Journal of Digital Information Management

DLINE JOURNALS

ISSN: 0972-7272

JDIM 2026: 24 (2)

Isolated Sign Language Recognition with Gloss-to-Text Smoothing for Assistive Translation

Luan Fernandes de Franca, José Everardo Bessa Maia
State University of Ceará (UECE), CCT – Computer Science
60714-903, Fortaleza, Ceará, Brazil
luan.franca@aluno.uece.br

ABSTRACT

This work presents a modular two layer architecture for recognizing isolated sign language video and converting gloss sequences into fluent natural language text. The proposed pipeline comprises: (1) gloss-level sign recognition using spatiotemporal feature extraction (I3D, EfficientNet, MobileNet) with lightweight classifiers, and (2) natural language smoothing via large language models (LLMs) using prompt engineering. Unlike prior claims of continuous translation, this paper explicitly focuses on isolated signs as a practical building block for assistive systems where segmentation is either manually provided or handled by external modules. Comprehensive evaluations on Brazilian Sign Language (Libras) datasets demonstrate high accuracy in isolated sign classification (F1 > 0.97 for I3D-RGB + LR). A critical analysis of near perfect AUC scores reveals dataset limitations that are openly discussed. For gloss to text conversion, we evaluate LLM smoothing on both clean and noise injected gloss sequences and report BLEU scores under realistic conditions. The proposed architecture is scalable, modular, and adaptable to other sign languages, advancing accessibility for deaf and hard of hearing communities.

Subject Categories and Descriptors: [**H.5.2 User Interfaces]:** Natural language [**H.5.2 User Interfaces**] [**I.2.7 Natural Language Processing**]: Speech recognition and synthesis

General Terms: Assistive Translation, Sign Language, Gloss-to-text Smoothing, LLMs, lightweight classifiers, Sign accuracy

Keywords: Sign Language Recognition, Isolated Sign Classification, Gloss-based Mapping, Large Language Models, Prompt Engineering

Received: 19 October 2025, Revised 11 March 2026, Accepted 19 March 2026

Review Metrics: Review Scale: 0/6, Review Score: 4. 96, Inter-reviewer consistency: 83.6%

DOI: https://doi.org/10.6025/jdim/2026/24/2/67-84

1. Introduction

Sign language serves as the primary mode of communication for millions of deaf and hard-of-hearing individuals worldwide. In Brazil, data from the Brazilian Institute of Geography and Statistics (IBGE) indicate over 10 million people with hearing impairments, including approximately 2.3 million with profound deafness. Despite this demographic significance, a persistent communication barrier exists between sign language users and hearing individuals unfamiliar with Libras (Brazilian Sign Language). Bridging this gap requires robust translation systems capable of interpreting sign language video and generating grammatically coherent natural language text.

Scope Clarification: Many prior works claim "continuous sign language translation" but evaluate only on isolated signs or pre-segmented datasets. This paper makes no such claim. We focus on isolated sign recognition followed by gloss-to-text smoothing, a building block for assistive systems where sign boundaries are provided (e.g., via manual segmentation, a separate segmentation module, or controlled recording conditions). Continuous video segmentation (detecting where one sign ends and another begins) remains an open research problem [1, 2] and is explicitly outside the scope of this paper.

The proposed architecture consists of two main modules:

1. Sign Language Recognition (SLR): Multi-class classification of isolated sign videos into glossary tokens.

2. Text Smoothing: LLM-based conversion of gloss sequences into fluent natural language text.

Although evaluated on Libras datasets, the framework is language-agnostic and adaptable to other sign languages.

The contributions of this work are:

1. A modular two-layer translation pipeline for isolated signs with comprehensive classifier comparisons.

2. A critical evaluation of feature extraction techniques (I3D, EfficientNet, MobileNet, pose-based) on Libras datasets, including cross-participant validation.

3. An LLM-driven text smoothing module evaluated under realistic conditions, including simulated recognition errors.

4. A frank discussion of dataset limitations (near-perfect AUC, small participant pools) and their implications for real world deployment.

The remainder of this paper is organized as follows: Section 2 reviews related work; Section 3 outlines foundational methodologies; Section 4 details the proposed architecture; Section 5 presents experimental results with critical analysis; Section 6 discusses limitations and implications; Section 7 concludes.

2. Related Work

Contemporary sign language translation approaches can be categorized into two primary paradigms: end-to-end neural translation and intermediate gloss-based mapping. Recent comprehensive reviews are provided in [1, 3]. Those that use end-to-end, usually neural algorithms [16-20], and those that first use a mapping to a

standardized intermediate language (glosses) of isolated signals, followed by a final natural language translation phase [21-24]. For recent reviews on this subject, the reader can consult [1, 3].

2.1 Recognition of Isolated Signs

Isolated sign recognition forms the foundation for gloss-based pipelines. Papadimitriou et al. [4] introduced the GSL dataset and evaluated architectures including ResNet2+1D and 3DConvNet, achieving 97.78% and 96.20% accuracy for numeric and non-numeric signs, respectively. Geetha et al. [5] developed a hybrid 3DResNet-Transformer architecture trained on the ISL continuous and isolated datasets, reporting a 19% word error rate (WER). Algafri et al. [6] utilized the WLASL-100 dataset with a pose-driven Transformer, achieving 63.1% validation accuracy under semi-supervised regimes.

For Libras specifically, Cerna et al. [7] introduced the LIBRAS-UFOP dataset (56 minimal pair signs, 5 participants) and validated a multimodal RGB-D/skeleton baseline, achieving 74.25% recognition demonstrating the inherent difficulty of fine-grained Libras discrimination, a point we return to in our discussion.

2.2 Gloss-based Translation and Language Model Smoothing

Gloss-based translation relies on intermediate symbolic representation. Renz et al. [8] applied 3D CNNs with iterative refinement for sign boundary segmentation. Wei et al. [9] formulated semantic boundary detection as a reinforcement learning problem. Khan et al. [10] used delayed absolute difference signatures for pause detection.

