Phonological Perception of Sign Language Models

Kayo Yin (kayoyin@berkeley.edu)[1], Jessica Carter[2], Alex X. Lu[3] & Annemarie Kocab[2]
[1]University of California, Berkeley
[2]Johns Hopkins University
[3]Microsoft Research

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 SLR model phonological perception by probing sensitivity using minimal pairs and measuring representational alignment with human behavioral data. Our results reveal emergent phonological sensitivity with clear architectural trade-offs: pose-based models are more 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 overcome their architectural inductive biases[1].

Keywords: sign language; phonology; deep learning.

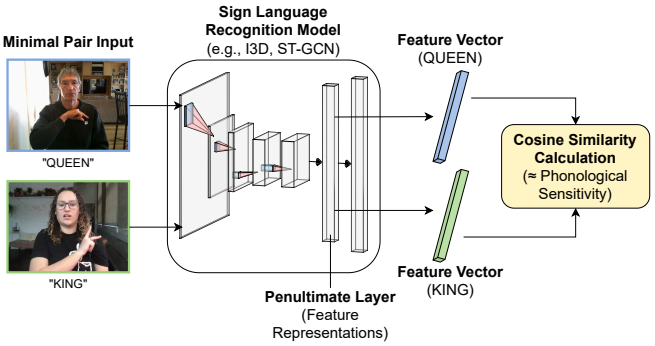

Figure 1: **Auditing phonological sensitivity with minimal pairs**. Feature representations of each sign in a minimal pair (e.g. *QUEEN* vs. *KING*) are extracted from the model's penultimate layer. We quantify the model's sensitivity to the phonological contrast by calculating the cosine similarity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

Introduction

A central question in language and cognition is how sensory systems abstract meaningful linguistic units from raw perceptual signals. 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; Guan et al., 2025; Yin & Read, 2020), it remains unclear whether they acquire human-like phonological representations. Current evaluations focus on translation accuracy on narrow datasets (Desai et al., 2024). However, high accuracy does not guarantee linguistic competence: high scores may stem from overfitting to supplementary cues like mouthing, which, unlike core phonological parameters, is optional and varies between signs and signers. Given that distinguishing these phonological parameters is critical for human lexical access (Lieberman & Borovsky, 2020; Williams et al., 2017), model reliance on such spurious features would suggest limitations in generalizability and a fundamental divergence from human processing.

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 differ only in handshape). This framework allows us to explicitly measure whether the model's latent space distinguishes signs along their phonological dimensions. 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 organizes handshapes compared 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

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

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

models, different visual processing architectures yield different phonological sensitivities, highlighting the critical importance of cognitively grounded benchmarks to guide the development of robust computational sign systems.

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 and 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. 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

Most Sign Language Recognition (SLR) models prioritize end-to-end translation (Camgoz et al., 2020; Yin & Read, 2020), often treating models as "black boxes" with unexamined internal representations. Existing attempts to analyze phonology typically rely on explicit supervision, training classifiers to predict specific features (Bilge et al., 2022; Tavella 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 performance is possible by predicting positives when a correlated parameter occurs, as opposed to 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 neural networks spontaneously acquire phonological structure. 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.

Methodology

We analyze latent representations of sign pairs differing by a single phonological parameter, requiring no explicit phonological supervision.

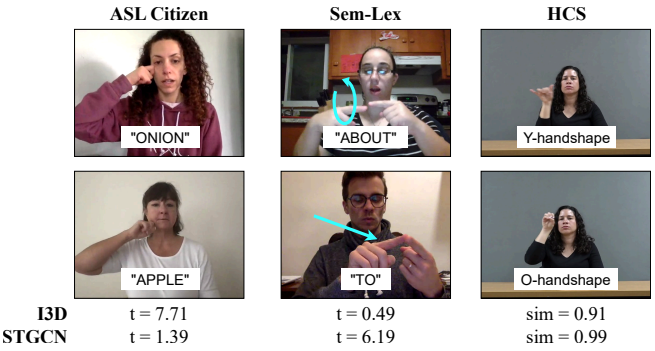

Figure 2: **Examples of minimal pair data and model sensitivity metrics.** (Left, Center) Naturalistic minimal pairs from ASL Citizen and Sem-Lex contrasting in Handshape and Location, respectively. Below: *t*-test results indicate the statistical significance of SLR models' ability to distinguish these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

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, Kocab, et al., 2026). Each minimal pair includes one change in either handshape, location, movement, orientation (Battison, 1978) or non-manual markers. 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 human perceptual baseline, we enlisted a hearing novice signer to annotate pairs as "very similar" (51 pairs), "somewhat similar" (98 pairs), or "not similar" (103 pairs). We use the standard test split: 4,427 videos from 11 signers for the selected signs.

