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

Dmitriy Sazonov, Sevgi Z. Gurbuz, Evie Malaia
University of Alabama, North Carolina State University, University of Alabama
Tuscaloosa, AL, USA, Raleigh, NC, USA, Tuscaloosa, AL, USA
desazonov@crimson.ua.edu, szgurbuz@ncsu.edu, eamalaia@ua.edu

Abstract

Incorporation of non-manual information is one of the most challenging aspects of Sign Language Understanding (SLU), as these features contribute to the semantic, syntactic, and pragmatic structure of signed communication as a critical feature of compositional meaning at sign, phrase and sentence level. Despite their key linguistic role, non-manuals are often an afterthought in SLU model and dataset design, with many recent models still neglecting to implement non-manual analysis or evaluate how articulators beyond the hands are contributing to the model prediction. In this work, we identify and analyze the challenges relating to recognition of non-manuals and generalization of their linguistic roles encountered by SLU models, offering new explanations for failures to properly model non-manual behavior. We perform a case study on the subtasks of Continuous Sign Language Recognition and Sign Language Translation by applying the Uni-Sign model to Isharah-1000, a Saudi Sign Language dataset. Using controlled partitioning and feature attribution, we further analyze model behavior and failure cases. With this work we hope to set the stage for the creation of diagnostic frameworks for generalization of non-manuals.

Keywords: Sign Language Translation, Continuous Sign Language Recognition, Non-manual Markers, Feature Attribution, Saudi Sign Language

1. Introduction

Historically, the field of Sign Language Understanding (SLU) has progressed from lower level tasks toward more complex and holistic tasks as technological capability has advanced. Early work was limited to lower-level tasks such as static sign recognition, where the objective is image classification focused on manual articulators(Ong and Ranganath, 2005). While this was a necessary step in the development of SLU systems, the field has largely remained focused on manual articulators since. The goal of these systems should ultimately be to model the full range of sign language articulation, including lexical, grammatical, and prosodic markers articulated via face, head, and body movement (Borneman et al., 2018; Malaia et al., 2018; Krebs et al., in press). Yet, non-manual markers (NMMs) such as facial expression or body movement remain underutilized in comparison to manual articulators (Kim et al., 2024). NMMs play an important role in sentence-level meaning by marking clause types, prosodic boundaries, and semantic distinctions, as part of a distributed system optimizing information transfer across articulators (Wilbur, 2021; Krebs et al., 2025; Malaia et al., 2018; Malaia and Wilbur, 2020).

Broadly, SLU includes (but is not limited to) Sign Language Recognition (SLR), which focuses on recognizing lexical units from sign language sequences, and Sign Language Translation (SLT), which is concerned with generating textual translations. SLR is typically further divided into Isolated (ISLR) and Continuous (CSLR) forms, addressing individual signs and sign sequence recognition, respectively. SLT can be further subdivided into gloss-based and gloss-free prediction, depending on whether glosses are predicted in addition to text (Alyami et al., 2026). Recent advances in machine learning have enabled substantial progress in ISLR, shifting attention toward CSLR and SLT. These tasks more directly demand integration of non-manual cues, motivating the development of multi-channel SLU models (Li et al., 2025; Ko et al., 2019; Camgoz et al., 2021).

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. This is due in part to common evaluation practices that treat performance gains from adding facial features as evidence of faithful NMM use. This 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 remaining unable to process the full range of grammatical and lexical meaning in native signing, including clause- and sign-modifying non-manuals (Tanzer et al., 2024). For example, in SSL, yes/no questions are marked by raised eyebrows and a forward head tilt, with no manual question particle (Altamimi and Alsager, 2023); similarly, the non-manuals "intense face" (combination of brow furrow, eye-squint, and downturned mouth) function as a non-manual adverb of degree modifying

Proceedings of the 12th Workshop on the Representation and Processing of Sign Languages, pages 438–449
16 May 2026.

verbs, adjectives, and classifier predicates (Morris and Schneider, 2012) without a manual component carrying the same meaning. Thus, a model that ignores these features cannot reliably distinguish a declarative from a question, or a neutral predicate from an intensified one.

This paper diagnoses systematic challenges in handling NMMs that arise throughout the SLU pipeline. Through a Saudi Sign Language (SSL) case study spanning CSLR and SLT, we show how comparative analysis using diagnostic measures and feature attribution can reveal failures to reliably integrate non-manual cues. We intentionally avoid restricting analysis to idealized conditions. Instead, we evaluate a representative end-to-end SLU pipeline and treat errors and occlusions as informative observations for future non-manual modeling practices.

2. Related Work

2.1. Challenges in CSLR and SLT

In order to create SLU models that are viable in real-world conditions, models must train and generalize on samples representative of real world signing conditions. Dataset structure can often be manipulated for evaluation purposes. Signer-independent evaluation tests whether models generalize beyond signer identity (Sincan et al., 2021). Unseen sentence evaluation is also used to test compositionality (Alyami et al., 2026). This is particularly critical for SLT, where the composition of lexical, semantic, and prosodic information from multiple articulators is a complex task. Both of these issues are closely linked to NMM utilization, where facial features may make identity obfuscation more difficult, and compositional generalization relies on faithful use of NMMs (Krebs et al., 2025).

