Phonological Perception of Sign Language Models

Kayo Yin*
University of California, Berkeley

Jessica Carter
Johns Hopkins University

Alex X. Lu
Microsoft Research

Annemarie Kocab
Johns Hopkins University

Abstract

Sign languages are compositional systems where meaning arises by combining sublexical phonological parameters, such as handshape, location, and movement. While deep learning models for Sign Language Recognition (SLR) have achieved increased performance on translation benchmarks, it remains unclear whether these models distinguish abstract phonological features or merely rely on low-level statistical correlations. This work evaluates the phonological perception of SLR models trained on American Sign Language (ASL) by probing phonological sensitivity using minimal pairs and evaluating representational alignment with human behavioral data. Our results reveal that SLR models exhibit emergent phonological sensitivity, but with clear architectural trade-offs: pose-based models are sensitive to handshape contrasts, while pixel-based models better capture location changes. Furthermore, pose-based models learn latent representations that correlate with human perceptual similarity judgments ($r \approx 0.49$). These findings suggest that while SLR models exhibit emergent phonology, current training paradigms are insufficient to scale them beyond their architectural inductive biases.[1]

1 Introduction

In signed languages, signs are constructed from sublexical phonological parameters: handshape, location, movement, orientation, and non-manual markers (Stokoe, 1960). While modern Sign Language Recognition (SLR) models achieve high translation fidelity (Camgoz et al., 2020; Yin & Read, 2020; Guan et al., 2025), it remains unclear if they acquire robust phonological representations. Current evaluations focus on translation accuracy on narrow datasets (Desai et al., 2024). However, high accuracy does not guarantee linguistic competence; it could mask a limited ability to generalize beyond training distributions. High scores may instead stem from overfitting to supplementary cues such as mouthing, which—unlike core phonological parameters—is optional and varies across signs and signers. The model may not generalize to the significant variability in signing across signers, or overfit to the recording conditions in the training data. This reliance on spurious features poses significant limitations for using SLR models as general-purpose tools for linguistic analysis, especially for low-resource sign languages. Given that distinguishing phonological parameters is critical for human lexical access (Lieberman & Borovsky, 2020; Williams et al., 2017), if models are not sensitive to the true underlying phonology, they lack the robust representational quality required for cross-linguistic transfer or language understanding.

In this work, we evaluate whether SLR models exhibit phonological sensitivity by probing them with **minimal pairs** of signs (Figure 1). These signs differ by one phonological element (e.g., *QUEEN* vs. *KING*, which share the same articulation location and sign movement and differ only in their handshape). This framework allows us to explicitly measure whether the model's latent space distinguishes signs along their phonological dimensions.

*Correspondence: kayoyin@berkeley.edu

[1]Code and data: https://github.com/kayoyin/sign-phonology.

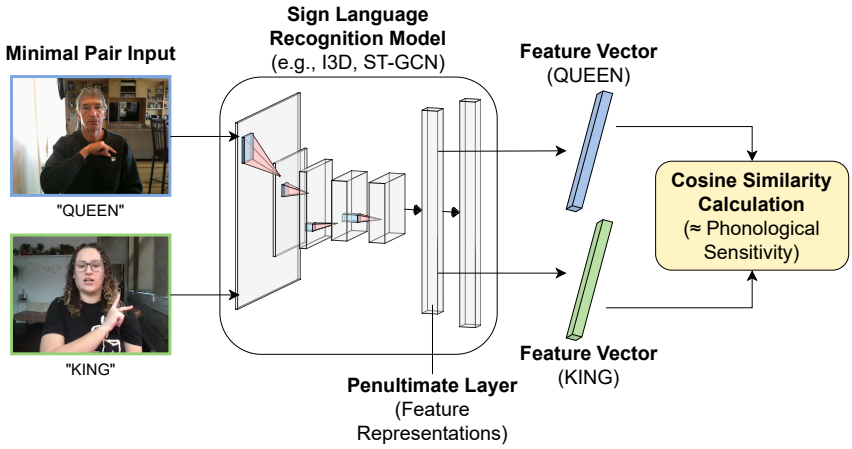


Figure 1: **Auditing phonological sensitivity with minimal pairs**. Feature representations of each sign in a minimal pair (e.g. *QUEEN* vs. *KING*) are extracted from the model's penultimate layer. We quantify the model's sensitivity to the phonological contrast by calculating the cosine similarity between these latent representations.

Furthermore, by studying latent representations instead of model outputs, our framework directly compares various architectures and data sources without requiring explicit training on phonological labels.

We apply this framework in two experiments. First, we assess discriminability in naturalistic American Sign Language (ASL) data. We compare distinct architectures (pixel-based vs. pose-based) and training domains (sign language vs. general action recognition) to determine which features facilitate phonological sensitivity.

Second, we investigate the alignment between model representations and human perception. We use controlled synthetic data to study how the latent space of models organize handshapes compares to theoretical models of phonology based on human perception data.

Our results reveal trade-offs in model sensitivity governed by architectural constraints. While pixel-based models better capture broader spatial contrasts like sign location, pose-based models demonstrate superior sensitivity to fine-grained handshape distinctions. We also find that the latent space of pose-based models significantly aligns with human perceptual judgments, partially reproducing the confusion patterns and hierarchical groupings found in human signers. These findings suggest that while emergent phonology exists in SLR models, models designed with different visual processing show different phonological sensitivities, highlighting the critical importance of cognitively grounded benchmarks to guide the development of robust computational sign systems.

