Lost in Expression: Diagnosing Systemic Challenges with Non-Manual Generalization in Sign Language Understanding

Dmitriy Sazonov[1], Evie Malaia[1],Sevgi Gurbuz[2]
[1]University of Alabama [2]North Carolina State University

Background

Sign Language Understanding (SLU) Sign Language Understanding (SLU) is a field of machine learning concerned with interpreting sign language signals into linguistic outputs such as gloss sequences, translations, or semantic labels. ICCV SignEval 2025 [1] saw the use of multimodal, multichannel models on Isolated Sign Language Recognition (ISLR) reaching exceptional accuracy, pointing future efforts to more difficult subtasks such as Continuous Sign Language Recognition (CSLR) and Sign Language Translation (SLT).

Non-manual markers (NMMs) are linguistic cues in sign language expressed through the face, head, and body that modify syntactic, prosodic and semantic meaning [2]. Although NMM utilization is crucial for real-world deployment of SLU systems, it has received limited focus and is often not regarded as a primary challenge. Though some work has cited performance gain in the presence of facial features, this is not sufficient to declare faithful non-manual usage [3].

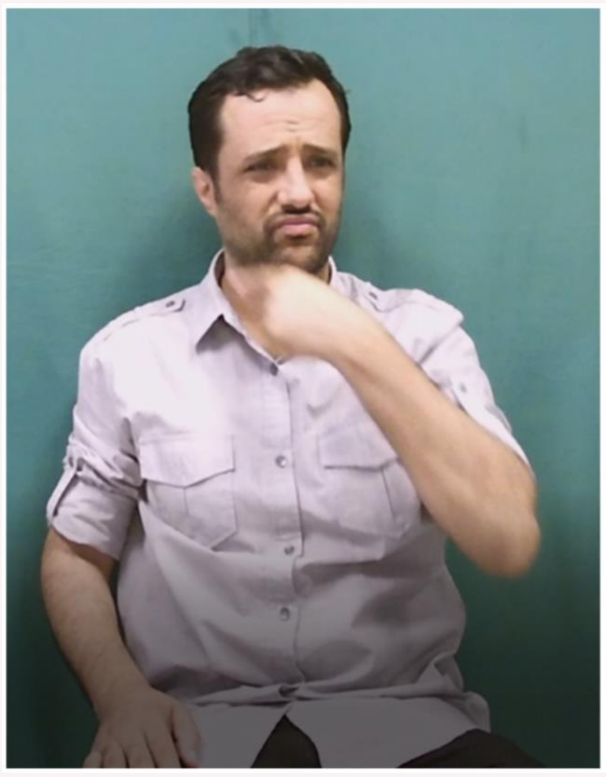
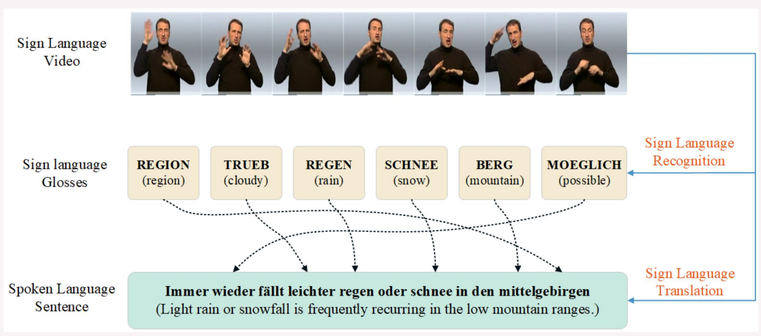


Fig. 1: ArSL sign for "UGLY" — note NMM (facial expression) integral to meaning.

Some work has cited performance gains in the presence of facial features, but this alone does not constitute *faithful* non-manual usage.

Methods: Controlled Case Study



Continuous Sign Language Recognition (CSLR) and Sign Language Translation (SLT) operate on the same video signal but produce different outputs. Both are evaluated here.