Sample collection procedures also affect outcomes in real-world scenarios. Directed, controlled sample collection has been shown to degrade model prediction quality compared to samples collected in a natural dialogue setting (Kurtoğlu et al., 2024; DeHaan et al., 2025). Similarly, discourse context has been shown to be critical for SLT (Tanzer et al., 2024), yet most available datasets remain restricted to sentence-level clips. This directly translates into a limitation for non-manual modeling, as NMMs play a crucial role in shaping sentence-level context. Other qualities, such as environment and appearance variability, have been shown to limit model robustness when applied in different settings, although this can mostly be abstracted away for pose-based models.

Significantly, different sensors and modalities have varying efficacy in perceiving linguistic parameters. For example, while video and pose is advantageous for extracting spatial information (e.g. hand shape and NMMs), radio frequency (RF) sensors extract micro-Doppler signatures showing the velocity versus time of the backscatter during signing, providing a unique representation for characterizing kinematics (Gurbuz et al., 2020; Malaia et al., 2024), while also enabling higher resolution in the depth dimension. Thus, the modality through which linguistic information is extracted is inherent to the model's representation of sign language.

2.2. Multi-Channel Sign Language Understanding

To process the coarticulation of manual and non-manual markers, multi-channel SLU architectures use separate feature extractors for the hands, head, and sometimes upper body. This approach has been applied across ISLR (Pu et al., 2016; Ko et al., 2019), CSLR (Camgoz et al., 2021; Jiao et al., 2023; Mukushev et al., 2020), and SLT (Camgoz et al., 2020; Gueuwou et al., 2025; Li et al., 2025). However, the intended role of non-manual markers differs by subtask, changing the implicit role for NMMs towards the objective, particularly when lacking direct annotation (Mukushev et al., 2020). While in ISLR they might exclusively be used for lexical cues such as mouthing, in CSLR they may additionally function as prosodic signals for boundary cues between glosses. SLT has the strongest theoretical dependence on NMMs, as translation requires the composition of semantic, syntactic, and pragmatic information. The breadth of these challenges suggests that partial feature usage isn't enough for linguistically faithful modeling, motivating closer analysis of non-manual utilization.

2.3. Non-Manual Modeling

Multiple studies report modest gains when adding NMMs to manual features ($\sim$1.5–13%) across sign languages (Aran et al., 2009; Luqman and El-Alfy, 2021; Brock et al., 2020; Mejía-Peréz et al., 2022; von Agris et al., 2008; Elons et al., 2014). However, some other studies report a drop in performance when using facial keypoints (Ko et al., 2019; Johnny et al., 2025). Ko et al. (2019) hypothesized that the performance drop was due to overparameterization caused by an excessive number of facial keypoints (70). However, Johnny et al. (2025) report a similar reduction despite using only 19 facial keypoints, suggesting that keypoint count alone may not be the sole cause of this degradation. Notably, these studies quantify the contribution of non-manuals primarily through overall task performance, rather than through metrics that faithfully reflect their linguistic function. As a result, it remains unclear whether the NMMs are truly being used for semantic and grammatical sentence-level information, as

opposed to cue-based memorization of samples.

2.4. Interpretability and Diagnostic Methods

Recent work has introduced interpretability tools, such as joint-wise feature importance (Holmes et al., 2024), temporal attention visualization (Dal Bianco et al., 2024; Zelezny et al., 2025), and saliency maps (Nam Pham and Avramidis, 2025). These methods provide insight into which features influence model predictions, allowing for further linguistic analysis of SLU model performance.

Despite this, the quantification of the functional contribution of non-manuals within SLU models has received limited attention. While Holmes et al. (2024) and Nam Pham and Avramidis (2025) examine contribution of non-manuals to performance, feature attribution alone does not prove faithful incorporation.

3. Dataset and Preprocessing

3.1. Dataset Overview

Isharah (Alyami et al., 2026) is a large multi-scene and multi-signer dataset for CSLR and SLT, divided into subsets Isharah-500, Isharah-1000, Isharah-2000 based on the number of unique sentences. The data consists of RGB videos of Saudi Sign Language (SSL) sentences with gloss annotation and transcription. The dataset was collected through video imitation of prompted sentences, so each target sentence is generally consistent across signers, with only slight variation in annotation when productions differ. As SSL remains underresourced in sign language technology, datasets such as Isharah provide an important setting for examining sentence-level SLU in a less-studied sign language. SSL, like other sign languages, uses non-manuals for grammatical marking (Morris and Schneider, 2012; Altamimi and Alsager, 2023), making their representation in Isharah a linguistic necessity. We use Isharah-1000, consisting of RGB videos of 18 signers performing up to 1,000 unique sentences for a total of 15,000 samples. With a sample count comparable to other popular CSLR and SLT datasets, such as Phoenix2014-T (Forster et al., 2014) (8,257 unique sentences across 8,257 samples) and CSL-Daily (Zhou et al., 2021) (6,598 unique sentences across 21,000 samples), the dataset provides a strong foundation for sentence-level CSLR and SLT experiments in an underexplored sign language. The dataset also features signer-independent (SI) and unseen sentence (US) partitions, allowing for controlled evaluation. The SI partition tests the model's accuracy on signers that were not seen in training, evaluating its ability to generalize across signing patterns. The US partition tests the model's ability to predict sentence combinations not seen in training, evaluating compositionality. This dataset was originally shared upon request by the authors, but has since been made publicly available.

