Beyond BLEU: Linguistic Invisibility and Interactional Repair Sequence in End-to-End Sign Language Translation

Zirui Wang, Mayumi Bono

SOKENDAI (The Graduate University for Advanced Studies), National Institute of Informatics
Shonan Village, Hayama, Kanagawa 240-0193 JAPAN, 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430 JAPAN
{wzr, bono}@nii.ac.jp

Abstract

Recent advances in end-to-end sign language translation (SLT) have achieved benchmark performance, yet little is known about whether these systems preserve the multi-channel linguistic structures that are essential for real-world communication. We argue that current optimization and evaluation practices create a form of linguistic invisibility, where interactionally decisive non-manual signals (NMS) are systematically underrepresented despite high translation scores.To empirically examine this issue, we analyze an interactional repair sequence from a Japanese Sign Language (JSL) conversational corpus as a diagnostic probe. Combining qualitative interactional analysis with kinematic measurements, we demonstrate a consistent manual–mouth decoupling pattern in which semantic resolution is carried primarily by mouthing while manual articulation remains largely constant. We show that such cross-channel contrast is unlikely to be preserved under current end-to-end training objectives that prioritize global motion similarity. Based on these findings, we argue that progress in SLT should be evaluated not only by sequence-level accuracy but also by the preservation of linguistically contrastive structures, motivating the development of diagnostic, multi-channel evaluation protocols for future SLT benchmarks. We therefore propose incorporating multi-channel diagnostic evaluation sets and decoupling-sensitive metrics into future SLT benchmarking frameworks, providing a pathway toward models that achieve both high performance and linguistic structural visibility.

Keywords: interactional repair sequence, non-manual signals, end-to-end model

1. Introduction

In the current era of Sign Language Processing (SLP), the dominance of end-to-end (E2E) neural architectures has led to a paradoxical state of "successful blindness." Driven by deep visual encoders and temporal alignment objectives, state-of-the-art models consistently achieve record-breaking BLEU scores on standardized benchmarks (Zhou et al., 2023; Wong et al., 2024; Hwang et al., 2025). However, we argue that this metric-driven progress masks a systematic "linguistic invisibility" where the fundamental multi-channel and contrastive structures of sign language are optimized away as noise to satisfy global sequence mapping.

The transition from modular, linguistics-heavy pipelines to data-driven E2E mapping has fundamentally redefined sign language within AI research (Camgoz et al., 2018). By treating signing as a continuous visual signal to be mapped onto linear text, the field has largely bypassed the "bottleneck" of linguistic annotation (e.g., glosses). While computationally efficient, this paradigm risks reducing a complex natural language to mere "gestures with labels," ignoring the simultaneous grammar such as non-manual signals (NMS) and spatial organization that distinguishes natural signing from pantomime (Bragg et al., 2019; De Meulder, 2021).

The core of this invisibility lies in the optimization target. Current models are trained to maximize surface-level similarity (BLEU/WER), which often rewards models for guessing context while missing the specific contrastive units that resolve ambiguity. This creates a dangerous gap: a model may achieve a high BLEU score by matching generic vocabulary, yet remain incapable of detecting the subtle interactional repair sequences such as a shift in mouthing or eye gaze that are decisive for real-world communication (Bono, 2017). Consequently, a "successful" translation on paper may fail entirely in a functional, human-centric context.

This paper challenges the E2E consensus by bridging a critical survey with micro-level kinematic evidence. We first analyze recurring design patterns in contemporary SLT pipelines that reinforce structural invisibility. We then present a pivot case study from a Japanese Sign Language (JSL) corpus, utilizing kinematic data to quantify how a signer resolves communicative "trouble" through manual-mouth decoupling. Finally, we provide empirical evidence that current E2E models tend to overlook these decisive pulses, and we outline a methodological roadmap toward maintaining linguistic structural visibility in future sign language resources.

Our contribution is threefold: (i) a diagnostic interactional case study revealing cross-channel repair mechanisms, (ii) quantitative kinematic evidence demonstrating manual–mouth decoupling that current E2E models are likely to miss, and (iii) a methodological proposal for evaluation frameworks that incorporate linguistic structural visibility.

Proceedings of the 12th Workshop on the Representation and Processing of Sign Languages, pages 491–500

2. Scope and Method

We adopt a critical survey approach to analyze methodological trends in end-to-end (E2E) sign language processing, moving beyond performance comparison to examine recurring design patterns and their implicit linguistic assumptions, following the methodological perspective advocated by Yin et al. (2021). Our methodology integrates a longitudinal review of literature with a focused kinematic case study to provide empirical grounding for theoretical critiques.

2.1 Criteria for Inclusion

Our survey primarily covers Sign Language Translation (SLT), where E2E methods have progressed most rapidly and where "representational weakening" is especially visible. We concentrate on recent representative works, prioritizing models that have defined state-of-the-art benchmarks on datasets such as PHOENIX-2014T and How2Sign (Hwang et al.,2025; Forster et al., 2012; Duarte et al., 2021). We also include seminal papers that established the current evaluation protocols (e.g., BLEU, WER) to trace the origins of current metric-driven biases (Papineni et al., 2002; Snover et al., 2006).

