Mathematical Linguistics and Scalable Modeling for Real-Time ASL Translation

Alex Hernandez Juarez

September 2025

Abstract

We present a mathematically grounded framework for large-vocabulary American Sign Language (ASL) translation that scales to 5,000–10,000 signs while supporting real-time deployment on edge and web platforms. The approach factorizes signs into phonological subunits, models spatial discourse with a referential algebra, constrains sequences via automata/CFG and WFST decoding, and uses information theory to quantify the contribution of multimodal cues. We show how these ingredients reduce sample complexity, enable compositional generalization, and keep inference tractable for continuous signing.

1 Introduction

Real-time ASL translation remains challenging due to its multimodality, spatial grammar, and signer variability. Prior work often targets hundreds of isolated signs; scaling to 5k–10k signs and continuous discourse requires structure beyond larger neural nets. We argue for a *mathematical linguistics* core that (i) formalizes signs and spatial reference, (ii) enforces well-formedness with automata/grammars, and (iii) composes these with efficient decoders. We integrate this with a practical pipeline based on MediaPipe landmarks and lightweight sequence models suitable for edge and cloud.

Contributions. (1) A formal phonological alphabet and observation/quantization map with invariance guarantees; (2) a spatial discourse algebra with probabilistic uniqueness for pointing-based reference; (3) a scalable WFST decoding architecture ($H \circ C \circ L \circ G$) adapted from ASR; (4) an information-theoretic analysis linking added modalities to achievable error; (5) an evaluation protocol emphasizing vocabulary growth and deployment metrics.

2 ASL as a Formal Language

Let $X_t \in (\mathbb{R}^3)^m$ be m landmarks (hands/face/body) at frame t.

Definition 1 (Phonological Alphabet)**.** *Let $\Sigma_H, \Sigma_L, \Sigma_O, \Sigma_M, \Sigma_N$ denote finite alphabets for handshape, location, orientation, movement, and non-manual markers, respectively. A phonological sign is*

$$s = (H, L, O, M, N) \in \Sigma_H \times \Sigma_L \times \Sigma_O \times \Sigma_M \times \Sigma_N \ =: \ \Sigma.$$

Group action and equivariance. Let $G \subset \mathrm{Sim}(3)$ act on landmarks by $g \cdot X$. A feature map $\phi : (\mathbb{R}^3)^m \to \mathbb{R}^k$ is G-equivariant up to a representation $R : G \to \mathrm{GL}_k$ if

$$\phi(g \cdot X) = R(g)\, \phi(X).$$

A quantizer $q : \mathbb{R}^k \to \Sigma$ is R-invariant if $q(R(g)z) = q(z)$. Then $q \circ \phi$ is G-invariant.

Proposition 1 (Noise robustness)**.** *Suppose ϕ is L-Lipschitz and q is a nearest-prototype quantizer with margin $\gamma(\phi(X))$. If landmarks satisfy $\|X' - X\| \le \varepsilon$ and $L\varepsilon < \gamma(\phi(X))$, then $q(\phi(X')) = q(\phi(X))$.*

Quantizer consistency. If q_n is a k-means quantizer learned from n samples of $\phi(X)$, then under standard regularity $q_n \to q^\star$ a.s., and downstream risk gap satisfies $\mathcal{E}(q_n) - \mathcal{E}(q^\star) = \mathcal{O}_p(n^{-1/2})$.

Temporal features. Let $f_t = \phi(X_t)$; define $\Delta f_t = f_t - f_{t-1}$, $\Delta^2 f_t = \Delta f_t - \Delta f_{t-1}$ to capture velocity/acceleration.

MediaPipe-grounded observation model (addendum to §2)

Raw tensors. MediaPipe returns $L_t \in \mathbb{R}^{21\times 3}$ (left hand), $R_t \in \mathbb{R}^{21\times 3}$ (right), $F_t \in \mathbb{R}^{468\times 3}$ (face mesh), $B_t \in \mathbb{R}^{33\times 3}$ (pose). Concatenate $X_t = \mathrm{vec}([L_t; R_t; F_t; B_t]) \in \mathbb{R}^{1623\times 3}$.