A unique feature of Isharah is that several of the signers featured in the videos wear veils. The majority of these veils (and the only one from signers included in Isharah-1000) is the Niqab, which obscures all facial features aside from the eyes. This presents both a linguistic and logistical challenge for the model when predicting sign glosses from veiled signers, since many relevant non-manual markers (e.g. lip contour) are totally obscured by the veil, while other potentially informative ones (e.g., head motion) may not be effectively inferred.

3.1.1. NMMs in Saudi Sign Language

NMMs perform both grammatical (clause-level) and morphological (sign-level) functions in SSL, making non-manual modeling a necessity for any SLU system for this language. At the clause level, non-manual cues are the primary means of distinguishing sentence types: yes/no questions are marked only by raised eyebrows and a forward head tilt, with no manual question particle, while wh-questions rely on facial expression whose form varies with context (Morris and Schneider, 2012). Thus, NMMs determine clause type independently of word order: a model that ignores facial features would be unable to distinguish a question from a statement.

At the sign and phrase level, SSL uses several non-manual morphemes as adverbs of degree. The "intense face" (a co-occurring combination of brow furrow, eye squint, and downturned mouth) functions as an obligatory modifier that can scope over individual adjectives, verbs, and classifier predicates (Morris and Schneider, 2012).

Tongue movements constitute another class of non-manual morphemes in SSL: forward tongue protrusion functions both as an intensifier and as an obligatory component of certain manual signs (e.g., the sign for red); side tongue protrusion appears as an intensifier restricted to negative affect and taboo contexts, and tongue flap marks small size (Morris and Schneider, 2012). These morphemes are obligatory: their omission changes the meaning of a signed sentence. The obligatory, multi-level linguistic functions of NMMs in SSL mean that they cannot be treated as auxiliary markers.

3.2. Pose Keypoint Extraction

We chose to use RTMPose (Jiang et al., 2023), a recent pose estimation model, to process Isharah-1000 into skeletal data. RTMPose is used in the original Uni-Sign model (as detailed in Section 4.1), making it an optimal choice for fine-tuning model

Figure 1: Uni-Sign Region Segmentation

weights pretrained on major datasets such as CSL-News (Li et al., 2025). Out of the 133 keypoints extracted by RTMPose, 69 are used by Uni-Sign: 42 for hands, 9 for body, and 18 for the face (representing facial orientation, jawline, lip contour, and nose). These keypoints and their corresponding regions can be seen in Figure 1. One linguistically critical articulator for NMMs (Wilbur, 2021) not included in this analysis is the eyebrow/eye area, being excluded from Uni-Sign's selected features.

4. Experimental Setup

4.1. Model Architecture

We use the pose-only Uni-Sign (Li et al., 2025) framework for both CSLR and SLT (gloss-free), fine-tuning publicly available weights pretrained on CSL-News (Li et al., 2025). This model introduced a pretraining framework that unifies downstream SLU tasks by treating them as variants of SLT, achieving state-of-the-art performance across ISLR, CSLR, and SLT datasets alike. This makes the model ideal for analytical comparison between tasks, minimizing any gaps between pretraining and downstream tasks that may bias the result. By sharing the same pose encoder and mT5-base decoder architecture between tasks (Xue et al., 2021), direct probing can reveal comparative differences in learned representations. This pose-only model reached nearly equivalent performance to the standard model, with additional allowance for joint-wise interpretability across regions of the body.

4.2. Tasks and Metrics

Several metrics are used to compare the two tasks, each serving a distinct purpose. We use Word Error Rate (WER) to evaluate CSLR, and BiLingual Evaluation Understudy (BLEU) to evaluate SLT, both of which are standard practice for the respective tasks.

Partition	Train	Test
Control	Random	Random
SI	—	Signers not in Train
US	—	Sentences not in Train

Table 1: Training-testing data partitions (US: Unseen Sentence, SI: Signer-Independent)

Additionally, we also evaluate exact-match accuracy, defined as the proportion of predictions that exactly match the ground truth. This was done to account for possible score inflation due to sentence memorization. We use the proportion of predicted samples within the sentences seen during training as an indirect measure for the model's ability to generalize. The difference between previously seen rate and exact-match accuracy indicates whether models learn surface-level pattern matching versus compositional semantics: a critical distinction for evaluating NMM learning. Finally, we use unigram overlap between the prediction and ground truth as a measure of sentence-level content overlap.

4.3. Dataset Partitions

We evaluate Uni-Sign in three data partitions (Table 1): Control, a standard random split; Signer-Independent, which evaluates generalization to unseen signers, distinguishing reliance on signer-specific versus linguistic cues; and Unseen Sentences, which evaluates on unseen sentences to test compositionality.

4.4. Integrated Gradients

To analyze how different body regions influenced model predictions, we used the Integrated Gradients (IG) method (Sundararajan et al., 2017) for feature attribution between the major keypoint regions used by Uni-Sign: hands, face, and body (including arms and head orientation). IG quantifies the contribution of each input feature by integrating the gradient of the model's output along a straight path from a neutral baseline pose to the actual input, providing a principled measure of how each region influences the prediction. The continuous form of Integrated Gradients is defined as:

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

where F denotes the model's output, x is the input sequence (skeleton keypoints), and $\tilde{x}$ is the baseline (zero-pose). The gradient $\frac{\partial F}{\partial x_i}$ measures the sensitivity of the model output to the i-th input feature, which is later aggregated over keypoints belonging to the same body region.