2.2 Analytical Framework: A Dual-Track Approach

Rather than an exhaustive catalog, we summarize widely reused pipelines and extract recurring modeling choices specifically in feature fusion, temporal alignment (e.g., CTC-style objectives; Graves et al., 2006), and loss function design. We evaluate these choices against a "Visibility Criterion": Does the model architecture allow for the explicit detection and separation of simultaneous linguistic channels? Recent JSL-focused work has shown that mouth actions form linguistically motivated and structured categories rather than incidental articulatory noise (Regmi, 2025). A recent study explicitly classifies mouth actions into mouthing, mouth gesture, other mouth movement, and no mouth movement, and demonstrates that these categories can be reliably modeled even with limited data. The linguistic literature summarized in that work further reports that a large proportion of signs in languages such as Auslan, BSL, and JSL co-occur with mouth activity, and that such activity plays a crucial role in lexical disambiguation and grammatical marking. This supports our assumption that mouth actions constitute a contrastive non-manual channel whose suppression risks erasing linguistically relevant structure in SLT.

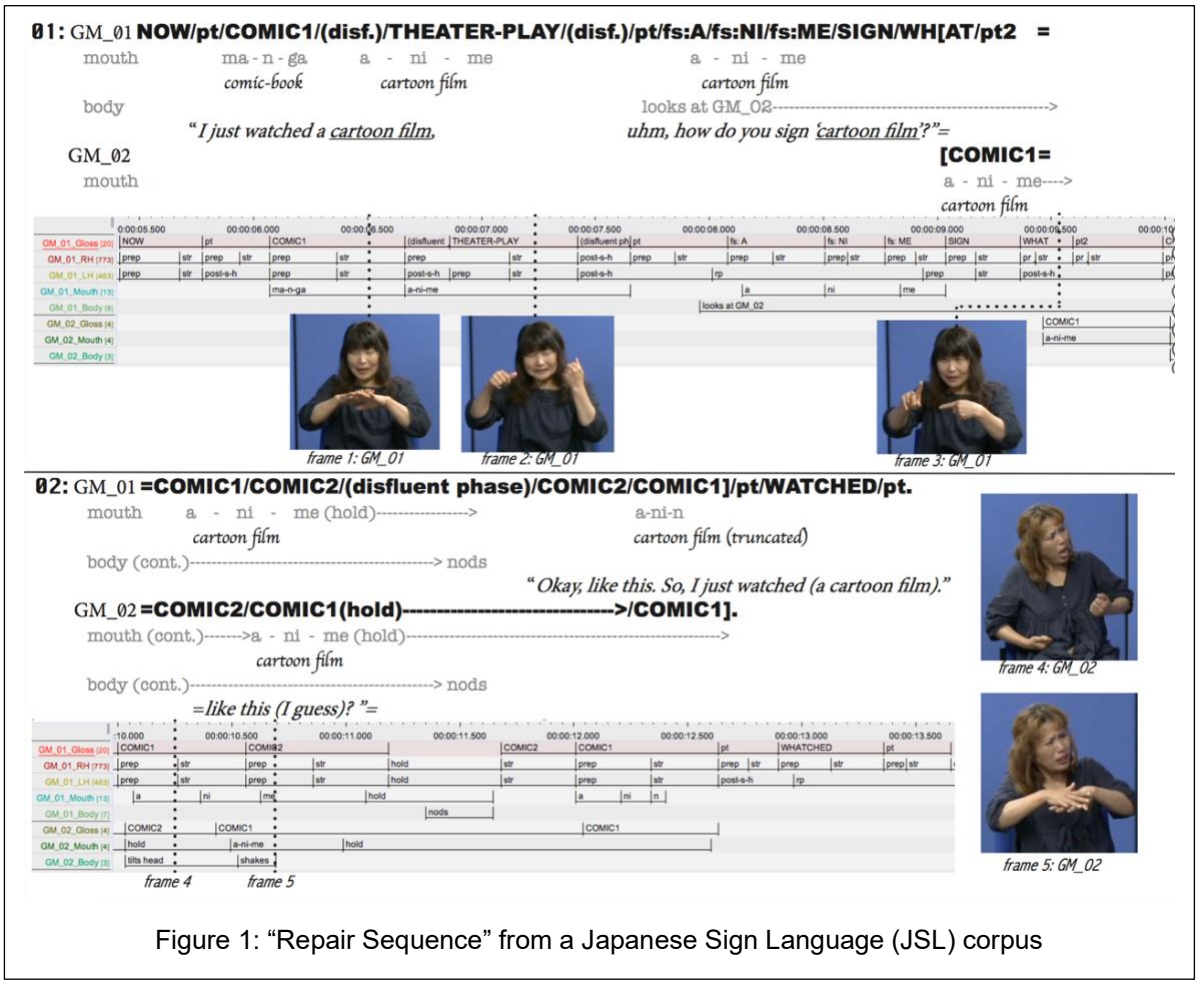

Figure 1: "Repair Sequence" from a Japanese Sign Language (JSL) corpus

To validate our findings, we supplement the literature survey with a Micro-level Kinematic Analysis. We utilize a specific interactional sequence from a Japanese Sign Language (JSL) conversational corpus (Bono, 2017), which serves as a "diagnostic probe." By calculating the decoupling between manual and mouth signals (Mouth Opening Magnitude vs. Hand Movement Energy), we quantify the precise linguistic structures that are systematically filtered out by current E2E optimization objectives.

2.3 Research Questions

Our analysis is guided by three central questions:

Representational Collapse: How do E2E visual encoders compress multi-channel signals (NMS vs. Manual) into flattened latent spaces?

