

Findings of the Shared Tasks at the Second Workshop on Sign Language Processing (WSLP 2026)

Sanjeet Singh¹, Abhinav Joshi¹, Pokala Dattatreya¹, Keren Artiaga²,
Khushi Jaydeep Pujara¹, Tomáš Železný³, Ivan Gruber³, Jakub Honzík³,
Sabyasachi Kamila⁴, Facundo Manuel Quiroga⁵, Marek Hruz³,
Mohammed Hasanuzzaman², Ashutosh Modi¹

¹IIT Kanpur, India, ²ADAPT Centre, MTU, Cork, Ireland,

³University of West Bohemia in Pilsen, Czech Republic,

⁴Manipal Institute of Technology Bengaluru, India, ⁵UNLP, Argentina

(sanjeet, ajoshi, pokala24, khushi.jp24, ashutoshm)@cse.iitk.ac.in
keren.artiaga@adaptcentre.ie, sabyasachi.kamila@gmail.com,
facundoq@gmail.com, zeleznyt@fav.zcu.cz, grubiv@fav.zcu.cz,
honzikj@fav.zcu.cz, mhruz@fav.zcu.cz, hasanuzzaman.im@gmail.com

Abstract

WSLP-2026 hosted three shared tasks on Indian Sign Language: sentence-level translation from video or pose into English (SLT), isolated recognition over a large one-shot vocabulary (ISLR), and word/sign presence prediction in continuous sentence video (WPP). Shared tasks were hosted on Codabench. Twenty-three teams registered for translation and five submitted; thirteen registered for WPP and five submitted; twenty-two registered for recognition and nine submitted. On the lexical tracks, the systems that report usable official scores treat the problem as retrieval or temporal alignment within a pose-embedding space, often with an external exemplar bank. Translation submissions almost uniformly map pose sequences into a pretrained T5 decoder. They obtain fluent English more readily than scores that can be trusted at the current accuracy level: official BLEU can sit at zero, while chrF remains small and sensitive to output length. The edition shows that ISL can be measured by automatic evaluation metrics, but that current scores still say more about retrieval recipes and decoder fluency than about signed grammar. A later edition should mark pose-tracker failures and place native-signer judgments alongside BLEU and chrF.

1 Introduction

Sign languages (SL) are the primary means of communication for Deaf and Hard-of-Hearing communities. SLs are not a gestural code for the surrounding spoken language. Handshape, movement, location, and orientation combine with facial expression, eye gaze, and body posture into a simultaneous signal, and much of the grammar is carried by that simultaneity rather than by linear word order (Brentari, 2019; Yin et al., 2021;

Sinha, 2017; Núñez-Marcos et al., 2023). Video is expensive to store and to annotate; fluent annotators are scarce; pose estimators drop the face and sometimes the hands; and automatic metrics borrowed from text translation say little about whether a system used signing space, classifiers, or non-manuals correctly (Moryossef et al., 2020; Joshi et al., 2025; Müller et al., 2023).

A second difficulty is uneven infrastructure. Continuous translation became a standard experiment once corpora such as RWTH-PHOENIX-Weather 2014T and CSL-Daily were available (Camgöz et al., 2018; Zhou et al., 2021). Isolated recognition and fingerspelling followed a similar path on a handful of high-resource languages (Li et al., 2020; Sincan and Keles, 2020). Most other sign languages entered the literature late, with smaller public splits and few agreed test sets (Yin et al., 2021). Without a shared evaluation protocol, research or systems in the same language can be impossible to compare, and it is hard to tell whether a reported gain is an architectural result or a change in data hygiene.

We organized the WSLP shared tasks to put one such protocol on the table for Indian Sign Language (ISL). ISL is widely used in the Indian subcontinent, yet until recently it lacked publicly available sentence-level resources at a scale sufficient to make translation experiments reproducible (Joshi et al., 2024, 2022). The shortage of certified interpreters makes automated tools both practically relevant and scientifically interesting (Joshi et al., 2023, 2024). The 2026 edition of the shared task, hosted at the Second Workshop on Sign Language Processing at EMNLP 2026 in Budapest (29 October), following the inaugural edi-

tion (Singh et al., 2025), hosted three problems on the public training resources.

