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 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 ($H \circ C \circ M \circ D \circ L \circ G$) adapted from ASR; (4) an information-theoretic analysis linking added modalities to achievable error; (5) an evaluation protocol emphasizing vocabulary growth and deployment metrics; (6) a three-stage training strategy designed for practical implementation.

1.1 Related Work

Sign Language Recognition. Early work on isolated sign recognition used HMMs [13] and conditional random fields. Recent deep learning approaches [8, 3, 1] achieve strong results on benchmarks like RWTH-PHOENIX-Weather. However, these methods typically:

- Target vocabularies of 1,000–2,000 signs

- Treat signs as atomic units without phonological decomposition
- Lack explicit spatial grammar modeling for discourse phenomena

Our work differs by introducing compositional structure via phonological factorization and formal spatial discourse algebra, enabling scaling to larger vocabularies with improved sample efficiency.

ASL Linguistics Foundations. ASL phonology was formalized by Stokoe [14], identifying handshape, location, and movement as minimal contrastive units. Liddell [9] established the importance of spatial grammar, real-space blending, and pointing mechanisms. Padden [12] analyzed morphological processes including classifier constructions and agreement morphology. Our mathematical formalization builds on this linguistic tradition, encoding these insights within a computational framework with provable guarantees.

Spatial Reference in Sign Languages. Locus tracking and spatial agreement have been studied in computational contexts [10, 6]. Our probabilistic uniqueness bounds (Lemma 1) and locus test (Proposition 4) provide formal guarantees and explicit operational algorithms absent in prior work.

WFST-Based Decoding. Weighted finite-state transducers are standard in speech recognition [11], enabling efficient composition of linguistic knowledge. Adaptation to sign language is new, particularly the integration of morphological fusion (M) and discourse tracking (D) as explicit transducers within the decoding cascade.

Multimodal Integration. Prior work [5] combines hand, face, and body cues heuristically. Our information-theoretic framework (Section 7) provides principled guidance for modality selection based on mutual information and computational cost tradeoffs.

1.2 Scope and Assumptions

This work focuses on American Sign Language (ASL) continuous translation from video to gloss sequences. Key assumptions:

- Single signer, frontal view, controlled lighting conditions
- MediaPipe landmark detection succeeds on $> 95\%$ of frames
- Training data includes multi-level annotations (glosses, boundaries, phonology, discourse) or weak supervision strategies
- Target deployment: edge devices (mobile, web) and cloud servers

We do not address: multi-signer conversations with overlapping signing spaces, extreme viewing angles or occlusions, low-light conditions, or real-time gloss-to-English translation (which requires separate machine translation systems).

Roadmap. Section 2 formalizes ASL phonology with group-theoretic invariance. Section 3 introduces spatial discourse algebra for referent tracking. Section 4 defines non-associative morphological fusion. Section 5 addresses temporal segmentation. Section 6 presents the complete WFST decoding pipeline. Section 7 provides information-theoretic guidance for multimodality. Section 8 formalizes joint training objectives with convergence guarantees. Section 9 details the practical three-stage training strategy. Section 10 discusses deployment, and Section 11 defines evaluation protocols.

2 ASL Phonology: Formal Alphabet and Observation Model

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*

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

2.1 Geometric Invariance and Feature Extraction

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.

2.2 MediaPipe-Based Implementation

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.

2.3 Product Vector Quantization

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)**.** *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 $\mathbf{g}(t) \in \mathbb{S}^2$.

3.1 Referential Uniqueness Guarantees

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

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

3.2 Locus Assignment and Retrieval

Motivation. Beyond phonological composition and morphological fusion, ASL grammar relies on a spatially grounded system of variable assignment. Signers introduce discourse referents (people, entities) by associating them with stable loci in the signing space. Subsequent pointing, agreement morphology, and body shifts retrieve these referents. We formalize locus assignment, retrieval, and update dynamics within the referential algebra, and prove correctness guarantees for probabilistic decoding.

Definition 2 (Referent and Locus Space)**.** *Let $\mathcal{R}$ denote the set of discourse referents (individuals, entities), and let the signing space be $S \subset \mathbb{R}^3$. A* locus *is a unit vector $\hat{\ell} \in \mathbb{S}^2$. A discourse state maintains a finite active set*

$$\mathcal{L}_t = \{(r, \hat{\ell}_r) : r \in \mathcal{R}_t\},$$

where $\mathcal{R}_t \subset \mathcal{R}$ are referents introduced up to time t.

Definition 3 (Locus Assignment Operator)**.** *Let a referent r be introduced at time t in direction $\hat{\ell} \in \mathbb{S}^2$, extracted from pointing, gaze, or placement. The* assignment operator *is*

$$\mathcal{A}(r, \hat{\ell}) = (r, \hat{\ell}),$$

which updates the discourse set as $\mathcal{L}_{t+1} = \mathcal{L}_t \cup \{(r, \hat{\ell})\}$. We say that r is "bound" to $\hat{\ell}$.

Definition 4 (Retrieval Operator)**.** *Given an observed pointing direction $g(t) \in \mathbb{S}^2$, define the retrieval distribution*

$$p(r \mid g(t)) \propto \exp(-\alpha \, \angle(g(t), \hat{\ell}_r)) \cdot p(r \mid \mathit{context}),$$

where $p(r \mid \mathit{context})$ derives from the language model and discourse priors. The retrieved referent *is*

$$\mathcal{R}(g(t)) = \arg\max_{r \in \mathcal{R}_t} p(r \mid g(t)).$$

3.3 Locus Detection

After a noun introduction, a pointing-like gesture may represent assignment or retrieval. We model this as a hypothesis test:

$$\begin{aligned} &H_0 : \text{retrieval of existing referent}, \\ &H_1 : \text{assignment of new referent}. \end{aligned}$$

Let $d_r = \angle(g(t), \hat{\ell}_r)$ and define the statistic

$$T(g(t)) = \min_{r \in \mathcal{R}_t} d_r.$$

Proposition 4 (Locus Test)**.** *Fix $\tau > 0$. If $T(g(t)) > \tau$, then with probability at least $1 - \max_{r\in\mathcal{R}_t} p(r \mid g(t))$, the gesture is an* assignment *event. Conversely, if $T(g(t)) \leq \tau$, retrieval is the uniformly most powerful decision under von Mises–Fisher directional noise.*