2 Background & Related Work

Phonological structure of sign languages. Sign languages contain sublexical structure where meaningful signs are constructed from combinatorial parameters. We adopt the standard five-parameter model established by Battison (1978) and Liddell & Johnson (1989), which expands Stokoe (1960)'s original classification to include: **handshape**, **location**, **movement**, **palm orientation**, and **non-manual markers**. This compositional structure allows us to define *minimal pairs* – signs that differ by exactly one parameter – to measure discriminability. While originally described for ASL, this parameter model also generalizes across different sign languages (Sandler, 2012). Psycholinguistic research demonstrates that signers activate these sublexical representations incrementally during lexical access (Meade et al., 2018), while behavioral studies show that phonological similarity influences sign production and recognition speeds (Carreiras et al., 2008; Williams et al., 2017).

Sign language recognition. SLR research has largely prioritized end-to-end translation accuracy, often treating models as "black boxes" whose internal representations remain unexamined (Camgoz et al., 2020; Yin & Read, 2020). Existing attempts to analyze phonology typically rely on explicit supervision, training classifiers to predict specific features (Tavella et al., 2022; Bilge et al., 2022). However, this approach is limited by data scarcity for rare features and statistical confounds (e.g., specific handshapes correlate with specific locations, so a classifier can score well by predicting a feature whenever a correlated parameter occurs, rather than by detecting the parameter of interest). To avoid these validity issues, we assess *implicit* phonological emergence using a minimal pair strategy. This method controls for confounding variables and requires no phonological labels, allowing us to probe the model's latent structure directly.

Phonological representations in spoken language models. Analogous work in speech processing demonstrates that in neural networks, phonological structure is emergent without explicit supervision. Research shows that phonological information is encoded in the lower to middle layers of speech recognition models (Belinkov & Glass, 2017; Pasad et al., 2021) and can be retrieved via linear geometry (Gauthier et al., 2025). We extend this line of inquiry to the visual-spatial domain, investigating whether sign language models exhibit similar emergent sensitivity to phonological constraints.

3 Methodology

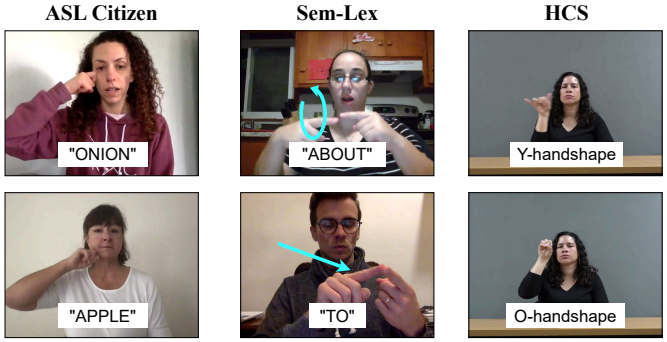
To assess the phonological sensitivity of SLR models, we use a **minimal pair** framework. Rather than relying on final classification outputs, we analyze the models' latent representations of sign pairs that differ by a single phonological parameter. This approach allows us to quantify sensitivity to specific phonological changes in models trained without explicit phonological supervision.

3.1 Minimal pairs and datasets

We curate minimal pairs from three datasets: ASL Citizen (Desai et al., 2023), Sem-Lex (Kezar et al., 2023), and Handshapes in Context Stimuli (Carter et al., 2026). We adopt the parameter inventory of Battison (1978) and Liddell & Johnson (1989) for minimal-pair decisions: each minimal pair contains exactly one change in handshape, location, movement, orientation, or non-manual markers, verified by a Deaf author. Our experiments focus on handshape, location, and movement contrasts, as orientation and non-manual contrasts are too sparse in the available naturalistic corpora to evaluate reliably. We provide samples from each dataset in Figure 2.

ASL Citizen. ASL Citizen is a large-scale, crowd-sourced dataset featuring 52 deaf or hard-of-hearing (DHH) signers. A Deaf author curated 259 minimal pairs spanning 360 unique signs to cover diverse phonological changes: 137 pairs differ in handshape, 31 in location, 73 in movement, and 18 in orientation or non-manual markers. To establish a baseline for human perceptual similarity, we enlisted a hearing novice signer to annotate the pairs as "very similar" (51 pairs), "somewhat similar" (98 pairs), or "not similar" (103 pairs). We use the standard test split, which includes 4,427 videos from 11 signers for the selected signs.

Sem-Lex. As all SLR models in this study were trained on ASL Citizen, to test for generalization, we also curated pairs from the Sem-Lex Benchmark. Unlike ASL Citizen, where users mimicked a seed video, Sem-Lex prompted signers with concepts, which resulted in natural variability in sign production. We mapped our ASL Citizen minimal pairs to Sem-Lex by manually verifying that each video sharing a gloss matched the intended sign, and excluding lexical variants whose production differed from its pair partner in more than one parameter. This ensures our framework evaluates robust phonological sensitivity rather than being confounded by mislabeled videos or lexical variation. This yielded 178 pairs across 263 signs (8,934 videos from 42 signers).



	ASL Citizen	Sem-Lex	HCS
I3D	t = 7.71	t = 0.49	sim = 0.91
STGCN	t = 1.39	t = 6.19	sim = 0.99

Figure 2: **Examples of minimal pair data and model sensitivity metrics.** (Left, Center) Naturalistic minimal pairs from ASL Citizen and Sem-Lex contrasting in Handshape and Location, respectively. Below: *t*-test results indicate the statistical significance of SLR models' ability to distinguish these pairs. (Right) A controlled minimal pair from HCS contrasting Handshape. Below: Cosine similarity scores quantify the distance between the two nonce signs in the models' latent space (lower similarity = higher sensitivity).

