The Princeton Journal of Interdisciplinary Research
ISSN 3069-8200

SignBridge: ASL Translation with Pose-based Recognition, LLM-driven Semantic Coherence, and Composite Translation Quality Index

Kavin Sai Kumar (author)[1]

[1]Eastside Catholic School, Sammamish, Washington, United States of America

Abstract

Deaf individuals use sign language to communicate through hand shape, motion and position rather than speech. Sign languages have their own grammar and word order – no structural resemblance to spoken languages in the same regions. No framework currently measures how translated words preserve the intended meaning of the sentence. Current automated captioning systems largely fail to handle this complexity, leaving 80% of digital content inaccessible to over 70 million Deaf individuals worldwide. This study developed SignBridge, an end-to-end American Sign Language-to-English translation system and introduced the Composite Translation Quality Index (CTQI) to address both the translation problem and the measurement gap. SignBridge enhanced skeletal pose tracking from a baseline of 27 to 83 landmarks, applied 50× training data augmentation, and trained a transformer encoder

to generate ranked Top-K sign candidates. A large language model then performed contextual recovery, selecting semantically coherent signs based on sentence-level meaning rather than individual confidence scores. CTQI was developed through four human-validated iterations, and it evaluated sign accuracy, content coverage, and plausibility. Top-1 accuracy increased from 41.23% to 80.97.0% ($z=18.23$, $p < 0.0001$), with Top-3 reaching 91.6%. CTQI improved from 45.5 to 74.0 ($p < 0.001$, Cohen's $d=1.53$), with 92% of sentences improving. CTQI achieved a correlation of $r=0.94$ with independent human evaluators (ICC=0.96), topping standard translation metrics. These results confirm that semantic coherence-based selection outperforms confidence-only approaches for ASL translation. CTQI provides a validated, reproducible evaluation framework that no prior work has established. Vocabulary expansion and continuous signing remain open for future work.

Keywords: ASL translation, pose estimation, sign language recognition, large language models, translation quality metrics

Introduction

Background and Significance

Deaf individuals use sign language to communicate through hand shape, motion and position – a fundamentally different modality - rather than speech. Sign languages have their own grammar and word order and do not resemble other spoken languages in those same regions. Right now, there are more than 300 different sign languages, connecting about 70 million Deaf people across the globe. However, less than 20% of digital content for the Deaf community comes with reliable automated captions. So, people get left out. Patients cannot

explain their symptoms to doctors. Students get lost in class. In court, people do not know what is going on. At workplace, people miss out on jobs and opportunities. Just in healthcare, Deaf patients hit a block in emergency settings where interpreters are not always around and the clinical staff have limited awareness of Deaf culture. These contribute to slower care and adverse health outcomes (Tannenbaum-Baruchi et al., 2025). Without qualified interpreters, Deaf patients face more misdiagnosis, don't always give truly informed consent, and their safety takes a hit (Sanchez, 2025). Fewer than one in five Deaf patients who require ASL interpreters get one every time they visit a doctor. These gaps persist despite federal mandates under the Americans with Disabilities Act (Char et al., 2024).

ASL is not just English with hand signs—it is a full language on its own, with its own rules and structure (Stokoe, 1960; Klima & Bellugi, 1979; Valli & Lucas, 2000). ASL often uses topic-comment word order, while English uses subject-verb-object. Translating word for word results in something like "STORE I GO YESTERDAY," which does not magically turn into "I went to the store yesterday". This linguistic gap requires explicit bridging.

Prior Work and State of the Field

Sign language recognition (SLR) has evolved through three broad phases. Initially, classic approaches like Hidden Markov Models and k-Nearest Neighbors relied on handcrafted features extracted from video or sensor data. These methods were lightweight but struggled to generalize across signers, lighting conditions, and backgrounds. Deep learning marked a breakthrough: Convolutional Neural Networks (CNNs) extract hand and body shapes directly

from video frames without requiring handcrafted features; recurrent Neural Networks (RNNs) introduced the capability to process signs as sequences over time, and transformer architecture brought self-attention mechanisms to model the entire sign sequence at once, outperforming earlier approaches. Recent approaches combine standard video, depth sensors and skeletal joint tracking (Pose) to improve predictions further (Koller et al.; Jian2g et al., 2021).

Gap 1 – Pose-Based Keypoint Density and Targeted Augmentation

Input representation is a critical methodological choice. Video-based methods can achieve high accuracy like the multi-stream CNN SOTA: 81.38% Top-1 on WLASL-100 (Maruyama et al., 2024). However, these are computationally prohibitive for real-time consumer hardware. Pose-based methods instead extract skeletal keypoint sequences — reducing video to the body's essential motion geometry — making models faster, more robust to signer and background variation, and more amenable than raw video to mathematical augmentation (rotation, scaling, mirroring). Prior skeleton-augmented approach - Ours2 - achieved 71.07% Top-1 accuracy on WLASL-100 dataset using 27-keypoint pose configuration (Maruyama et al., 2024). The OpenHands framework (Selvaraj et al., 2022), which standardized pose-based recognition earlier across multiple sign languages using a similar 27-keypoint MediaPipe configuration, reported results on WLASL-2000 but did not evaluate on WLASL-100. These prior accomplishments stop at word-level sign recognition, and the 27 keypoints omit individual finger joints critical for distinguishing confusable signs. No existing pose-based framework combines expanded keypoint density with pre-generated augmentation optimized for data-scarce word-level recognition.

Gap 2 – Unexplored LLM Role Beyond End-End Video Translation

On the translation side, Gong et al. (2024) demonstrated that LLMs can improve sign language translation by transforming signed videos into grammatically correct output — establishing that the linguistic knowledge encoded in LLMs is directly useful for the sign-to-text task. However, their SignLLM framework operates end-to-end on video as input format, converting signs to token representations before invoking LLM. This approach lacks transparency into intermediate sign-selection decisions and has not been evaluated in resource-constrained or real-time deployment settings where pose-based input have practical advantages. No prior work uses an LLM as a contextual selector across a Top-K pool of pre-recognized gloss candidates with confidence scores. Structured prompt engineering for this use case — prioritizing semantic coherence over per-position confidence scores and enforcing grammatical constraints to produce natural-sounding translations — has not been formally investigated.

Gap 3 – Absence of a composite, human-evaluated evaluation metric

Evaluation presents a third gap. Current translation systems rely on one-dimensional metrics: BLEU measures n-gram overlap but misses semantic meaning and grammatical correctness; BERTScore captures semantic similarity and ignores grammatical correctness; Word Error Rate measures how many changes are needed but not naturalness or fluency. These metrics, borrowed from spoken-language machine translation, treat translation quality as one-dimensional, making it impossible to determine whether an apparent improvement reflects genuine holistic quality or optimizes a single-metric at the expense of other dimensions. No prior work establishes a unified composite metric that integrates gloss accuracy, semantic

preservation, grammatical correctness, and information coverage into a single validated, human-aligned measure.

Research Questions and Corresponding Hypotheses

This study addresses these three gaps through the following research questions.

RQ1 - Can tracking more body points (vs. baseline 27) and targeted data augmentation improve word-level sign recognition accuracy when training data is limited?

Existing pose-based frameworks that use 27 keypoints systematically miss the fine-grained finger movements that distinguish many confusable sign pairs. With only ~3-6 training samples per class in WLASL-100 dataset, standard models overfit and cannot compensate for the missing hand detail. We hypothesize (H1) that an expanded keypoint representation, combined with pre-generated augmentation, can achieve competitive word-level accuracy.

RQ2 - Can an LLM improve translation by selecting signs based on sentence-level semantic coherence rather than individual confidence scores?

Sign recognition systems produce isolated glosses that lack grammatical structure. These result in incomprehensible output such as “MOTHER HELP DEAF” instead of “The mother helps the deaf person.” Confidence-only selection compounds this by greedily picking the highest-scored sign at each position independently, thereby missing cases where the contextually correct sign has a second or third highest score. We expect (H2) that an LLM analyzing the full Top-K candidate pool across all positions simultaneously — guided by sentence-level semantic

coherence and grammatical constraints rather than per-position confidence scores — can produce meaningfully better English output.