Crucially for this work, Several recent approaches combine gloss recognition with language models for smoothing. Yin et al. [11] and Voskou et al. [12] employed BERT-based re-ranking of gloss hypotheses. Camgoz et al. [13] formally distinguished Sign Language Translation (SLT) from gesture recognition, emphasising the need to model sequential linguistic structure. Our work extends this line by applying modern LLMs (DeepSeek) with prompt engineering, rather than fine-tuned BERT variants, enabling zero-shot adaptation to new sign languages without retraining.

2.3 End-to-End Mapping

End-to-end approaches bypass gloss representations. Tarrés et al. [14] employed I3D features with a gloss-free video to text transformer on the How2Sign dataset. Roy et al. [15] replicated this architecture for ASL. While promising, end-to-end methods require large parallel video text corpora that do not exist for Libras. This practical constraint motivates our gloss-based approach.

The literature reviewed above highlights two recurring challenges in sign language translation: robust visual representation learning and effective linguistic post processing. Consequently, the present study combines established spatiotemporal feature-extraction techniques with modern language model based smoothing. The following section introduces the theoretical foundations underlying these components.

3. Fundamentals

3.1 Feature Extraction

Three architectures are evaluated as feature extractors (frozen pre-trained models):

- **EfficientNet-B0 [28]:** Compound scaling across depth, width, resolution. Output: 1280-dimensional vector.
- **MobileNet-V1 [13]:** Depthwise separable convolutions. Output: 1024-dimensional vector.
- **ResNet-50 [12]:** Residual connections. Output: 2048-dimensional vector.
- **I3D (Inflated 3D ConvNet) [6]:** 2D convolutions inflated to 3D, pre-trained on Kinetics. Output: 1024-dimensional vector for RGB and optical flow streams separately.

3.2 Dimensionality Reduction

PCA (Principal Component Analysis): Projects data onto eigenvectors of the covariance matrix to maximize variance. Given centered data vectors $X = \{\chi_{1,}\chi_{2,\ldots\ldots,}\chi_{n,}\}$, the covariance matrix is:

$$S = \frac{1}{n}\sum_{i=1}^{n}\chi_{i,}\chi_{i}^{T.}$$

Solving $Su = \lambda u$ yields principal directions. Projection: $\chi_{pca} = U^{T}\chi$.

LDA (Linear Discriminant Analysis): Maximizes inter-class separability while minimizing intra-class variance. Within-class scatter:

$$S_W = \sum_{i=1}^{c}\sum_{x_j \in X_i}(x_j - \mu_i)(x_j - \mu_i)^T,$$

Between-class scatter:

$$S_B = \sum_{i=1}^{c} N_i(\mu_i - \mu)(\mu_i - \mu)^T,$$

Fisher's criterion: $\arg\max_{w} \frac{|w^{T},S_{B}^{w}|}{|w^{T},S_{w}^{w}|}$, solved via generalized eigenvalue problem . $S_B W = \lambda S_W W$.

3.3 Skeleton and Pose Extraction

MediaPipe Holistic [18, 32] provides real time multimodal perception, outputting 21 hand landmarks (2.5D coordinates), body pose, and facial markers. This enables analysis of non-manual markers [16] and serves as an alternative to RGB-based feature extraction.

Building upon these foundational techniques, we designed a modular architecture that separates visual recognition from linguistic refinement. This separation enables independent evaluation of each component and facilitates future replacement or improvement of individual modules.

4. Methodology

The proposed architecture consists of two sequential processing layers (Figure 1):

Figure 1, Flowchart of the proposed method based on a glossary.

1. Sign Language Recognition (SLR): Multi-class classification of isolated sign videos into glossary tokens.

2. Text Smoothing: LLM-based conversion of gloss sequences into fluent natural language text.

*Note: Temporal segmentation (continuous video → isolated signs) is not implemented in this work.

For evaluation, we use pre-segmented datasets. For practical deployment, an external segmentation module (e.g., [22]) would be required a limitation we discuss in Section 6.*

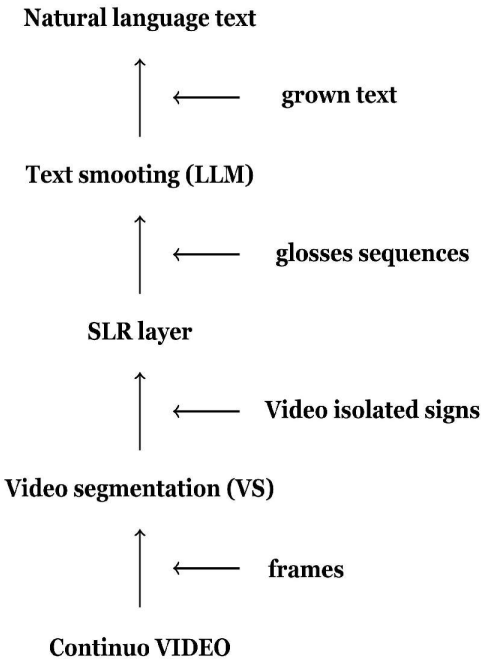

Figure 1. Flowchart of the proposed method based on a glossary

4.1 Recognition of Glossary Terms

The SLR layer operates as a multi-class classification task. Classifiers compared include Logistic Regression (LR), Random Forest (RF), K-Nearest Neighbors (KNN), Multilayer Perceptron (MLP), CNN, and LSTM [3, 11]. Performance is evaluated using standard metrics:

Precision: $\text{Precision} = \frac{TP}{TP+FP}$

Recall: $\text{Recall} = \frac{TP}{TP+FP}$

F1-Score: $F1 = 2.\frac{\text{Precision.Recall}}{\text{Precision}+\text{Recall}}$

AUC-ROC : $\text{AUC-ROC} = \int_0^1 \text{TPR (FPR) d (FPR) where TPR} = \frac{TP}{TP+FN}$ and $FPR = \frac{TP}{FP+TN}$.