Handshapes in Context Stimuli (HCS). Real-world vocabularies often lack systematic coverage of phonological combinations. To address this, we use HCS, a set of controlled experimental stimuli consisting of nonce signs designed to systematically sample variations in handshape and phonological context. The design of these stimuli was guided by prior phonological inventories (Lane et al., 1976; Stungis, 1981), while the specific handshape inventory was expanded from 20 to 36 shapes based on frequency scores from a lexical database (Sehyr et al., 2021). HCS includes 972 videos covering 36 handshapes in 27 phonological contexts (3 locations × 3 movements × 3 orientations). This systematic design allows for precise minimal pair contrasts even for features rare in natural vocabulary. Unlike ASL Citizen and Sem-Lex, HCS features one Deaf signer with one video per nonce sign, recorded in a controlled laboratory setting.

3.2 Models

We evaluate variants of two distinct video classification architectures, I3D and STGCN, which represent the predominant modeling paradigms [2] in the current SLR literature (i.e. vision versus pose-based) (Desai et al., 2023). By comparing these standard backbones, we investigate how *training domain* (general human action vs. sign language recognition) and *input modality* (raw pixels vs. pose estimations) influence phonological perception. The two input modalities are illustrated in Appendix A.1.

I3D (Pixel-based). The Inflated 3D ConvNet (I3D; Carreira & Zisserman (2017)) is a two-stream architecture operating directly on RGB video frames. We evaluate three variants: (1) **I3D-Rand**, initialized with random weights to serve as a baseline; (2) **I3D-Kine**, trained on the Kinetics dataset for general action recognition; and (3) **I3D-ASL**, trained on ASL Citizen for sign language recognition.

STGCN (Pose-based). The Spatio-Temporal Graph Convolutional Network (STGCN; Yan et al. (2018)) operates on a skeletal graph representation of the body. We evaluate: (1) **STGCN-Rand**, a randomly initialized baseline; and (2) **STGCN-ASL**, trained on ASL Citizen. We excluded an STGCN model trained on Kinetics because existing pre-trained models use action recognition pose graphs that lack fine-grained hand landmarks.

[2] I3D and STGCN are the dominant open-weight representatives of pixel- and pose-based SLR at the time of publication. Although more advanced systems have been proposed in more recent work, these largely do not make their weights openly available, which precludes systematic comparison.

Table 1: **Model sensitivity to phonological changes on ASL Citizen.** Results are grouped by linguistic feature (Handshape, Location, Movement) and human perception of similarity. t-stat represents the mean t-statistic calculated across all minimal pairs in that category. % wins (vs. Baseline) denotes the percentage of pairs where the trained model had a significant t-statistic ($p < 0.05$) that was higher than its randomly initialized counterpart. Head-to-Head columns show the percentage of pairs where one trained model (I3D-ASL or STGCN-ASL) significantly outperformed the other ($p < 0.05$). Note that Head-to-Head percentages do not sum to 100% as pairs with no significant t-stat are excluded. STGCN is generally more sensitive to phonological changes than I3D, except for changes in location.

	Baseline	**Pixel-based (I3D)**				**Baseline**	**Pose-based (STGCN)**		**Head-to-Head**	
	I3D-Rand	I3D-Kinetics		**I3D-ASL**		STGCN-Rand	**STGCN-ASL**		**I3D-ASL**	**STGCN-ASL**
Category	t-stat	t-stat	% wins	t-stat	% wins	t-stat	t-stat	% wins	% wins	% wins
Total ($N = 259$)	0.13	0.12	1.93	5.00	69.50	0.23	**6.08**	**81.08**	33.20	**56.37**
Control ($N = 259$)	0.06	0.76	—	**10.98**	—	0.44	**12.66**	—	—	—
By Phonological Parameter										
Handshape ($N = 137$)	0.14	0.12	0.73	4.80	69.34	0.18	**6.30**	**86.13**	29.20	**64.96**
Location ($N = 31$)	0.02	1.13	6.45	**7.73**	**80.65**	0.59	5.67	67.74	**64.52**	16.13
Movement ($N = 73$)	0.19	0.20	1.37	4.72	73.97	0.30	**5.81**	**80.82**	38.36	**54.79**
By Human Perception										
Not similar ($N = 103$)	0.15	0.35	0.97	5.01	66.99	0.10	**6.98**	**89.32**	28.16	**64.08**
Somewhat similar ($N = 98$)	0.17	0.06	2.04	5.08	76.53	0.43	**5.64**	**79.59**	37.76	**56.12**
Very similar ($N = 51$)	0.07	0.17	3.92	4.26	56.86	0.03	**4.62**	**64.71**	31.37	**43.14**

Strict exclusion of test signs. To ensure our analysis measures generalized phonological sensitivity rather than lexical memorization, we re-trained both ASL-trained models (I3D-ASL and STGCN-ASL) on a dataset split that explicitly excludes all signs appearing in our minimal pair test sets.

Table 2: **Out-of-domain phonological sensitivity results on Sem-Lex.** Models trained on ASL Citizen were evaluated on minimal pairs from the Sem-Lex dataset to test domain generalization. Both models demonstrate a reduced sensitivity to phonological differences in this out-of-domain setting compared to ASL Citizen.