RQ3 - Can a composite metric measuring multiple quality dimensions correlate better with human judgment than any single metric?

Single-metric optimization can produce misleading apparent improvements that do not reflect genuine quality gains. None can detect when a translation preserves content words but fails grammatically or is grammatically fluent but still semantically wrong. We predict (H3) that a composite metric integrating gloss accuracy, content coverage, grammatical fluency and semantic coherence — validated against independent human raters — can provide a more reliable and holistic measure of translation quality than any individual metric alone.

Goal

This study will develop SignBridge — an end-to-end ASL-to-English pipeline integrating enhanced pose recognition, LLM-based linguistic bridging, and a novel evaluation framework. It will validate the evaluation framework as a reproducible, human-aligned metric addressing the specific quality dimensions of cross-linguistic sign translation.

Methodology

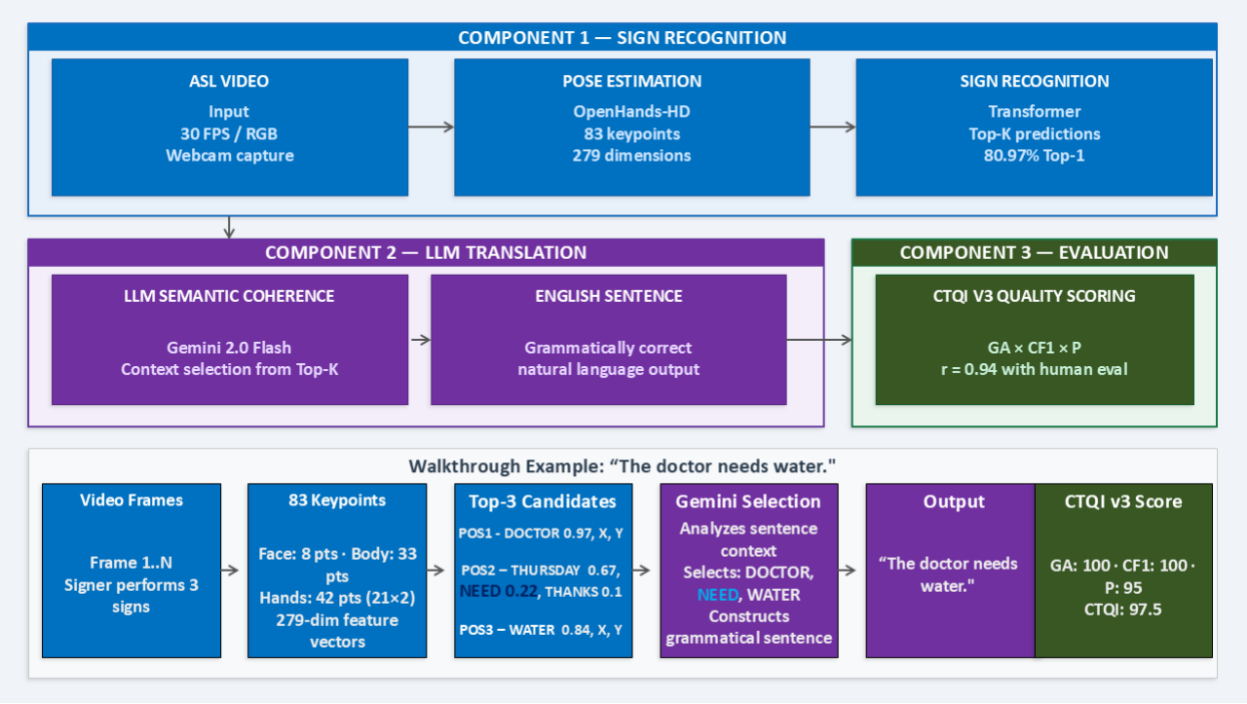
SignBridge Architecture, Dataset and Evaluation Corpus

SignBridge integrates a three-component architecture (see Figure 1) into a sequential pipeline. The first component produces a ranked list of candidate glosses for each position in the

input sentence. The LLM semantic coherence module then selects the most contextually appropriate combination across all positions simultaneously. Finally, a novel framework evaluates the resulting translation across multiple quality dimensions to produce a composite score. All experiments used the WLASL-100 benchmark dataset (Li et al., 2020), comprising 100 sign classes and 342 raw video clips for training. For evaluation, 53 sentence pairs (2–5 signs each) were constructed from the test set, each pairing a concatenated ASL gloss sequence with a reference English translation based on established ASL-to-English linguistic rules. The same corpus was used for all comparisons between the confidence-only baseline and the full SignBridge pipeline.

Figure 1

SignBridge Three-Component Architecture



Component 1: OpenHands-HD Sign Recognition

The first component of SignBridge takes a raw ASL video as input and applies four sign segmentation strategies in parallel to determine sentence boundaries: fixed time window, pause detection, confidence drop and semantic boundary detection, with a manual trigger available as an override for demonstration purposes. It then uses OpenHands-HD, which is built on the original OpenHands framework (Selvaraj et al., 2022) due to its extensibility, for isolated gloss recognition. OpenHands pioneered pose-based transformer recognition for sign language using a compact 27-keypoint MediaPipe representation, a fixed small transformer architecture, and runtime augmentation. While it demonstrated strong results on Indian Sign Language, several limitations became apparent for word-level ASL: the 27-point skeleton excludes individual finger joints critical for distinguishing confusable signs; runtime-only augmentation provides insufficient diversity for data-scarce training; and the fixed architecture lacks flexibility to adjust capacity for different dataset scales.

OpenHands-HD training pipeline (see Figure 2) addresses each of these limitations. It expanded the pose representation to 83 keypoints: 33 body pose landmarks, 21 keypoints per hand (42 total), and 8 face landmarks for mouth-vs-forehead sign disambiguation (see Figure 3). Each keypoint with x, y, z coordinates plus 30 derived finger features yielded 279-dimensional feature vectors per frame. The 83-keypoint configuration was selected after systematic ablation across 27, 33, 54, 75, and 83 keypoints; a full-face mesh (468 points) added noise rather than signal.

To address dataset scarcity, a 50× pre-generated augmentation pipeline expanded 342 clips to 17,100 training samples via speed variation (±10–30%), random frame subsampling, interpolation, rotation (±10°), scaling (0.8–1.2×), and noise injection. A transformer encoder was trained on 30-frame pose sequences. Due to the limited dataset, a smaller 3-layer architecture (128 hidden, 8 heads) generalized best — the larger 6-layer model overfit.