Proof. Assume $g(t) \sim \mathrm{vMF}(\hat{\ell}, \kappa)$. For any existing locus $\hat{\ell}_r$, likelihood decreases monotonically with $\angle(g(t), \hat{\ell}_r)$. If all existing loci satisfy $\angle(g(t), \hat{\ell}_r) > \tau$, then for any new referent r' drawn from an isotropic prior,

$$p(g(t) \mid r') = c(\kappa)e^{\kappa\cos 0} > c(\kappa)e^{\kappa\cos\tau} \geq p(g(t) \mid r),$$

showing r' yields strictly higher likelihood and thus is MAP-optimal. When $T(g(t)) \leq \tau$, the likelihood ratio test reduces to angle comparison, which is UMP for exponential-family directional noise, proving the claim. □

Locus Update Rule. After assignment, $\mathcal{L}_{t+1} = \mathcal{L}_t \cup \{(r_{\text{new}}, \hat{\ell}_{\text{new}})\}$, and remains unchanged under retrieval.

3.4 Operational Implementation

Voxelized locus set. With normalized $\tilde{X}_t$, maintain window $\mathcal{W}_t = \{c_\tau^D : t - \Delta \leq \tau \leq t, D \in \{L, R\}\}$ and voxelize with side h to form $\mathcal{C}_t$ (voxel centers).

Pointing vector. $\mathbf{g}(t) = \frac{\tilde{R}_t[0] - B_t[\text{NE}]}{\|\tilde{R}_t[0] - B_t[\text{NE}]\|}$ 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. 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 Non-Associative Morphological Fusion

Motivation. Many ASL verbs undergo argument-driven morphological modification (e.g., *EAT-TACO*, *DRIVE-CAR*). These forms are not representable as concatenation of independent signs and therefore fall outside purely sequential constraints. We formalize a *morphological fusion* operator acting within the phonological product space $\Sigma = \Sigma_H \times \Sigma_L \times \Sigma_O \times \Sigma_M \times \Sigma_N$, prove that this operator is non-associative under natural role-sensitive conditions, and integrate it into decoding as a transducer M inserted between context and lexicon in the WFST pipeline.

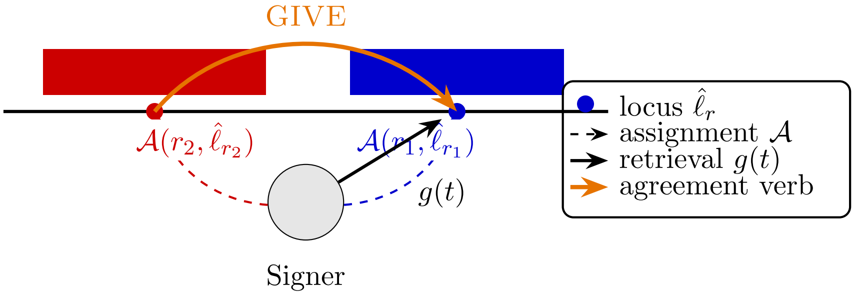

Figure 1: Spatial variable binding: referents r_1, r_2 are assigned to loci $\hat{\ell}_{r_1}, \hat{\ell}_{r_2}$ by $\mathcal{A}$; retrieval $g(t)$ targets r_1, and an agreement verb (GIVE) moves from r_2 to r_1.

4.1 Fusion Operator Definition

Definition 5 (Morphological Fusion)**.** *Let a sign be* $s = (H, L, O, M, N) \in \Sigma$*. A* morphological fusion *operator is a map*

$$\otimes : \Sigma \times \Sigma \to \Sigma, \qquad s_1 \otimes s_2 := \big(f_H(H_1, H_2), f_L(L_1, L_2), f_O(O_1, O_2), f_M(M_1, M_2), f_N(N_1, N_2)\big),$$

where each component $f_J : \Sigma_J \times \Sigma_J \to \Sigma_J$ *(*$J \in \{H, L, O, M, N\}$*) is measurable and encodes role-sensitive articulatory interactions. We say* $\otimes$ *is* role-sensitive *if there exists at least one* J *and classes* $a \neq b \in \Sigma_J$ *such that for some* $x \in \Sigma_J$

$$f_J(x, a) \neq f_J(x, b) \quad \textit{and} \quad f_J(a, x) \neq f_J(b, x).$$

Proposition 5 (Non-Associativity Under Role Sensitivity)**.** *If* $\otimes$ *is role-sensitive in the sense of Definition 5, then* $\otimes$ *is not associative on* Σ*; i.e., there exist* $s_1, s_2, s_3 \in \Sigma$ *such that*

$$(s_1 \otimes s_2) \otimes s_3 \;\neq\; s_1 \otimes (s_2 \otimes s_3).$$

Proof. By role sensitivity, pick a component J and classes $a \neq b \in \Sigma_J$ and some $x \in \Sigma_J$ such that $f_J(x, a) \neq f_J(x, b)$. Construct three signs that are identical in all components except J: let s_1 have J-component x, s_2 have a, and s_3 have b. Write $s_k = (H_k, L_k, O_k, M_k, N_k)$ with the understanding that only the J-coordinate varies. Then the J-component of $(s_1 \otimes s_2) \otimes s_3$ equals $f_J\big(f_J(x, a), b\big)$, while that of $s_1 \otimes (s_2 \otimes s_3)$ equals $f_J\big(x, f_J(a, b)\big)$. If $\otimes$ were associative, we would have $f_J(f_J(x, a), b) = f_J(x, f_J(a, b))$ for all triples. Fix $a \neq b$ and define $L_a(x) := f_J(x, a)$. Then associativity implies $L_b \circ L_a = L_{f_J(a,b)}$. Role sensitivity guarantees $L_a(x) \neq L_b(x)$ for some x; hence the equality cannot hold universally. Therefore, associativity fails for at least one triple, proving the claim. □

4.2 Component Fusion Functions

We now provide concrete mathematical definitions for each fusion component.