Metric Blindness: To what extent do sequence-level metrics (BLEU/WER) mask catastrophic semantic errors in the repair sequence?

The Interactional Gap: Why is the "success" of current models hard to translate into usability in real-world, high-context Deaf communication?

3. Case Study: The Invisibility of Repair Sequence

To empirically examine the visibility criterion introduced in the previous section, we turn to a micro-level case study of an interactional repair sequence from a Japanese Sign Language (JSL) corpus involving pairs of Deaf adults from Gunma (Bono, 2017). Rather than serving as an isolated anecdotal example, the sequence functions as a diagnostic probe that allows us to observe how linguistic contrastivity is maintained in real interaction and to test whether such contrastive structures remain detectable under current end-to-end modeling assumptions.

3.1 The Interactional Challenge: Seven Trials of Repair Sequence

The session involves a lexical search for the term 'anime'. As the signer struggles to recall the specific lexical unit, they undergo seven distinct trials, shifting strategies between regional signs, similar phonological signs, and fingerspelling. Critically, the trouble is resolved in the final trial: the signer returns to a previous manual form but alters the mouthing from ma-n-ga to a-ni-me.

From an interactional linguistic perspective, this sequence is not merely an instance of lexical search difficulty but a structured process of collaborative meaning negotiation. In natural sign language interaction, repair sequences provide a privileged analytic window into how linguistic contrastivity is actively maintained in real time. What makes the present case analytically significant is that the decisive contrast does not occur in the manual channel, which remains largely constant across trials, but in the mouthing channel, which carries the final semantic disambiguation. This interactional configuration allows us to isolate a situation where linguistic success depends precisely on the model's ability to detect cross-channel decoupling rather than global motion similarity. Consequently, the repair sequence functions as a diagnostic probe: if a translation model fails to recognize the mouthing-based correction, it does not merely produce a lexical error but fails to capture the very mechanism through which meaning is stabilized in dialogue.

Table 1 provides a qualitative interactional summary of the repair trajectory across the seven trials, complementing the visual sequence shown in Figure 1. While the manual lexical form repeatedly converges on the regional sign COMIC1, the decisive semantic shift occurs in the mouthing channel, where the production changes from *ma-n-ga* to *a-ni-me*. This cross-channel asymmetry reveals that the communicative resolution of the repair sequence is not achieved through the introduction of a new manual sign, but through the stabilization of the mouth signal while the manual form remains largely constant. The figure therefore illustrates a key diagnostic property of repair sequence: meaning may be re-established through subtle cross-channel contrast rather than through large-scale manual movement, precisely the type of signal configuration that current end-to-end optimization pipelines are likely to underrepresent.

While the qualitative interactional analysis demonstrates where the decisive semantic contrast is located, a quantitative examination is required to understand why this contrast is likely to be suppressed in current optimization pipelines. We therefore turn to kinematic analysis to measure the degree of manual–mouth decoupling underlying the repair sequence.

TRIAL	WHO	HAND	MOUTH
1st	GM_01	Lexical sign	ma-n-ga
2nd	GM_01	Lexical sign	a-ni-me
3rd	GM_01	Finger spelling	a-ni-me
4th	GM_02	Lexical sign	a-ni-me
5th	GM_02	Lexical sign	↓ (hold)
6th	GM_02	Lexical sign:	a-ni-me
7th	GM_02	Lexical sign:	↓ (hold)

Table 1: Trials for sharing the semantic meaning

3.2 Kinematic Evidence: Manual-Mouth Decoupling

To quantify why this linguistic nuance is invisible to current models, we analyze the kinematic features of the sequence.

From a technical perspective, recent work has also shown that mouth movements exhibit temporal boundaries that do not necessarily align with manual gestures. Sandermann (2025) formulates mouth movement detection in JSL as a temporal segmentation problem and demonstrates that manual-focused pipelines often miss or misplace non-manual events. This work further reports difficulties in distinguishing interactionally relevant mouth movements from incidental facial activity, underscoring that mouth dynamics cannot be treated as a simple by-product of hand motion. This technical evidence closely aligns with our kinematic finding of manual–mouth decoupling and reinforces the claim that single-stream E2E encoders face a systematic blind spot for non-manual signals.

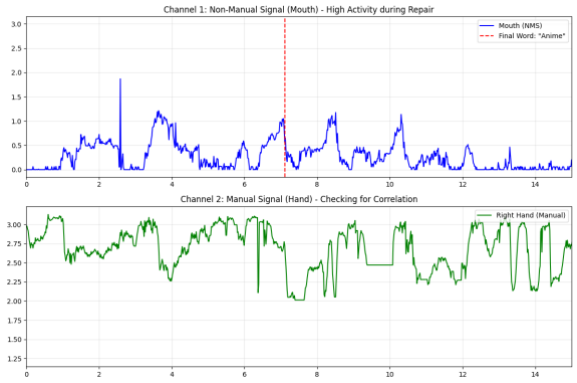

Figure 2: Kinematic decoupling between manual and non-manual signals
(A) Upper panel; (B) Lower panel

Figure 2 displays the kinematic contrast within the narrator's performance. In Panel A, we plot the Mouth Opening Magnitude (MAR) against the Right-Hand Movement Energy to observe a prominent “manual-mouth decoupling”. Between 0s and 7.11s, while the hand maintains a high-amplitude, repetitive oscillation, representing the ‘stagnant’ manual sign COMIC1, the mouth signal exhibits intense, pulsing activities. This suggests that the mouth independently carries the semantic load of the repair sequence. Under current E2E optimization schemes, which are biased toward globally dominant motion patterns (i.e., manual movement), such localized, high-frequency bursts of mouth activity are unlikely to be preserved and may instead be treated as temporal jitter or noise, despite being interactionally decisive.