Figure 2

OpenHands-HD Training Pipeline and Sample Pose Estimation

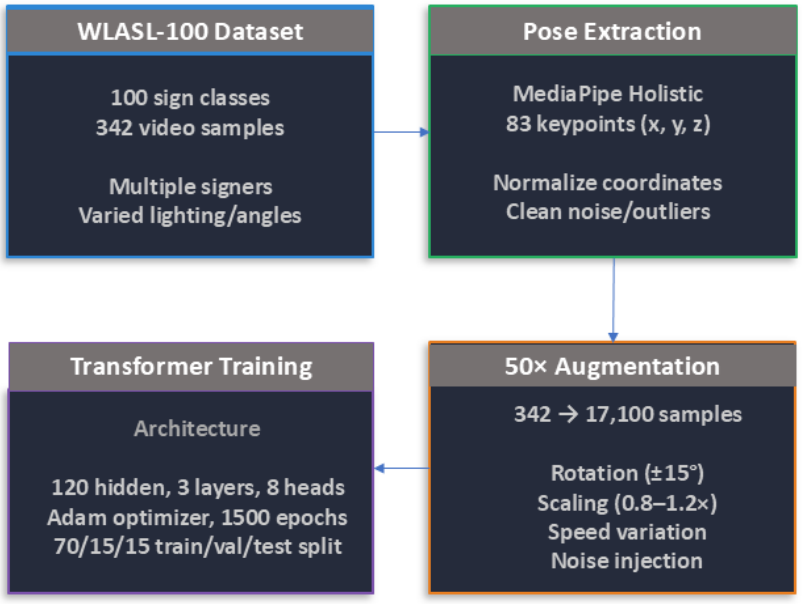
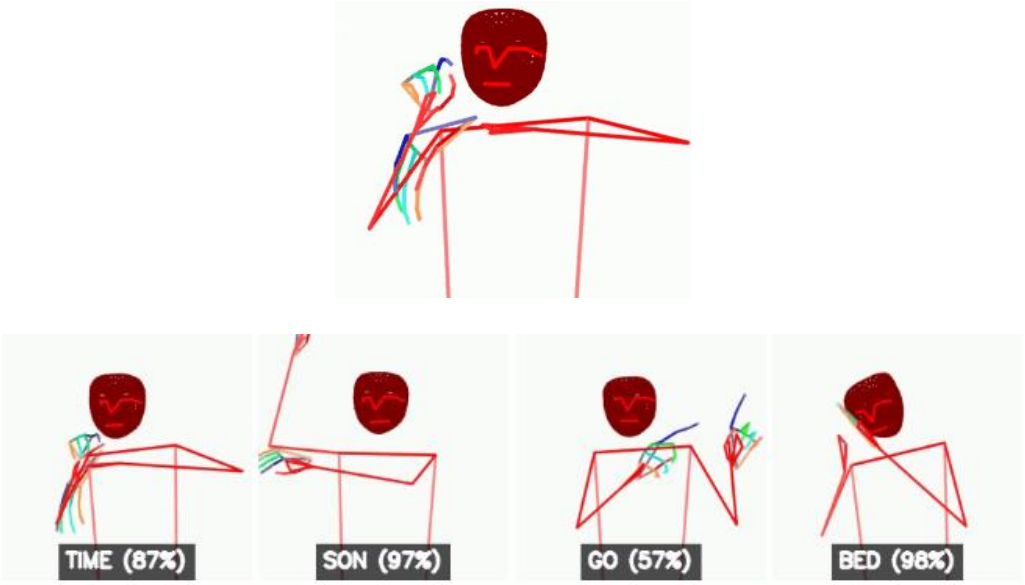


Figure 3

Sample Pose Extraction for intended sentence "It is time for my son to go to bed"



Component 2: LLM Semantic Coherence Translation

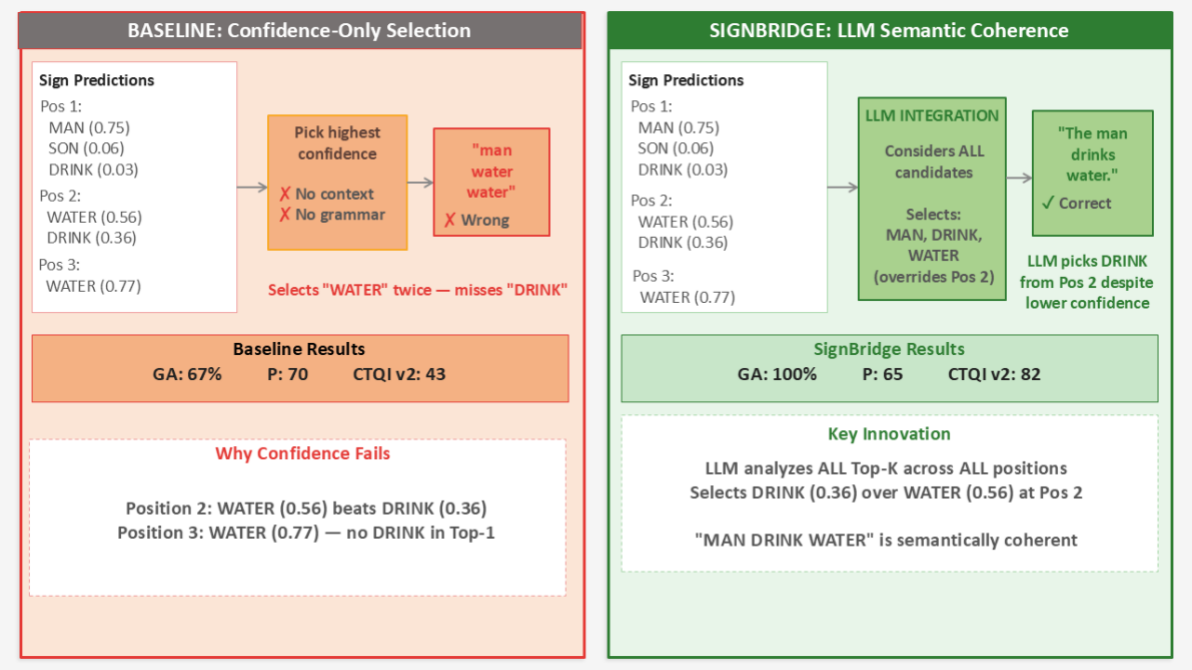
Gemini 2.0 Flash (gemini-2.0-flash) was selected as the LLM component after evaluating alternatives. Key criteria were inference latency (target <500 ms for real-time use), instruction-following reliability for structured JSON output, and zero cost during the experimental API tier — a practical constraint for an independent research setting, that eliminated GPT-4o and Claude Sonnet. Gemini 2.0 Flash achieved ~200–400 ms per request with strong structured-output compliance, making it the optimal tradeoff for the real-time pipeline requirements. A local confidence-only fallback activates if the API is unavailable.

Instead of greedily selecting the highest-confidence sign at each position, Component-2 uses the full Top-3 candidate pool with confidence scores (see Figure 4). The LLM simultaneously selects the semantically optimal sign from each pool based on sentence-level meaning and restructures ASL topic-comment order into English SVO structure. SignBridge's prompt

constrains the model to only select signs from provided candidates (preventing hallucination), using JSON-structured output with schema enforcement to guarantee parseable responses.

Figure 4

Baseline Confidence-Only vs. SignBridge LLM Semantic Coherence



Component 3: Composite Translation Quality Index (CTQI) Evaluation Framework

CTQI evolved through four versions. The table below summarizes the progression, formula, and the specific finding that drove each revision (see Table 1).

Table 1

CTQI Design Iteration — Four Versions Driven by Human Validation Feedback

Version	Formula	Problem Identified

v0	CTQI = α×BLEU + β×BERTScore + γ×Quality	Additive: Introduced Quality (based on GPT-2 Perplexity), but one high score masks a low one
v1	CTQI = α ×GA + β ×Quality + γ ×PTR	Still additive; introduced Gloss Accuracy (GA) and Perfect Translation Rate (PTR) but partial scores still compensate for failures
v2	CTQI = (GA/100) × (CF1/100) × (α + (α-1) ×P/100) × 100	Multiplicative; Introduced Coverage F1 (CF1) and Plausibility (P, LLM-driven) with a floor of 'α ', but human validation showed v2 correlated worse than v1: high P masked low GA
v3	CTQI = (GA/100) × (CF1/100) × (α + (α-1)×P/100×GA/100) × 100	P contribution scaled by GA; r=0.9427 with human raters — highest of all versions

CTQI v3 = (GA/100) × (CF1/100) × (α + (α-1) × P/100 × GA/100) × 100 where GA = Gloss Accuracy (% signs correctly recognized), CF1 = Coverage F1 (content word overlap between output and reference), P = Plausibility (LLM-assessed grammatical and semantic quality, 0–100), and α = floor parameter (validated at 0.50, robust across 0.2–0.6). The multiplicative structure ensures all components must perform well; any single failure collapses the composite score. Critically, scaling P by GA ensures that grammatically fluent but

semantically wrong translations — the failure mode that v2 missed — cannot artificially inflate the composite score.

The table below shows a representative subset of the n=53 evaluation sentences, selected to illustrate the range of outcomes: high-performing translations, cases where recognition errors limit the score despite fluent output, and the minority of cases where the baseline holds or ties (see Table 2).