	I3D-ASL		**STGCN-ASL**		**Head-to-Head**	
Category	t-stat	% wins	t-stat	% wins	I3D	STGCN
Total ($N = 71$)	2.47	45.07	**2.76**	**49.30**	25.35	**33.80**
Control ($N = 71$)	**8.39**	—	**10.40**	—	—	—
By Phonological Parameter						
Handshape ($N = 26$)	3.29	57.69	**3.93**	**61.54**	30.77	**38.46**
Location ($N = 8$)	3.52	**75.00**	**4.79**	62.50	25.00	**50.00**
Movement ($N = 25$)	**2.47**	36.00	1.63	**48.00**	32.00	32.00
By Human Perception						
Not similar ($N = 23$)	**3.92**	**60.87**	3.07	56.52	34.78	**39.13**
Somewhat sim. ($N = 31$)	2.05	41.94	**2.39**	**45.16**	**29.03**	25.81
Very similar ($N = 17$)	1.29	29.41	**2.99**	**47.06**	5.88	**41.18**

4 Are Models Sensitive to Phonological Changes?

4.1 Experiment

We evaluate whether SLR models learn latent representations that preserve the phonological distinctions found in minimal pairs. By focusing on the latent feature space rather than final prediction accuracy, we establish a unified evaluation framework that allows for direct comparison across diverse model architectures and training objectives. Furthermore, we posit that robust phonological separability in the latent space is a strong indicator of a model's capacity for zero-shot transfer—a critical capability for analyzing underrepresented sign languages where training data are scarce.

To analyze these representations, we extract feature vectors from the model's penultimate layer (preceding the classification head) for videos in our naturalistic minimal pair datasets,

where there are multiple samples per sign (ASL Citizen and Sem-Lex). We operationalize "phonological sensitivity" by comparing intra-sign similarity against inter-sign similarity. Our hypothesis is that if a model is sensitive to phonological change, the cosine similarity between two *minimally different* signs should be **significantly lower** than two instances of the *same* sign (produced by different signers).

Formally, for each minimal pair of signs (s_1, s_2) we construct two distributions of similarity scores. **(1) Intra-sign similarity (D_{same}):** The cosine similarity between the representation of sign s_1 produced by signer A and an instance of the same sign s_1 produced by a different signer B. **(2) Inter-sign similarity (D_{diff}):** The cosine similarity between the representation of sign s_1 produced by signer A and the phonologically distinct sign s_2 produced by signer B. **(3) Sampling:** We repeat these computations for 10 random pairs of signers (A, B) to populate the distributions D_{same} and D_{diff}.

To quantify phonological sensitivity, we perform a two-tailed independent samples t-test to determine if the distribution of similarities of D_{same} is statistically significantly higher than that of D_{diff}:

$$t(s_1, s_2) \;=\; \frac{\bar{D}_{same} - \bar{D}_{diff}}{\sqrt{s^2_{same}/N + s^2_{diff}/N}}, \tag{1}$$

where $\bar{D}$ and s^2 denote the sample mean and variance of each similarity distribution, and N is the number of sampled signer pairs.

As a control, we repeat this procedure on random non-minimal sign pairs sampled from the same gloss pool, providing an upper bound on the t-statistics that trained models should trivially achieve (*Control* rows in Tables 1–2).

4.2 Results

We present the results of t-tests on ASL Citizen and Sem-Lex in Table 1 and Table 2, respectively. For each phonological category, we report the **mean t-statistic** (magnitude of separation) and the **% wins** (frequency with which the trained model significantly outperforms the randomly initialized baseline, $p < 0.05$). Additionally, we provide head-to-head comparisons indicating how often one ASL-trained model proved significantly more sensitive than the other. In Appendix A.2, we also provide qualitative examples of model sensitivity results to minimal pairs in ASL-Citizen.

First, **training on sign language data is prerequisite for phonological sensitivity,** which notably deviates from human learning. Across all categories, models trained on general action recognition (I3D-Kinetics) failed to distinguish minimal pairs, achieving a negligible 1.93% win rate over random baselines on ASL Citizen. In contrast, both ASL-trained models demonstrated significant emergent sensitivity (I3D-ASL: 69.50%; STGCN-ASL: 81.08%). The *Control* row confirms that the ASL-trained models separate random non-minimal pairs far more strongly than minimal pairs (e.g., $t = 12.66$ vs. 6.08 for STGCN-ASL), while the untrained and Kinetics baselines remain at chance.

This failure of the general action model is surprising given that phonological sensitivity builds upon general visual capabilities: prior work demonstrates that non-signers can reliably distinguish phonological contrasts in ASL and there is a strong correlation between signer and non-signer perceptual confusion (Stungis, 1981). Our results suggest that current video models are poor proxies for human perception. Unlike the human visual system, which can transfer general object and motion recognition to identify sign contrasts, these models require domain-specific training to acquire phonological sensitivity. We attribute this to the training signal itself: general action datasets like Kinetics lack the fine-grained phonological contrasts found in sign language, and so never pressure the model to encode the subtle distinctions that sign-specific training captures.

Second, **architectures exhibit distinct biases in what aspects of phonology they are sensitive to.** On ASL Citizen, the pose-based STGCN-ASL model demonstrates superior

sensitivity to **handshape** over I3D-ASL (86.13% vs. 69.34%) and **movement** (80.82% vs. 73.97%). This suggests that explicit skeletal modeling better captures fine-grained configuration and trajectory than raw pixels. Conversely, the pixel-based I3D-ASL model significantly outperformed STGCN-ASL on **location** contrasts (80.65% vs. 67.74%), winning the head-to-head comparison in 64.52% of location pairs. This indicates that pixel-based models may better retain the spatial context relative to the frame (e.g., forehead vs. chin) that graph-based representations – which are often normalized for position – can lose.