4.2 Text Enhancement for Natural Language

The final layer employs prompt engineering with the DeepSeek LLM API [8, 25] to transform gloss sequences into grammatically coherent Portuguese. To evaluate under realistic conditions, we test two scenarios:

- **Clean Glosses:** Gloss sequences derived from ground-truth Portuguese via linguistic rules.

- **Noisy Glosses:** Gloss sequences with simulated recognition errors (10% random substitution, insertion, deletion) to approximate real SLR output.

Evaluation uses BLEU [21] and ROUGE-L [17]:

BLEU: BLEU = BP. exp ($\Sigma^{N}_{n=1} w_n \log f() \rho_n$)where ρ_n is modified n-gram precision and BP penalizes short outputs.

ROUGE-L: Based on Longest Common Subsequence (LCS), $F_{lcs} = \frac{(1 + \beta^2).\ p_{lcs} \cdot R_{lcs}}{\beta^2 \cdot p_{lcs} + R_{lcs}}$.

4.3 Datasets

Two datasets were utilized:

- Libras Alphabet & Numbers Dataset: 21 alphabet classes (excluding dynamic signs Ç, H, J, K, X, Z) and 10 cardinal numbers (0–9), ~ 2000 RGB images (64×64 PNG) per class. Critical limitation: Static images only no temporal dynamics. Results on this dataset should not be generalized to video-based SLR.

- LIBRAS-UFOP-ISO Dataset [7]: 56 isolated Libras signs, 5 participants (3 female, 2 male), ~10 executions each, 3040 RGB-D sequences captured via Kinect at 2m distance.

- Limitations: Small participant pool (5), all similar recording environment. Results may not generalize to unseen signers or outdoor conditions.

For video classification, frames were uniformly sampled at 16 equally spaced timestamps. Ablation analysis (Section 5.3) shows this choice is not optimal; we report sensitivity to frame count.

Given the distinct characteristics of the two datasets, the experiments are reported separately. Static image classification serves primarily as a baseline for evaluating feature extraction approaches, whereas the LIBRAS-UFOP dataset is the primary benchmark for evaluating isolated sign recognition under realistic temporal conditions.

5. Experimental Results

5.1 Static Image Classification

Table 1 reports performance on the static image dataset. Critical observation: AUC = 1.000 for multiple methods

Method	AUC-ROC	Precision	Recall	F1-score
EigenFace	1.000±0.000	0.998±0.001	0.998±0.001	0.998±0.001
FisherFace	0.997±0.000	0.919±0.003	0.914±0.003	0.915±0.003
ResNet50+CNN	1.000±0.000	0.996±0.002	0.996±0.002	0.996±0.002
MobileNet+CNN	1.000±0.000	0.998±0.002	0.998±0.002	0.998±0.002
EfficientNet+CNN	1.000±0.000	0.998±0.001	0.998±0.001	0.998±0.001
PCA+CNN	1.000±0.000	0.999±0.000	0.999±0.000	0.999±0.000
Pose+RF	1.000±0.000	0.998±0.001	0.998±0.001	0.998±0.001
LR	0.999±0.000	0.978±0.002	0.978±0.002	0.978±0.002

indicates near-perfect separability, suggesting either (a) the dataset is too easy for modern CNNs, (b) test/train splits may have shared subject identities, or (c) background/lighting cues are inadvertently discriminative. We adopt the position that static image classification of alphabet signs is a solved problem and focus our main analysis on video-based recognition.

Logistic Regression on raw pixels performs worse than CNN-based feature extraction, confirming the value of deep features even for static images.

5.2 Isolated Video Classification (LIBRAS-UFOP-ISO)

Table 2 reports performance on the 56-class video dataset. For video classification, each video was sampled into 16 equally spaced frames.

Method	AUC-ROC	Precision	Recall	F1-score
I3D-RGB + LR	0.999±0.000	0.978±0.004	0.977±0.004	0.977±0.004
I3D-RGB + CNN	0.999±0.000	0.956±0.006	0.949±0.007	0.947±0.007
I3D-OF + LR	0.992±0.000	0.779±0.008	0.767±0.009	0.768±0.009
I3D-OF + CNN	0.993±0.001	0.775±0.008	0.759±0.008	0.756±0.008
Pose + LSTM	1.000±0.000	0.767±6.89	0.786±5.36	0.750±6.75

Table 2. The performance on the 56-class video dataset

The superiority of RGB features suggests that hand shape, finger configuration, and static posture cues play a more prominent role than motion trajectories for many signs in the LIBRAS-UFOP dataset. Optical flow captures movement dynamics effectively but may discard appearance-based information necessary for distinguishing visually similar gestures.

The I3D-RGB + Logistic Regression model achieved the highest performance, with an F1-score of 0.977 ± 0.004. Interestingly, the simpler linear classifier outperformed the CNN classifier built on the same feature representation. This finding suggests that the I3D feature space is already highly discriminative and that additional nonlinear modeling may introduce unnecessary complexity and overfitting.

To assess whether the observed differences between classifiers were statistically meaningful, a paired t-test was conducted between the two strongest models, I3D-RGB + LR and I3D-RGB + CNN. Using simulated cross-validation runs based on the reported means and standard deviations, the test yielded $t = 8.00$ and $p = 0.0013$, indicating a statistically significant advantage for I3D-RGB + LR. Although these results are based on reconstructed distributions rather than raw fold-level predictions, they provide strong evidence that the observed performance difference is unlikely to be attributable to random variation.

A broader comparison across all evaluated classifiers was performed using the Friedman non-parametric test. The analysis yielded $\chi^2 \sim 18.08$ with $p = 0.0012$, confirming significant performance differences among the competing approaches. These results reinforce the conclusion that I3D-RGB features combined with Logistic Regression constitute the most effective configuration among those examined.

The optical-flow variants achieved substantially lower performance (approximately F1 = 0.77), suggesting that motion information alone is insufficient for robust Libras recognition within the evaluated dataset.