Table 2

Representative Subset of CTQI v3 Evaluation (n=12 of 53)

#	ASL Glosses	Model Output	GA	CF1	P	Raw	Model	CTQI Δ	W
1	MAN DRINK WATER	The man drinks water.	100	100	65	45	82	+37	M
2	SON PLAY BASKETBALL	My son plays basketball.	100	100	87	68	93	+25	M
6	GIVE APPLE MAN	Give the apple to the man.	100	100	99	26	100	+73	M
11	AFRICA HOT	Africa is hot.	100	100	79	18	90	+71	M

9	MAN DRINK WATER THURSDAY	The man drinks water on Thursday.	100	100	74	42	87	+45	M
3	FAMILY ENJOY THANKSGIVING	I enjoy basketball on Thanksgiving.	67	75	56	30	39	+8	M
4	DOCTOR GIVE HELP	Give time to help.	67	67	98	34	44	+9	M
16	AFRICA YEAR BEFORE	Africa needs help before.	67	40	41	0	19	+18	M
32	TALL MAN DRINK WATER	The man drinks water from the water.	25	86	89	22	20	−1	R
17	BASKETBALL PLAY TIME	Basketball play time.	100	100	30	87	65	−22	R
7	STUDY BEFORE PLAY	Study before play.	100	100	60	80	80	0	T
5	WATER HOT DRINK	The hot water is water.	67	80	90	50	51	+0	T

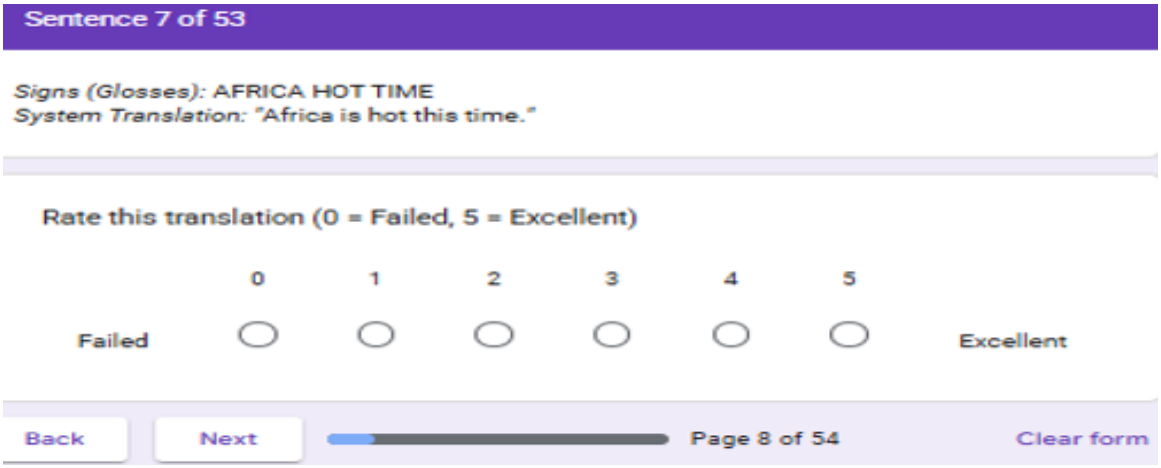
The Princeton Journal of Interdisciplinary Research
ISSN 3069-8200

Human Validation Study

This study validated CTQI v3 against human judgment, where five independent raters (ages 15–60, varying ASL familiarity — from no prior knowledge to intermediate) each evaluated all 53 sentence pairs. For each pair, raters were shown the original ASL gloss sequence alongside the SignBridge-generated English sentence and asked to rate the quality on a 0–5 scale (0 = completely failed translation, 5 = excellent). The survey was administered through a structured online form with one sentence pair per page to prevent rater fatigue and cross-contamination between items (see Figure 5).

Figure 5

SignBridge Human Validation Survey



Results

The charts below place OpenHands-HD in context against existing methods on WLASL-100 (see Figure 6) and show the per-sentence CTQI v3 improvement distribution across all 53 test sentences (see Figure 7). The baseline OpenHands replication trained on the raw WLASL-

100 dataset — averaging approximately 3 training samples per class after splitting — achieved 41.23% Top-1 accuracy, consistent with expected overfitting under severe data scarcity. OpenHands-HD's 50× pre-generated augmentation strategy expanded the effective training set from 342 to 17,100 samples, yielding 80.97% Top-1 accuracy and 91.62% Top-3 accuracy using only skeletal pose data — matching video-based SOTA while running in real time on consumer hardware. In Figure 7, 49 of 53 sentences (92%) sit above the diagonal, confirming the LLM improves translation quality across nearly the full test set. The mean CTQI shifted from 45.5 to 74.0 (gold star). The 4 grey points below the diagonal are failure cases analyzed in the error taxonomy.

Figure 6

OpenHands-HD Baseline Comparison (80.97% Top-1 matches video SOTA (81.38%) using only pose data)

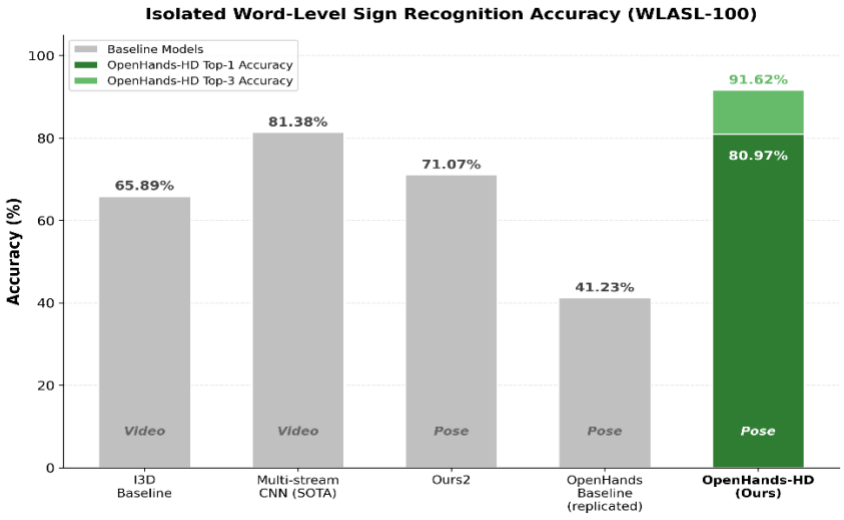
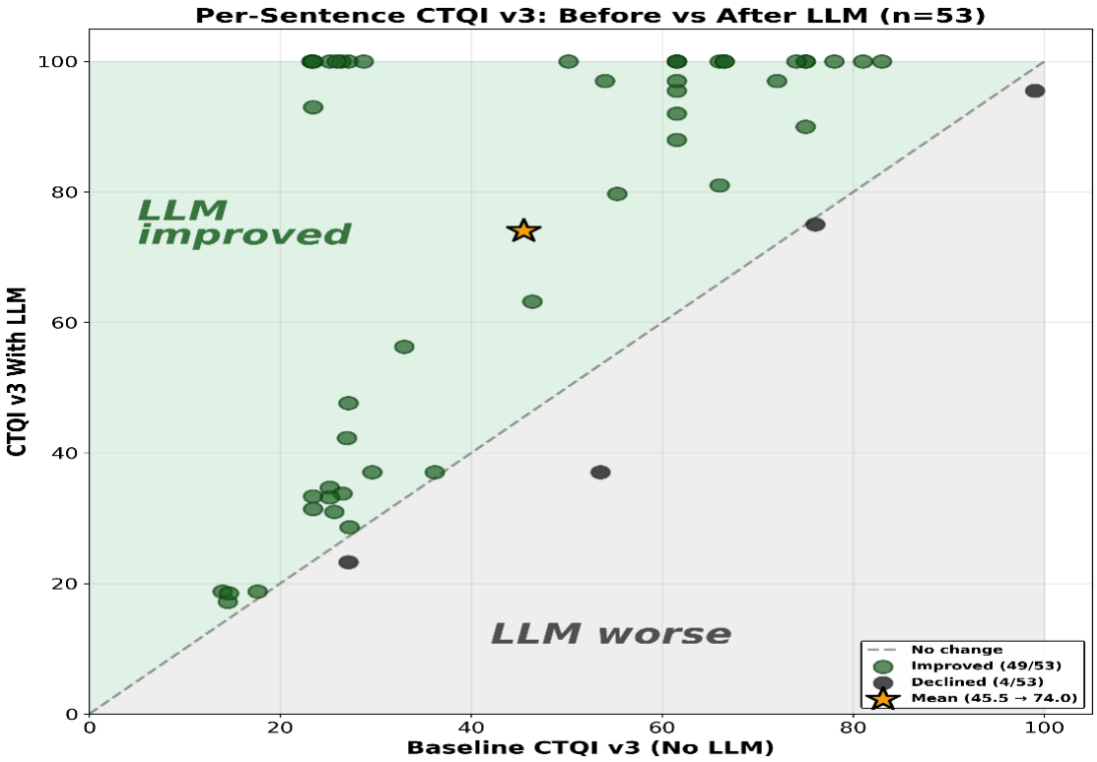


Figure 7

CTQI v3 Improvement on 49 out of 53 sentences in test dataset



The ablation study below tracks recognition accuracy as each component is added to the pipeline, isolating the individual contribution of keypoint expansion, augmentation scale, and fingermark integration. Starting from the replicated OpenHands baseline (27 keypoints, 41.23% Top-1), each configuration adds one variable. The final OpenHands-HD configuration (83 keypoints + 50× augmentation) reaches 80.97% Top-1 and 91.62% Top-3 — matching video-based SOTA using only skeletal pose data. The corresponding data is shown in the table below the chart (see Figure 8 and Table 3).