Third, **model sensitivity aligns with human perception.** For both architectures, performance degrades for minimal pairs rated by a human non-signer as perceptually more similar. A Kruskal-Wallis H-test confirms a statistically significant difference in STGCN-ASL sensitivity across the "Not / Somewhat / Very Similar" human ratings ($H(2) = 12.92, p = 0.002$), validating that the models struggle most with the distinctions humans may find subtle. In contrast, this difference was not statistically significant for the I3D-ASL model or the non-ASL trained variants.

Finally, **representations are brittle to domain shifts.** When evaluated on Sem-Lex, which introduces variations in prompts and filming conditions, sensitivity drops substantially (I3D-ASL: 45.07%; STGCN-ASL: 49.3%). This suggests that while current models learn phonological distinctions, their representations remain highly brittle to the specific production constraints of their training distribution.

5 How Does Model Phonological Sensitivity Compare to Human Perception?

5.1 Experiment

Beyond assessing whether models *can* distinguish phonological features, we investigate whether their internal representation space captures meaningful phonological structure. We use the HCS data to compare model feature distances with subjective human perception and articulatory geometry. While naturalistic datasets like ASL Citizen offer ecological validity, they suffer from sparse phonological coverage. HCS allows us to control for non-phonological variance (e.g., lighting, signer identity) while systematically sampling the full spectrum of handshape contrasts.

First, we quantify **human perception** using confusion matrices collected by Stungis (1981). This data quantifies how frequently human signers and non-signers confuse different ASL handshapes. We treat the confusion frequency as a proxy for perceptual similarity: handshapes that are confused more often are considered more perceptually similar.

Second, we calculate the **handshape distance (HD)** between handshape pairs. HD, a computational measure of articulatory geometry adapted from Yin et al. (2024), is defined as the mean angular difference between their corresponding joints. For each pair of handshapes in HCS data, we computed the HD for the handshape pair in 27 phonological contexts and took the median.

Third, we measure **model sensitivity** to the phonological contrast in each pair of handshapes in HCS by calculating the cosine similarity between their corresponding video representations in the model's latent space. We then compute the Pearson correlation coefficient (r) between the vector of human confusion frequencies, handshape distances, and model feature similarities. We summarize the results of our correlation analysis in Table 3.

To visualize the structural organization of the learned feature spaces, we performed U-statistic hierarchical clustering (D'Andrade, 1978) of handshapes according to HD and model feature similarity, following the methodology in Stungis (1981). In Figure 3, we plotted dendrograms for four representations: (1) the **Lane-Boyes-Braem II (LBB2) Model** (a theoretical model based on human perceptual studies (Lane et al., 1976; Stungis, 1981)), (2) **Handshape Distance** (our geometric reference metric), (3) **I3D-ASL**, and (4) **STGCN-ASL**. This allows us to inspect whether models recover high-level phonological features that humans rely on, or if they use different features for distinguishing handshapes.

Table 3: **Correlation between perceptual similarity reported by humans (Stungis, 1981), handshape distance (HD), and cosine similarity of model features of handshapes.** Significant correlations ($p < 0.05$) are bolded. Models trained on ASL data, especially STGCN, have higher alignment with human perception.

	Reference			I3D			STGCN	
	Non-signer	Signer	HD	Rand	Kine	ASL	Rand	ASL
Non-signer	–	**0.89**	**0.48**	0.04	**0.19**	**0.32**	0.10	**0.48**
Signer	**0.89**	–	**0.53**	0.03	**0.19**	**0.31**	0.14	**0.49**
HD	**0.48**	**0.53**	–	0.10	**0.19**	**0.20**	0.06	**0.55**

5.2 Results

Correlation analysis. First, we established reference bounds for our analysis. Consistent with prior literature, human signers and non-signers have high agreement ($r = 0.89$). The theoretical HD metric showed moderate positive correlation with human perception, suggesting that human perception is grounded in articulatory geometry ($r = 0.53$ for signers). As expected, models with randomly initialized weights (*Rand*) yielded insignificant correlations, confirming that untrained networks do not inherently encode phonologically meaningful features.

Second, we find that **training on ASL data significantly improved alignment with human perception**. For the I3D architecture, the model trained on general human action recognition (*Kine*) has a weak correlation with human signers ($r = 0.19$). Training on ASL data increased this correlation to $r = 0.31$ for the I3D model, suggesting that exposure to sign language data drives the feature space closer to human perceptual representations.

Third, we observe clear **architectural differences in handshape perception** between models. STGCN-ASL is the model with the highest alignment with human signers ($r = 0.49$). The correlation between I3D-ASL and STGCN-ASL is itself notably low ($r = 0.19$). This indicates that while both models successfully distinguish some handshapes, their architectures induce different structures in their latent spaces. STGCN-ASL also shows a strong correlation with handshape distance ($r = 0.55$), substantially higher than I3D-ASL ($r = 0.20$). I3D, lacking the skeletal prior of pose-based models, seems to learn a visual abstraction that correlates less with joint geometry and the human perception of handshape. We also note that the alignment between STGCN-ASL and humans is not higher than the alignment between HD and humans, which may suggest that training on sign language allows for similarities already present in input space to trickle through the model, but does not enhance perceptual similarity past what is expected from joint geometry.