Sentence-level translation (SLT) asks a system to produce English text from continuous ISL sign videos. Isolated sign language recognition (ISLR) asks it to name a short citation-form clip when the target vocabulary is large and almost every class has a single training example. Word/sign presence prediction (WPP) asks whether a queried sign occurs anywhere inside a continuous sentence video. Taken together, they cover the lexical, grounding, and sentence levels that a working sign-language system actually needs. Public training data come from iSign for translation and for word/sign presence prediction (Joshi et al., 2024), and from CISLR for isolated recognition (Joshi et al., 2022). This paper is the shared-task report. Here, we present the task designs, document official participation, and compare the submitted system papers.

2 Task Design

Table 1 summarises the three tracks. Tables 2, 3, and 4 report the official test-phase rankings on translation, isolated recognition, and word/sign presence prediction. Eleven system papers were submitted in the shared-task catalog: six were accepted into the program, and five were not included. In this paper, we will discuss all the papers and technical reports submitted.

2.1 Formal task definitions

We denote a signed input as a variable-length sequence $x = (x_1, \dots, x_T)$, where each x_t is either an RGB frame or a pose vector (body, hand, and, when available, face landmarks). Both modalities were released; participants were free to use either or both. Let $\mathcal{X}$ be the space of such sequences, $\mathcal{Y}$ the space of English strings, and $\mathcal{V}$ a finite inventory of word or gloss labels.

Sign language translation. Given a continuous sentence-level sequence $x \in \mathcal{X}$, the system must generate a token sequence $\hat{y} = (\hat{y}_1, \dots, \hat{y}_M) \in \mathcal{Y}$. The intended mapping is

$$f_{\text{SLT}} : \mathcal{X} \rightarrow \mathcal{Y}, \quad \hat{y} = f_{\text{SLT}}(x). \quad (1)$$

Training pairs (x, y) come from iSign (Joshi et al., 2024). At test time x is drawn from a hidden collection of source videos that is disjoint from the public training set. Official scores are corpus-level BLEU, ROUGE, and chrF between $\hat{y}$ and English reference y (Papineni et al., 2002; Lin, 2004; Popović, 2015).

Isolated sign language recognition. Given a short citation-form clip $x \in \mathcal{X}$ that contains one lexical sign, the system must return a label $\hat{v} \in \mathcal{V}$,

$$f_{\text{ISLR}} : \mathcal{X} \rightarrow \mathcal{V}, \quad \hat{v} = f_{\text{ISLR}}(x). \quad (2)$$

The natural decision rule is nearest-exemplar retrieval in an embedding space (Joshi et al., 2022; Selvaraj et al., 2022). Evaluation is top- k accuracy on a 526-clip blind test.

Word/sign presence prediction. Given a continuous sentence sequence $x \in \mathcal{X}$ and a query word $w \in \mathcal{W}$ drawn from the official presence set built on iSign sentence video (Joshi et al., 2024), the system must decide whether the corresponding sign occurs anywhere in x :

$$f_{\text{WPP}} : \mathcal{X} \times \mathcal{W} \rightarrow \{0, 1\}, \quad \hat{z} = f_{\text{WPP}}(x, w). \quad (3)$$

The official metrics are accuracy, precision, recall, and F1 on $\hat{z}$ against a binary label z .

The three maps are related but not interchangeable. ISLR names an isolated clip; WPP asks whether that lexical item is realized inside a longer utterance; SLT requires a fluent spoken-language string for the whole utterance. A system that solves one does not automatically solve the others.

3 Translation

Two single-task translation papers were cataloged, plus one multi-task paper that reports a translation number.

Shared-Encoder Contrastive Grounding projects pose sequences into a pretrained T5 encoder-decoder and, during training, encodes the target sentence with the same T5 encoder. An auxiliary contrastive loss aligns pose and text representations beside the usual translation objective. Clip-level normalization is used to reduce scale artifacts. On a video-held-out iSign development split (Joshi et al., 2024), contrastive grounding raises chrF (Popović, 2015) from 0.102 to 0.121, and normalization takes it to 0.128. The official submission reports chrF 0.16. The same paper shows that official chrF moves with output length and with English content that does not depend on the input.