- **Model:** Pose-only Uni-Sign
- **Dataset:** Isharah-1000 (Saudi Sign Language)
- **Subtasks:** CSLR / SLT
- **Partitions:** Control / Signer-Independent (SI) / Unseen-Sentences (US)

Diagnostics:

- Task metrics (WER ↓ / BLEU ↑)
- Prediction similarity (exact match, previously seen, unigram overlap)
- Integrated Gradients region attribution

Integrated Gradients formula:

$$\mathrm{IG}_i(x) = (x_i - \tilde{x}_i) \times \int_0^1 \frac{\partial F\big(\tilde{x} + \alpha(x - \tilde{x})\big)}{\partial x_i}\, d\alpha$$

Case Study: Isharah-1000 Results

Overall Task Performance
Uni-Sign achieves competitive CSLR/SLT scores, but US/SI splits reveal substantially weaker generalization.

Method	Partition	CSLR WER ↓	SLT BLEU ↑
Uni-Sign	Control	17.78 %	81.56 %
	SI	49.56 %	52.19 %
	US	90.74 %	2.13 %
Swin-MSTP	SI	26.6 %	—
SMKD	US	48.0 %	—
GFSLT-VLP	SI	—	39.4 %
MMTLB*	SI	—	42.5 %

Table 1: CSLR/SLT performance across partitions.

Hidden Weaknesses
Across splits, predictions suggest reliance on peripheral cues, pipeline failures, and subtask convergence.

Feature Attribution (Integrated Gradients)
Hands > Body ≫ Face
Evidence indicates face is used for *signer identification* rather than compositional generation.

Joint	CSLR (%)	SLT (%)
8:right_elbow	10.38	7.50
6:right_shoulder	7.13	6.70
7:left_elbow	6.87	2.70
2:right_eye	3.07	1.18
1:left_eye	2.88	1.36
0:nose	2.71	1.38
5:left_shoulder	2.70	1.81
4:right_ear	2.45	1.57
3:left_ear	1.58	1.03
Total body share	**39.76**	**25.23**

(a) Body

Joint	CSLR (%)	SLT (%)
R:124:right_middle_finger4	2.51	3.13
R:129:right_pinky_finger1	1.89	3.00
R:116:right_thumb4	1.92	2.98
R:132:right_pinky_finger4	2.01	2.88
R:120:right_forefinger4	2.48	2.87
R:119:right_forefinger3	2.53	2.61
L:98:left_forefinger3	1.50	2.56
R:131:right_pinky_finger3	2.51	2.50
L:99:left_forefinger4	1.30	2.40
R:117:right_forefinger1	1.59	2.33
Total hands share	**54.77**	**71.94**

(b) Hands

Joint	Control		SI		US	
	CSLR (%)	SLT (%)	CSLR (%)	SLT (%)	CSLR (%)	SLT (%)
32:face-9	0.41	0.22	0.55	0.81	0.20	0.22
36:face-13	0.49	0.25	0.38	0.69	0.19	0.18
23:face-0	0.63	0.12	0.34	0.50	0.41	0.13
31:face-8	0.34	0.13	0.57	0.60	0.21	0.12
29:face-6	0.58	0.27	0.37	0.29	0.46	0.13
30:face-7	0.55	0.18	0.45	0.39	0.21	0.10
24:face-1	0.43	0.07	0.14	0.33	0.16	0.07
27:face-4	0.20	0.23	0.23	0.29	0.40	0.19
25:face-2	0.22	0.17	0.36	0.09	0.27	0.12
38:face-15	0.23	0.20	0.15	0.07	0.36	0.16
Total head share	**5.47**	**2.83**	**4.95**	**5.80**	**4.39**	**2.24**

(c) Head/Face

Table 2: Joint attribution share by region. Total head share ≈5%; hands dominate (>54% CSLR, >71% SLT).

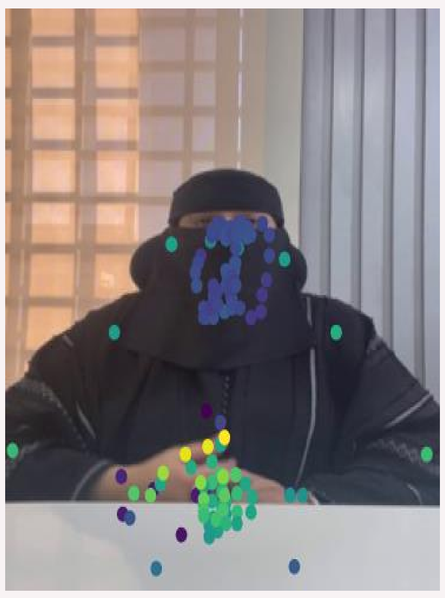


Fig. 3: Keypoint estimation failure on veiled signer — facial information rendered inaccessible.