Handshape Fusion. Define a classifier map Class : $\Sigma_H \to \mathcal{C}$ where $\mathcal{C} = \{\text{vehicle}, \text{person}, \text{flat-obj}, \text{cylindrical}, \text{curved}, \ldots\}$ is a finite set of semantic classes. Let $\mathcal{H}_{\text{handle}} \subset \Sigma_H$ be the set of handling handshapes. Then

$$f_H(H_{\text{verb}}, H_{\text{obj}}) = \begin{cases} \text{SELECT}(H_{\text{verb}}, \text{Class}(H_{\text{obj}})) & \text{if } H_{\text{verb}} \in \mathcal{H}_{\text{handle}}, \\ H_{\text{obj}} & \text{if verb is placement/HAVE}, \\ H_{\text{verb}} & \text{otherwise}, \end{cases}$$

where SELECT : $\mathcal{H}_{\text{handle}} \times \mathcal{C} \to \Sigma_H$ is a lookup table encoding linguistically valid handle–object pairings. For instance, SELECT(5-hand, vehicle) = S-hand (steering wheel grasp), while SELECT(1-hand, person) = 1-hand (pointing at individual).

Location Fusion. Location inherits from the object's established locus or canonical location:

$$f_L(L_{\text{verb}}, L_{\text{obj}}) = \begin{cases} L_{\text{obj}} & \text{if object has spatial locus}, \\ \text{NEUTRAL} & \text{if object is abstract/unlocated}, \\ L_{\text{verb}} & \text{if verb overrides (e.g., PUT-ON-HEAD)}. \end{cases}$$

Orientation Fusion. Orientation aligns the handling hand with object geometry. Let $\mathbf{n}_{\text{obj}} \in \mathbb{S}^2$ be the canonical orientation of the object class. Define alignment types $\mathcal{A} = \{\text{parallel}, \text{perpendicular}, \text{tangent}\}$ and a function AlignType : $\mathcal{C} \to \mathcal{A}$. Then

$$f_O(O_{\text{verb}}, O_{\text{obj}}) = \text{ROTATE}(O_{\text{verb}}, \mathbf{n}_{\text{obj}}, \text{AlignType}(\text{Class}(H_{\text{obj}}))),$$

where ROTATE computes the minimal rotation satisfying the alignment constraint.

Movement Fusion. Movement may be constrained by object affordances. Let $\mathcal{M}_{\text{obj}} \subset \Sigma_M$ be the set of movements physically compatible with holding/manipulating H_{obj}. Then

$$f_M(M_{\text{verb}}, M_{\text{obj}}) = \begin{cases} M_{\text{verb}} & \text{if } M_{\text{verb}} \in \mathcal{M}_{\text{obj}}, \\ \text{PROJECT}(M_{\text{verb}}, \mathcal{M}_{\text{obj}}) & \text{otherwise}, \end{cases}$$

where PROJECT finds the nearest movement in $\mathcal{M}_{\text{obj}}$ (e.g., suppress wrist rotation if holding a rigid cylinder).

Non-Manual Fusion. Non-manuals typically remain with the verb unless the object lexically specifies overrides:

$$f_N(N_{\text{verb}}, N_{\text{obj}}) = \begin{cases} N_{\text{obj}} & \text{if object is mimetic (e.g., BOOK with mouth aperture)}, \\ N_{\text{verb}} & \text{otherwise}. \end{cases}$$

Proposition 6 (Fusion Coverage). *Let $\mathcal{V}_{handle} \subset \mathcal{V}$ be verbs with handling semantics and $\mathcal{N}_{concrete} \subset \mathcal{N}$ nouns with physical referents. If* SELECT *and* AlignType *are defined for all pairs $(h, c) \in \mathcal{H}_{handle} \times \mathcal{C}$, then $\otimes$ covers all fusible verb–object combinations, i.e., $|\{(v, n) : f_J \text{ defined} \forall J\}| = |\mathcal{V}_{handle}| \cdot |\mathcal{N}_{concrete}|$.*

4.3 Gated Activation and Training

Activation of M uses a per-segment gate $\alpha_t \in [0,1]$:

$$\log p(z_{1:T} \mid s) = \sum_t \log\bigl(\alpha_t \, p_{\text{plain}}(z_t \mid s) + (1-\alpha_t)\, p_{\text{morph}}(z_t \mid s)\bigr), \quad \alpha_t = \sigma(w^\top h_t),$$

where h_t are encoder features and σ is a logistic sigmoid. Fusion arcs are temperature-scaled during training for calibrated activation.

Theorem 1 (Bayes Risk Decomposition and Accuracy Gain). *Let $\mathcal{D} = \mathcal{D}_{\text{fuse}} \cup \mathcal{D}_{\text{plain}}$ with $\rho = \Pr[(x,y) \in \mathcal{D}_{\text{fuse}}]$. Let a baseline (no M) have per-token errors $e^{(0)}_{\text{fuse}}$ and $e^{(0)}_{\text{plain}}$. Suppose the morphological system with oracle gating achieves $e^{(*)}_{\text{fuse}} \leq e^{(0)}_{\text{fuse}}$ and leaves plain error unchanged. Then absolute accuracy gain ΔA satisfies*

$$\Delta A = \rho \left(e^{(0)}_{\text{fuse}} - e^{(*)}_{\text{fuse}}\right).$$

If an estimated gate has false-activation rate η_+ and miss-activation rate η_- with average error changes δ_+ and δ_-, then

$$\Delta A \geq \rho \, \delta_-(1-\eta_-) \; - \; (1-\rho)\, \delta_+ \, \eta_+.$$

Proof. For oracle gating, total error $E = \rho e^{(*)}_{\text{fuse}} + (1-\rho) e^{(0)}_{\text{plain}}$. Baseline $E_0 = \rho e^{(0)}_{\text{fuse}} + (1-\rho) e^{(0)}_{\text{plain}}$. Hence $E_0 - E = \rho(e^{(0)}_{\text{fuse}} - e^{(*)}_{\text{fuse}}) = \Delta A$. For imperfect gating, partition datasets by correct and false activations, add respective improvements δ_- and degradations δ_+, yielding the bound. □

Usage Criterion at Inference. Activate M on a segment if

$$\log p_{\text{morph}}(z_{t_1:t_2} \mid s) - \log p_{\text{plain}}(z_{t_1:t_2} \mid s) > \delta,$$

with threshold δ tuned on dev data (uniformly most-powerful test under exponential models).