Sim(3) normalization. Let $s_t = \|B_t[\mathrm{RS}] - B_t[\mathrm{LS}]\|_2$ (bi-shoulder width), R_t the yaw that aligns $(B_t[\mathrm{RS}] - B_t[\mathrm{LS}])$ with the x-axis, and $T_t = B_t[\mathrm{NE}]$ (neck). Define

$$\tilde{X}_t = \frac{(X_t - T_t)R_t^\top}{s_t}.$$

Any feature composed of differences, cross products, and distance ratios is G-equivariant (hence $q \circ \phi$ is invariant as above).

Hand/face/body primitives. With $\tilde{L}_t, \tilde{R}_t$:

$$c_t^L = \tfrac{1}{5} \sum_{j\in\{0,5,9,13,17\}} \tilde{L}_t[j], \quad c_t^R = \tfrac{1}{5} \sum_{j\in\{0,5,9,13,17\}} \tilde{R}_t[j],$$

$$n_t^L = \frac{(\tilde{L}_t[5] - \tilde{L}_t[0]) \times (\tilde{L}_t[17] - \tilde{L}_t[0])}{\|(\cdot)\|_2}, \quad n_t^R = \frac{(\tilde{R}_t[5] - \tilde{R}_t[0]) \times (\tilde{R}_t[17] - \tilde{R}_t[0])}{\|(\cdot)\|_2}.$$

Finger flexion (index/middle/ring/pinky $k = 1..4$):

$$\theta_{k,t} = \angle\big(\tilde{L}_t[4k+1] - \tilde{L}_t[0],\, \tilde{L}_t[4k+4] - \tilde{L}_t[4k+1]\big), \quad \theta_t^{\mathrm{th}} = \angle(\tilde{L}_t[4] - \tilde{L}_t[0],\, \tilde{L}_t[5] - \tilde{L}_t[0]).$$

Non-manuals via face:

$$g_t = \tfrac{1}{2}(F_t[33] + F_t[133]) - \tfrac{1}{2}(F_t[362] + F_t[263]), \quad a_t = \|F_t[61] - F_t[291]\|_2.$$

Velocities $\Delta c_t^{L/R}, \Delta a_t, \Delta g_t$ form Σ_M candidates.

Product quantization aligned to Σ (addendum to §2)

Partition f_t into $u^H \in \mathbb{R}^{10}$ (angles), $u^L \in \mathbb{R}^6$ (palm centers), $u^O \in \mathbb{R}^6$ (normals), $u^M \in \mathbb{R}^9$ (vel/acc), $u^N \in \mathbb{R}^5$ (gaze/mouth). Each sub-vector is vector-quantized with codebooks of sizes $(64, 128, 32, 64, 32)$; the token is $Z_t = (z_t^H, z_t^L, z_t^O, z_t^M, z_t^N) \in \Sigma$.

Proposition 2 (Joint margin robustness (refines noise robustness))**.** *If ϕ is L-Lipschitz and each head uses nearest-prototype with margins $\gamma_H^*, \dots, \gamma_N^*$, then for landmark perturbation $\|\eta\| \le \varepsilon$,*

$$(q \circ \phi)(X_t + \eta) = (q \circ \phi)(X_t) \quad \textit{whenever} \quad L\varepsilon < \min_j \gamma_j^*.$$

Proposition 3 (Product VQ sample complexity)**.** *Let sub-dimensions d_j and sizes k_j. For n i.i.d. frames,*

$$\mathbb{E}[\mathrm{dist}(q_n)] - \mathrm{dist}(q^\star) = \tilde{\mathcal{O}}\Big(\sum_j \sqrt{\tfrac{d_j \log k_j}{n}}\Big).$$

3 Spatial Grammar and Discourse Algebra

Let the signing space be $S \subset \mathbb{R}^3$. A referent is a pair (ℓ, s) with locus $\ell \in S$ and semantic label s. A pointing vector is $\mathbf{g}(t) \in \mathbb{S}^2$.