Sem-Lex. Since all SLR models were trained on ASL Citizen, to test generalization, we curated pairs from Sem-Lex. Unlike ASL Citizen, where users mimicked a seed video, Sem-Lex prompted signers with concepts, resulting in natural production variability. We mapped our ASL Citizen minimal pairs to Sem-Lex by manually verifying that each video with the same gloss matched, and excluding variants of minimal pairs that differed in more than one parameter. This ensures our framework evaluates robust phonological sensitivity rather than being confounded by mislabeled videos. This yielded 178 pairs across 263 signs (8,934 videos from 42 signers).

Handshapes in Context Stimuli (HCS). Real-world vocabularies often lack systematic coverage of phonological combinations. Furthermore, signer variation in the above datasets

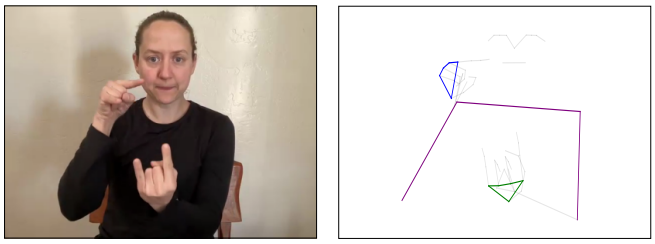

Figure 3: **Comparison of model input modalities.** (Left) Raw RGB video frames processed by the I3D model. (Right) Skeletal pose graph estimation processed by the STGCN model. Both inputs depict a frame from the sign "HIPPO."

may introduce near-minimal pairs. To address this, we use HCS, controlled experimental stimuli of nonce signs that systematically sample handshape and phonological context, containing only strict minimal pairs. 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.

Models

We evaluate two video classification architectures, I3D and STGCN, representing the predominant modeling paradigms in SLR literature (Desai et al., 2023), and are the publicly available state-of-the-art. 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 Figure 3.

I3D (Pixel-based). The Inflated 3D ConvNet (I3D; Carreira and 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-Kine variant as standard action recognition pose graphs lack the fine-grained hand landmarks necessary for sign phonology.

Strict exclusion of test signs. To ensure we measure phonological sensitivity rather than lexical memorization, we retrained I3D-ASL and STGCN-ASL on a split excluding all signs in our minimal pair test sets.

Are Models Sensitive to Phonological Changes?

Experiment

We evaluate whether SLR models learn latent representations that preserve the phonological distinctions found in minimal pairs. Focusing on latent representations rather than prediction accuracy enables direct comparison across diverse architectures and training objectives.

We extract feature vectors from the model's penultimate layer for videos in ASL Citizen and Sem-Lex, which have multiple samples per sign. We operationalize "phonological sensitivity" by comparing intra-sign similarity against inter-sign similarity. If a model is sensitive to phonological change, cosine similarity between two *minimally different* signs should be **significantly lower** than between instances of the *same* sign 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.

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 an independent samples t-test to determine whether D_{same} is significantly higher than D_{diff}.

As a control, we repeat the above procedure on random non-minimal sign pairs sampled from the same gloss pool, providing an upper bound on t-statistics that trained models should trivially achieve.

Results

We present t-test results on ASL Citizen and Sem-Lex in Tables 1 and 2, respectively. Per category, we report the mean t-statistic, % wins (frequency of outperforming the random baseline, $p < 0.05$), and head-to-head comparisons.

First, **training on sign language data is a prerequisite for phonological sensitivity,** which notably deviates from human learning. Models trained on general action recognition (I3D-Kine) 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 shows that ASL-trained models distinguish random sign pairs with larger effects than

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-Kine		**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**

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**

minimal pairs, while untrained and Kinetics baselines remain at chance, situating minimal-pair sensitivity within a broader scale where non-minimal pairs with more phonological differences are correspondingly more separated. While phonological sensitivity may build upon general visual capabilities — prior work demonstrates that non-signers can reliably distinguish phonological contrasts in ASL, with strong correlation between signer and non-signer perceptual confusion (Stungis, 1981) — our results suggest that current video models are poor proxies for human perception. General action datasets like Kinetics lack the fine-grained phonological contrasts found in sign language, failing to provide the model with the specific signal needed 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.

How Does Model Phonological Sensitivity Compare to Human Perception?

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 4, 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.

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.