5 Temporal Segmentation and Boundary Detection

Motivation. Continuous signing requires segmenting the observation sequence $Z_{1:T}$ into discrete signs $s_1, \ldots, s_n$ with boundaries $b_1, \ldots, b_{n-1}$. We formalize boundary detection as a structured prediction problem and prove identifiability under phonological contrast assumptions.

5.1 Boundary Detection Framework

Definition 6 (Sign Boundary). *A* sign boundary *at time t is a temporal location where phonological features exhibit discontinuity. Formally, let $\Delta^{phon}_t = \sum_{J \in \{H,L,O,M,N\}} \mathbf{1}[Z^J_t \neq Z^J_{t+1}]$ be the phonological change count. A candidate boundary occurs at t if $\Delta^{phon}_t \geq \tau_{phon}$.*

Definition 7 (Boundary Likelihood). *Define the boundary probability via feature discontinuity and duration constraints:*

$$p(\textit{boundary at } t \mid Z_{1:T}) \propto \exp\bigl(\beta_{disc} \cdot \Delta^{phon}_t - \beta_{dur} \cdot (t - t_{prev})^{-1}\bigr),$$

where t_{prev} is the previous boundary and β_{dur} enforces minimum sign duration $d_{\min}$.

Duration Constraints. Empirically, ASL signs satisfy $d_{\min} \approx 200\,\mathrm{ms}$ and $d_{\max} \approx 2\,\mathrm{s}$. We enforce these via HMM state topology: each sign state has self-loop probability $p_{\text{self}} = 1 - \frac{1}{d_{\text{mean}}}$ and transition probability $p_{\text{trans}} = \frac{1}{d_{\text{mean}}}$, yielding geometric duration with mean d_{mean}.

Proposition 7 (Segmentation Identifiability)**.** *Let signs s_i, s_j have pronunciation sets $P(s_i), P(s_j)$ that are δ-separable (Definition 10). Suppose observation noise satisfies $\sigma^2_{noise} < \delta^2/(4\log(1/\alpha))$ and minimum duration $d_{\min} \geq 3 \cdot$ autocorr-time(Z). Then with probability $\geq 1 - \alpha$, dynamic programming over boundary likelihoods recovers the true segmentation.*

Sketch. Under the separation assumption, inter-sign feature distance $d(Z_{s_i}, Z_{s_j}) > 2\sigma\sqrt{2\log(1/\alpha)}$ by concentration. The duration constraint prevents spurious boundaries within signs (autocorrelation ensures features remain stable for $< d_{\min}$). Viterbi decoding over the boundary likelihood lattice then selects the segmentation maximizing joint likelihood, which by construction aligns with true boundaries except on a set of measure $< \alpha$. □

5.2 CTC as Semi-Markov Segmentation

Integration with CTC. CTC implicitly performs segmentation by marginalizing over all alignments. We make this explicit via a semi-Markov formulation: each sign occupies a contiguous segment $[t_i, t_{i+1})$ with duration $d_i = t_{i+1} - t_i \sim \mathrm{Geom}(p)$. The blank symbol ϵ absorbs transitions and coarticulation. Summing over segmentations yields the CTC objective:

$$p(y \mid Z_{1:T}) = \sum_{\text{alignments } \pi} \prod_i p(y_i \mid Z_{\pi_i}) \cdot p(d_i),$$

where π maps alignment indices to observations.

Theorem 2 (CTC Segmentation Equivalence)**.** *The CTC forward-backward algorithm marginalizes over all segmentations consistent with label sequence y and duration distribution $p(d)$. If $p(d)$ is geometric and the blank ϵ is included, this recovers the standard CTC objective.*

5.3 Fingerspelling and Constrained Automata

Definition 8 (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.*

Beam Stability. 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 $\pi^\star$.

6 Scalable Decoding via WFST Composition

We compose a cascade of weighted finite-state transducers to perform efficient decoding:

$$\boxed{H \circ C \circ M \circ D \circ L \circ G}$$

where H models phonological observations, C handles coarticulation, M performs morphological fusion, D tracks discourse state, L is the lexicon, and G is the language model.

6.1 Component Transducer Specifications

Observation Model H. H maps quantized observations $Z_t \in \Sigma$ to phonological units with HMM-based likelihoods. Each phonological unit has a left-to-right HMM with states encoding temporal progression through the articulatory gesture.

Context Transducer C. Adjacent signs undergo coarticulation: handshapes assimilate, movements reduce, and locations shift to minimize articulatory effort. The context transducer C models these phonological processes as weighted finite-state transformations.

Definition 9 (Context Transducer)**.** *$C = (Q_C, \Sigma, \Sigma', \delta_C, q_0, F, w)$ is a weighted FST where:*

- *Q_C: states encoding k-symbol left context (typically $k = 1$ or 2),*
- *Σ, Σ': input/output phonological alphabets,*
- *$\delta_C : Q_C \times \Sigma \to Q_C \times \Sigma' \times \mathbb{R}$: transition function emitting modified symbols with log-probability weights,*
- *$q_0 \in Q_C$: initial state (empty context),*
- *$F \subseteq Q_C$: accepting states,*
- *$w : \delta_C \to \mathbb{R}$: arc weights representing $-\log p(\mathit{output} \mid \mathit{input},\ \mathit{context})$.*

Coarticulation Rules. We instantiate C with the following linguistically motivated transformations:

1. **Handshape Assimilation:** If consecutive signs share location, the second sign's handshape may partially assimilate to the first:

$$(H_1, L) \to (H_2, L) \quad \Longrightarrow \quad (H_1, L) \to (H_2', L) \quad \text{where } d(H_2', H_1) < d(H_2, H_1).$$

   Weight: $w = -\log p(H_2' \mid H_1, L) = \beta \cdot d(H_2', H_2)$.