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 2: CSLR/SLT performance across test set of various partitions (* indicates gloss-based SLT model)

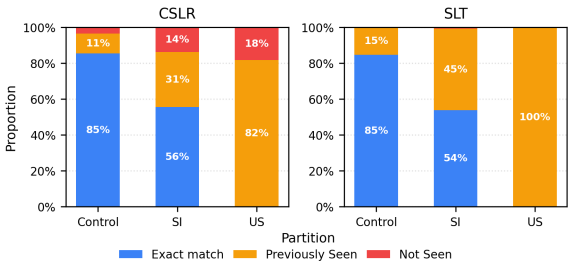

Figure 2: Proportion of predictions that are an exact match to the reference or a previously seen sentence, by partition for CSLR and SLT

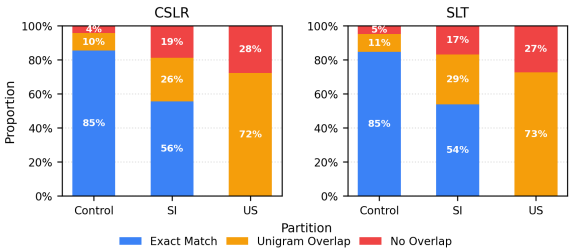

Figure 3: Proportion of predictions that are an exact match to the reference or have unigram overlap, by partition for CSLR and SLT

5. Results and Analysis

5.1. Overall Task Performance

Table 2 shows the performance of Uni-Sign across partitions for both CSLR and SLT, alongside the previous highest performing model benchmarks on Isharah-1000 (note that better performance is indicated by lower WER and higher BLEU). Across subtasks, Uni-Sign achieves similar results for each partition, with a moderate drop in performance for the signer-independent (SI) partition, and significant degradation for unseen sentences (US).

Comparison with previous benchmarks, such as Swin-MSTP (Alyami and Luqman, 2025), SMKD (Hao et al., 2021), GFSLT-VLP (Zhou et al., 2023), and MMTLB (Chen et al., 2022), shows that for CSLR Uni-Sign underperforms in both signer-independent and unseen sentence partitions. However, in SLT Uni-Sign performs higher than previous model benchmarks. Uni-Sign reaches 55.69 BLEU, significantly higher than 39.4 BLEU for GFSLT-VLP, or even 42.5 BLEU for MMTLB, a model that uses both gloss and text representations in training. In fact, the results for SLT would be regarded as highly competitive.

5.2. Sample Prediction Similarity

Figure 2 depicts the proportion of predictions across partitions that are an exact-match to the true sentence, along with the proportion of predictions that are an exact-match to any sentence seen in training. The figure shows much less incorrect non-exact-match sentences than expected, with CSLR having a small amount of generated sentences unseen previously, and SLT generating almost entirely previously seen sentences. For unseen sentences in particular, a large proportion of generated sentences were previously seen, yet none of the predicted sentences were exact matches. It's important to note that in both cases, predictions can often not match previously seen sentences but still be near accurate.

Figure 3 illustrates the presence of any non-zero unigram overlap between prediction and reference gloss annotation / sentences, representing shared sentence content. These values are almost identical between CSLR and SLT, as is exact-match accuracy.

5.3. Feature-Level Attribution

Tables 3, 4, and 5 reveal the most salient joints across each region contributing to prediction using the Integrated Gradients method. The figure reveals a pattern that holds across all partitions: contribution from the hands is the highest, closely followed by the body, while facial contribution is lowest.

In particular, Table 5 shows how the exact value of facial contribution differs between subtasks and partitions. Across partitions for SLT, the signer-independent partition achieves a significantly higher attribution value compared to unseen sentences and control. This reveals a connection between evaluation of signers not seen in training and facial feature contribution.

5.4. Observed Tracking Failures

As shown in Figure 4, face keypoint estimation often fails catastrophically for veiled signers. In some cases, the predicted facial keypoints are not merely imprecise, but spatially incoherent, failing to align

Figure 4: Keypoint Estimation Failure on Veiled Signer

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**

Table 3: Top joints in Body by global attribution share (control partition).

even approximately with the visible eye region or a plausible facial configuration. By contrast, the hands are still being inferred with reasonable accuracy, despite the partial obstruction by surrounding objects.

6. Discussion

We identify three key sources of systemic error in non-manual modeling: model-level, dataset-level, and modality-level.

6.1. Model-Level Error

6.1.1. Sentence Memorization

The results reveal several otherwise hidden failures in the model that couldn't be seen solely from the control partition's WER and BLEU metrics. Across the Isharah benchmarks, Uni-Sign provided reasonable performance for CSLR and state-of-the-art results for SLT. Despite this, the models were shown to have poor compositionality, instead preferring to memorize previously seen sentences, with SLT in particular making virtually no attempts to compose

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**

Table 4: Top joints in Hands by global attribution share (control partition).

new, unseen sentences in evaluation.