Lemma 1 (Deterministic uniqueness)**.** *If any two distinct loci $\ell_1 \neq \ell_2$ in $\mathcal{C}_t$ satisfy $\angle(\widehat{\ell_1}, \widehat{\ell_2}) > 2\tau$, then the candidate set $\Gamma_t(\tau) = \{(\ell, s) \in \mathcal{C}_t : \angle(\mathbf{g}(t), \widehat{\ell}) \le \tau\}$ satisfies $|\Gamma_t(\tau)| \le 1$.*

Probabilistic uniqueness. If m loci directions are i.i.d. uniform on $\mathbb{S}^2$, then

$$\Pr[|\Gamma_t(\tau)| \le 1] \;\ge\; 1 - \binom{m}{2} \tfrac{1 - \cos(2\tau)}{2}.$$

Cue integration. For cues C (point, gaze, non-manual), define likelihoods $\ell_c(r) \propto p(c \mid r)$ over referents r. Then

$$p(r \mid C) \propto \prod_{c \in C} \ell_c(r).$$

Operational instantiation (addendum)

Voxelized locus set. With normalized $\tilde{X}_t$, maintain window $\mathcal{W}_t = \{c_\tau^D : t - \Delta \le \tau \le t, D \in \{L, R\}\}$ and voxelize with side h to form $\mathcal{C}_t$ (voxel centers).
Pointing vector. $\mathbf{g}(t) = \frac{\tilde{R}_t[0] - B_t[\mathrm{NE}]}{\|\tilde{R}_t[0] - B_t[\mathrm{NE}]\|}$ for the hand with larger $\|\Delta c_t^D\|$.
Angular budget. For voxel side h at radius r, $\tau \approx \arctan(h/(2r))$ guarantees separation in Lemma 1. With h=8 cm, r≈1 m, $\tau \simeq 2.3°$.

Bayesian sensor fusion and calibration

Let $C = \{\text{point}, \text{eye-gaze}, \text{eyebrow}, \text{mouthing}\}$. With calibrated (temperature-scaled) likelihoods,

$$\hat{r}_t = \arg\max_{r \in \mathcal{C}_t} \sum_{c \in C} \log p(c_t \mid r) + \log p(r \mid r_{t-1}).$$

By Neyman–Pearson, this log-likelihood test is uniformly most powerful at fixed false-alarm rate for pairwise r vs. r'.

4 Automata, Grammars, and CTC Collapse

Let $Z_t = (q \circ \phi)(X_t) \in \Sigma$ and ϵ a blank.

Definition 2 (Fingerspelling automaton)**.** *A DFA $\mathcal{A}_{FS} = (Q, \Lambda, \delta, q_0, F)$ over letters Λ constrains legal transitions and pairs with per-letter HMMs mapping $Z_{1:T}$ to letters.*

Semi-Markov view. CTC is a special case of an HSMM with geometric durations and a blank symbol.

Proposition 4 (CTC as HSMM)**.** *If segment durations are geometric and include a blank ϵ, marginalizing boundaries recovers the CTC path-sum objective.*

Beam stability from margin (addendum)

Let $\pi^\star$ be the Viterbi path and Δ_t the per-frame log-score gap between the top-B and $(B+1)$-th partial hypotheses. If $\sum_{t=1}^{T} \Delta_t \geq \delta$ and per-frame spread is bounded by β, then any beam with $B \geq e^{\delta/\beta}$ preserves $\pi^\star$.

5 Scalable Decoding via WFST Composition

We compose

$$H \circ C \circ L \circ G$$

with H subunit HMMs, C context, L lexicon, G gloss LM.

Proposition 5 (Soundness)**.** *If (z, v) is accepted by $H \circ C \circ L \circ G$, then $v \in \mathcal{L}(G)$ and z is a pronunciation of v in L.*

Complexity. Beam decoding cost is $\mathcal{O}(TB\bar{d}c)$ with beam B, average out-degree $\bar{d}$, and per-arc cost c.