Figure 8

Ablation Study — Progressive Component Contributions

The Princeton Journal of Interdisciplinary Research
ISSN 3069-8200

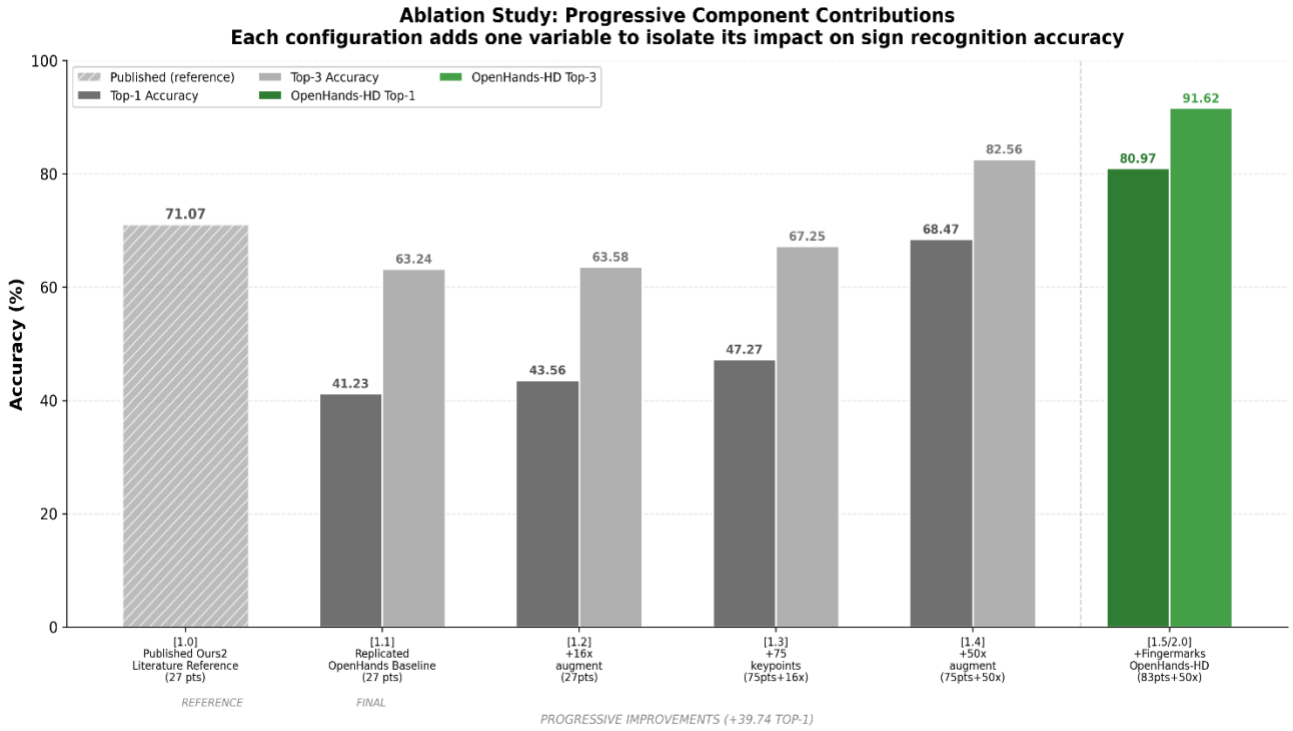


Table 3

Ablation Study Data — Progressive Contributions of Keypoint Expansion, Augmentation Scale, and Fingermark Integration

Config	Keypoints	Augmentation	Architecture	Top-1	Top-3
[1.0] Published Ours2 (Literature Reference)	27	Runtime	Fixed small	71.07%	—
[1.1] OpenHands Baseline (replicated)	27	16×	3-layer	41.23%	63.24%
[1.2] + 16× augmentation	27	16×	3-layer	43.56%	63.58%

[1.3] + 75 keypoints	75	16×	3-layer	47.27%	67.25%
[1.4] + 50× augmentation	75	50×	3-layer	68.47%	82.56%
[1.5/2.0] + Fingermarks (OpenHands-HD)	83	50×	3-layer	80.97%	91.62%

Statistical Analysis

Results are organized around the eight research questions tested in this study. Each finding is presented with the supporting statistical chart.

Q1: Is the 83-keypoint, fingermarks and 50x augmentation improvement collectively significant?

Figure 9(a) shows that expanding from 27 to 83 keypoints, along with fingermarks and 50x data augmentation, improved Top-1 accuracy from 41.23% replicated baseline to 80.97% (+39.74 pp, $z=18.23$, $p<0.0001$). This also outperformed with statistical significance over Ours2 (literature reference) model at 71.07% (+9.9 pp, $z=4.04$, $p<0.001$). See the ablation study above (Table 3) for the per-configuration breakdown.

Q2: Is Top-3 prediction accuracy significantly higher than Top-1?

Figure 9(b) shows that the model identifies the correct sign in its Top-3 predictions 91.62% of the time vs. 80.97% for Top-1 (McNemar's test, $p<0.001$). This 10.7 pp gap is the recovery window exploited by LLM contextual selection.

Figure 9

Keypoint Expansion Accuracy and Top-3 Recoverability

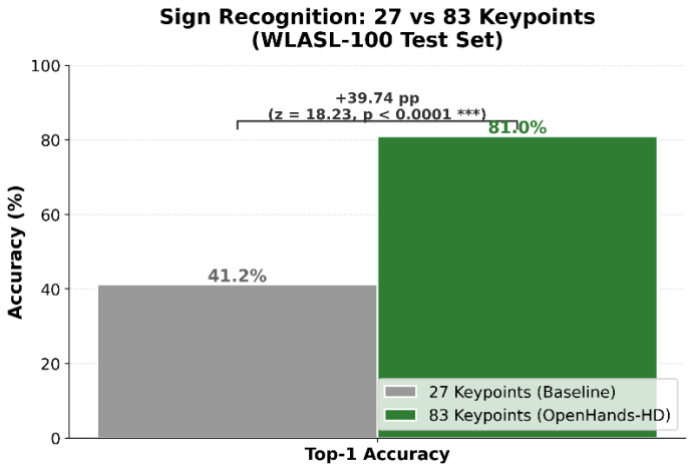


Figure 9a: 27 vs. 83 Keypoints — Top-1 accuracy (z=18.23, p<0.0001)

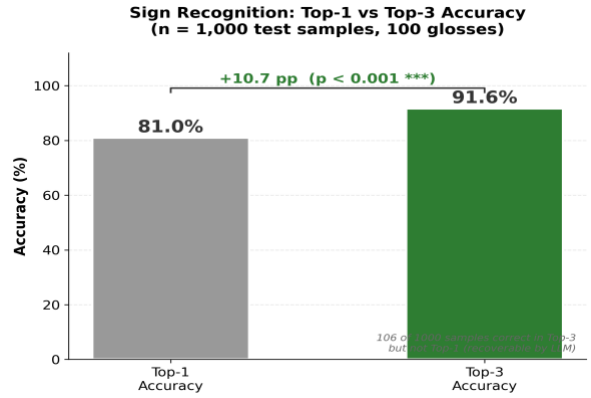


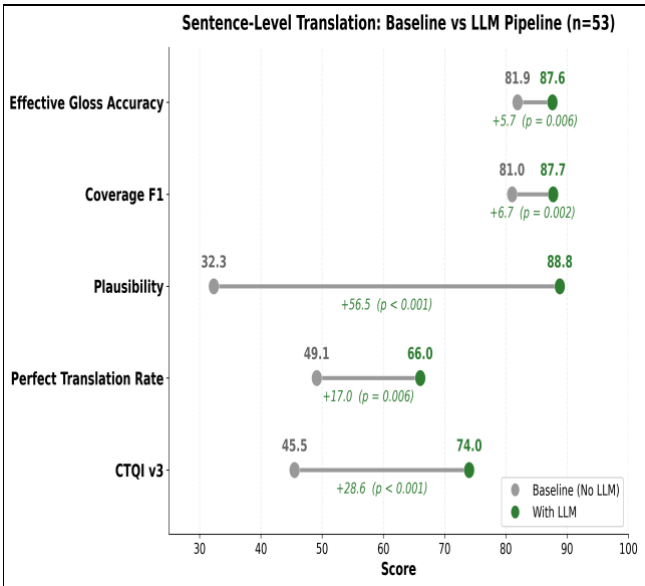
Figure 9b: Top-1 vs. Top-3 Accuracy — 106/1,000 samples recoverable by LLM

Q3: Will semantic coherence significantly improve translation quality?