Investigating Temporal Motion Features feeds pose through a lightweight encoder into T5-small, T5-base, and T5-large (Raffel et al., 2022), and adds a T5-small variant that explicitly models frame-to-frame pose differences. T5-small is

Task	Input → output	Public data	Metrics
SLT	ISL video/pose → English sentence	iSign (118k pairs) + hidden set	BLEU, ROUGE, chrF
ISLR	isolated clip → word label	CISLR + official recognition set	Accuracy, top- <i>k</i>
WPP	(sentence video, word) → present/absent	official presence set	Accuracy, P/R/F1

Table 1: WSLP-2026 shared tasks. Official scoring was hosted on Codabench.

#	Team	chrF
1	slp_sign	0.17
2	gyanendra312	0.16
3	anhnamxtanh	0.15
4	thekraken	0.14
5	manav_dhamecha	0.13

Table 2: SLT test phase, ranked by chrF.

#	Team	Top-1
1	slp_sign	0.728
2	Wazz	0.700
3	varaprabha	0.683
4	ijsboy	0.671
5	Noob_Squad	0.671
6	WSLP-Task2	0.662
7	anhnamxtanh	0.656
8	gauandjosh	0.574
9	iiitdmk-nlp	0.196

Table 3: ISLR test phase, ranked by Top-1 accuracy.

strongest on BLEU and ROUGE among spatial-only models; T5-large is strongest on chrF; motion features give the largest BLEU gain in the authors’ development study. On the official test set, the submitted system reports chrF 0.13, ROUGE-L 0.07, and BLEU-4 0.00, and states a fifth-place leaderboard finish. A BLEU-4 of zero with non-zero chrF is evidence that n-gram overlap with the references is vanishingly rare.

Different Scales, Different Signals (multi-task) uses a language-conditioned generator for translation and reports chrF 0.17. The technical report proposes separate pathways for recognition, spotting, and translation rather than a single pooled representation.

Observations: Every translation system we received is pose-to-T5. No submitted paper describes a competitive RGB video translator on the official hidden split. Motion features and contrastive grounding are the two concrete attempts to stop the decoder from ignoring the sign stream. Both help on development data. Neither produces translation quality that the authors themselves treat as reliable. We therefore read SLT 2026 as a diagnostic year: the task is well specified, participation is thin relative to registration, and the main result is that current automatic scores sit in a range where ranking systems are fragile.

#	Team	F1
1	slp_sign	0.980
2	apurbo57	0.720
3	anhnamxtanh	0.669
4	Yaswitha	0.605
5	thekraken	0.546

Table 4: WPP test phase, ranked by macro-F1.

4 Isolated Recognition

Whitening and Fine-Tuning Are Substitutes uses MediaPipe keypoints (MediaPipe, 2023), a graph-attention backbone pretrained on a different ISL corpus, PCA whitening, and nearest-exemplar cosine retrieval against an enriched gallery. The official scores reported in the paper are 0.6825 top-1 (3rd) and 0.7357 top-10 (2nd). The analytical claim is sharper than the ranking. PCA whitening of a frozen cross-corpus encoder is worth +18.85 points on a leave-one-out gallery evaluation; end-to-end fine-tuning is worth about the same; the two do not add. Once the encoder is fine-tuned, whitening contributes +0.38. Fifteen other interventions are reported as null or harmful, with McNemar tests (McNemar, 1947) on a 2,030-query held-out split. The authors argue that the 526-clip official test set is too small for the comparisons participants want to make on it.

Metric-Learning Retrieval for Cross-Signer ISLR pretrains a metric encoder on multi-signer American Sign Language, then fine-tunes on every responsible cross-signer ISL recording the authors can add. Multiple performances inside one recording are split into separate samples. A missing-keypoint interpolation method, originally developed for classification, is adapted here for retrieval. Official test scores in the paper are 67.11% top-1, 69.20% top-5, and 70.34% top-10.

CARE (Coverage-Aware Retrieval Ensemble) keeps a complete local visual dictionary and adds a CISLR (Joshi et al., 2022) reference bank for labels that exist in both vocabularies. A frozen Video-Swin encoder (Liu et al., 2022) ranks local labels; frozen I3D (Carreira and Zisserman, 2017) and SignRep encoders supply external evidence after orthogonal Procrustes alignment on shared-class centroids. External evidence is masked to supported labels. No target-specific backbone

fine-tuning and no large-vocabulary classification head are used. CARE reports 74.31/76.15/76.15 on the validation split and 65.59/67.11/68.25 on the official test split. CARE’s authors read the 8.72-point top-1 drop as a caution against model selection on only 109 validation queries.