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**

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 ($r = 0.49$). Also, the correlation between I3D-ASL and STGCN-ASL is 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 with human perception of handshape.

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 have 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.

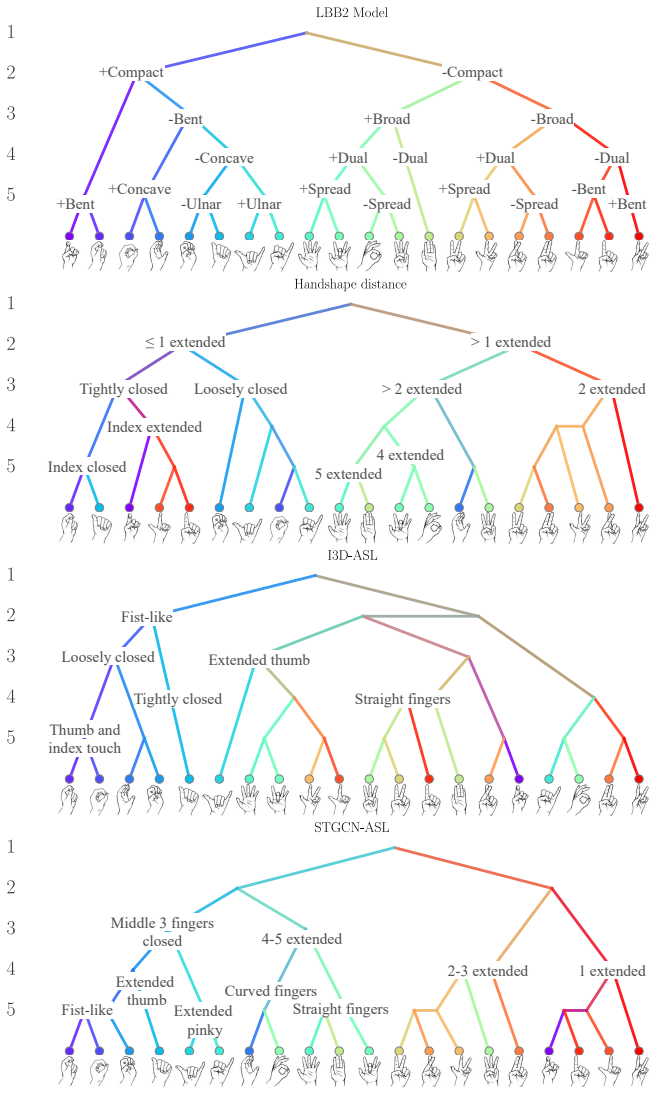

Figure 4: **Structural organization of handshape representations.** From top to bottom: the LBB2 Model (theoretic model based on human perception), handshape distance (geometric reference), I3D-ASL, STGCN-ASL.

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.

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 STGCN's high correlation with human perception (LBB2) and the theoretical HD metric.

Discussion & Future Work

Phonological sensitivity emerges without explicit supervision in SLR models, but is fractionated by architectural inductive biases: pose-based STGCNs are more sensitive to fine-grained handshape contrasts, whereas pixel-based I3Ds are better at location distinctions. This functional split suggests that current training methods and architectures are insufficient; a cognitively plausible model likely requires a dual-stream integration of skeletal abstraction and holistic scene processing.

The strong correlation between STGCN and human perception indicates skeletal graphs facilitate human-like representation structure, validating their potential as proxies for sign perception. However, these representations remain brittle. The significant drop on out-of-domain Sem-Lex exposes a tendency to overfit production variables (e.g., recording conditions) rather than acquiring robust phonological abstractions.

Our study has three primary limitations. First, our current analysis focused on three of the five phonological parameters — handshape, location, and movement — due to the sparsity of annotations for palm orientation and non-manual markers in existing data. Future work should evaluate whether models capture the full grammar of sign beyond manual features. Second, we rely on established supervised baselines (I3D, STGCN) due to the unavailability of open-source weights for state-of-the-art self-supervised sign language models. Future work should verify whether self-supervised objectives, which leverage massive unlabeled datasets, induce distinct phonological representations. Third, we focus on isolated sign recognition. It remains unclear whether our findings transfer to continuous sign language recognition models, where models must resolve co-articulation and syntactic non-manual markers. These sentence-level constraints may pressure models to learn different phonological abstractions.