6.1.2. Signer Identification

Feature-level attribution shows that the hands and body are most responsible for the prediction. The face share is small, yet consistent, indicating that face features are likely still being used as identification cues for the sake of memorization. Analyzing the region prediction contribution across partitions shows that face share is significantly higher for signer-independent in SLT, the partition where train/test is split based on signer identity. This suggests that facial features are used as peripheral signer identification cues rather than as sources of compositional information.

6.1.3. Subtask Convergence

Though unified architectures can provide efficient and effective performance across downstream tasks, our results suggest that they may exhibit convergence of subtask behavior, limiting task-specific reasoning. In the case of CSLR and SLT, despite different reference text (gloss vs. sentence), the tasks exhibit highly similar memorization patterns to the point of having near identical exact-match accuracy. In addition, SLT does not demonstrate increased reliance on non-manual features compared to CSLR as expected from linguistic theory.

One possible explanation for this is that gloss and sentence targets occupy closely aligned regions of the decoder's embedding space, particularly when gloss-sentence mappings are nearly 1:1. This offers limited incentive to differentiate between differing linguistic demands of non-manuals, resulting in similar behavior. As a result, the model does not develop a functional distinction between recognition and translation.

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**

Table 5: Top facial joints by global attribution share across partitions.

6.2. Dataset-Level Error

6.2.1. Gloss-Sentence Mapping

The dataset's high full-sentence similarity (1000 sentences with 1:1 gloss-sentence mapping) likely prevented the model from learning compositional semantics at all, including that contributed by NMMs. This highlights a critical limitation in dataset design for SLT/CSLR: models require high training data variability at the phrase and sentence level to learn compositional semantics.

This redundancy presents both benefits and limitations: it broadens coverage of articulatory variation within individual sentences, which was seen by the relative success of the signer-independent partition, yet also encourages memorization of a narrow set of reference sentences, undermining the model's ability to compose new sentences.

6.2.2. Low Non-Manual Diversity

The structure of this dataset also limits opportunities for non-manual learning by reducing instances in which syntactic or semantic disambiguation is required. Since each gloss sequence maps to a single reference translation, features vital to interpretation, such as non-manuals, may be underutilized. Although lexical content is recognized and utilized in this process, other linguistically critical features are not.

6.2.3. Directed Sample Collection

A broader issue is that of SLU data collection, where directed wording has been shown to produce movement patterns different from naturally occurring signing. In contrast, natural signing contexts allow greater linguistic variation, providing the depth and diversity necessary to capture non-manuals and other features of sign language communication. In this dataset, signers were prompted to sign by imitating a video, which may introduce such issues.

6.3. Modality-Level Error

6.3.1. Pose Estimation Error

As seen in Figure 4, the model encounters practical pose estimation errors, where facial content is obscured and facial keypoints are inaccurately estimated. In this context, aside from head orientation, minimal facial expression information can be inferred. The failure of the pose extraction model to even approximate the position of keypoints renders facial information largely unusable, further diminishing opportunities for non-manual learning. These cases require robustness from both pose estimation and SLU models to missing or degraded facial features, while still enabling the extraction of rich non-manual information when it is present. Challenges such as these are crucial to address as sign language understanding advances toward greater accessibility for all signers.

6.3.2. Incomplete Input Pose Data

Out of 68 available facial keypoints in COCO-WholeBody 2D, only 18 are used in Uni-Sign, meaning that eye and eyebrow movements are excluded, despite being known articulators of non-manuals (Wilbur, 2021). Such reductions are often implemented to decrease input size and improve training efficiency, but they may limit long-term and holistic learning of non-manual features. However, the omission does not preclude broader generalizability of the analysis of NMMs, although it limits its linguistic interpretability.

6.3.3. Pose-Only Data Limitations

Even when all available facial keypoints are included, facial movement remains highly complex. Subtle variations in facial morphology and prediction noise can cause keypoint trajectories to deviate from the true visual signal. As a result, skeletal

representations may fail to capture the full expressive nuance of facial articulation, highlighting the inherent limitations of the pose modality alone for non-manual analysis.

One might ask if evaluating veiled and non-veiled signers separately may have better isolated the effect of facial occlusion. However, framing occlusion as a condition to be controlled underestimates what a linguistically adequate CSLR or SLT model should be able to do. One meaningful comparison, or benchmarking bar, may be with a fluent human signer processing a sign language under similar conditions (e.g. when the signing partner is veiled). Research on neural bases of sign language comprehension shows that Deaf signers rely on predictive processing: rather than reconstructing meaning sign-by-sign from bottom-up visual input, fluent signers maintain hierarchical internal models of linguistic structure (unfolding syntactic structure, prior information about participants/arguments and likely event structure) that allow them to infer partially observable content from prior information contained in the visual signal (Malaia et al., 2021, 2023; Borneman et al., 2025). A signer who acquired SSL to native proficiency would have no difficulty communicating with veiled signers, because obligatory grammatical markers — including those carried by head movement — are recoverable by humans from the broader motion dynamics through predictive inference. It is, then, notable, that the current model appears to have an analogous capacity for the hands: as shown in Figure 4, even when a signer's hand passes behind a table, RTMPose infers plausible joint positions from the prior context. The limitation, then, is not that an occlusion occurs, but that no equivalent predictive capacity is acquired by the model for facial articulators. The more productive model benchmarking that we argue for should target human-comprehension-level performance, and evaluate if trained models can recover linguistically meaningful NMM information under naturalistic conditions including partial occlusion.