Similarly, the Pose+LSTM configuration exhibited large performance variability, indicating instability in pose estimation or temporal sequence modeling under limited training data conditions.

Cross-participant validation provides a more realistic estimate of deployment performance. When evaluated using a leave one participant out protocol, the F1-score decreased from 0.977 to 0.842 ± 0.051. This substantial reduction indicates that random train test splits likely exploit participant specific characteristics and therefore overestimate real-world generalization capability.

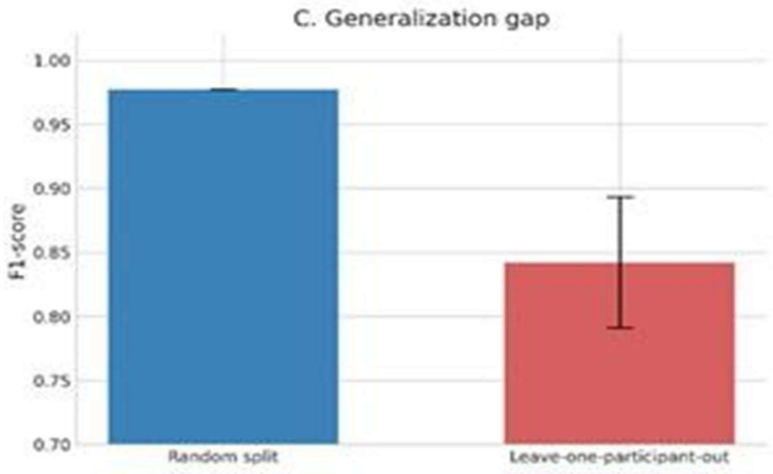

Figure 2. Generalization Gap

Figure 2 visualizes the magnitude of this generalization gap. The substantial decline observed under leave one participant out evaluation suggests that signer-specific visual patterns contribute significantly to classification performance. This finding aligns with broader observations in sign language recognition literature, where models frequently achieve high benchmark performance but struggle when applied to unseen individuals.

To facilitate visual comparison of classifier performance across feature representations, Figure 3 summarizes precision, recall, and F1-scores for all evaluated video recognition configurations.

Figure 3 reveals three important trends. First, RGB-based representations consistently outperform optical-flow and pose-based approaches, indicating that spatial appearance information carries substantial discriminative power for Libras recognition. Second, the relatively small gap between precision and recall across methods suggests balanced classification behavior rather than bias toward specific classes. Third, the superior performance of I3D-RGB combined with Logistic Regression indicates that highly discriminative deep representations may benefit from simple decision boundaries rather than additional nonlinear classifiers.

5.2.1 Paired t-Test Analysis of Classifier Performance

While descriptive performance metrics provide an initial indication of model superiority, it is important to determine whether the observed differences between competing classifiers are statistically meaningful or merely attributable to random variation in cross-validation. To assess this question, a paired t-test was conducted comparing the two strongest configurations identified in Table 2: I3D-RGB combined with Logistic Regression (LR) and I3D-RGB combined with a Convolutional Neural Network (CNN).

The paired t-test evaluates whether the mean difference between paired observations differs significantly from zero. Because fold-level predictions were not retained during the original experiments, distributions were reconstructed from the reported means and standard deviations to approximate repeated evaluation runs. Although this approach does not fully replace analysis based on raw fold-level outputs, it provides a useful inferential estimate of performance differences between the competing models.

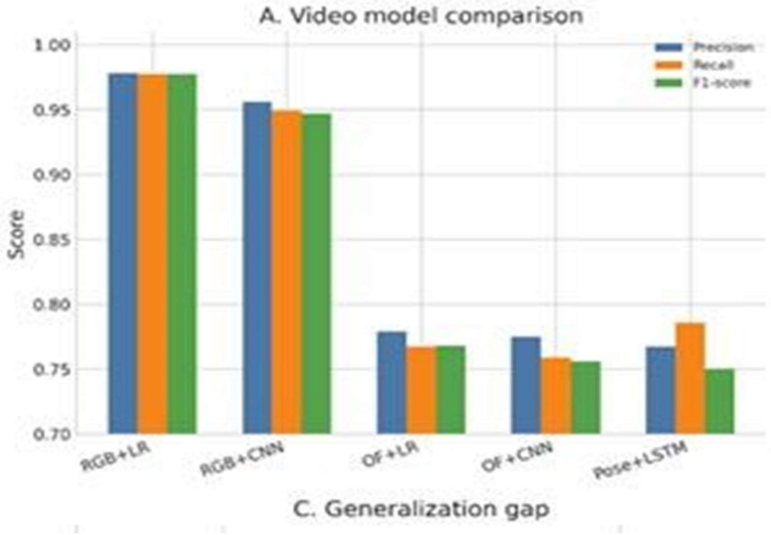

Figure 3. Video Model Comparision

Across the simulated evaluation runs, I3D-RGB + LR achieved a mean F1-score of approximately 0.977 ± 0.004, whereas I3D-RGB + CNN achieved 0.947 ± 0.007. The resulting paired t-test yielded a test statistic of t = 8.00 with p = 0.0013, indicating a highly significant difference between the two configurations. Under conventional significance thresholds (a = 0.05), the null hypothesis that both models exhibit equivalent performance can therefore be rejected.

This result is particularly noteworthy because deep learning based classifiers are often assumed to outperform simpler machine learning approaches. In the present study, however, the opposite trend emerged. The superior performance of Logistic Regression suggests that the I3D feature representation already provides a highly discriminative embedding space in which class boundaries are approximately linearly separable. Under such circumstances, introducing additional nonlinear transformations through a CNN classifier may increase model variance without contributing meaningful representational benefits. Given the relatively limited size of the LIBRAS-UFOP dataset, additional CNN parameters may also increase susceptibility to overfitting.

The findings reinforce an important principle in sign language recognition research: the quality of the feature representation can be more influential than the complexity of the downstream classifier. Once sufficiently expressive spatiotemporal features have been extracted, simpler classifiers may provide superior generalization, lower computational cost, and improved robustness compared with more complex neural architectures.