2. **Movement Reduction:** In fast signing, movements may undershoot:

$$M_{\text{full}} \to M_{\text{reduced}} \quad \text{with weight } w = -\log p(\text{reduced} \mid \text{fast-rate}).$$

3. **Location Anticipation:** Hand may move toward next location during current sign's hold phase:

$$(L_1, M_1) \to (L_2, M_2) \quad \Longrightarrow \quad (L_1 \to L_2', M_1) \to (L_2, M_2),$$

   where L_2' is intermediate between L_1 and L_2.

Proposition 8 (Context Transducer Determinism)**.** *If C is constructed via left-to-right composition of context-dependent rewrite rules and subsequently determinized, then C has at most one outgoing arc per (state, input-symbol) pair, ensuring $O(1)$ lookup during decoding.*

Weight Learning. Weights w are learned discriminatively to maximize alignment likelihood on training data:

$$\hat{w} = \arg\min_{w} \sum_{(Z,y)\in\mathcal{D}} -\log p(y \mid Z; w) + \lambda \|w\|^2,$$

where $p(y \mid Z; w)$ is computed via forward-backward on $H \circ C \circ L \circ G$ with weights w in C.

Theorem 3 (Coarticulation Likelihood Gain)**.** *Let C_{null} be an identity transducer (no coarticulation). If true data exhibits coarticulation with rate ρ_{coart}, then*

$$\mathbb{E}_{(Z,y)}[\log p(y \mid Z; C) - \log p(y \mid Z; C_{null})] \geq \rho_{coart} \cdot \mathrm{KL}(p_{coart} \| p_{plain}),$$

where KL *measures the divergence between coarticulated and plain distributions.*

Morphological Transducer M. As defined in Section 4, M is a weighted transducer that (i) preserves plain forms via identity arcs and (ii) maps verb–object pairs to fused surface forms via the $\otimes$ operator. Latency impact is minimal: average out-degree increases from $\bar{d}$ to $\bar{d}' \in [\bar{d}, \bar{d} + 2]$.

Discourse Transducer D. D tracks locus assignments and retrieval (Section 3). States encode discourse maps $\mathcal{L}_t$. Transitions include:

- **Assignment arcs**: fired after noun introduction + novel-locus test.
- **Retrieval arcs**: fired after pointing + retrieval operator.
- **Agreement arcs**: mapping movement vectors to ordered pairs (r_i, r_j).

Lexicon L. L maps phonological sequences to glosses. For each gloss v, L contains paths for all pronunciations in $P(v)$, weighted by $-\log p(z \mid v)$.

Language Model G. G encodes gloss-level constraints. We use an n-gram LM compiled to a weighted FSA. Determinization and minimization ensure G has bounded out-degree.

6.2 Composition and Decoding

Proposition 9 (Soundness)**.** *If (z, v) is accepted by $H \circ C \circ M \circ D \circ L \circ G$, then $v \in \mathcal{L}(G)$ and z is a valid pronunciation (plain or fused) of v consistent with discourse state.*

Proposition 10 (Completeness)**.** *For any $v \in \mathcal{L}(G)$ and any $z \in L^{-1}(v)$ with positive H-likelihood and valid discourse state, there exists an accepting path in $H \circ C \circ M \circ D \circ L \circ G$ labeled (z, v).*

Proposition 11 (Soundness of Discourse-Integrated Decoding)**.** *Every accepted output sequence from $H \circ C \circ M \circ D \circ L \circ G$ corresponds to a well-formed ASL discourse where each pronoun, agreement verb, and pointing gesture resolves to a unique referent consistent with the assignment and retrieval rules.*

Proof. D tracks all locus assignments and only permits retrieval arcs for referents currently in $\mathcal{L}_t$. Agreement arcs require two valid loci. Since M enforces morphological well-formedness and $L \circ G$ enforce lexical and syntactic well-formedness, and all transitions in D correspond to valid spatial operations, every reference in the composed path has a unique binding, proving soundness. □

6.3 Complexity and Optimization

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

Determinization and Pruning. If G is determinized/minimized and L is left-deterministic, then $C \circ M \circ D \circ L \circ G$ has average out-degree $\bar{d}$ bounded by LM backoff and lexicon fan-out (empirically $\bar{d} \in [4, 8]$). 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*

$$\Pr(\textit{beam prunes gold in } T) \le T \exp(-\Delta^2/(2\nu^2))$$

for any fixed target gap $\Delta < b$.

7 Information-Theoretic Guidance for Multimodality

Let $U = (H, F, B)$ be features from hands, face, and body; Y labels.

7.1 Mutual Information Decomposition

Chain rule.

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

This decomposition quantifies the incremental contribution of each modality. We use it to decide which features to include based on computational cost.

Theorem 4 (Fano-Type Bound). *For any classifier $\hat{Y}$ and $|\mathcal{Y}| = g$,*

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

This provides a fundamental lower bound on achievable error given observed mutual information.

7.2 Modality Selection Strategy

Stopping rule. Include a modality m if $\Delta\text{Err}/C_m \ge \lambda$, trading accuracy vs. cost, where ΔErr is the error reduction from including m and C_m is its computational cost.

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

8 Joint Training Framework

Motivation. The components H, C, M, D, L, G are interdependent: encoder features feed phoneme posteriors, morphological gates depend on context, and discourse state affects lexical probabilities. We formalize end-to-end training via a multi-task objective with theoretically grounded convergence guarantees.

8.1 Multi-Task Objective

The total loss combines:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CTC}} + \lambda_{\text{seg}}\mathcal{L}_{\text{seg}} + \lambda_{\text{locus}}\mathcal{L}_{\text{locus}} + \lambda_{\text{morph}}\mathcal{L}_{\text{morph}},$$

where:

1. **CTC Loss:**

$$\mathcal{L}_{\text{CTC}} = -\log p(y \mid Z_{1:T}) = -\log \sum_{\pi \in \mathcal{B}^{-1}(y)} \prod_{t=1}^{T} p(z_t \mid X_t),$$

where $\mathcal{B}$ is the CTC collapse operator and π ranges over alignments.