Completeness, determinization, and pruning (addendum)

Proposition 6 (Completeness). *For any $v \in \mathcal{L}(G)$ and any $z \in L^{-1}(v)$ with positive H-likelihood, there exists an accepting path in $H \circ C \circ L \circ G$ labeled (z, v).*

If G is determinized/minimized and L is left-deterministic, then $C \circ L \circ G$ has average out-degree $\bar{d}$ bounded by LM backoff and lexicon fan-out (empirically $\bar{d} \in [4, 6]$). With integer log-weights, static memory is $\mathcal{O}(|\text{arcs}|)$.

Lemma 2 (Pruning-loss tail). *If log-score gaps Δ_t are sub-exponential (ν, b), then $\Pr($beam prunes gold in T $T \exp(-\Delta^2/(2\nu^2)))$ for any fixed target gap $\Delta < b$.*

6 Information-Theoretic Guidance for Multimodality

Let $U = (H, F, B)$ be features; Y labels.

Chain rule.

$$\mathrm{I}(U;Y) = \mathrm{I}(H;Y) + \mathrm{I}(F;Y \mid H) + \mathrm{I}(B;Y \mid H, F).$$

Theorem 1 (Fano-type bound). *For any classifier $\hat{Y}$ and $|\mathcal{Y}| = g$,*

$$\Pr[\hat{Y} \neq Y] \;\geq\; \frac{\mathrm{H}(Y) - \mathrm{I}(U;Y) - 1}{\log g}.$$

Stopping rule. Include a modality m if $\Delta\mathrm{Err}/C_m \geq \lambda$, trading accuracy vs. cost.

DPI and calibrated priors (addendum)

Data processing gives $I(Z;Y) \leq I(\phi(X);Y) \leq I(X;Y)$. For skewed gloss priors $\{p_y\}$, use $H(Y) = -\sum_y p_y \log p_y$ in Fano for a tighter achievable error lower bound. Estimating ΔI_m with CLUB/MINE and dividing by device cost C_m yields an operational Pareto rule.

7 Practical Pipeline (Edge and Web)

Steps: Keypoint extraction (MediaPipe), normalization, features, encoder (BiLSTM/TCN), segmentation (CTC), decoder (WFST beam), translation, deployment.

Latency budget. Per-frame latency

$$\mathrm{Lat} = \mathrm{KP} + \mathrm{Enc} + \mathrm{Dec} + \mathrm{Post},$$

with $\mathrm{Dec} \approx TB\bar{d}c/F$. Require $\mathrm{Lat} < 200\,\mathrm{ms}$.

Measurable knobs and guarantees (addendum)

Let F be FPS, B the beam, $\bar{d}$ measured post-composition, c micro-benchmarked per-arc cost, κ_{KP} and κ_{Enc} profiled on-device. Then

$$\mathrm{Lat/frame} = \kappa_{\mathrm{KP}} + \kappa_{\mathrm{Enc}} + B\bar{d}\,c + \kappa_{\mathrm{Post}},$$

and Lemma 5 (pruning-loss tail) informs safe B increases to suppress search errors exponentially without touching κ_{KP} or κ_{Enc}.

8 Evaluation Protocol

Datasets: WLASL, PHOENIX. Metrics: top-k accuracy, WER, BLEU, segmentation F1, referent resolution.

Definition 3 (Pronunciation separability)**.** *Glosses $v \neq v'$ with pronunciation sets $P(v), P(v')$ are δ-separable if $\min_{z\in P(v), z'\in P(v')} \mathrm{DTW}(z, z') \geq \delta$.*

Proposition 7 (Identifiability)**.** *If variance σ^2 and separation margin $\delta > 2\sigma\sqrt{2\log(1/\alpha)}$, then with prob. $\geq 1-\alpha$ MAP decoding distinguishes v from v'.*

9 Discussion and Limitations

Productive morphology and regional variants need multiple pronunciations or templates. Annotation costs shift to subunits; mitigate with weak supervision.

10 Conclusion

A mathematical linguistics foundation—phonological factorization, discourse algebra, automata/CFG, and information theory—enables scaling ASL translation to 10k signs with tractable inference and real-time deployment.