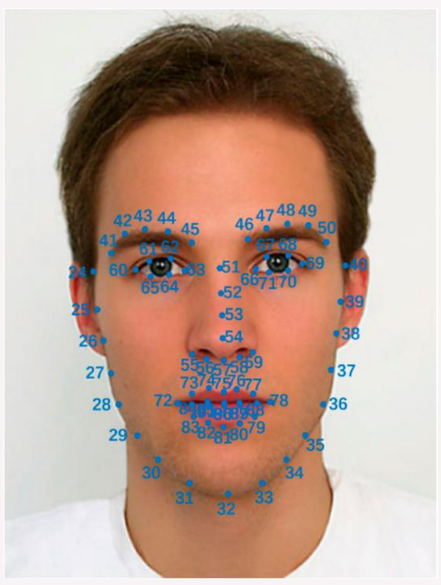


Fig. 4: Of 68 COCO-WholeBody keypoints, only 18 are used by the pose encoder — eyes and eyebrows excluded.

References

[1] Luqman et al. 2025. The SignEval 2025 Challenge at the ICCV Multimodal Sign Language Recognition Workshop: Results and discussion. In *Proceedings of the IEEE ICCV Workshops, 5027–5036.* [2] Altamimi & Alsager. 2023. Argument Structure and Word Order in Saudi Sign Language. *Journal of Language Teaching and Research,* 14(1):203–214. [3] Luqman & El-Sayed. 2021. Towards Hybrid Multimodal Manual and Non-Manual Arabic Sign Language Recognition: mArSL Database and Pilot Study. *Electronics*. 10, 1739.

Challenges in Non-Manual Learning

Dataset-level

Low NMM Diversity
The variation and diversity of non-manuals is not considered in sentence selection, despite being vital to translation.

Directed Sample Collection
Signers were prompted by imitating a video, which may introduce movement patterns different from naturally occurring signing.

Gloss–Sentence Mapping
High sentence-level similarity prevented the model from learning compositional semantics at all, including that contributed by NMMs.

Modality-level

Pose Estimation Error
The failure of the pose extraction model to even approximate the position of keypoints renders facial information largely unusable.

Incomplete Input Data
Out of 68 provided facial keypoints in COCO-WholeBody 2D, only 18 are used by the pose encoder, excluding features such as eyes and eyebrows.

Pose-only Limitations
Skeletal representations may fail to capture expressive nuance or temporal resolution of facial articulation.

Model-level

Sentence Memorization
The models were shown to have poor compositionality, preferring to memorize previously seen sentences rather than generate new ones.

Signer Identification
Facial features may have been used as peripheral signer identification cues rather than as sources of compositional information.

Representation Convergence
The subtasks exhibit highly similar memorization patterns to the point of having near identical exact-match accuracy.

Conclusions

- **Competitive metrics can mask compositional failure**. Both CSLR and SLT achieve strong benchmark performance, yet fail to exhibit linguistically correct behavior.
- **Non-manuals are not used compositionally**. All evidence points to facial signals being used minimally, at most as peripheral cues rather than semantic/syntactic sources of information.
- **Addressing these issues will require deliberate redesign across the SLU pipeline** to explicitly account for the linguistic role of non-manuals and to frame NMM learning as a distinct training objective in SLU systems.
- **Future work must develop benchmarks that directly test non-manual function** in the context of CSLR/SLT (clause-type marking, etc) rather than inferring NMM learning from accuracy gains alone.
- **These issue underscores the need to amplify voices of native signers in SLU research,** as they are ultimately the communities affected by this work. Without proper consideration for this issue, SLU research may optimize for benchmark performance while failing to serve the lived reality of native signers.

Acknowledgements

We thank the AraSLP team for providing access to the Isharah-1000 dataset in video form.

For a copy of the poster or paper, contact: eamalaia@ua.edu