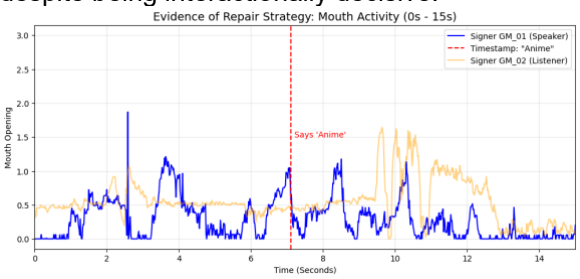

Figure 3: Interactional coordination and signal validation

Figure 3 illustrates the interactional relevance of the NMS signal through the listener’s response. Around the critical 7.11s mark, the narrator’s peak mouth activity (blue line) is followed by a temporally aligned change in the listener’s behavior (yellow line). While this temporal coupling does not by itself establish causality, it provides interactional evidence that the mouth signal constitutes a salient cue for the repair sequence, coinciding with the moment at which the communicative trouble is resolved in the dialogue. This cross-participant temporal alignment suggests that the NMS activity is not merely articulatory noise, but a socially consequential signal that is attended to by the interlocutor.

3.3 The Architectural Blind Spot: Visual Noise and Dyadic Interference

While the previous sections have shown that end-to-end models tend to compress multi-channel signals into flattened representations, the more critical question is what this implies for real signed interaction. Figure 4 provides a consolidated view of the architectural limitation in current end-to-end (E2E) frameworks by jointly displaying the manual and non-manual channels of both interlocutors. Based on this kinematic evidence, we identify three recurrent architectural blind spots through which standard visual encoders are likely to obscure interactionally decisive signals in the repair sequence.

Visual Noise Dominance (t = 2.5–4.0 s). At the initiation of the repair sequence, the linguistically relevant information is localized in the non-manual channel (the right signer’s mouth), whereas the manual channel exhibits high-frequency jitter with larger overall motion energy. Given the typical bias of visual encoders toward globally dominant motion patterns, the mouth-based cue is therefore likely to be attenuated in the learned representation in favor of the more salient manual signal, despite its central role in resolving the interactional trouble.

Channel Conflict and Attention Misalignment. This cross-channel imbalance creates a systematic challenge for attention-based mechanisms. Architectures that allocate attention weights primarily based on motion magnitude tend to prioritize the manual channel, which carries stronger but linguistically less informative variation in this segment, over the non-manual channel, which carries the decisive contrastive cue. As a consequence, the signal that is critical for the repair sequence is at risk of being treated as secondary or negligible within the model’s internal representation.

Diarization Failure (t ≈ 5.0 s, interaction onset). A further source of error arises from the temporal overlap between the two participants, where the left signer’s activity decreases while the right signer initiates a response. In the absence of

explicit speaker diarization or spatially grounded attention, global encoders may conflate these two streams into a single feature trajectory. This conflation increases the risk of feature-level interference and, ultimately, of collapsing distinct interactional turns into an undifferentiated semantic representation.

Together, these three mechanisms illustrate how interactionally decisive non-manual signals in the repair sequence can be systematically overshadowed by globally stronger but linguistically less informative motion patterns in current E2E architectures.

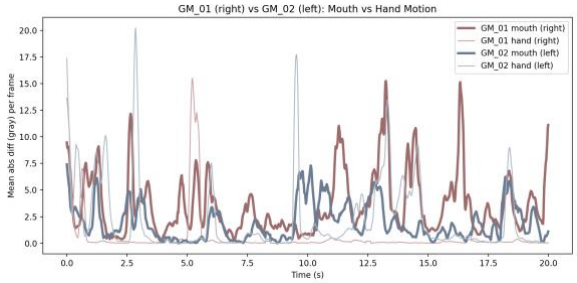

Figure 4: Distractor effect in dual-signer input In contrast to Figure 3, which isolates the mouth channel, this plot jointly displays all four channels: mouth (thick lines) and hand (thin lines) motion for both signers

3.4 Empirical Shortcomings of E2E Models

A quantitative analysis of the JSL repair sequence reveals a substantial kinematic divergence, raising questions about how current E2E models account for such discrepancies. We quantify the similarity between the kinematic time series using Pearson's correlation coefficient (r) computed over aligned temporal windows (±30 frames) around each target time point. When comparing the initial error (Trial 1, ma-n-ga) with the successful repair sequence (Trial 7, a-ni-me), we found that the Manual Signal Correlation is only 0.3448, while the Mouth Signal (MAR) Correlation drops to -0.1685.

This low correlation highlights a dual challenge for state-of-the-art architectures like GFSLT-VLP. First, the weak manual correlation indicates that the "stagnant" hand signal is not a static posture but a temporally jittered repetitive motion, which E2E encoders often struggle to align across different trials. Second, the near-zero mouth correlation confirms that the semantic shift is driven entirely by independent mouth activity.