SignaFusion establishes a primary vision prediction, then calls auxiliary appearance, pose, and vision-language experts only when the primary model is uncertain. A disagreement-gated Qwen2.5-VL (Bai et al., 2025) module is one of those auxiliaries. Fusion variants are combined with reciprocal rank fusion. Official blind-test scores in the paper are 66.16% top-1, 72.43% top-5, and 73.19% top-10.

ContextVQ-Mamba trains a pose-only discrete tokenizer (articulator-aware graph encoder, bidirectional Mamba blocks (Gu and Dao, 2023), EMA codebook) and then recognises with frozen representations only. Because almost every official class is a singleton, the authors build a 25-class multiplicity-aware probe from the few labels that have genuine repeats. Token-plus-canonical-pose fusion reaches 28.57% top-1 on that probe. The paper is an analysis of functions as an analysis of representation complementarity rather than a full-vocabulary official submission.

Spatiotemporal CNN Ensembles compare MediaPipe k NN with pretrained R3D-18, R(2+1)D-18, and MC3-18 RGB networks. Landmark k NN obtains 4.56% top-1; R3D-18 reaches 16.54%; ensembles reach 17.68–18.63%. The strongest zero-shot submission compares MediaPipe k NN (MediaPipe, 2023) with pretrained R3D-18, R(2+1)D-18, and MC3-18 RGB networks (Tran et al., 2018).

The multi-task paper reports 0.7281 top-1 for isolated recognition, with protocol notes intended to separate visual transfer from gallery overlap.

Observations: The working ISLR systems in this edition are retrievers, not 4,361-way classifiers. External data, another ISL corpus, ASL pre-training, and CISLR as a reference bank are used by every accepted paper. Appearance and vision-language experts help when they are gated; they do not replace pose or local galleries. We also observe a spread in reported official top-1 roughly from the high teens (RGB ensembles without target adaptation) to the high 60s (adapted retrieval). That spread is large enough to be informative even if the 526-clip test set is too small for fine rank-

ings.

5 Word/Sign Presence Prediction

Multiple-Instance Contrastive Alignment is a 2.86M-parameter pose Siamese encoder. Isolated query and sentence clip are embedded in one space, a frame-to-frame cosine map is formed, and multiple-instance pooling reduces the map over temporal windows. Training combines binary cross-entropy with MIL-NCE (Miech et al., 2020). The paper reports 72.69 accuracy and 71.95 macro-F1 on the blind test set, an improvement over the pose baseline’s 61.66 accuracy and 58.42 macro-F1 on the same data. Ablations attribute the gain to contrastive alignment: without MIL-NCE the two clips do not share a metric space. The same paper documents a label-balance property of the released splits that, on the authors’ analysis, fixes 76.3% of blind-test labels without a model, and asks future editions to change the split construction.

Active-Frame Filtering and Cross-Architecture Ensembling starts from a diagnostic: isolated query videos miss a hand detection in 49.3% of cases against 2.8% for continuous sentence videos. The system filters to active frames and ensembles a recurrent encoder with a cross-attention Transformer. Official test macro-F1 is 60.51%. Two of four cumulative design decisions are statistically robust on the authors’ validation split; two are not, and the paper reports both.

The multi-task paper reports 0.9801 macro-F1 for presence prediction and flags evaluation identities that can inflate that number if protocols are not separated.

Observations: WPP is the task where pre-processing and split design matter as much as the encoder. Missing hands on isolated queries, not sentence video, dominate one diagnostic. Label balance in the released splits dominates another. A lightweight pose aligner with an explicit similarity map outperforms a generic sequence classifier in the accepted paper, which is consistent with the retrieval view that already governs ISLR.