Hierarchical clustering analysis. The dendrograms reveal distinct topological differences between models that corroborate the correlation results. The **LBB2 Model** organizes handshapes primarily by the 'Compact' feature (whether the three middle fingers are closed). Within the '+Compact' group, it further distinguishes based on the number of bent fingers, while the '-Compact' group is organized by the number of extended fingers. Similarly, the **HD dendrogram** organizes handshapes by finger count, creating a primary split between handshapes with ≤ 1 finger extended (with the exception of two handshapes that has the thumb and another finger extended) versus those with > 1. It further sub-groups the first cluster by specific articulators (thumb vs. index extension) and the second by the count of extended fingers.

The **I3D-ASL dendrogram** diverges the most from the LBB2 Model. It primarily separates by handshapes that visually resemble a fist from those that do not. Within the 'fist' cluster, it sorts by closure tightness. The non-fist cluster separates a group with extended thumbs but leaves the remaining handshapes in sub-clusters lacking clear patterns. This suggests I3D relies primarily on global visual texture (e.g., "compact fist" vs. "open hand with protruding thumb") rather than fine-grained articulatory features.

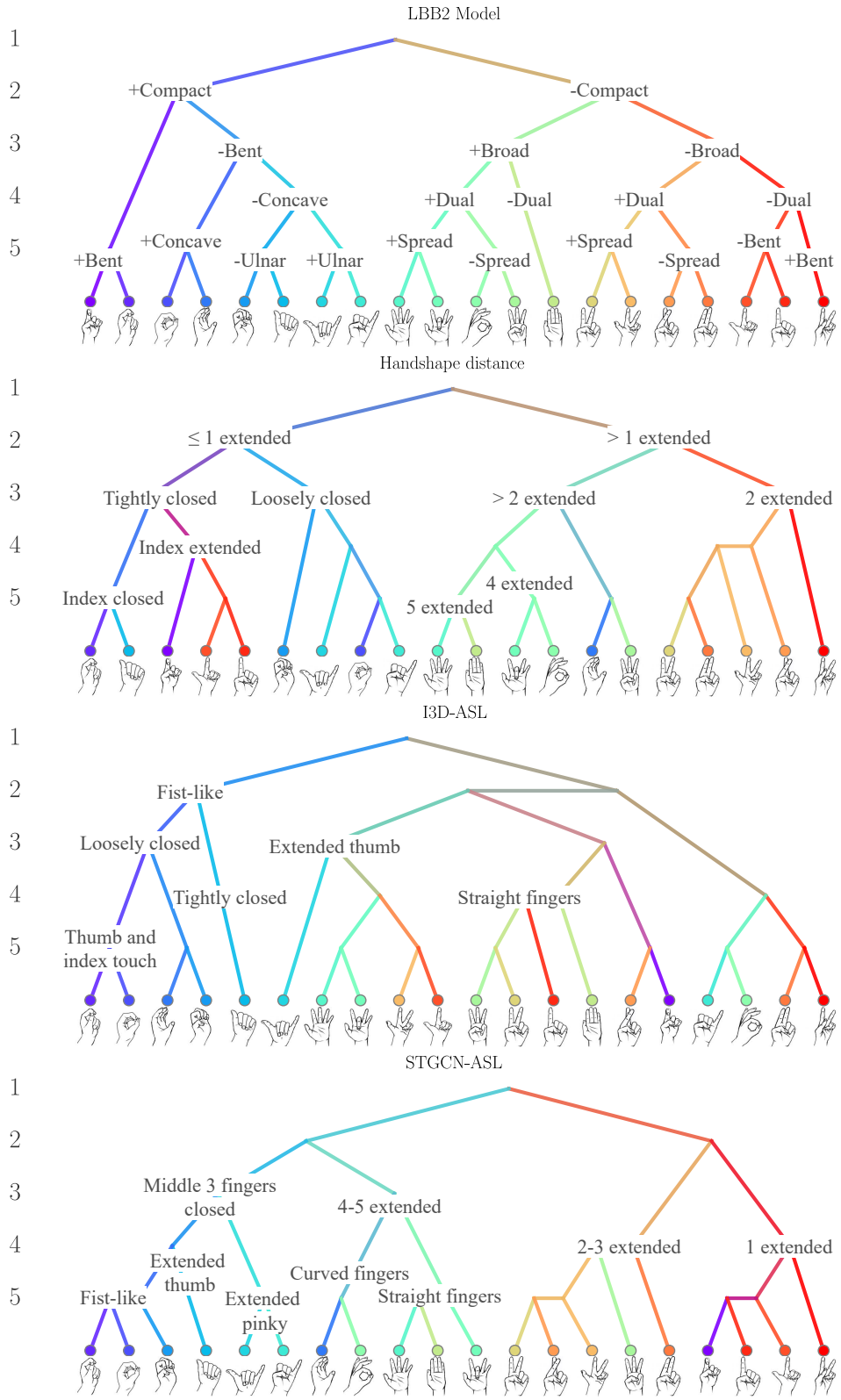


Figure 3: **Structural organization of handshape representations.** From top to bottom: the LBB2 Model (theoretic model based on human perception), handshape distance (geometric reference), I3D-ASL, STGCN-ASL.