Current models, which rely on global pooling or fixed-window attention, essentially "drown" these negative-correlation mouth signals in the noise of the physically more dominant manual channel. As a result, under current objectives, the models are biased toward overlooking the repair sequence, leading to a semantic collapse in which two distinct events are mapped to the same incorrect class, "Manga."

The empirical data in Table 2 provides quantitative evidence that helps explain the limitations of current E2E architectures. Semantically, the manual form remains constant (COMIC1), but kinematically it exhibits jittered repetition, which results in only a weak correlation (r = 0.34). The mouth signal, by contrast, exhibits a near-zero negative correlation (r = -0.1685). In a robust linguistic system, this lack of correlation in mouth activity should trigger a semantic shift (from "manga" to "anime"). However, although r = 0.3448 is conventionally interpreted as a "weak correlation" in general statistics, under the global optimization objective of end-to-end models it can still become a dominant cue. This is because the physical motion energy of the manual channel is substantially larger than the subtle and localized mouth signal, leading the optimization process to implicitly assign greater weight to the manual channel in terms of representation learning and gradient contribution. As a result, models such as GFSLT-VLP (Zhou et al., 2023) are likely to

Feature Channel	Signal Type	Correlation (r)	Linguistic Significance
Manual Channel	Right Hand Energy	0.3448	Weak correlation indicating temporal jitter in repetitive motion.
Non-Manual signal (NMS)	Mouth Opening (MAR)	-0.1685	Strong negative/zero correlation confirming distinct mouthing patterns.
Expected Model Behavior	Latent Representation (conceptual)	-	Given global-motion-dominant optimization, E2E models are likely to preserve manual similarity while attenuating the contrastive mouth signal, leading to semantic collapse.

Table 2: Kinematic correlation and semantic alignment analysis

prioritize the weak but high-energy manual correlation over the linguistically decisive mouth signal, and erroneously cluster both trials into the same semantic category, effectively rendering the model "blind" to the repair sequence shown in Figure 3.

3.5 Why This Invisibility Matters

Linguistic units are not merely "intermediate layers" to be optimized away for the sake of efficiency. When we flatten the simultaneity of signing into a single sequential mapping, we lose the contrastivity that defines meaning.

As demonstrated in the Gunma case, the resolution of communicative trouble relies on the entire multimodal repertoire. If an E2E pipeline filters out the NMS channel to improve the signal-to-noise ratio of manual signs, it effectively renders the most critical semantic corrections invisible. We argue that "progress" in SLT must be redefined: a model has not learned the language if it cannot perceive the very signals that human signers use to negotiate shared meaning. If a model can match 90% of a sentence's tokens but misses at the 10% that constitutes the repair sequence, it has not merely made a "minor error", it has fundamentally failed to participate in the communicative act.

4. The Linguistic Disconnect: What is Being Lost?

The shortcomings of E2E models to capture the JSL repair sequence is not an isolated technical glitch but a symptom of a systematic misalignment between machine learning objectives and sign language linguistics. In this section, we analyze the specific linguistic structures that become "invisible" when SLT is treated as a monolithic signal-to-text mapping task.

4.1 What Linguistics Demands: Contrastivity and Disentanglement

In linguistic theory, sign languages are defined by their simultaneous, multi-channel architecture. Meaning is not merely a sequence of symbols but a co-constructed product of manual signs and non-manual signals (NMS) (Bragg et al., 2019).

The principle of contrastivity is important. Linguistic units are defined by their differences. As demonstrated in the Gunma case, the "minimal pair" distinguishing ma-n-ga from a-ni-me resides entirely in mouthing, within the non-manual channel, rather than in the manual channel. When E2E encoders collapse these channels into a single latent vector, they violate the principle of contrastivity, effectively treating semantic shifts as intra-class variance (noise).

Interactional alignment implies Sign language is inherently social. The "repair sequence" proves that meaning is negotiated in real-time. The fact that the listener's response triggers exactly at the moment of the mouth pulse (Figure 2, Lower panel) underscores that NMS are not "auxiliary" but are the primary drivers of interactional resolution.

4.2 The Metric Mirage: Why BLEU and WER are "Linguistically Inadequate"

Current evaluation protocols systematically penalize linguistic depth by rewarding surface-level approximation.

This first influence is performance Inflation. Metrics like BLEU and ChrF are based on n-gram overlap (Papineni et al., 2002; Popović, 2015). In the JSL example, if a model correctly translates "I am reading a..." and "...it is interesting," it receives a high score even if it fails the core repair sequence (translating 'anime' as 'manga'). This creates a false sense of progress where models are "guessing" the context rather than "translating" the language.

Then sometimes it will cause the area-weight bias. In E2E loss functions (e.g., Cross-Entropy on pixels or keypoints), large-scale movements like arm swings carry more gradient weight than subtle facial or mouth activity. Consequently, the "linguistic signal" (the mouth) is drowned out by the "physical noise" (the stagnant hand). Current metrics provide no mechanism to penalize a model for missing a high-weight semantic correction if the global sequence looks "plausible."

The case study therefore establishes an empirical observation: interactionally decisive linguistic information may be concentrated in localized non-manual channels even when global motion patterns remain largely unchanged. This observation raises a broader methodological question: if such structures are systematically attenuated by current modeling and evaluation practices, which linguistic dimensions of sign language are being lost at scale? The next section addresses this question by examining the specific linguistic structures that disappear under current SLT paradigms.