7. Conclusion

In this paper, we argued that modern SLU research remains systemically misaligned with the linguistic role of non-manual markers. Using Uni-Sign (Li et al., 2025) on SSL as a controlled comparison between CSLR and SLT, we evaluated three partitions (control, signer-independent, unseen sentences) and applied prediction-similarity analysis and integrated gradients to probe feature use.

We found that despite the fact that both models reached competitive prediction accuracy, they encountered several challenges in proper non-manual representation. Although non-manual markers were shown to contribute to the prediction, analysis suggested this was in the form of peripheral identification cues rather than semantic and syntactic compositional contributions. Additionally, tracking failures in the model pipeline highlighted broader challenges for reliable integration of non-manual markers.

This work highlights critical pitfalls in current approaches to non-manual modeling and compositional learning. 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.

7.1. Future Work

Prior work has primarily evaluated the addition of non-manuals in terms of performance improvements, but this neglects to examine whether non-manual features are incorporated in a manner consistent with their linguistic function. Instead, benchmarks that explicitly evaluate non-manual learning across SLU subtasks should be developed. This should assess whether models meaningfully extract semantic and syntactic information from non-manual features. Design choices for such benchmarks must be deliberate to avoid the structural errors identified in this paper. This may be particularly challenging as many of these errors have opposing constraints. For example, the pose-based models are more effective at reducing reliance on identity cues compared to RGB video, but may sacrifice fine details of facial expression. Addressing these trade-offs will require broad and extensive experimentation.

In addition, this work has shown that there is no singular unimodal solution in sign language understanding. Multi-modal sensing has the potential to overcome the limitations of a single sensor, as shown by the 2025 ICCV Multi-modal Sign Language Recognition (MSLR) Workshop, which featured the first Italian Sign Language (LIS) dataset with both video and radar data (Luqman et al., 2025). While the work done at this workshop successfully achieved state-of-the-art performance (Islam et al., 2025; Juranek, 2025; Manjur et al., 2025; Sazonov et al., 2025), current approaches espouse a purely computational perspective to feature extraction that does not as of yet incorporate any linguistic perspectives in the data analysis, such as selecting the optimal sensing data for estimating task-cognizant linguistic parameters (Malaia et al., 2024). Consequently, multi-modal sensing for SLU remains an open area of research, with great potential to enrich the robustness and accuracy of SLU in real-world conditions.

8. Acknowledgements

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

9. Bibliographical References

Haya Altamimi and Haroon Alsager. 2023. Argument structure and word order in saudi sign language. *Journal of Language Teaching and Research*, 14:203–214.

Sarah Alyami and Hamzah Luqman. 2025. Swin-MSTP: Swin transformer with multi-scale temporal perception for continuous sign language recognition. *Neurocomputing*, 617:129015.

Sarah Alyami, Hamzah Luqman, Sadam Al-Azani, Maad Alowaifeer, Yazeed Alharbi, and Yaser Alonaizan. 2026. Isharah: A Large-Scale Multi-Scene Dataset for Continuous Sign Language Recognition. *IEEE Transactions on Multimedia*.

Oya Aran, Thomas Burger, Alice Caplier, and Lale Akarun. 2009. A belief-based sequential fusion approach for fusing manual signs and non-manual signals. *Pattern Recognition*, 42(5):812–822.

Joshua Borneman, Evguenia Malaia, and Ronnie Wilbur. 2018. Motion characterization using optical flow and fractal complexity. *Journal of Electronic Imaging*, 27:1.

Sean C. Borneman, Julia Krebs, Ronnie B. Wilbur, and Evie A. Malaia. 2025. Decoding predictive inference in visual language processing via spatiotemporal neural coherence. NeurIPS 2025 Workshop on Foundation Models for the Brain and Body.

Heike Brock, Iva Farag, and Kazuhiro Nakadai. 2020. Recognition of Non-Manual Content in Continuous Japanese Sign Language. *Sensors*, 20(19):5621.

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)*, pages 10020–10030.

Necati Cihan Camgoz, Oscar Koller, Simon Hadfield, and Richard Bowden. 2021. Multi-channel transformers for multi-articulatory sign language translation. In *Computer Vision – ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part IV*, page 301–319, Berlin, Heidelberg. Springer-Verlag.

Yutong Chen, Fangyun Wei, Xiao Sun, Zhirong Wu, and Stephen Lin. 2022. A simple multi-modality transfer learning baseline for sign language translation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5110–5120.

Pedro Alejandro Dal Bianco, Oscar Agustín Stanchi, Facundo Manuel Quiroga, Franco Ronchetti, and Enzo Ferrante. 2024. SignAttention: On the Interpretability of Transformer Models for Sign Language Translation.

Kenneth DeHaan, Emre Kurtoğlu, Sabyasachi Biswas, Caroline Kobek Pezzarossi, Darrin J. Griffin, Chris Crawford, Ali C. Gurbuz, Evie A. Malaia, Abraham Glasser, Raja Kushalnagar, and Sevgi. Z Gurbuz. 2025. Rf -chesssign: Radar-enabled human-computer interaction in a real- time sign language-controlled game. In *2025 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 5000–5010.