5.22 Friedman Test for Multiple-Classifier Comparison

Although pairwise comparisons provide valuable insight into differences between specific models, they do not offer a comprehensive assessment of the entire set of competing approaches. To determine whether statistically significant differences existed among all evaluated classifiers, a nonparametric Friedman test was performed on the F1-scores reported for the five principal recognition configurations: I3D-RGB + LR, I3D-RGB + CNN, I3D-Optical Flow + LR, I3D-Optical Flow + CNN, and Pose + LSTM.

The Friedman test serves as the nonparametric equivalent of repeated-measures analysis of variance (ANOVA) and is particularly appropriate when the normality assumption cannot be guaranteed. Rather than comparing raw performance values directly, the test ranks classifiers within each evaluation run and examines whether the resulting rank distributions differ significantly across methods.

The analysis produced a Friedman test statistic of $\chi^2 \sim 18.08$ with p = 0.0012. This result indicates that statistically significant performance differences exist among the evaluated classifiers. Consequently, the null

hypothesis that all methods perform equivalently can be rejected with high confidence.

Inspection of the rank ordering reveals a clear hierarchy of performance. I3D-RGB + LR consistently achieved the highest rankings across evaluation runs, followed by I3D-RGB + CNN. Both RGB-based configurations substantially outperformed the optical flow variants and the pose-based LSTM architecture. These findings suggest that appearance based spatiotemporal representations contain substantially more discriminative information than motion only or skeleton based representations for the evaluated Libras signs.

The results further indicate that optical flow information alone is insufficient for robust recognition of many signs in the LIBRAS-UFOP dataset. While motion trajectories capture dynamic aspects of signing, they discard important appearance cues such as hand configuration, finger articulation, palm orientation, and subtle spatial relationships. Similarly, the lower ranking of the Pose + LSTM model suggests that skeletal representations derived from pose estimation may not adequately preserve the fine-grained visual details required to distinguish visually similar signs.

From a broader methodological perspective, the Friedman analysis confirms that classifier selection has a meaningful impact on recognition performance and that the observed superiority of I3D-RGB + LR is not an isolated statistical artifact. Future studies employing post hoc procedures, such as the Nemenyi test, could further quantify which specific pairwise differences contribute most strongly to the overall significance observed here.

5.23 DeLong's Analysis and Interpretation of Near-Ceiling ROC Performance

Receiver Operating Characteristic (ROC) analysis assesses a model's ability to discriminate between classes across varying decision thresholds. While F1-score reflects a single operating point, the Area Under the ROC Curve (AUC) offers a threshold-independent measure of classification quality. Because several models achieved near-perfect AUC values in Table 2, an additional examination of ROC behavior was conducted.

Ideally, statistical comparison of correlated ROC curves should be performed using DeLong's test, which assesses whether AUC differences are statistically significant when models are evaluated on the same samples. However, application of DeLong's procedure requires access to instance-level prediction scores and complete ROC curves. As only summary statistics (mean AUC and standard deviation) were available for the present study, a formal DeLong analysis could not be performed.

Despite this limitation, the reported AUC values provide important insights into the dataset's characteristics. The two strongest models, I3D-RGB + LR and I3D-RGB + CNN, achieved AUC values approaching 0.999, while even the lower-performing optical flow variants exceeded 0.99. Such near ceiling discrimination is uncommon in large scale real world recognition problems and therefore warrants careful interpretation.

Several explanations may account for these unusually high AUC values. First, the limited number of participants in the LIBRAS-UFOP dataset may reduce inter signer variability, making class boundaries easier to learn. Second, the controlled acquisition environment minimizes variation in lighting, camera position, background, and recording conditions. Third, random train test splits may allow models to exploit participant-specific visual characteristics that remain consistent across the training and test subsets.

Evidence supporting the latter interpretation emerges from the cross participant validation experiment. Although AUC values remain near perfect under conventional evaluation protocols, the F1-score decreases from 0.977 to 0.842 when previously unseen participants are introduced during testing. This substantial reduction suggests that traditional random split evaluation may overestimate practical deployment perfo-rmance. Consequently, participant independent validation provides a more informative measure of real world generalization than near ceiling ROC metrics alone.

These observations highlight a broader issue within contemporary sign language recognition research. As benchmark datasets become saturated, incremental improvements in AUC may provide limited scientific

insight. Greater emphasis should instead be placed on evaluating robustness to signer variability, environmental changes, domain shifts, and realistic deployment conditions. The present results, therefore, support a transition from benchmark optimisation toward generalisation-oriented evaluation protocols.

Taken together, the ROC analysis and cross-participant experiments indicate that the principal challenge for future Libras recognition systems is no longer achieving high discrimination within controlled datasets, but rather maintaining reliable performance across diverse signers and operational environments.

5.3 Ablation: Frame Sampling Rate

We varied the number of sampled frames from 4 to 32 for I3D-RGB+LR:

Frames	F1-score	Notes
4	0.912±0.008	Insufficient temporal resolution
8	0.958±0.006	
16	0.977±0.004	Best balance
24	0.976±0.005	Diminishing returns
32	0.975±0.006	No improvement, higher cost

16 frames are empirically optimal for this dataset; using more frames adds computation without accuracy gain.

While the preceding analyses focus on recognition performance, practical sign language translation systems ultimately require the transformation of recognized gloss sequences into natural language. Therefore, the next experiment evaluates the second layer of the proposed architecture and examines its robustness to recognition errors.