5. Implications: Rethinking "Progress" in SLT

The "linguistic invisibility" observed in our JSL case study is not merely a technical limitation but a methodological consequence. As the field moves toward larger models and higher-order benchmarks, we must re-evaluate the criteria for successful translation. Representation design determines not only model performance but also which linguistic structures survive the pipeline.

Our analysis can be understood as a bridge between three strands of work: interactional analyses of repair sequence in sign language, linguistically motivated categorization of mouth actions, and technical efforts toward modeling the

temporal structure of non-manual signals. Recent JSL-focused studies have shown that mouth actions form structured linguistic categories, and technical work has demonstrated that their temporal boundaries often diverge from manual gestures. Yet, current end-to-end SLT pipelines still tend to absorb these signals into a single flattened visual stream.

Our L1–L3 framework is designed to make this tension explicit: it allows us to examine whether cross-channel contrast, which may be masked at isolated and local levels (L1/L2), becomes observable only at the interactional level (L3), where repair sequence is functionally resolved.

5.1 Linguistics as a Diagnostic Coordinate

End-to-end modeling does not render linguistics unnecessary. On the contrary, linguistic structure provides the essential coordinates for observing, diagnosing, and comparing model behavior (Bragg et al., 2019; Yin et al., 2021, see also Bono, 2026). Without structural visibility, it becomes impossible to judge whether a model has truly learned the language or has simply memorized a signal-to-text mapping. Our JSL analysis proves that if we do not know to look for manual-mouth decoupling, we cannot explain why a "high-performance" model is at risk of failing to resolve a basic communicative repair sequence.

5.2 Representation as a Survival Mechanism

In SLT, the choice of representation determines which structures are locatable and which are absorbed as noise. As emphasized in overviews of multimodal corpora (Bono, 2026), representational design directly constrains what kinds of linguistic structure can be observed, analyzed, and compared across modalities. Skobov and Bono (2023) propose a three-dimensional normalization of body movement that preserves analyzable structure while remaining compatible with E2E objectives. This trend is further exemplified by recent gloss-free architectures such as SpaMo (Hwang et al., 2025), which explicitly refines motion dynamics to integrate with large language models, avoiding the pitfalls of a single flattened visual stream. This illustrates that structural visibility is not incompatible with data-driven SLT; rather, it requires explicit representational interfaces. Ensuring identifiable "places" for key structures such as dedicated channels for NMS, enables the analysis of internal model logic.

5.3 Evaluation as the "Invisible Hand"

Evaluation protocols act as an invisible hand: what cannot be evaluated will eventually disappear from the research loop (De Meulder, 2021). Incorporating structural visibility into evaluation does not require abandoning BLEU or WER, but it does require recognizing their coverage limits. We therefore suggest a shift from single-score improvements toward diagnostic test sets (akin to NLP "CheckLists"), including repair-sequence benchmarks based on naturalistic dialogues (e.g., the Gunma JSL data with annotated "trouble" and "resolution"), as well as contrastive sensitivity tests that probe minimal pairs differing only in NMS (e.g., mouthing).

Taken together, these considerations point to a methodological shift: progress in SLT should be assessed not only by benchmark scores, but by whether models preserve, express, and support key linguistic structures (Yin et al., 2021). They also motivate diagnostic analyses that go beyond single-score evaluation and directly probe how models organize linguistic structure. To this end, we introduce a small-scale hierarchical context probe using a representative E2E model to examine whether successful and failed repair sequences exhibit different patterns of contextual integration in the model's internal representations.

5.4 A Hierarchical Context Probe: Computational Evidence for Context Alignment

To complement the interactional and kinematic analysis, we introduce a hierarchical context probe using a representative E2E model (GFSLT-VLP) to examine how successful and failed repair sequences are organized under different levels of contextual embedding. We consider three conditions: (L1) isolated segments (Trial 1 vs. Trial 7), (L2) local-context segments extracted around each trial (Trial 1_local vs. Trial 7_local), and (L3) the full 20-second interaction, which serves as a proxy for discourse-level "natural flow".

We analyze these conditions using two diagnostic measures derived from the last-layer self-attention: attention entropy (H), which reflects the degree of temporal information integration, and self-attention focus (F), which reflects the degree of frame-wise independence. In addition, we assess representation stability by comparing isolated and embedded representations, and compute the distance of each condition to the L3 baseline in the (H, F) space. These measures are used as indicators of contextual integration rather than as claims about human cognition.

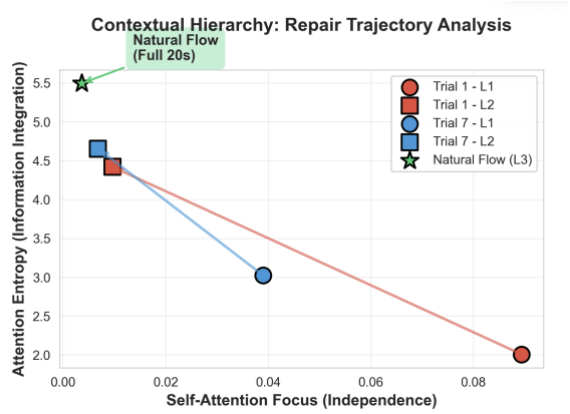

Figure 5: H–F trajectory across contextual levels

Trajectories of Trial 1 and Trial 7 in the attention entropy (H)–self-attention focus (F) space from isolated (L1) to local context (L2), with the full 20-second interaction (L3) shown as the natural-flow baseline

Figure 5 shows the trajectories of Trial 1 and Trial 7 in the (H, F) space from L1 to L2, relative to the L3 baseline, and Figure 6 reports the corresponding entropy and focus values. Two contrasts emerge. First, Trial 7 starts closer to the discourse-level baseline and undergoes a smoother transition from L1 to L2, whereas Trial 1 requires a substantially larger representational shift, suggesting that the successful repair sequence is already structurally closer to discourse-level organization. Second, across both L1 and L2, Trial 7 consistently exhibits higher entropy and lower self-attention focus than Trial 1, indicating a stronger reliance on contextual integration rather than frame-wise, citation-like independence.