6 Cross-Task Analysis

Two patterns cut across tracks. **1) Pose is the default input; appearance is an expert, not a base model.** Translation and WPP systems in the catalog are pose-first. Recognition systems that score well on the official vocabulary use pose or mixed-gallery settings and call RGB or vision-language models only for uncertain queries or for lexically

matched external references. The RGB ensemble that stays within generic video pretraining remains far below that of adapted pose retrieval. **2) Alignment objectives beat pooled classification when supervision is thin.** Contrastive grounding in SLT, cosine retrieval in ISLR, and MIL-NCE in WPP are the same idea at three timescales: put the signed input and the linguistic target in one space, then read a similarity rather than a 4,361-way or open-vocabulary softmax. Where teams kept a standard classification or generation head and hoped scale would help, official numbers are weaker or, in translation, hard to trust.

Participation itself is uneven. Recognition produced the most completed systems; translation produced the fewest relative to registrations. We believe that reflects compute and data movement more than community interest.

7 Lessons and Future Directions

The three tasks were meant to serve as a point of measurement, not a complete account of Indian Sign Language. What the submissions actually returned, once the leaderboard rank is set aside, is a short list of places where the measurement itself got in the way, and a longer list of linguistic problems that no official metric touched.

Start with translation. Several systems could produce English that reads as English. That is not the same as reading the signed sentence. A pretrained decoder will happily finish a fluent clause from a few content words, or from habits it learned on text, and BLEU-4 (Papineni et al., 2002) near zero with a non-zero chrF (Popović, 2015) is what that looks like on a hidden test set. If we run SLT again, the official suite has to stop treating decoder fluency as success. Length-controlled chrF (Popović, 2015) and content-word recall would already make small ranking gaps less theatrical. A small public subset scored by native signers would do something the automatic numbers cannot: say whether the output is an utterance a signer would accept, not only a string that shares characters with a reference. Until that exists, we should treat SLT scores as diagnostic, not as a claim that sentence-level ISL translation has been solved.

In recognition, a 4,361-class problem with almost one clip per class cannot be tuned on 109 validation queries and then ranked, to two decimal places, on 526 test clips. A larger public development gallery, or a frozen-encoder track with

the backbone held fixed, would let the argument shift toward retrieval design whitening, galleries, and external references rather than overfitting the smallest split.

Presence prediction exposed the release as much as the models. If label balance can settle most of the blind-test items before a frame is read, the split is doing part of the classification. Isolated query clips and sentence clips also do not arrive in the same condition under current pose estimators (MediaPipe, 2023; Cao et al., 2019): hands are far more often missing in the queries. Shipping detection-quality flags with the data is not a courtesy. Without them, a spotting paper cannot tell a one-handed sign from a tracker failure, and a linguistic claim about dominance or classifiers is not available to make. Future WPP splits should be balanced so that the video is necessary, and they should mark when the pose is simply not there.

In the next edition, it should look less like three isolated contests and more like a small map of that grammar. Keep pose; do not pretend it is the whole signal. Give rare lexical items and productive constructions a place in the evaluation, not only the frequent content words that a decoder can guess.

8 Conclusion

The 2026 WSLP hosted three shared tasks on Indian Sign Language: Translation, Recognition, and word/gloss presence prediction. Fifty-eight registrations produced nineteen official submissions. Translation is still expensive to finish; recognition is the track most teams can complete. The system papers and technical reports, in the program, tell a consistent story about the method. ISLR and WPP work when the signed input and the linguistic target are placed in one space and read by similarity: nearest-exemplar retrieval, coverage-aware galleries, frame-to-frame maps, multiple-instance contrastive loss. SLT work in this edition feeds the pose modality into T5 (Raffel et al., 2022), using contrastive grounding or explicit motion as the main attempt to prevent the decoder from ignoring the hands. The official translation numbers remain in a range where the ranking systems are fragile. Pose is still the right default input for privacy and cost. It is not a full linguistic representation, and the official data should say so: mark tracker failures, keep non-manual channels in view, and put a small native-signer study next to BLEU and chrF.