A.S. Elons, Menna Ahmed, and Hwaidaa Shedid. 2014. Facial expressions recognition for Arabic Sign Language translation. In *2014 9th International Conference on Computer Engineering & Systems (ICCES)*, pages 330–335. IEEE.

Jens Forster, Christoph Schmidt, Oscar Koller, Martin Bellgardt, and Hermann Ney. 2014. Extensions of the sign language recognition and translation corpus RWTH-PHOENIX-weather. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1911–1916, Reykjavik, Iceland. European Language Resources Association (ELRA).

Shester Gueuwou, Xiaodan Du, Greg Shakhnarovich, and Karen Livescu. 2025. SignMusketeers: An efficient multi-stream approach for sign language translation at scale. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 22506–22521, Vienna, Austria. Association for Computational Linguistics.

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, Trevor Macks, and Engin Ozcelik. 2020. A linguistic perspective on radar micro-doppler analysis of american sign language. In *2020 IEEE International Radar Conference (RADAR)*, pages 232–237.

Aiming Hao, Yuecong Min, and Xilin Chen. 2021. Self-Mutual Distillation Learning for Continuous Sign Language Recognition. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11283–11292.

Ruth M. Holmes, Ellen Rushe, and Anthony Ventresque. 2024. The key points: Using feature importance to identify shortcomings in sign language recognition models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15970–15975, Torino, Italia. ELRA and ICCL.

Md. Milon Islam, Md Rezwanul Haque, S M Taslim Uddin Raju, and Fakhri Karray. 2025. Fusionensemblenet: An attention-based ensemble of spatiotemporal networks for multimodal sign language recognition. In *2025 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 4983–4989.

Tao Jiang, Peng Lu, Li Zhang, Ningsheng Ma, Rui Han, Chengqi Lyu, Yining Li, and Kai Chen. 2023. Rtmpose: Real-time multi-person pose estimation based on mmpose.

Peiqi Jiao, Yuecong Min, Yanan Li, Xiaotao Wang, Lei Lei, and Xilin Chen. 2023. Cosign: Exploring co-occurrence signals in skeleton-based continuous sign language recognition. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 20619–20629.

Samuel Ebimobowei Johnny, Blessed Guda, Andrew Blayama Stephen, and Assane Gueye. 2025. AutoSign: Direct Pose-to-Text Translation for Continuous Sign Language Recognition . In *2025 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 5071–5078, Los Alamitos, CA, USA. IEEE Computer Society.

Jakub F. Juranek. 2025. Multimodal italian sign language recognition with radar-video late fusion on the multimedalis dataset. In *2025 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 5079–5085.

Jung-Ho Kim, Changyong Ko, Mathew Huerta-Enochian, and Seung Yong Ko. 2024. Shedding Light on the Underexplored: Tackling the Minor Sign Language Research Topics. In *Proceedings of the LREC-COLING 2024 11th Workshop on the Representation and Processing of Sign Languages: Evaluation of Sign Language Resources*, pages 147–158, Torino, Italia. ELRA and ICCL.

Sang-Ki Ko, Changjo Kim, Hyedong Jung, and Choongsang Cho. 2019. Neural sign language translation based on human keypoint estimation. *Applied Sciences*, 9:2683.

Julia Krebs, Dietmar Roehm, Ronnie B. Wilbur, and Evie A. Malaia. in press. Visual-Linguistic Cues in Manual Communication: Neural and Behavioral Responses to Mouthings in Sign-Naïve Perception. *Brain and Language*. Special Issue: Gesture, Cognition and Language.

Julia Krebs, Ronnie B Wilbur, Dietmar Roehm, and Evie A Malaia. 2025. The interaction of syntax, non-manuals, and prosodic cues as potential topic markers in Austrian Sign Language. *Sign Language & Linguistics*, 28(1):1–48.

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

Zecheng Li, Wengang Zhou, Weichao Zhao, Kepeng Wu, Hezhen Hu, and Houqiang Li. 2025. Uni-sign: Toward unified sign language understanding at scale. In *The Thirteenth International Conference on Learning Representations*.

Hamzah Luqman and El-Sayed M. El-Alfy. 2021. Towards Hybrid Multimodal Manual and Non-Manual Arabic Sign Language Recognition: marsl Database and Pilot Study. *Electronics*, 10(14):1739.

Hamzah Luqman, Raffaele Mineo, Murtadha Aljubran, Ahmed Abul Hasanaath, Amelia Sorrenti, Sarah Alyami, Sadam Al-Azani, Maad Alowaifeer, JiHwan Moon, Vaclav Javorek, Tomas Zelezny, Marek Hruz, Gaia Caligiore, Silvio Giancola, Senya Polikovsky, Motaz Alfarraj, Sabina Fontana, Mufti Mahmud, Muhammad Haris Khan, Kamrul Islam, Sevgi Gurbuz, Egidio Ragonese, Giovanni Bellitto, Federica Proietto Salanitri, Concetto Spampinato, and Simone Palazzo. 2025. The SignEval 2025 Challenge at the ICCV Multimodal Sign Language Recognition Workshop: Results and Discussion . In *2025 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 5086–5095, Los Alamitos, CA, USA. IEEE Computer Society.