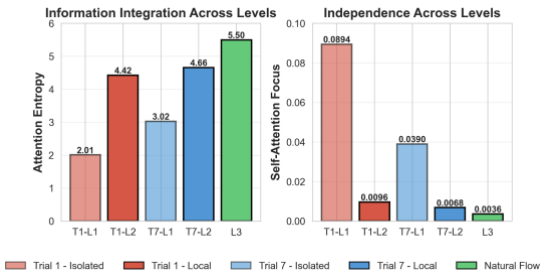

Figure 6: Information integration and independence across contextual levels

Attention entropy (H; left) and self-attention focus (F; right) for Trial 1 and Trial 7 across isolated (L1), local-context (L2), and full-sequence (L3) conditions

This contrast becomes especially clear in representation stability. When comparing isolated and embedded conditions, Trial 7 preserves 75% similarity between its representations, while Trial 1 drops to 18%. Likewise, the amplification of attention entropy from L1 to L2 reaches 120% for Trial 1 but only 54% for Trial 7. Together, these results indicate that the successful repair sequence is already pre-adapted to discourse-level processing, whereas the failed repair sequence relies on a more context-independent form that requires substantial reconfiguration when embedded into interaction.

Taken together, this hierarchical probe provides converging computational evidence for the context alignment hypothesis: successful repair sequence is achieved by producing forms that are structurally aligned with the ongoing discourse and therefore integrate smoothly into the interactional flow.

6. Conclusion

This paper has shown that linguistic invisibility in end-to-end sign language translation does not arise primarily from insufficient model capacity, but from optimization and evaluation regimes that privilege global signal similarity over cross-channel linguistic contrast. Through an interactional and kinematic analysis of a Japanese Sign Language repair sequence, we demonstrated that communicative resolution may depend on localized non-manual signals that remain largely undetected by current modeling pipelines despite high benchmark performance.

These findings suggest that translation accuracy alone is an incomplete indicator of linguistic competence in SLT systems. When interactionally decisive structures lack representational and evaluative visibility, models may appear successful while remaining functionally inadequate in real communicative settings. Ensuring structural visibility is therefore not a peripheral linguistic concern but a methodological requirement for reliable human-centered sign language technologies.

Future research should move beyond single-score benchmarking toward evaluation frameworks that explicitly test cross-channel contrastivity and interactional robustness. Developing diagnostic datasets, decoupling-sensitive metrics, and representations that preserve identifiable non-manual channels will be essential steps toward SLT systems that both perform well on benchmarks and faithfully capture the linguistic organization of natural signed interaction. Making these structures visible will be essential for ensuring that future SLT systems learn not only to translate sequences, but to participate reliably in human signed interaction.

7. Ethics Statement

This work is intended as a diagnostic and methodological analysis of current end-to-end (E2E) sign language translation (SLT) paradigms, rather than as a contribution toward deploying a production-ready translation system. Our goal is to make linguistically consequential structures, particularly non-manual signals (NMS) in interactional repair sequence visible to modeling and evaluation, and to highlight risks that arise when such structures are optimized away by metric-driven pipelines.

The empirical case study is based on an existing Japanese Sign Language (JSL) conversational corpus (Bono, 2017). Data collection and storage were conducted under the ethical protocols of the original project, including informed consent from participants (Bono et al., 2023). The present study performs secondary analysis only. We do not release any personally identifiable information, and all visualizations are presented in abstracted, de-identified kinematic form (e.g., motion energy and mouth opening measures), rather than raw video, to minimize risks of re-identification and to respect participant privacy.

We are mindful that SLT technologies can have direct social consequences for Deaf communities, especially if systems that achieve high benchmark scores nevertheless fail in real communicative settings. A central motivation of this work is therefore harm prevention: we argue that models optimized solely for surface-level metrics risk producing translations that are misleading or interactionally inappropriate. Our analysis does not advocate the deployment of such systems in high-stakes or user-facing scenarios; instead, it calls for linguistically informed, human-centered evaluation practices, and for the inclusion of Deaf perspectives in the design and assessment of SLT technologies, in line with prior ethical and participatory design discussions in the field. In short, this paper positions linguistics-based diagnostic evaluation as a means to improve the responsibility and reliability of future SLT research, rather than as a step toward replacing human interpreters or supporting unsupervised real-world deployment.

8. Limitations

This study has several important limitations that should be acknowledged.

First, the empirical analysis is based on a single interactional repair sequence from a Japanese Sign Language (JSL) corpus. The case is used as a diagnostic probe to demonstrate how linguistically decisive cross-channel contrasts can be obscured by current E2E optimization and evaluation practices. As such, our findings are not intended to provide statistical generalization about the frequency of such phenomena, but rather to establish the existence and methodological relevance of this failure mode.