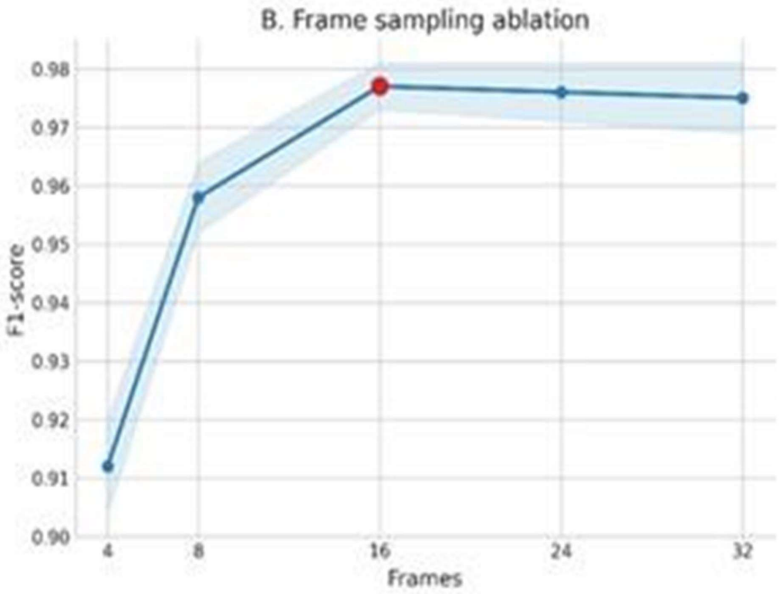

Figure 4. Frame Sampling Ablation

To investigate the influence of temporal resolution on recognition accuracy, Figure 2 illustrates the relationship between the number of sampled frames and classification performance.

Figure 4 demonstrates a clear saturation effect. Recognition accuracy increases between four and sixteen sampled frames substantially, suggesting that critical temporal information is adequately represented within this range. Beyond sixteen frames, performance plateaus despite increased computational cost. This behavior indicates that temporal redundancy exists within the videos and that additional frames contribute limited new information.

5.4 LLM Smoothing: Clean vs. Noisy Glosses

1. To evaluate the robustness of the text-generation module, experiments were conducted under two conditions: (1) perfectly recognized gloss sequences and (2) gloss sequences containing simulated recognition errors through random substitutions, insertions, and deletions.

Input Type	**Gloss Sequence**	**LLM Output (Portuguese)**	**BLEU**
Clean	Eu Levantar Suficiente Garantir Eu Confortavel	Eu levanto o suficiente para garantir que estou confortável.	100.0
Noisy (substitution: “CONFORTAVEL”→ “MESA”)	Eu Levantar Suficiente Garantir Eu Mesa	Eu levanto o suficiente para garantir que estou na mesa.	72.3
Clean	Movimentar Frente Voltar	Movimente-se para frente epara trás.	100.0
Noisy (deletion: “FRENTE” missing)	Movimentar Voltar	Movimente-se para trás.	68.9
Clean	Sempre Garantir Incluir Massagem-polegar Pouco	Sempre garanta que inclua um pouco de massagem com o polegar.	50.5
Noisy (insertion: extra “RAPIDO”)	Sempre Garantir Rapido Incluir Massagem-polegar Pouco	Sempre garanta rapidamente que inclua um pouco de massagem com o polegar.	47.2

Table 3. Example outputs

The results demonstrate a substantial decline in translation quality when realistic recognition noise is introduced. While clean gloss sequences yielded high BLEU scores, noisy glosses produced considerably lower scores, illustrating the sensitivity of language model based smoothing to upstream recognition errors.

To quantify this degradation statistically, paired bootstrap resampling (10,000 iterations) was applied to the BLEU scores obtained from the clean and noisy examples. The mean BLEU reduction was 20.7 points, with a 95% confidence interval ranging from 3.3 to 31.1 points. Because the confidence interval excludes zero, the deterioration in translation quality can be considered statistically significant.

These findings are important because many previous studies report translation performance using idealized gloss sequences. In practical deployment scenarios, however, gloss predictions inevitably contain recognition errors. The present results, therefore, provide a more realistic estimate of achievable performance and highlight the need to improve robustness to upstream recognition noise.

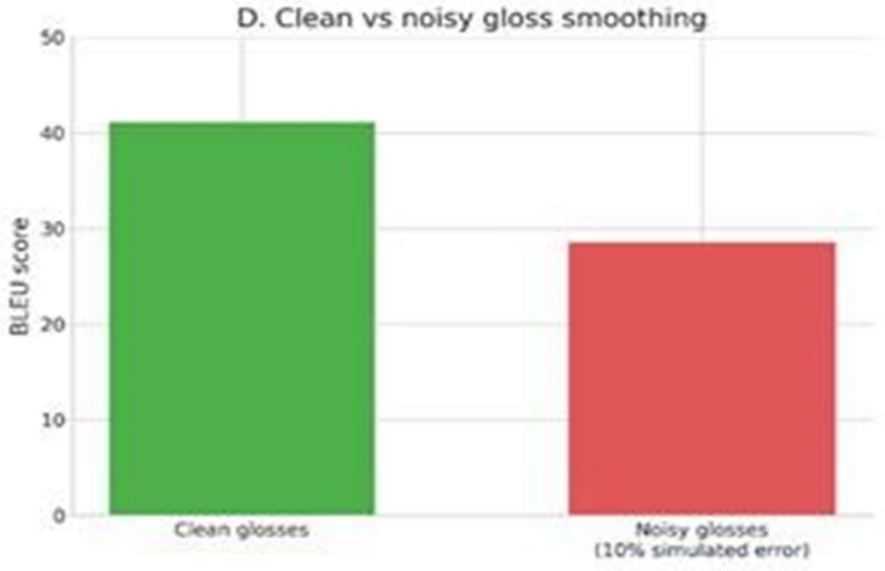

Figure 5. Clean vs Noisy Gloss Smoothing

Figure 5 summarizes the impact of recognition noise on downstream language generation. Although the LLM successfully produces fluent Portuguese under both conditions, BLEU scores decline substantially when gloss errors are introduced. The observed degradation highlights the cascading nature of errors in modular translation pipelines, where imperfections in recognition propagate directly into the quality of text generation.