All five translation quality metrics improved significantly with LLM contextual selection (all $p<0.01$). CTQI v3 increased from 45.5 to 74.0 ($p<0.001$, Cohen's $d=1.53$). Plausibility saw the largest gain (+56.5 pts), reflecting the transformation from ASL gloss sequences to grammatical English sentences (see Figure 10).

Figure 10

Translation Metrics: Baseline vs. SignBridge Pipeline



Metric	Baseline	SignBridge	Change	p
CTQI v3	45.5	74.0	+28.5	<0.001
Gloss Accuracy	81.9	87.6	+5.7	0.006
Plausibility	32.3	88.8	+56.5	<0.001
Coverage F1	81.0%	87.7%	+6.7pp	0.002
Model wins	4/53	49/53	92%	

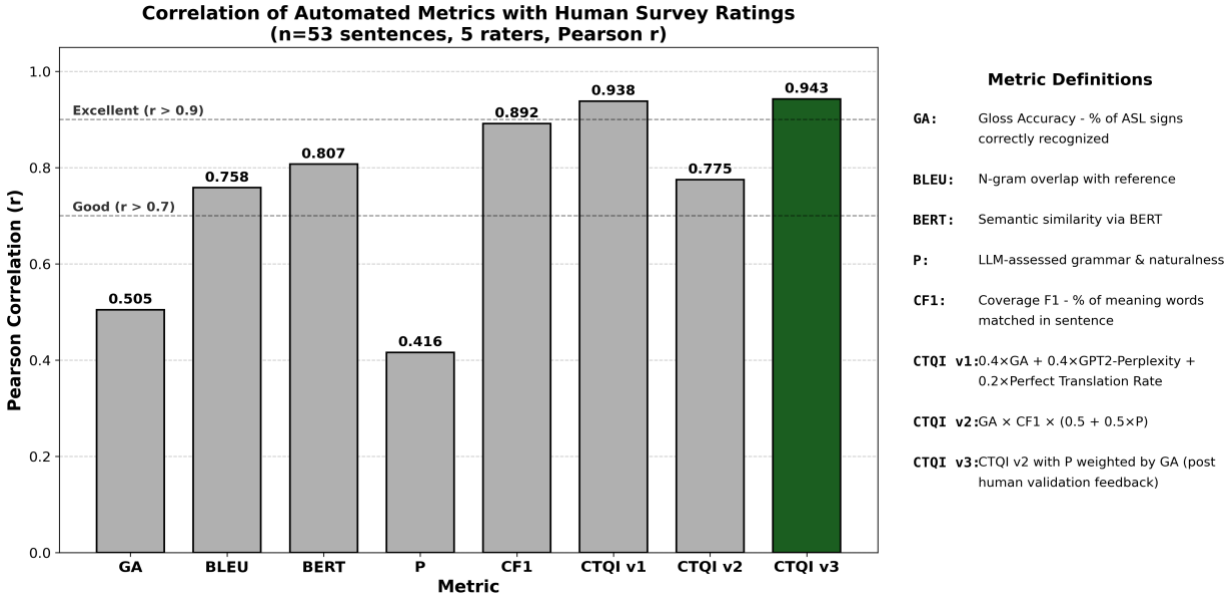
Q4: Does CTQI v3 correlate better with human ratings than BLEU/BERTScore?

CTQI v3 achieved r=0.94 with human raters, significantly outperforming both BLEU (r=0.76) and BERTScore (r=0.81) via Hotelling-Williams test (both p<0.0001) (see Figure 11).

Figure 11

Automated Metric Correlation with Human Survey Ratings

The Princeton Journal of Interdisciplinary Research
ISSN 3069-8200

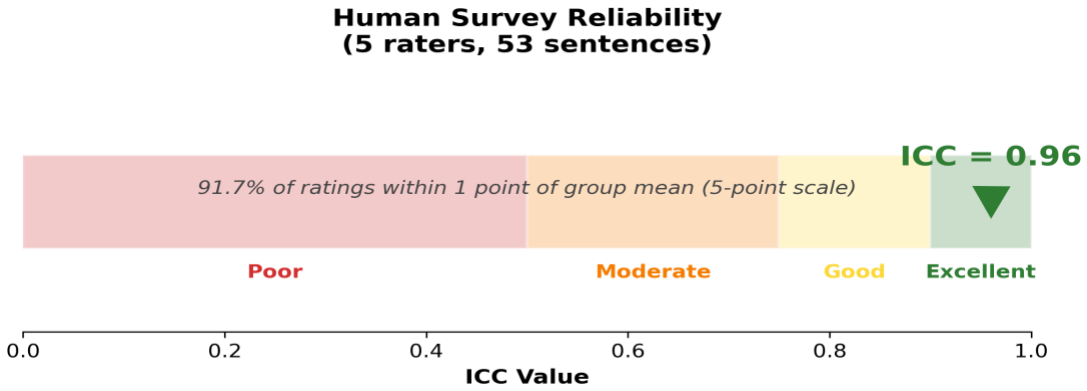


Q5: Does inter-rater agreement support using mean rating as ground truth?

Five independent raters achieved ICC=0.96 (“Excellent”), with 91.7% of individual ratings within 1 point of the group mean on a 5-point scale, validating averaged human scores as reliable ground truth (see Figure 12).

Figure 12

Inter-Rater Reliability

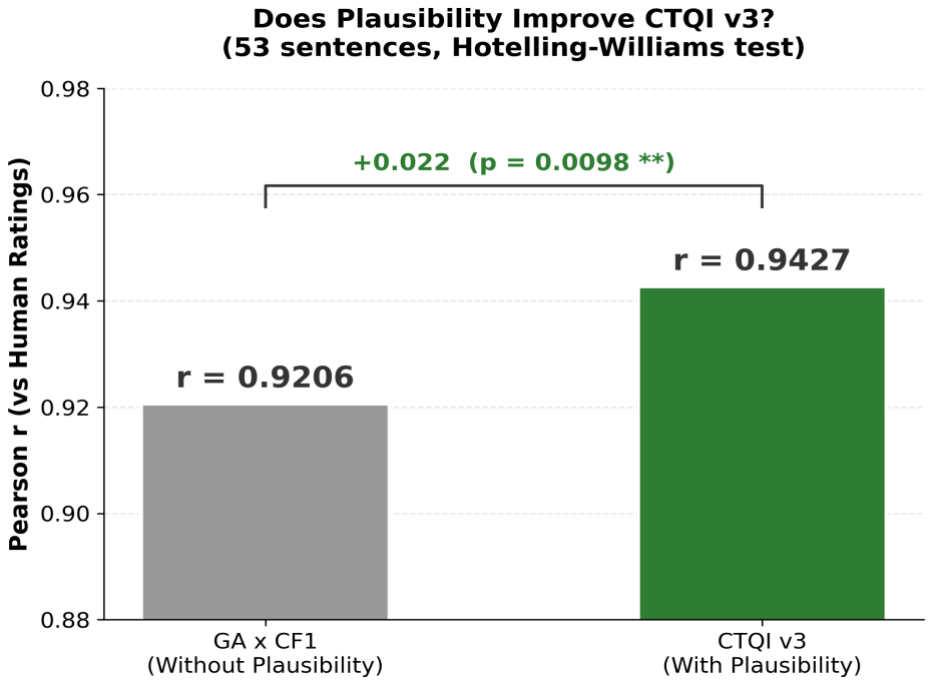


Q6: Does Plausibility improve CTQI v3?