Second, our analysis is language-specific in its empirical grounding. While the methodological concern, namely, the risk of collapsing multi-channel linguistic structure into global motion-dominant representations is likely to be relevant across sign languages, we do not claim that the exact same repair sequence dynamics or kinematic patterns will hold for other languages such as ASL, DGS, or CSL. Broader cross-linguistic validation remains an important direction for future work.

Third, the computational probe mainly focuses on representative contemporary E2E model (e.g., GFSLT-VLP) and related architectural assumptions. We do not claim that all possible future architectures will necessarily exhibit the same behavior. Rather, our analysis targets a dominant modeling paradigm in current SLT research, especially end-to-end models and highlights a structural risk inherent to this paradigm under common training objectives and evaluation metrics.

Finally, the kinematic measures used in this study (e.g., mouth opening magnitude and hand movement energy) should be understood as operational proxies for linguistically meaningful non-manual and manual activity. While these measures are sufficient for demonstrating manual–mouth decoupling and for constructing a diagnostic contrast, they do not exhaust the full richness of non-manual signals such as eye gaze, facial expression, or head movement. Future work should incorporate richer, linguistically grounded annotations and multi-channel measurements to refine and extend the present analysis.

9. Bibliographical References

Bono, M. 2017. Improvisational signing in sign language interaction through a lens of repair sequence. *The Japanese Journal of Language in Society*, 19(2), 20–31. DOI: 10.19024/jajls.19.2_59

Bono, M. 2026. Multimodal corpora. In H. Nesi and P. Milin (eds.), *International Encyclopedia of Language and Linguistics*, 3rd ed. Elsevier.

Bono, M., Okada, T., Kikuchi, K., Sakaida, R., Skobov, V., Miyao, Y., and Osugi, Y. 2023. Utterance unit annotation for the Japanese Sign Language Dialogue Corpus: Towards a method for detecting interactional boundaries in spontaneous sign language dialogue. In Ella Wehrmeyer (ed.), *Advances in Sign Language Corpus Linguistics*, SCL 108, pp. 353–382. John Benjamins. URL: https://cir.nii.ac.jp/crid/1360580230632859264

Bragg, D., Koller, O., Bellard, M., Berke, L., Boudreault, P., Braffort, A., Caselli, N.,

Huenerfauth, M., Kacorri, H., Verhoef, T., Vogler, C., and Morris, M. R. 2019. Sign Language Recognition, Generation, and Translation: An Interdisciplinary Perspective. In *Proceedings of the 21st International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '19)*, pages 16–31, New York, NY, USA. Association for Computing Machinery. DOI: 10.1145/3308561.3353774

Camgoz, N. C., Koller, O., Hadfield, S., and Bowden, R. 2018. Neural Sign Language Translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

De Meulder, M. 2021. Is "Good Enough" Good Enough? Ethical and Responsible Development of Sign Language Technologies. In *Proceedings of Machine Translation Summit XVIII: Workshop on Automatic Translation for Signed and Spoken Languages* (AT4SSL). URL: https://aclanthology.org/2021.mtsummit-at4ssl.2/

Duarte, A., et al. 2021. How2Sign: A Large-scale Multimodal Dataset for Continuous American Sign Language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* DOI: 10.1109/CVPR46437.2021.00276

Forster, J., Schmidt, C., Hoyoux, T., Koller, O., Zelle, U., Piater, J., and Ney, H. 2012. RWTH-PHOENIX-Weather: A Large Vocabulary Sign Language Recognition and Translation Corpus. In *Proceedings of LREC 2012*, Istanbul, Turkey. URL: https://aclanthology.org/L12-1503/

Graves, A., Fernández, S., Gomez, F., and Schmidhuber, J. 2006. Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks. In *Proceedings of ICML 2006*.

Hwang, E. J., Cho, S., Lee, J., and Park, J. C. 2025. An Efficient Gloss-Free Sign Language Translation Using Spatial Configurations and Motion Dynamics with LLMs. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3901–3920, Albuquerque, New Mexico, USA. Association for Computational Linguistics.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of ACL 2002*.

Popović, M. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*.

Regmi, B. S. 2025. Mouth Movement Classification in Japanese Sign Language. Internship Report, National Institute of Informatics, Tokyo, Japan.

Sandermann, T. H. 2025. Temporal Mouth Movement Segmentation in Japanese Sign Language. Internship Report, National Institute of Informatics, Tokyo, Japan.

Skobov, V. and Bono, M. 2023. Making Body Movement in Sign Language Corpus Accessible for Linguists and Machines with Three-Dimensional Normalization of MediaPipe. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1913–1926, Singapore. Association for Computational Linguistics.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., and Makhoul, J. 2006. *A study of translation edit rate with targeted human annotation*. In *Proceedings of AMTA*.

Wong, R., Camgoz, N. C., and Bowden, R. 2024. Sign2GPT: Leveraging Large Language Models for Gloss-Free Sign Language Translation. In *Proceedings of the International Conference on Learning Representations*.

Yin, K., Kordjamshidi, P., and Roth, D. 2021. Including Signed Languages in Natural Language Processing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, pages 5774–5787.

Zhou, B., Zhang, S., Chen, X., Liu, M., and Wu, Y. 2023. *Gloss-Free Sign Language Translation: Improving from Visual-Language Pretraining*. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.