Our minimal pairs framework opens several broader avenues for research. For example, we could probe geometric compositionality of model features: verifying whether these vectors support algebraic operations over phonological parameters (e.g., analogical reasoning) would distinguish true linguistic abstraction from statistical clustering. In addition, testing for cross-linguistic universals by evaluating zero-shot transfer of SLR models to other sign languages (e.g., BSL, newly emerging sign languages) could isolate phonological constraints grounded in human body biomechanics from language-specific learning, and inform data-efficient modeling. Finally, analyzing the ontogeny of representations during training, such as comparing the emergence of phonological sensitivity in model learning curves to the "coarse-to-fine" trajectory of child language acquisition, may yield insights into the biases that learners bring to language learning.

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

Battison, R. (1978). *Lexical borrowing in american sign language*. ERIC.

Belinkov, Y., & Glass, J. (2017). Analyzing hidden representations in end-to-end automatic speech recognition systems. *Advances in Neural Information Processing Systems*, *30*.

Bilge, Y. C., Cinbis, R. G., & Ikizler-Cinbis, N. (2022). Towards zero-shot sign language recognition. *IEEE transactions on pattern analysis and machine intelligence*, *45*(1), 1217–1232.

Camgoz, N. C., Koller, O., Hadfield, S., & Bowden, R. (2020). Sign language transformers: Joint end-to-end sign language recognition and translation. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10023–10033.

Carreira, J., & Zisserman, A. (2017). Quo vadis, action recognition? a new model and the kinetics dataset. *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6299–6308.

Carreiras, M., Gutiérrez-Sigut, E., Baquero, S., & Corina, D. (2008). Lexical processing in spanish sign language (lse). *Journal of Memory and Language*, *58*(1), 100–122.

Carter, J., Kocab, A., et al. (2026). *Manuscript in preparation* [Manuscript in preparation].

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

Desai, A., Berger, L., Minakov, F. O., Milan, V., Singh, C., Pumphrey, K., Ladner, R. E., Daumé III, H., Lu, A. X., Caselli, N., & Bragg, D. (2023). Asl citizen: A community-sourced dataset for advancing isolated sign language recognition. *arXiv preprint arXiv:2304.05934*.

Desai, A., De Meulder, M., Hochgesang, J. A., Kocab, A., & Lu, A. X. (2024). Systemic biases in sign language ai research: A deaf-led call to reevaluate research agendas. *arXiv preprint arXiv:2403.02563*.

Gauthier, J., Breiss, C., Leonard, M. K., & Chang, E. F. (2025). Emergent morpho-phonological representations in self-supervised speech models. *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 28055–28074.

Guan, M., Wang, Y., Ma, G., Liu, J., & Sun, M. (2025). Mska: Multi-stream keypoint attention network for sign language recognition and translation. *Pattern Recognition*, *165*, 111602.

Kezar, L., Thomason, J., Caselli, N., Sehyr, Z., & Pontecorvo, E. (2023). The sem-lex benchmark: Modeling asl signs and their phonemes. *Proceedings of the 25th International ACM SIGACCESS Conference on Computers and Accessibility*.

Lane, H., Boyes-Braem, P., & Bellugi, U. (1976). Preliminaries to a distinctive feature analysis of handshapes in american sign language. *Cognitive Psychology*, *8*(2), 263–289.

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

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

Meade, G., Lee, B., Midgley, K. J., Holcomb, P. J., & Emmorey, K. (2018). Phonological and semantic priming in american sign language: N300 and n400 effects. *Language, Cognition and Neuroscience*, *33*(9), 1092–1106.

Pasad, A., Chou, J.-C., & Livescu, K. (2021). Layer-wise analysis of a self-supervised speech representation model. *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 914–921.

Sehyr, Z. S., Caselli, N., Cohen-Goldberg, A. M., & Emmorey, K. (2021). 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.

Stokoe, J., William C. (1960). 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. https://doi.org/10.1093/deafed/eni001

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

Tavella, F., Galata, A., & Cangelosi, A. (2022). Phonology recognition in american sign language. *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 8452–8456.

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

Yan, S., Xiong, Y., & Lin, D. (2018). Spatial temporal graph convolutional networks for skeleton-based action recognition. *Proceedings of the AAAI conference on artificial intelligence*, *32*(1).

Yin, K., & Read, J. (2020, December). Better sign language translation with STMC-transformer. In D. Scott, N. Bel, & C. Zong (Eds.), *Proceedings of the 28th international conference on computational linguistics* (pp. 5975–5989). International Committee on Computational Linguistics. https://doi.org/10.18653/v1/2020.coling-main.525

Yin, K., Regier, T., & Klein, D. (2024). Pressures for communicative efficiency in american sign language. *Annual Conference of the Association for Computational Linguistics (ACL)*.