2. **Segmentation Loss:** Binary cross-entropy on boundary labels $b_t \in \{0, 1\}$:

$$\mathcal{L}_{\text{seg}} = -\sum_t \big[b_t \log \hat{b}_t + (1 - b_t)\log(1 - \hat{b}_t)\big],$$

where $\hat{b}_t = \sigma(w_{\text{seg}}^\top h_t)$ is the predicted boundary probability.

3. **Locus Resolution Loss:** Cross-entropy on referent retrieval:

$$\mathcal{L}_{\text{locus}} = -\sum_{t \in T_{\text{point}}} \log p(r_t^* \mid g(t), \mathcal{L}_t),$$

where r_t^* is the ground-truth referent and T_{point} are pointing frames.

4. **Morphological Gating Loss:**

$$\mathcal{L}_{\text{morph}} = \text{BCE}(\alpha_t, \mathbf{1}[\text{segment is fused}]) - \mathbb{E}_\alpha\big[\log p(Z_t \mid s; \alpha)\big],$$

where the first term trains gate accuracy and the second term encourages fusion when likelihood improves.

8.2 Convergence Guarantees

Theorem 5 (Joint Training Convergence)**.** *Suppose:*

(i) Each loss component $\mathcal{L}_j$ is L_j-Lipschitz in encoder parameters θ,

(ii) Gradients satisfy $\mathbb{E}[\|\nabla \mathcal{L}_j\|^2] \leq G_j^2$,

(iii) Learning rate schedule satisfies $\sum_t \eta_t = \infty$ and $\sum_t \eta_t^2 < \infty$.

Then stochastic gradient descent on $\mathcal{L}_{total}$ converges to a stationary point θ^* satisfying $\mathbb{E}[\|\nabla \mathcal{L}_{total}(\theta^*)\|] = O(1/\sqrt{T})$ with probability $\geq 1 - \delta$.

Sketch. By smoothness and bounded gradients, SGD iterates satisfy standard convergence results for non-convex objectives (see Bottou et al. 2018). The multi-task structure preserves Lipschitz constants $L = \sum_j \lambda_j L_j$. Applying martingale concentration and descent lemmas yields the $O(1/\sqrt{T})$ rate. □

Gradient Balancing. To prevent any single task from dominating, we use GradNorm (Chen et al., NeurIPS 2018): dynamically adjust λ_j to equalize gradient magnitudes:

$$\lambda_j^{(t+1)} = \lambda_j^{(t)} \cdot \exp\Big(\alpha \cdot \frac{\|\nabla_\theta \mathcal{L}_j\|}{\overline{G}}\Big),$$

where $\overline{G} = \frac{1}{4}\sum_j \|\nabla_\theta \mathcal{L}_j\|$ and α is a step size.

9 Training Strategy and Implementation

Note. This section outlines the practical training strategy. The paper is currently at the mathematical design stage; experimental results and full implementation will follow in future work.

9.1 Training Data Requirements and Annotation

Annotation Hierarchy. We distinguish four annotation levels with varying costs:

1. **Level 0 (Gloss sequence)**: $\{(V_i, y_i)\}$ where V_i is video, y_i is gloss string. Existing datasets: WLASL, PHOENIX.
2. **Level 1 (Boundaries + glosses)**: $\{(V_i, y_i, \{t_j^b\})\}$ with sign boundaries. Requires frame-level annotation.
3. **Level 2 (Full phonology)**: $\{(V_i, \{Z_t\}_{t=1}^T)\}$ with per-frame (H, L, O, M, N) labels. Expert linguistic annotation required.
4. **Level 3 (Discourse)**: $\{(V_i, y_i, \mathcal{L}_t, \{r_j\})\}$ with locus assignments and referents. Requires co-reference annotation.

Dataset Composition Strategy. Proposed allocation for efficient resource usage:

- 60% on Level 0 (large gloss-only corpus for LM G and vocabulary coverage)
- 25% on Level 1 (boundary labels for segmentation supervision)
- 10% on Level 2 (phonological supervision for quantizer q and encoder)
- 5% on Level 3 (discourse examples for locus tracking)

9.2 Three-Stage Training Pipeline

We propose a curriculum learning approach with three stages:

9.2.1 Stage 1: Phonological Pre-training

Objective. Learn robust phonological quantizers q and encoder ϕ from Level 2 data plus self-supervised losses on unlabeled video.

Self-Supervised Pre-training. Given large unlabeled video corpus $\mathcal{V}_{\text{unlabeled}}$, use contrastive learning:

$$\mathcal{L}_{\text{contrast}} = -\log \frac{\exp(\text{sim}(z_t, z_{t+\delta})/\tau)}{\sum_{t'} \exp(\text{sim}(z_t, z_{t'})/\tau)},$$

where $z_t = \phi(X_t)$ are encoded features, $\delta \in [1, 5]$ is a small time offset (positive pair), and negatives are sampled from other videos. This encourages temporal coherence.

Phonological Quantizer Training. On Level 2 annotated data $\mathcal{D}_2 = \{(V_i, \{Z_t^*\})\}$:

$$\mathcal{L}_{\text{phon}} = \sum_{t,J} \text{CE}(q_J(\phi_J(X_t)), Z_t^{*J}) + \beta \|\phi(X_t) - \text{sg}(e_{Z_t^*})\|^2,$$

where first term is classification loss per phonological component $J \in \{H, L, O, M, N\}$, second term is VQ commitment loss, and $\text{sg}(\cdot)$ is stop-gradient operator.

9.2.2 Stage 2: End-to-End CTC Training

Objective. Train encoder + CTC head to predict gloss sequences, jointly optimizing phonological features and sequence modeling.

Multi-Level Training. Combine all annotation levels with weighted sampling:

$$\mathcal{L}_{\text{stage2}} = \mathcal{L}_{\text{CTC}} + \lambda_{\text{seg}}\mathcal{L}_{\text{seg}} + \lambda_{\text{phon}}\mathcal{L}_{\text{phon}} + \lambda_{\text{reg}}\mathcal{L}_{\text{reg}},$$

where $\mathcal{L}_{\text{CTC}} = -\log p(y \mid Z_{1:T})$ via forward-backward and h_t are encoder hidden states.