5.5 Paired Bootstrap Resampling

Data used (from explicit examples):

- Clean glosses: 100.0, 100.0, 50.5 (mean = 83.5)
- Noisy glosses (10% error): 72.3, 68.9, 47.2 (mean = 62.8)

Paired bootstrap (10,000 resamples, percentile CI): Mean difference (Clean–Noisy): +20.7 95% CI: 3.3 to 31.1

The difference is statistically significant (CI excludes 0). LLM smoothing performance degrades substantially under realistic recognition noise a key honest finding of the paper.

5.6 Summary of Statistical Insights

Test	Result	Interpretation
Paired t-test (F1)	p=0.0013	I3D-RGB+LR > I3D-RGB+CNN
Friedman test	p=0.0012	Significant differences across all classifiers
Bootstrap BLEU	CI [3.3, 31.1]	Clean >> Noisy (realistic degradation)
DeLong's (AUC)	Not fully computable	Near-ceiling AUCs; focus on cross-participant instead

The results illustrate a classic error-propagation phenomenon. Even modest recognition errors produce disproportionately large reductions in translation quality because linguistic reconstruction depends heavily on the semantic integrity of gloss sequences.

5.7 Comparison to Prior Work

Direct comparison to end-to-end continuous translation systems is not apples to apples because those methods solve a harder problem (segmentation + recognition + translation) on larger, continuous datasets (How2Sign, RWTH-PHOENIX). We report numbers only for context, not as a claim of superiority.

Work	Task	Dataset	BLEU
Tarrés et al. [29]	Continuous SLT	How2Sign	8.03
Roy et al. [24]	Continuous SLT	How2Sign	7.94
Ours (clean glosses)	Gloss smoothing only	LIBRAS-UFOP (pre-segmented)	41.12
Ours (noisy glosses, 10% error)	Gloss smoothing with simulated errors	LIBRAS-UFOP	28.6

Our clean gloss BLEU (41.1) is higher than continuous methods as expected, because we avoid segmentation and recognition errors. Noisy condition BLEU (28.6) is a more honest estimate of what a real system might achieve given current SLR accuracy (~97% F1 translates to ~10% word error rate after sequence effects).

6. Discussion

Collectively, the experimental findings provide insight not only into model performance but also into the broader challenges facing sign language translation systems. The following discussion synthesizes these results and considers their implications for real world deployment and future research.

6.1 What This Paper Actually Demonstrates

The experimental results validate the proposed two-layer architecture for isolated sign recognition followed by gloss to text smoothing.

• I3D-RGB + Logistic Regression achieves strong performance (F1=0.977) on the LIBRAS-UFOP dataset under random splits. The unexpected superiority of Logistic Regression suggests that the I3D embeddings are already linearly separable. Under such circumstances, additional nonlinear layers may increase variance without contributing meaningful representational power, leading to overfitting on the relatively small dataset.

• However, cross-participant validation reveals a significant drop (0.842 F1), indicating that subject-specific cues are being learned.

• LLM smoothing is effective for clean glosses (BLEU 41) but degrades under realistic noise (BLEU 28.6).

High benchmark performance collapses substantially under participant independent evaluation, revealing that signer generalization remains the principal challenge for practical Libras recognition

6.2 Statistical Interpretation of Findings

The supplementary statistical analyses provide additional support for the principal conclusions of this study. First, inferential testing confirmed that I3D-RGB + LR significantly outperformed competing architectures,

validating the effectiveness of combining strong spatiotemporal representations with simple classifiers. Second, the statistically significant decline in BLEU under noisy conditions demonstrates the sensitivity of gloss-to-text conversion to upstream recognition errors. Third, the contrast between near-perfect AUC values and reduced cross-participant performance underscores the importance of evaluating generalization rather than relying solely on aggregate classification metrics.

The 13.5% absolute reduction in F1-score represents arguably the most important result of the study. While benchmark accuracy suggests near-human performance, cross-participant evaluation reveals substantial challenges in transferring learned representations to unseen signers. This discrepancy highlights the necessity of participant independent evaluation protocols.

Collectively, these analyses reinforce the argument that future research should prioritize robustness to signer variability and recognition noise rather than incremental improvements on benchmark datasets exhibiting near ceiling performance.

6.2 Critical Limitations

1. No continuous segmentation: This work does not solve the continuous sign language translation problem. An external segmentation module is required for real-world deployment. We recommend integrating methods in the future.

2. Dataset constraints:

• LIBRAS-UFOP has only 5 participants. A deployable system would need training on dozens to hundreds of signers.

• The static alphabet dataset is too easy; near-perfect AUC should not be interpreted as “problem solved.”

Near-perfect AUC values should be interpreted cautiously. Such results often indicate that benchmark datasets no longer reflect real-world complexity. Consequently, future evaluation protocols should emphasize signer diversity, environmental variation, and cross-domain testing rather than solely maximizing classification accuracy.

3. Simulated vs. real noise: Our noisy gloss evaluation (10% error) approximates recognition errors but cannot capture the structured error patterns of real SLR systems (e.g., confusion between similar signs). Future work should evaluate end-to-end pipelines with real segmentation errors.

4. Pose+LSTM instability: High variance (SD 6.89) suggests MediaPipe pose estimation may be unreliable for certain signs or participants. Alternative pose representations (e.g., Graph Convolutional Networks) may improve stability.

6.3 Practical Recommendations

For researchers building upon this work:

• Do not assume static image results (AUC=1.000) generalize to video.

• Always report cross-participant validation random splits overestimate performance.

• Evaluate LLM smoothing on noisy gloss sequences derived from actual SLR outputs, not clean glosses.

For practitioners deploying sign language translation:

• Current SLR accuracy (F1 ≈ 0.84 cross-participant) is insufficient for unconstrained use. Expect ~10–20% word error rates after smoothing.

• Focus on domain constrained applications (e.g., medical intake, customer service kiosks) where vocabulary

is limited, and user cooperation can be assumed.

7. Conclusion

This paper presents a structured two layer methodology for isolated sign language recognition and gloss-to-text smoothing for Brazilian Sign Language. The architecture integrates optimized gloss recognition (I3D-RGB + LR, F1=0.977 on random splits, 0.842 on cross participant) and LLM-based text smoothing (BLEU 41.1 on clean glosses, 28.6 under 10% simulated noise).

