Vision (ECCV)*.
- Patrick Buehler, Mark Everingham, and Andrew Zisserman. 2009. [Learning sign language by watching TV \(using weakly aligned subtitles\)](#). In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Hannah Bull, Triantafyllos Afouras, Gül Varol, Samuel Albanie, Liliane Momeni, and Andrew Zisserman. 2021. [Aligning subtitles in sign language videos](#). In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.
- Necati Cihan Camgoz, Oscar Koller, Simon Hadfield, and Richard Bowden. 2020. [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 (CVPR)*.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186. Association for Computational Linguistics.
- Thomas G. Dietterich, Richard H. Lathrop, and Tomás Lozano-Pérez. 1997. [Solving the multiple instance problem with axis-parallel rectangles](#). *Artificial Intelligence*, 89(1–2):31–71.
- Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. 2020. [Shortcut learning in deep neural networks](#). *Nature Machine Intelligence*, 2(11):665–673.
- Hezhen Hu, Weichao Zhao, Wengang Zhou, and Houqiang Li. 2023. [SignBERT+: Hand-model-aware self-supervised pre-training for sign language understanding](#). *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- Maximilian Ilse, Jakub Tomczak, and Max Welling. 2018. [Attention-based deep multiple instance learning](#). In *Proceedings of the 35th International Conference on Machine Learning (ICML)*.
- Songyao Jiang, Bin Sun, Lichen Wang, Yue Bai, Kunpeng Li, and Yun Fu. 2021a. [Skeleton aware multi-modal sign language recognition](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*.
- Tao Jiang, Necati Cihan Camgöz, and Richard Bowden. 2021b. [Looking for the signs: Identifying isolated sign instances in continuous video footage](#). In *Proceedings of the British Machine Vision Conference (BMVC)*.
- Glenn Jocher, Ayush Chaurasia, and Jing Qiu. 2023. Ultralytics YOLOv8. <https://github.com/ultralytics/ultralytics>.
- Samuel Ebimobowei Johnny, Blessed Guda, Emmanuel Enejo Aaron, and Assane Gueye. 2025. Pose-based sign language spotting via an end-to-end encoder architecture. In *Proceedings of the Workshop on Sign Language Processing (WSLP)*, pages 71–76. Association for Computational Linguistics.

- Abhinav Joshi, Susmit Agrawal, and Ashutosh Modi. 2023. [ISLTranslate: Dataset for translating Indian sign language](#). In *Findings of the Association for Computational Linguistics: ACL 2023*. Association for Computational Linguistics.
- Abhinav Joshi, Ashwani Bhat, Pradeep S, Priya Gole, Shashwat Gupta, Shreyansh Agarwal, and Ashutosh Modi. 2022. [CISLR: Corpus for Indian sign language recognition](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.
- Oscar Koller, Jens Forster, and Hermann Ney. 2015. [Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers](#). *Computer Vision and Image Understanding*, 141:108–125.
- Ilya Loshchilov and Frank Hutter. 2019. [Decoupled weight decay regularization](#). In *International Conference on Learning Representations (ICLR)*.
- Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Ubaweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg, and Matthias Grundmann. 2019. [MediaPipe: A framework for building perception pipelines](#). *Computing Research Repository*, arXiv:1906.08172.
- Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. 2020. [End-to-end learning of visual representations from uncurated instructional videos](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Liliane Momeni, Gül Varol, Samuel Albanie, Triantafyllos Afouras, and Andrew Zisserman. 2020. [Watch, read and lookup: Learning to spot signs from multiple supervisors](#). In *Proceedings of the Asian Conference on Computer Vision (ACCV)*.
- Tomas Pfister, James Charles, and Andrew Zisserman. 2013. [Large-scale learning of sign language by watching tv \(using co-occurrences\)](#). In *Proceedings of the British Machine Vision Conference (BMVC)*.
- Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. [Hypothesis only baselines in natural language inference](#). In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics (*SEM)*, pages 180–191. Association for Computational Linguistics.
- Hiroaki Sakoe and Seibi Chiba. 1978. [Dynamic programming algorithm optimization for spoken word recognition](#). *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26(1):43–49.
- Prem Selvaraj, Gokul N C, Pratyush Kumar, and Mitesh Khapra. 2022. [OpenHands: Making sign language recognition accessible with pose-based pretrained models across languages](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics.
- Advaith Sridhar, Rohith Gandhi Ganesan, Pratyush Kumar, and Mitesh Khapra. 2020. [INCLUDE: A large scale dataset for Indian sign language recognition](#). In *Proceedings of the 28th ACM International Conference on Multimedia*.
- Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. [Representation learning with contrastive predictive coding](#). *Computing Research Repository*, arXiv:1807.03748.
- Gül Varol, Liliane Momeni, Samuel Albanie, Triantafyllos Afouras, and Andrew Zisserman. 2021. [Read and attend: Temporal localisation in sign language videos](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Gül Varol, Liliane Momeni, Samuel Albanie, Triantafyllos Afouras, and Andrew Zisserman. 2022. [Scaling up sign spotting through sign language dictionaries](#). *International Journal of Computer Vision*, 130(6):1416–1439.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. [Attention is all you need](#). In *Advances in Neural Information Processing Systems 30 (NeurIPS)*.
- Ruibin Xiong, Yunchang Yang, Di He, Kai Zheng, Shuxin Zheng, Chen Xing, Huishuai Zhang, Yanyan Lan, Liwei Wang, and Tie-Yan Liu. 2020. [On layer normalization in the transformer architecture](#). In *Proceedings of the 37th International Conference on Machine Learning (ICML)*.