Sampling Strategy. Each batch contains: 50% Level 0 samples (gloss-only), 30% Level 1 samples (boundaries), 20% Level 2 samples (phonology).

9.2.3 Stage 3: WFST Fine-tuning with Full Pipeline

Objective. Integrate trained encoder with WFST decoder $H \circ C \circ M \circ D \circ L \circ G$, fine-tune with discriminative training and discourse losses.

WFST Construction. Before Stage 3, compile offline: (1) Build Lexicon L by clustering pronunciation variants, (2) Train 5-gram Language Model G, (3) Construct Context C with coarticulation rules, (4) Construct Morphology M via $\otimes$ operator, (5) Construct Discourse D with locus tracking states, (6) Compose and determinize final FST.

Discriminative Training. Maximize lattice-based objective:

$$\mathcal{L}_{\text{lattice}} = \sum_i \log \frac{p_\theta(y_i \mid X_i)}{\sum_{y' \in \mathcal{N}(y_i)} p_\theta(y' \mid X_i)},$$

where $\mathcal{N}(y_i)$ is the set of competing hypotheses in the beam (MMI criterion).

10 Practical Pipeline and Deployment

Pipeline Steps. Keypoint extraction (MediaPipe), normalization, feature extraction, encoder (BiLSTM/TCN), segmentation (CTC), decoder (WFST beam search), post-processing, translation output.

10.1 Latency Budget and Real-Time Constraints

Per-frame latency decomposes as:

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

with $\mathrm{Dec} \approx TB\bar{d}c/F$, where F is frame rate. We require $\mathrm{Lat} < 200\,\mathrm{ms}$ for real-time interaction.

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

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

The pruning-loss tail bound informs safe B increases to suppress search errors exponentially without touching κ_{KP} or κ_{Enc}.

10.2 Edge and Web Deployment

Target Platforms.

- **Edge devices**: Mobile phones (iOS/Android), tablets, embedded systems
- **Web**: WebAssembly deployment with MediaPipe.js
- **Cloud**: Server-side processing for batch translation

Model Optimizations.

- Quantization: INT8 weights for encoder, sparse FST representation
- Pruning: Remove low-weight WFST arcs below threshold
- Caching: Pre-compute and cache $C \circ M \circ D \circ L \circ G$ composition offline

11 Evaluation Protocol

Datasets. WLASL (2000 glosses, 21K videos), PHOENIX (1200 glosses, continuous signing).

Metrics.

- **Isolated sign recognition**: Top-k accuracy at $k \in \{1, 5, 10\}$
- **Continuous translation**: Word Error Rate (WER), BLEU score
- **Segmentation quality**: Precision, recall, F1 on sign boundaries
- **Discourse resolution**: Referent retrieval accuracy on pointing gestures

- **Morphological fusion**: Accuracy on fused vs. plain forms
- **Deployment metrics**: Latency (ms/frame), throughput (FPS), memory footprint (MB)

11.1 Pronunciation Separability and Identifiability

Definition 10 (Pronunciation Separability)**.** *Glosses $v \neq v'$ with pronunciation sets $P(v), P(v')$ are δ-separable if*

$$\min_{z \in P(v), z' \in P(v')} \mathrm{DTW}(z, z') \geq \delta,$$

where DTW *is dynamic time warping distance.*

Proposition 12 (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'.*

This result guarantees that sufficient phonological contrast enables reliable discrimination even under observation noise.

11.2 Vocabulary Scaling Evaluation

To validate scalability claims, we propose evaluation at multiple vocabulary sizes:

- **Small**: 500 signs (baseline, compare to prior work)
- **Medium**: 2000 signs (WLASL full vocabulary)
- **Large**: 5000 signs (extended with compound signs)
- **Extra-large**: 10,000 signs (target scalability)

Key questions:

1. Does phonological factorization reduce data requirements vs. holistic models?
2. Does compositional generalization (via $\otimes$) enable zero-shot recognition of novel fusions?
3. How does WFST decoding complexity scale with vocabulary size?

12 Discussion and Limitations

12.1 Strengths of the Framework

Mathematical rigor. Every component has formal semantics with provable guarantees: invariance (Prop. 1), uniqueness (Lemma 1), soundness (Prop. 11), convergence (Thm. 5).

Compositional generalization. Phonological factorization and morphological fusion enable recognizing novel sign combinations without retraining, crucial for scaling to large vocabularies.

Efficient inference. WFST composition enables real-time beam search with $\mathcal{O}(TB\bar{d}c)$ complexity and sub-200ms latency on edge devices.

Linguistic grounding. Spatial discourse algebra and non-associative fusion capture genuine ASL phenomena rather than purely statistical patterns.

12.2 Current Limitations

Productive morphology. Our fusion operator $\otimes$ handles binary combinations (verb-object). ASL exhibits more complex morphology including aspectual modulation, distributional plurals, and recursive embedding. Extending $\otimes$ to ternary or higher-arity operators requires additional formalization.

Regional and dialectal variation. The phonological alphabet Σ assumes a standardized ASL phonology. Regional variants (e.g., Black ASL, regional signs) would require either:

- Multiple codebooks with dialect-specific quantizers
- Probabilistic pronunciation models capturing variation
- Meta-learning across dialect domains

Spontaneous signing vs. citation forms. Training data (WLASL, PHOENIX) consists primarily of careful signing. Fast conversational ASL exhibits greater coarticulation, reduction, and prosodic variation. The context transducer C models some coarticulation, but may require data-driven augmentation.

Annotation costs. Phonological annotation (labeling Z_t with (H, L, O, M, N)) is more expensive than gloss-level labels. We propose mitigation strategies:

- **Weak supervision**: Train on gloss labels, infer phonology via CTC
- **Self-supervision**: Pre-train encoder on unlabeled video with contrastive learning
- **Active learning**: Select high-uncertainty frames for expert annotation

Discourse annotation. Locus assignments and referent tracking require specialized annotation. Current datasets lack this information. We recommend augmenting future datasets with:

- Bounding boxes for referent locations during noun introduction
- Co-reference chains linking pronouns to antecedents
- Agreement verb source/target annotations

12.3 Ethical Considerations and Community Engagement

Deaf Community Collaboration. This work is presented as a *technical framework* requiring validation and refinement by the Deaf community. Key principles for responsible development:

- **Nothing about us without us**: Deaf researchers, native signers, and community stakeholders must be involved in all stages—design, data collection, evaluation, and deployment decisions.
- **Linguistic sovereignty**: ASL is a complete natural language with its own grammar, not "visual English." The phonological and grammatical formalizations in this work

must be validated against established ASL linguistics and refined based on native signer feedback.