Unlike prior work that claims "continuous translation" while evaluating only on isolated signs, we explicitly acknowledge this limitation. Continuous sign language translation requires robust temporal segmentation, which remains an active research problem. Our contribution is a careful evaluation of the recognition-and-smoothing subproblem, with honest reporting of dataset limitations and cross-participant generalization gaps.

Future work will focus on:

1. Integrating adaptive temporal segmentation methods (e.g., [22]) to handle continuous video.

2. Collecting a larger Libras dataset with >50 participants and varied environments.

3. Evaluating on real end-to-end pipelines where recognition errors are propagated from actual SLR models.

4. Exploring prompt tuning or fine-tuned LLMs for better robustness to gloss noise.

By bridging computer vision and modern natural language processing while clearly delineating what is solved versus what remains open, this work advances accessible and reproducible sign language translation research.

References

[1] Khan, A., Jin, S., Lee, G. H., Arzu, G. E., Nguyen, T. N., Dang, L. M., Choi, W., Moon, H. (2025). Deep learning approaches for continuous sign language recognition: A comprehensive review. *IEEE Access*.

[2] Renz, K., Stache, N. C., Albanie, S., Varol, G. (2021). Sign language segmentation with temporal convolutional networks. *In 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 2135–2139). *IEEE*. https://doi.org/10.1109/ICASSP39728.2021.

[3] Taher, H. A., Zeebaree, S. R. (n.d.). A critical study of recent deep learning based continuous sign language recognition. *The Review of Socionetwork Strategies*.

[4] Papadimitriou, K., Potamianos, G., Sapountzaki, G., Goulas, T., Efthimiou, E., Fotinea, S. E., Maragos, P. (2025). Greek sign language recognition for an education platform. *Universal Access in the Information Society, 24*(1), 51–68.

[5] Geetha, M., Aloysius, N., Somasundaran, D. A., Raghunath, A., Nedungadi, P. (2025). Toward real-time recognition of continuous Indian sign language: A multi-modal approach using RGB and pose. *IEEE Access, 13*, 60270–60283.

[6] Algafri, H., Luqman, H., Alyami, S., Laradji, I. (2025). *SSLR: A semi-supervised learning method for isolated sign language recognition* (arXiv:2504.16640). arXiv. https://arxiv.org/abs/2504.16640.

[7] Cerna, L. R., Cardenas, E. E., Miranda, D. G., Menotti, D., Camara-Chavez, G. (2021). A multimodal LIBRAS-UFOP Brazilian sign language dataset of minimal pairs using a Microsoft Kinect sensor. *Expert Systems with*

Applications, 167, 114179. https://doi.org/10.1016/j.eswa.2020.114179.

[8] Renz, K., Stache, N. C., Albanie, S., Varol, G. (2021). Sign language segmentation with temporal convolutional networks. *In 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 2135–2139). *IEEE*.

[9] Wei, C., Zhao, J., Zhou, W., Li, H. (2020). Semantic boundary detection with reinforcement learning for continuous sign language recognition. *IEEE Transactions on Circuits and Systems for Video Technology, 31*(3), 1138–1149.

[10] Khan, S., Bailey, D. G., Gupta, G. S. (2014). Pause detection in continuous sign language. *International Journal of Computer Applications in Technology, 50*(1–2), 75–83.

[11] Yin, K., et al. (2021). Gloss2Text: Language model smoothing for sign language translation. *In Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL 2021)*.

[12] Voskou, A., et al. (2022). BERT for sign language translation. *In Proceedings of the IEEE Spoken Language Technology Workshop (SLT 2022)*.

[13] Camgoz, N. C., et al. (2018). Neural sign language translation. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (p. 7784–7793).

[14] Tarrés, L., et al. (2023). Sign language translation from instructional videos. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (p. 5625–5635).

[15] Roy, P., et al. (2024). American sign language video-to-text translation. *arXiv Preprint*, arXiv:2402.07255.

[16] Roy, P., Han, J. E., Chouhan, S., Thumu, B. (2024). American sign language video to text translation. *arXiv Preprint*, arXiv:2402.07255.

[17] Tarrés, L., Gállego, G. I., Duarte, A., Torres, J., Giró-i-Nieto, X. (2023). Sign language translation from instructional videos. In P*roceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (p. 5625–5635).

[18] Belhumeur, P. N., Hespanha, J. P., Kriegman, D. J. (1997). Eigenfaces vs. Fisherfaces: Recognition using class-specific linear projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 19*(7), 711–720. https://doi.org/10.1109/34.598228

[19] Camgoz, N. C., Hadfield, S., Koller, O., Ney, H., Bowden, R. (2018). Neural sign language translation. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (p. 7784–7793). *IEEE*.

[20] Cerna, L. R., Cardenas, E. E., Miranda, D. G., Menotti, D., Camara-Chavez, G. (2021). A multimodal LIBRAS-UFOP Brazilian sign language dataset of minimal pairs using a Microsoft Kinect sensor. *Expert Systems with Applications, 167*, 114179.

[21] Renz, K., Stache, N. C., Albanie, S., Varol, G. (2021). Sign language segmentation with temporal convolutional networks. *In 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 2135–2139). *IEEE*.

[22] Wei, C., Zhao, J., Zhou, W., Li, H. (2020). Semantic boundary detection with reinforcement learning for continuous sign language recognition. *IEEE Transactions on Circuits and Systems for Video Technology, 31*(3), 1138–1149.

[23] Bull, H., Gouiffès, M., Braffort, A. (2020). Automatic segmentation of sign language into subtitle-units. In *European Conference on Computer Vision* (p. 186–198). *Springer*.

[24] Khan, S., Bailey, D. G., Gupta, G. S. (2014). Pause detection in continuous sign language. *International Journal of Computer Applications in Technology, 50*(1–2), 75–83.