Evie Malaia, Joshua D Borneman, and Ronnie B Wilbur. 2018. Information transfer capacity of articulators in American Sign Language. *Language and speech*, 61(1):97–112.

Evie A. Malaia, Joshua Borneman, and Sevgi Gurbuz. 2024. 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 384–389, Torino, Italy. ELRA Language Resources Association (ELRA) and the International Committee on Computational Linguistics (ICCL).

Evie A Malaia, Sean C Borneman, Joshua D Borneman, Julia Krebs, and Ronnie B Wilbur. 2023. Prediction underlying comprehension of human motion: an analysis of Deaf signer and non-signer EEG in response to visual stimuli. *Frontiers in Neuroscience*, 17:1218510.

Evie A Malaia, Sean C Borneman, Julia Krebs, and Ronnie B Wilbur. 2021. Low-frequency entrainment to visual motion underlies sign language comprehension. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 29:2456–2463.

Evie A Malaia and Ronnie B Wilbur. 2020. Syllable as a unit of information transfer in linguistic communication: The entropy syllable parsing model. *Wiley Interdisciplinary Reviews: Cognitive Science*, 11(1):e1518.

Sultan Mohammad Manjur, Sabyasachi Biswas, and Ali C. Gurbuz. 2025. A multimodal video and radar fusion framework for high-accuracy isolated sign language recognition. In *2025 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 5061–5070.

Kenneth Mejía-Peréz, Diana-Margarita Córdova-Esparza, Juan Terven, Ana-Marcela Herrera-Navarro, Teresa García-Ramírez, and Alfonso Ramírez-Pedraza. 2022. Automatic Recognition of Mexican Sign Language Using a Depth Camera and Recurrent Neural Networks. *Applied Sciences*, 12(11):5523.

Carla Morris and Erin Schneider. 2012. On selected morphemes in saudi arabian sign language. *Sign Language Studies*, 13:103–121.

Medet Mukushev, Arman Sabyrov, Alfarabi Imashev, Kenessary Koishybay, Vadim Kimmelman, and Anara Sandygulova. 2020. Evaluation of manual and non-manual components for sign language recognition. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6073–6078, Marseille, France. European Language Resources Association.

Dinh Nam Pham and Eleftherios Avramidis. 2025. The importance of facial features in vision-based sign language recognition: Eyes, mouth or full face? In *Adjunct Proceedings of the 25th ACM International Conference on Intelligent Virtual Agents*, IVA Adjunct '25, New York, NY, USA. Association for Computing Machinery.

Sylvie Ong and Surendra Ranganath. 2005. Automatic sign language analysis: A survey and the future beyond lexical meaning. *IEEE transactions on pattern analysis and machine intelligence*, 27:873–91.

Junfu Pu, Wengang Zhou, and Houqiang Li. 2016. Sign language recognition with multi-modal features. In *Advances in Multimedia Information Processing - PCM 2016*, pages 252–261, Cham. Springer International Publishing.

Dmitriy Sazonov, Kamrul Islam, Evie Malaia, and Sevgi Z. Gurbuz. 2025. Modality-specific benchmarks and radar range-doppler envelope classification for multimodal isolated sign language recognition. In *2025 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 5046–5053.

Ozge Mercanoglu Sincan, Julio C. S. Jacques Junior, Sergio Escalera, and Hacer Yalim Keles. 2021. ChaLearn LAP Large Scale Signer Independent Isolated Sign Language Recognition Challenge: Design, Results and Future Research.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3319–3328. PMLR.

Garrett Tanzer, Maximus Shengelia, Ken Harrenstien, and David Uthus. 2024. Reconsidering sentence-level sign language translation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6262–6287, Miami, Florida, USA. Association for Computational Linguistics.

Ulrich von Agris, Moritz Knorr, and Karl-Friedrich Kraiss. 2008. The significance of facial features for automatic sign language recognition. In *2008 8th IEEE International Conference on Automatic Face & Gesture Recognition*, pages 1–6. IEEE.

Ronnie B Wilbur. 2021. Non-manual markers: Theoretical and experimental perspectives. In *The Routledge handbook of theoretical and experimental sign language research*, pages 530–565. Routledge.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference*

of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 483–498, Online. Association for Computational Linguistics.

Tomas Zelezny, Jakub Straka, Vaclav Javorek, Ondrej Valach, Marek Hruz, and Ivan Gruber. 2025. Exploring pose-based sign language translation: Ablation studies and attention insights.

Benjia Zhou, Zhigang Chen, Albert Clapés, Jun Wan, Yanyan Liang, Sergio Escalera, Zhen Lei, and Du Zhang. 2023. Gloss-free sign language translation: Improving from visual-language pretraining. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 20814–20824.

Hao Zhou, Wengang Zhou, Weizhen Qi, Junfu Pu, and Houqiang Li. 2021. Improving sign language translation with monolingual data by sign back-translation. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1316–1325.

10. Language Resource References

Sarah Alyami, Hamzah Luqman, Sadam Al-Azani, Maad Alowaifeer, Yazeed Alharbi, and Yaser Alonaizan. 2026. *Isharah: A Large-Scale Multi-Scene Dataset for Continuous Sign Language Recognition*. PID https://doi.org/10.1109/TMM.2026.3664959.