Acknowledgments

We thank the reviewers for their comments and suggestions, the WSLP-2026 programme committee, the authors of the accepted papers, and the teams that took part in the shared tasks. Workshop travel and related costs for the IIT Kanpur organizers were supported in part by the Research-I Foundation of the Department of Computer Science and Engineering, IIT Kanpur, and in part by the Anusandhan National Research Foundation (ANRF) under Grant No. ANRF/ARG/2025/009207/ENS. We gratefully acknowledge the financial support of the Science Foundation Ireland ADAPT Centre for Digital Content Technology under Grant No. 13/RC/2106. The ADAPT Centre’s grant is co-funded under the European Regional Development Fund. The organizers affiliated with the University of West Bohemia in Pilsen have been exclusively supported by the grant of the University of West Bohemia, Project No. SGS-2025-011.

References

- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*.
- Diane Brentari. 2019. *Sign Language Phonology*. Key Topics in Phonology. Cambridge University Press.
- Necati Cihan Camgöz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. 2018. Neural sign language translation. In *IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT.
- Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. 2019. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- Joao Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.
- Albert Gu and Tri Dao. 2023. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*.
- 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*, Toronto, Canada. 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*, pages 10357–10366, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Abhinav Joshi, Romit Mohanty, Mounika Kanakanti, Andesha Mangla, Sudeep Choudhary, Monali Barbate, and Ashutosh Modi. 2024. *iSign: A benchmark for Indian Sign Language processing*. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 10827–10844, Bangkok, Thailand. Association for Computational Linguistics.
- Abhinav Joshi, Vaibhav Sharma, Sanjeet Singh, and Ashutosh Modi. 2025. *PoseStitch-SLT: Linguistically inspired pose-stitching for end-to-end sign language translation*. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 13834–13853, Suzhou, China. Association for Computational Linguistics.
- Dongxu Li, Cristian Rodriguez, Xin Yu, and Hongdong Li. 2020. *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*.
- Chin-Yew Lin. 2004. *ROUGE: A package for automatic evaluation of summaries*. In *Text Summarization Branches Out*, Barcelona, Spain. Association for Computational Linguistics.
- Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. 2022. Video swin transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3202–3211.
- Quinn McNemar. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157.
- MediaPipe. 2023. *MediaPipe Holistic*.
- 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)*, pages 9879–9889.
- Amit Moryossef, Ioannis Tsochantaridis, Roei Aharoni, Sarah Ebling, and Srini Narayanan. 2020. *Real-time sign language detection using human pose estimation*. In *European Conference on Computer Vision*. Springer.
- Mathias Müller, Zifan Jiang, Amit Moryossef, Annette Rios, and Sarah Ebling. 2023. *Considerations for meaningful sign language machine translation based*

- on glosses. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Toronto, Canada. Association for Computational Linguistics.
- Adrián Núñez-Marcos, Olatz Perez de Viñaspre, and Gorka Labaka. 2023. [A survey on sign language machine translation](#). *Expert Systems with Applications*, 213:118993.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. [Bleu: a method for automatic evaluation of machine translation](#). In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
- Maja Popović. 2015. [chrF: character n-gram F-score for automatic MT evaluation](#). In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2022. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21.
- Prem Selvaraj, Gokul Nc, 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)*, Dublin, Ireland. Association for Computational Linguistics.
- Ozge Mercanoglu Sincan and Hacer Yalim Keles. 2020. Autsl: A large scale multi-modal turkish sign language dataset and baseline methods. *IEEE Access*, 8.
- Sanjeet Singh, Abhinav Joshi, Keren Artiaga, Mohammed Hasanuzzaman, Facundo Manuel Quiroga, Sabyasachi Kamila, and Ashutosh Modi. 2025. [Overview of the first workshop on sign language processing \(WSLP 2025\)](#). In *Proceedings of the Workshop on Sign Language Processing (WSLP)*, pages 1–5, IIT Bombay, Mumbai, India (Co-located with IJCNLP–AACL 2025). Association for Computational Linguistics.
- Samar Sinha. 2017. *Indian Sign Language: An Analysis of Its Grammar*. Gallaudet University Press.
- Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. 2018. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6450–6459.
- Kayo Yin, Amit Moryossef, Julie Hochgesang, Yoav Goldberg, and Malihe Alikhani. 2021. [Including signed languages in natural language processing](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Online. Association for Computational Linguistics.
- Hao Zhou, Wengang Zhou, Weizhen Qi, Junfu Pu, and Houqiang Li. 2021. Improving sign language translation with monolingual data by sign back-translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1316–1325.