Including Plausibility in CTQI v3 achieved r=0.9427 correlation with human validation, compared to r=0.9206 for CTQI v3 without Plausibility. This difference +0.022 was statistically significant per Hotelling-Williams test (p=0.0098), thereby confirming that Plausibility adds incremental value to the CTQI evaluation framework (see Figure 13).

Figure 13

Does Plausibility Improve CTQI v3?

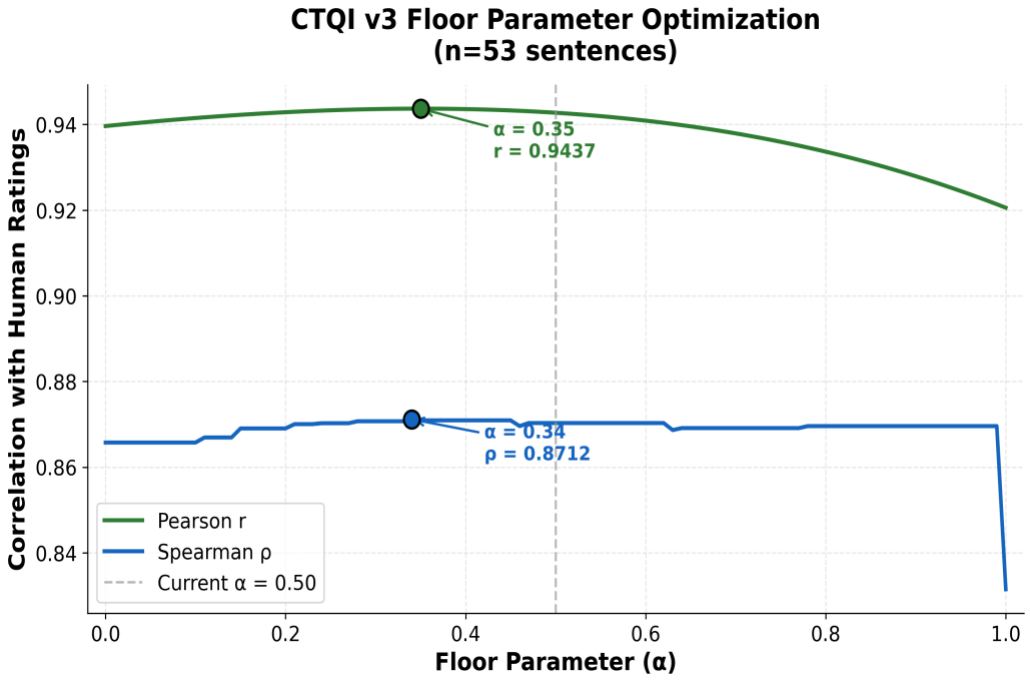


Q7: Is the floor parameter α=0.50 in the CTQI v3 formula robust?

A grid search across α ∈ [0,1] shows a broad plateau from α=0.2–0.6 in both Pearson r and Spearman ρ. The optimal α=0.35 (r=0.9437) improves on the selected α=0.50 by only +0.001, confirming the parameter choice is robust and not sensitive to small variations (see Figure 14).

Figure 14

CTQI v3 Floor Parameter Optimization



Q8: Is it possible to classify errors to learn and recover from?

Of 164 total gloss positions across the 53 sentences, 146 (89%) were correctly identified. The 18 errors split across four categories, directly mapping to targeted improvements: more training data (Cat. A), improved LLM prompting (Cat. B/C), and confidence-threshold tuning (Cat. D). (see Figure 15)

Figure 15

Gloss-Level Outcomes and Error Category Breakdown

Figure 13a:

Gloss-Level Outcomes (164 glosses, 89% correct)

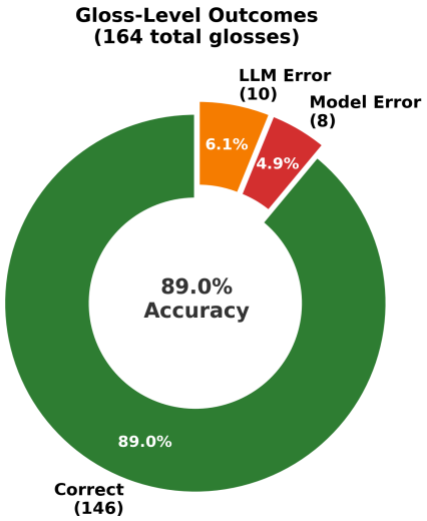
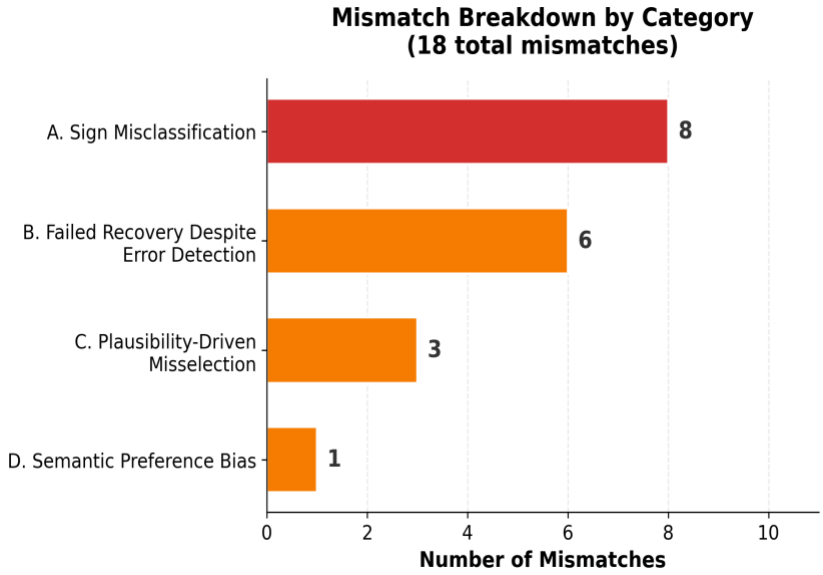


Figure 13b:

Mismatch Breakdown by Category (18 total)



The Princeton Journal of Interdisciplinary Research
ISSN 3069-8200

Discussion

Interpretation of Results

The results confirm all three hypotheses. For H1, the +39.74 pp keypoint improvement is consistent with the literature's finding that finger articulation and additional keypoints (along with data augmentation) are primary disambiguation signals — yet prior pose-based systems, including OpenHands, systematically omit them. SignBridge challenges the assumption that high accuracy requires computationally expensive video processing, as OpenHands-HD matches video-based SOTA (81.38%) using only skeletal data. For H2, the +28.5 CTQI and +56.5 Plausibility gains confirm that ASL's topic-comment grammar cannot be bridged to English SVO by confidence-only selection; it requires sentence-level reasoning. Gong et al. (2024) established LLMs' capacity for sign language translation end-to-end from video; SignBridge extends this to pose-based Top-K candidate disambiguation with structured prompting. For H3, CTQI v3's r=0.94 vs. BLEU's r=0.76 and BERTScore's r=0.81 quantifies the gap single-metric evaluation leaves: borrowed spoken-language benchmarks treat translation quality as one-dimensional, masking whether improvements are genuine or single-metric artifacts.

Failure Scenario Analysis (18 mismatches of 164 glosses, 11%)

Of 53 test sentences, 4 (8%) showed CTQI regression: 2 attributable to gloss-level errors, 2 reflecting trivial phrasing differences (−3.5 and −1.0 points). Errors were classified into four root categories:

A. Sign Misclassification (8 cases, 44.4%)

Correct sign never appears in Top-3 — unrecoverable by LLM. All 8: LATER misclassified as DRINK (45.1%). Example: LATER, GIVE, JACKET → "I drink and give a jacket" (P=100) instead of "Give the jacket later". This establishes that targeted data collection for confusable sign pairs is the single highest-leverage improvement available.