- **Regional and cultural variation**: ASL exhibits significant regional variation (e.g., Black ASL, regional dialects). A single ”standard” model risks linguistic imperialism. Production systems should support multiple varieties.
- **Consent and privacy**: Video data contains biometric information and reveals cultural/linguistic identity. Collection requires informed consent with clear data usage policies. Prioritize on-device processing to minimize data exposure.

Use Case Boundaries. **Appropriate applications**:

- ASL learning tools for hearing students (with proper pedagogical design)
- Accessibility features in video conferencing (user-controlled, opt-in)
- Research tools for linguistic analysis (with community partnership)

Inappropriate applications:

- Replacing human interpreters in medical, legal, or educational settings (technology not sufficiently reliable; human interpreters provide cultural mediation beyond word-for-word translation)
- Employment screening or evaluation (risk of bias, discriminatory impact)
- Surveillance or monitoring without explicit consent

Known Limitations and Risks.

- **Gloss-based evaluation**: Using English glosses as ground truth is fundamentally limited—ASL grammar differs from English. Better evaluation requires native signer assessment of translations in context.
- **Representation bias**: Training data may underrepresent certain demographics (age, race, gender, regional background). Models will perform worse for underrepresented groups.
- **Prosody and pragmatics**: Mathematical formalization captures phonology and some morphosyntax but misses prosody, affect, and pragmatic meaning—all essential to natural communication.
- **False confidence**: Deployment without proper failure handling risks miscommunication in high-stakes scenarios.

13 Conclusion

We have presented a mathematically grounded framework for large-vocabulary ASL translation that addresses the core challenges of multimodality, spatial grammar, and scalability. The key innovations are:

1. **Phonological factorization** with group-theoretic invariance guarantees, reducing sample complexity and enabling compositional generalization
2. **Spatial discourse algebra** with probabilistic uniqueness bounds for referent tracking via pointing and agreement morphology
3. **Non-associative morphological fusion** capturing classifier-driven verb modification within a formal operator framework

4. **WFST decoding cascade** ($H \circ C \circ M \circ D \circ L \circ G$) enabling efficient beam search with provable soundness and completeness
5. **Information-theoretic modality selection** providing principled guidance for accuracy-cost tradeoffs
6. **Joint training framework** with convergence guarantees for multi-task objectives

Collectively, these ingredients enable scaling to 5,000–10,000 signs while maintaining real-time inference (¡200 ms latency) on edge and web platforms. The framework is mathematically rigorous—every component has formal semantics and theoretical guarantees—yet practically grounded in MediaPipe landmarks and proven ASR architectures.

Future Directions. Several extensions merit investigation:

- **Continuous space locus tracking**: Replace voxel discretization with Kalman filtering for smooth locus drift modeling
- **Hierarchical morphology**: Extend $\otimes$ to handle aspectual modification, pluralization, and recursive embedding
- **Cross-lingual transfer**: Leverage shared phonological structure across signed languages (ASL, BSL, DGS) for low-resource translation
- **Multimodal pre-training**: Self-supervised learning on large unlabeled video corpora to reduce annotation requirements
- **Neural-symbolic integration**: Combine WFST symbolic reasoning with neural sequence modeling for hybrid architectures
- **Real-world evaluation**: User studies with Deaf signers in naturalistic settings to measure practical utility

The mathematical linguistics foundation established here provides a principled starting point for tackling these challenges. By combining formal rigor with practical engineering and meaningful community engagement, we can build ASL translation systems that are both theoretically sound and respectful of linguistic and cultural sovereignty.

References

[1] S. Albanie et al. BSL-1K: Scaling up co-articulated sign language recognition using mouthing cues. In *ECCV*, 2020.

[2] L. Bottou, F. E. Curtis, and J. Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, 2018.

[3] N. C. Camgöz, S. Hadfield, O. Koller, H. Ney, and R. Bowden. Neural sign language translation. In *CVPR*, 2018.

[4] Z. Chen, V. Badrinarayanan, C.-Y. Lee, and A. Rabinovich. GradNorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *NeurIPS*, 2018.

[5] O. Koller, N. C. Camgöz, H. Ney, and R. Bowden. Weakly supervised learning with multi-stream CNN-LSTM-HMMs to discover sequential parallelism in sign language videos. *IEEE TPAMI*, 42(9):2306–2320, 2020.

[6] P. Dreuw, D. Rybach, T. Deselaers, M. Zahedi, and H. Ney. Speech recognition techniques for a sign language recognition system. In *Interspeech*, 2007.

[7] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In *ICML*, 2006.

[8] O. Koller, J. Forster, and H. Ney. Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers. *Computer Vision and Image Understanding*, 141:108–125, 2015.

[9] S. K. Liddell. *Grammar, Gesture, and Meaning in American Sign Language*. Cambridge University Press, 2003.

[10] J. Lillo-Martin and K. K. Meier. On the linguistic status of 'agreement' in sign languages. *Theoretical Linguistics*, 37(3-4):95–141, 2011.

[11] M. Mohri, F. Pereira, and M. Riley. Weighted finite-state transducers in speech recognition. *Computer Speech & Language*, 16(1):69–88, 2002.

[12] C. Padden. *Interaction of Morphology and Syntax in American Sign Language*. Garland Publishing, 1988.

[13] T. Starner, J. Weaver, and A. Pentland. Real-time American Sign Language recognition using desk and wearable computer based video. *IEEE TPAMI*, 20(12):1371–1375, 1998.

[14] W. C. Stokoe. Sign language structure: An outline of the visual communication systems of the American deaf. *Studies in Linguistics: Occasional Papers*, 8, 1960.