In contrast, the **STGCN-ASL dendrogram** closely recovers the structure of the LBB2 Model. It organizes handshapes into four primary clusters: (1) all three middle fingers are closed, (2) 4-5 extended fingers, (3) 2-3 extended fingers, and (4) just one extended finger (with the exception of the 'K' handshape on the far right). STGCN also captures fine-grained sub-distinctions within these groups: in the first cluster, it distinguishes thumb/pinky extension; in the second, it separates curved fingers from straight extension. This structural alignment explains the STGCN's high correlation with human perception (LBB2) and the theoretical HD metric.

6 Discussion & Future Work

Our investigation reveals that while phonological sensitivity emerges without supervision in SLR models, it is fractionated by architectural inductive biases: pose-based STGCN are more sensitive to fine-grained handshape contrasts, whereas pixel-based I3D are better at location distinctions. This functional split suggests that current training methods and architectures are insufficient for holistic sign language processing, and that a more complete model likely requires dual-stream integration of the skeletal abstraction captured by pose-based models and the holistic spatial processing captured by pixel-based models.

Furthermore, the strong correlation of STGCN and human perception serves as a benchmark for the structural integrity of the latent space in skeletal models. However, these representations remain brittle to domain shifts. The significant drop in phonological sensitivity on out-of-domain data such as Sem-Lex exposes a tendency to overfit production variables (e.g., recording conditions) rather than acquiring robust phonological abstractions. Thus, scaling current training paradigms on naturalistic data may not be sufficient to induce the robust phonological representations required for real-world applications.

While our framework for phonological sensitivity advances SLR evaluation, our study has three primary limitations. First, data sparsity restricted our analysis to handshape, location, and movement, omitting palm orientation and non-manual markers. Second, we used only supervised baselines due to the unavailability of open-source models; future work should investigate whether self-supervised objectives induce different representations. Finally, our focus on isolated signs leaves the effects of co-articulation and syntax in continuous SLR for future study.

Beyond these, our approach enables broader exploration of model robustness and generalizability. Future research can probe compositionality of model features, similarly to how word embeddings enable analogical reasoning, and explore multimodal architectures that bridge the trade-off between I3D and STGCN models. The minimal pairs could also be used to improve training itself: they supply hard negatives for contrastive fine-grained phonological discrimination, and could pre-train models to first acquire phonological distinctions that then transfer to lower-resource sign languages. Future work could also test for cross-linguistic transfer between sign languages to see whether phonological representations learned from higher-resource sign languages could generalize to lower-resources sign languages; testing transfer across structurally unrelated sign languages could further isolate phonological constraints grounded in human body biomechanics from language-specific learning, which would inform data-efficient modeling.

Acknowledgements

KY is supported by the Vitalik Buterin Ph.D. Fellowship in AI Existential Safety. AK is supported by NSF STEM-APWD 136073-5130403.

References

Robbin Battison. *Lexical borrowing in American sign language.* ERIC, 1978.

Yonatan Belinkov and James Glass. Analyzing hidden representations in end-to-end automatic speech recognition systems. *Advances in Neural Information Processing Systems*, 30, 2017.

Yunus Can Bilge, Ramazan Gokberk Cinbis, and Nazli Ikizler-Cinbis. Towards zero-shot sign language recognition. *IEEE transactions on pattern analysis and machine intelligence*, 45 (1):1217–1232, 2022.

Necati Cihan Camgoz, Oscar Koller, Simon Hadfield, and Richard Bowden. Sign language transformers: Joint end-to-end sign language recognition and translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10023–10033, 2020.

Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6299–6308, 2017.

Manuel Carreiras, Eva Gutiérrez-Sigut, Silvia Baquero, and David Corina. Lexical processing in spanish sign language (lse). *Journal of Memory and Language*, 58(1):100–122, 2008.

Jessica Carter, Annemarie Kocab, et al. Manuscript in preparation. Manuscript in preparation, 2026.

Aashaka Desai, Lauren Berger, Fyodor O Minakov, Vanessa Milan, Chinmay Singh, Kriston Pumphrey, Richard E Ladner, Hal Daumé III, Alex X Lu, Naomi Caselli, and Danielle Bragg. Asl citizen: A community-sourced dataset for advancing isolated sign language recognition. *arXiv preprint arXiv:2304.05934*, 2023.

Aashaka Desai, Maartje De Meulder, Julie A Hochgesang, Annemarie Kocab, and Alex X Lu. Systemic biases in sign language ai research: A deaf-led call to reevaluate research agendas. *arXiv preprint arXiv:2403.02563*, 2024.

Roy G D'Andrade. U-statistic hierarchical clustering. *Psychometrika*, 43(1):59–67, 1978.

Jon Gauthier, Canaan Breiss, Matthew K Leonard, and Edward F Chang. Emergent morpho-phonological representations in self-supervised speech models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 28055–28074, 2025.

Mo Guan, Yan Wang, Guangkun Ma, Jiarui Liu, and Mingzu Sun. Mska: Multi-stream keypoint attention network for sign language recognition and translation. *Pattern Recognition*, 165:111602, 2025.

Lee Kezar, Jesse Thomason, Naomi Caselli, Zed Sehyr, and Elana Pontecorvo. The sem-lex benchmark: Modeling asl signs and their phonemes. *Proceedings of the 25th International ACM SIGACCESS Conference on Computers and Accessibility*, 2023.

Harlan Lane, Penny Boyes-Braem, and Ursula Bellugi. Preliminaries to a distinctive feature analysis of handshapes in american sign language. *Cognitive Psychology*, 8(2):263–289, 1976.

Scott K Liddell and Robert E Johnson. American sign language: The phonological base. *Sign language studies*, 64(1):195–277, 1989.