B. Failed Recovery Despite Error Detection (6 cases, 33.3%)

Correct sign in Top-3 at very low confidence; wrong candidate produces implausible output, but LLM still fails to recover. Example: NEED at 0.8% behind BOWLING at 91.5% → "The doctor is bowling to help" (P=7). This shows that, while the LLM architecture is sound, improved prompting strategies (chain-of-thought, multi-pass) are needed for recovery.

C. Plausibility-Driven Mis-selection (3 cases, 16.7%)

Wrong candidate produces an equally fluent sentence — no signal for correction. Example: COMPUTER (20.5%) loses to SON (67.2%) → "My son plays basketball" (P=100). This represents an inherent limit of plausibility-based disambiguation: when both candidates produce fluent sentences, no signal distinguishes correct from incorrect.

D. Semantic Preference Bias (1 case, 5.6%)

LLM overrides correct Top-1 for a concrete noun. Example: FINE (34.4%) overridden by PAPER (23.4%) → "The doctor has a chair and paper."

Notable Challenges and Limitations

The study was conducted on WLASL-100, which contains only 342 raw samples across 100 sign classes. This limited vocabulary is a proof-of-concept; real-world ASL communication requires thousands of signs. WLASL also has limited signer demographic diversity, and all

results are evaluated against researcher-constructed reference translations rather than Deaf community feedback at this point. Full-quality LLM output requires internet access; offline fallback to confidence-only selection degrades CTQI substantially. Several challenges shaped key design decisions during development, and some limitations remain open for future work.

Limited Training Data

With ~50 samples per sign, the 6-layer transformer overfit. The 3-layer model generalized best; 50× pre-generated augmentation was critical and could not be replaced by runtime-only variation.

Identifying Which Body Features Matter

Full 468-point face mesh added noise rather than signal. Empirical ablation across five keypoint configurations was required to identify the optimal 83-point set.

LLM Prompt Engineering

Early prompts hallucinated translations unrelated to the input signs. JSON schema enforcement and iterative prompt refinement were required to produce reliable, structured output.

Metric Validation Failure

CTQI v2 was theoretically sound but human validation showed it correlated worse with human judgment than v1. Empirical human feedback directly drove the redesign to v3 — a concrete example of human validation correcting theoretical assumptions.

Continuous Sign Segmentation

Determining where one sign ends and the next begins proved challenging. A four-strategy hybrid approach improved reliability, but variable signing speeds remain an active development area.

Future Research Roadmap

Delivered

This study makes the following primary contributions: (1) OpenHands-HD – enhanced pose recognition with 83-keypoint MediaPipe Holistic integration and 50× pre-generated augmentation, achieving 80.97% Top-1 and 91.62% Top-3 accuracy on WLASL-100; (2) LLM integration with Top-K contextual selection and ASL-to-English grammar restructuring produced sentence-level improvement in 92% of test sentences; (3) CTQI Evaluation Framework ($r=0.94$, $ICC=0.96$), outperforming BLEU and BERTScore; and (4) End-to-end real-time pipeline deployable on consumer-hardware, with API fallback for offline use.

Near-Term Improvements

Near-term development priorities include: (1) Targeted Data Collection for visually confusable sign pairs responsible for 44% of the errors; (2) Vocabulary Expansion from the 100-class proof-of-concept to 1,000 sign classes for meaningful real-world coverage; (3) Improved LLM prompt strategies using chain-of-thought and multi-pass prompting for edge cases; (4) Improved segmentation for continuous signing streams with variable speeds; (5) Facial Expression Recognition via non-manual grammatical markers — questions, negation, emphasis — which are grammatically significant in ASL; and (6) On-device recognition beyond desktop setup and formal usability testing with Deaf signers and educators across regional dialects.

Long-Term Vision

Long term goals include Bidirectional Translation with signing avatars and video output for full two-way communication between hearing and Deaf individuals, integration as a captioning plugin for video-conferencing platforms, and extension of the pipeline and frameworks to BSL, JSL, and 300+ sign languages worldwide without retraining the underlying recognition model.

Conclusions

SignBridge makes three distinct contributions to the sign language recognition field. First, the OpenHands-HD framework and its four-stage ablation study provide a reproducible methodology for expanding pose density under data-scarce conditions — applicable beyond ASL to any sign language with limited training resources. Second, the top-K LLM disambiguation architecture with JSON-constrained structured prompting offers a reusable design pattern for any recognition system that produces ranked candidates. Third, CTQI v3 and its four-version design evolution — with the human validation study showing that theoretically sound formulas can still diverge from human judgment, requiring empirical grounding — establishes a new standard for how ASL translation quality should be measured and reported. Beyond the research and laboratory, SignBridge is designed for low-barrier deployment on consumer hardware, with immediate applications in healthcare settings where interpreter unavailability creates critical communication gaps, and broader extensibility to legal, educational and employment contexts across 300+ sign languages worldwide.

References

Cao, Z., Hidalgo, G., Simon, T., Wei, S.-E., & Sheikh, Y. (2019). OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. IEEE TPAMI, 43(1), 172–186. arxiv.org/abs/1812.08008

Char, C. J., Miner, C. E., Wright, K., Carter, A., Akintunde, O., Cebulko, K., Jackson, J., & Lin, K. W. (2024). Understanding the health care needs of the Deaf community through medical interpreters. American Family Physician, 110(4), 349–350.

Gong, J., Foo, L. G., He, Y., Rahmani, H., & Liu, J. (2024). LLMs are Good Sign Language Translators. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 18362–18372.

Jiang, S., Sun, B., Wang, L., Bai, Y., Li, K., & Fu, Y. (2021). Skeleton Aware Multi-Modal Sign Language Recognition. IEEE CVPRW.

Klima, E. S., & Bellugi, U. (1979). The Signs of Language. Harvard University Press.

Koo, T. K., & Li, M. Y. (2016). A Guideline of Selecting and Reporting Intraclass Correlation Coefficients. Journal of Chiropractic Medicine, 15(2), 155–163.

Li, D., Rodriguez, C., Yu, X., & Li, H. (2020). Word-level Deep Sign Language Recognition from Video: A New Large-scale Dataset and Methods Comparison. IEEE WACV, 1459–1469. github.com/dxli94/WLASL

Maruyama, M., Tada, M., & Yanai, K. (2024). Word-level sign language recognition with multi-stream neural networks focusing on local regions and skeletal information. IEEE Access, 10, 85728–85742. https://doi.org/10.1109/ACCESS.2022.3198007

Moryossef, A., Müller, M., & Fahrni, R. (2021). pose-format: A Library for Viewing, Augmenting, and Handling Pose Files. github.com/sign-language-processing/pose

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). BLEU: A Method for Automatic Evaluation of Machine Translation. ACL 2002, 311–1318.

Valli, C., & Lucas, C. (2000). Linguistics of American Sign Language: An Introduction (3rd ed.). Gallaudet University Press.

Sanchez, C. E. (2025). Overcoming communication barriers to improve patient safety for American Sign Language users who are Deaf or hard of hearing. Patient Safety, 7(2). https://doi.org/10.33940/001c.138084

Selvaraj, P., Koller, O., Golber, L., Narayanan, S., & Anastasopoulos, A. (2022). OpenHands: Making Sign Language Recognition Accessible with Pose-Based Pretrained Models. arXiv:2110.05877. github.com/AI4Bharat/OpenHands

Stokoe, W. C. (1960). Sign Language Structure: An Outline of the Visual Communication Systems of the American Deaf. Studies in Linguistics, Occasional Papers 8.

Tannenbaum-Baruchi, C., Feder-Bubis, P., & Aharonson-Daniel, L. (2025). Communication barriers to optimal access to emergency rooms according to deaf and hard-of-hearing patients and health care workers: A mixed-methods study. Academic Emergency Medicine, 32(3), 246–259. https://doi.org/10.1111/acem.15037

World Federation of the Deaf. (2023). WFD Position Paper on Sign Language Rights and Technology. wfdeaf.org

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2020). BERTScore: Evaluating Text Generation with BERT. ICLR 2020. arxiv.org/abs/1904.09675