Amy M Lieberman and Arielle Borovsky. Lexical recognition in deaf children learning american sign language: Activation of semantic and phonological features of signs. *Language learning*, 70(4):935–973, 2020.

Gabriela Meade, Brittany Lee, Katherine J Midgley, Phillip J Holcomb, and Karen Emmorey. Phonological and semantic priming in american sign language: N300 and n400 effects. *Language, Cognition and Neuroscience*, 33(9):1092–1106, 2018.

Ankita Pasad, Ju-Chieh Chou, and Karen Livescu. Layer-wise analysis of a self-supervised speech representation model. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 914–921. IEEE, 2021.

Wendy Sandler. The phonological organization of sign languages. *Language and linguistics compass*, 6(3):162–182, 2012.

Zed Sevcikova Sehyr, Naomi Caselli, Ariel M Cohen-Goldberg, and Karen Emmorey. The asl-lex 2.0 project: A database of lexical and phonological properties for 2,723 signs in american sign language. *The Journal of Deaf Studies and Deaf Education*, 26(2):263–277, 2021.

Jr. Stokoe, William C. Sign Language Structure: An Outline of the Visual Communication Systems of the American Deaf. *The Journal of Deaf Studies and Deaf Education*, 10(1):3–37, 01 1960. ISSN 1081-4159. doi: 10.1093/deafed/eni001. URL https://doi.org/10.1093/deafed/eni001.

James Stungis. Identification and discrimination of handshape in american sign language. *Perception & Psychophysics*, 29(3):261–276, 1981.

Federico Tavella, Aphrodite Galata, and Angelo Cangelosi. Phonology recognition in american sign language. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8452–8456. IEEE, 2022.

Joshua T Williams, Adam Stone, and Sharlene D Newman. Operationalization of sign language phonological similarity and its effects on lexical access. *The Journal of Deaf Studies and Deaf Education*, 22(3):303–315, 2017.

Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Kayo Yin and Jesse Read. Better sign language translation with STMC-transformer. In Donia Scott, Nuria Bel, and Chengqing Zong (eds.), *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 5975–5989, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.525. URL https://aclanthology.org/2020.coling-main.525/.

Kayo Yin, Terry Regier, and Dan Klein. Pressures for communicative efficiency in american sign language. In *Annual Conference of the Association for Computational Linguistics (ACL)*, August 2024.

A Appendix

A.1 SLR model input modalities

In Figure 4, we show examples of the video input modality to the I3D and STGCN models.

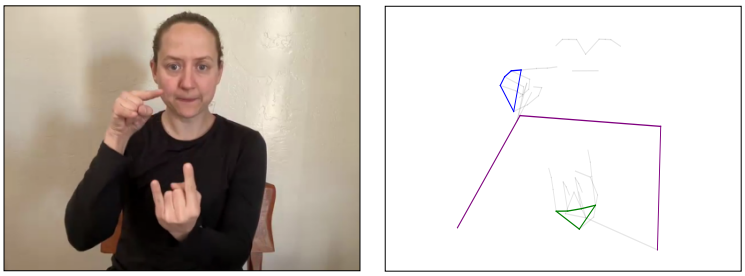


Figure 4: **Comparison of model input modalities.** (Left) Raw RGB video frames processed by the I3D model. (Right) Skeletal pose graph estimation processed by the STGCN model. Both inputs depict a frame from the sign "HIPPO."

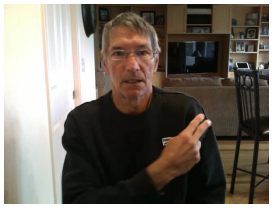
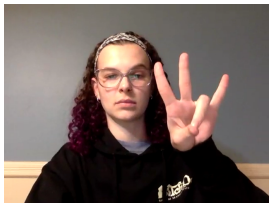
A.2 Qualitative analysis of minimal pair sensitivity

In Table 4, we provide qualitative examples of *t*-test results for I3D-ASL and STGCN-ASL models. The first three examples show cases where both models fail to represent the phonological change in minimal pairs of varying similarity as judged by humans. The first

pair (*HOSPITAL / PATIENT*) has a change in handshape, where the handshapes primarily differ by the position of the index finger. The second pair (*5DOLLARS / FIFTH*) has a change in movement, where *5DOLLARS* is signed by twisting the wrist once, while *FIFTH* is signed by twisting the wrist inward-outward in small movements. The third pair (*BEAVER / TABLE*) has a change in location, where the elbow of the dominant arm and wrist of the non-dominant arm stay in contact for *BEAVER*, but neither elbows and wrists stay in contact during the movements for *TABLE*.

The final two examples show the divergence between architectures' phonological sensitivity. In example 4, only STGCN fails to distinguish *FEEL / HAPPY*, where these signs mostly differ in whether the middle finger is bent or straight. In example 5, only I3D fails to distinguish *SEVEN / NINE*, where the handshapes differ in the finger that is in contact with the thumb.

Table 4: Qualitative examples of model sensitivity to minimal pairs in ASL Citizen.

Minimal Pair	**Sign 1**	**Sign 2**	**Human**	**I3D**	**STGCN**
HOSPITAL / PATIENT			Very similar	0.56	-0.26
5DOLLARS / FIFTH			Somewhat similar	0.10	-1.01
BEAVER / TABLE			Not similar	0.24	-0.81
FEEL / HAPPY			Somewhat similar	4.89	-1.26
SEVEN / NINE			Not similar	-0.72	10.71