Modeling Sign Language Phonology through Motion-Based Embeddings

Anonymous ACL submission

Abstract

Existing computational approaches to modeling sign languages often rely on representations not explicitly designed to capture phonological structure. As a result, key phonological components (handshape, movement, location) remain difficult to model systematically. We propose a motion-centric framework for learning phonological embeddings of sign language from visual input, capturing structured motion patterns and temporal dynamics. It is instantiated using Motion History Images (MHI) combined with a spatio-temporal encoder and temporal attention, resulting in representations structured in a compact embedding space using metric learning. We adopt a semi-hard triplet loss to organize signs according to motion-based similarity. For evaluation, we annotate phonological features, including number of hands, symmetry, and major location, and analyze if the resulting sign clusters share these properties and display coherent phonological patterns. For American Sign Language (ASL) and Peruvian Sign Language (LSP), we capture motion-based phonological properties, enabling meaningful clustering of signs. These results highlight the potential of our approach to model sign phonology directly from visual data, reducing reliance on manual annotation and enabling scalable analysis for sign languages. The code is publicly available at https://anonymous.4open.science/r/motion_based_sign_phonology-F0C2.

1 Introduction

Phonology plays a central role in the organization of all language lexicons, governing how form-based contrasts are structured to distinguish one element from another. In sign languages, this phonological structure is expressed through parameters such as movement, location, handshape, and non-manual features, which collectively define the internal composition of a sign.

Despite its importance, modeling phonological structure computationally remains challenging. Several works annotate different phonological parameters manually to compare signs (Naranjo-Zeledón et al., 2020; Crasborn et al., 2001; Cormier et al., 2016), a process that requires expert knowledge and does not scale to large datasets due to the time-consuming nature of fine-grained annotation. While resources such as ASL-LEX and SignBank provide curated phonological descriptions, they are limited in coverage and are unavailable for many sign languages, including Peruvian Sign Language (LSP). For these reasons, computationally derived embedding representations offer a scalable approach to capturing phonological structure.

To address this limitation, we propose a data-driven approach to learn phonological embeddings directly from visual input. Our method focuses on capturing motion and spatio-temporal dynamics, enabling the model to learn consistent patterns of sign formation. These representations are then structured in an embedding space using metric learning, encouraging signs with similar motion patterns to be closer while separating dissimilar ones. This approach enables the model to learn phonological representations without requiring manual annotation.

2 Related Work

Existing work on sign language representation includes linguistic studies of sign language phonology, annotation-based approaches, and computational methods for sign representation learning. In this section, we review these lines of work and discuss their limitations, followed by learning-based approaches for modeling motion in video.

2.1 Manual annotation for phonological representations

Phonology studies the features that distinguish linguistic units and how they combine. In spoken languages, phonology is expressed through sound, whereas in sign languages, it is expressed through parameters such as movement, location, handshape, and non-manual features.

The phonology of sign languages has been extensively studied following the proposal that their structures and systems are comparable to the grammars of spoken languages (Stokoe, 1960). Over time, the field has been enriched by diverse perspectives, including language acquisition (Newport et al., 1985; Lillo-Martin, 2009), language processing (Emmorey, 2020), psycholinguistics (Klima and Bellugi, 1979; Steinbach et al., 2019), and historical linguistics (Frishberg, 1975; Sandler and Lillo-Martin, 2006; Moita et al., 2023). Studying phonology in sign or oral languages requires notation systems that enable systematic analysis. The process of creating these notations is not without controversy, and various systems have been proposed.

Traditionally, phonological annotation has relied on manual methods using tools such as ELAN, iLex, Anvil, Transana or SignStream (Neidle, 2002), as well as conventions developed in projects such as ECHO (Schembri and Crasborn, 2010), which served as the basis for corpora in sign languages including those of Australia, Ireland, the Netherlands, and the United Kingdom. However, the lack of theoretical consensus limits standardization, and annotation remains resource-intensive, requiring significant time, expertise, and funding, which hinders scalability.

From a computational perspective, several methods have been proposed to automate annotation and feature extraction. These include semi-automatic annotation approaches based on tracking data (Dreuw and Ney, 2008) and pose-estimation methods to extract phonological features using normalized mathematical models (Şahin and Gökgöz, 2024). Despite these advances, such approaches remain limited: they often require human supervision (Dreuw and Ney, 2008), are sensitive to input quality (Şahin and Gökgöz, 2024), for instance if it loses track of a finger, the glossed phonological category may be inexact.

Similarly, computational methodologies have been proposed to quantify phonological similarity by representing signs through parametric encodings and measuring distances to identify phonological neighbors (Naranjo-Zeledón et al., 2020). While these approaches enable quantitative analysis, they rely on manually annotated phonological features, making them time-intensive and difficult to scale. As a result, large-scale, fine-grained phonological annotations, particularly those produced by expert annotators, remain scarce for most sign languages.

Recently, vector space models have been adopted to explore phonological phenomena. Annotated datasets have been used to study how phonological similarity relates to semantic organization by encoding phonological features such as handshape, location, and movement into vector representations (Martinez del Río et al., 2022).

2.2 Deep Learning Approaches for Sign Language Representation

Recent work in computer vision has explored learning motion representations directly from video using deep neural networks. Attention-based models help capture long-range temporal dependencies and identify informative temporal segments (Xue et al., 2025; Pondaven et al., 2025; Meinhardt et al., 2022). Feature extraction modules are commonly used to capture local motion patterns and are often combined with attention-based components to jointly model spatial and temporal dynamics for video generation (Wang, 2025; Liu et al., 2026; Ma et al., 2025) and classification (Zhang et al., 2023; Biswas and Le, 2024; Shin et al., 2023).

Metric learning approaches, such as triplet loss, have been widely used to structure embedding spaces by organizing representations according to similarity (Frome et al., 2007; Elezi et al., 2023). In the context of sign language, these methods have been applied to recognition tasks (Mopidevi et al., 2023; Damarjati et al., 2025; Sadeghzadeh and Islam, 2022). Several variants have been proposed, including semi-hard and hard negative mining strategies (Ge, 2018; Ott et al., 2023; Chung and Lee, 2023; Damarjati et al., 2025).

Existing work in sign language processing, including recognition (Hu et al., 2023; Wali et al., 2023; Tran et al., 2025) and translation (Liang et al., 2023; Gong et al., 2024; Zhang et al., 2024; Li et al., 2020), has largely relied on semantic supervision through gloss labels. Some approaches explore gloss-free sign language translation (Yin et al., 2023; Ye et al., 2024; Low et al., 2025), with the aim of reducing the dependence on explicit

semantic annotations.

In parallel, several works learn latent representation spaces for sign language through retrieval-based objectives (Jiang et al., 2024) and translation tasks (Li et al., 2020), which often result in representations aligned with semantic similarity. In contrast, our approach structures the embedding space to capture phonological similarities grounded in motion patterns.

3 Datasets

We use two sign language datasets containing 105 aligned concepts in Peruvian Sign Language (LSP) and American Sign Language (ASL). The LSP dataset consists of 50 native LSP signers, while the ASL dataset contains 40 native ASL signers. Each signer performs the same set of concepts across both datasets. The LSP videos were recorded at 60 fps using a green chroma background, whereas the ASL videos were recorded at 30 fps using a blue chroma background. Additionally, the LSP recordings were captured at 1K resolution, while the ASL recordings were captured at 4K resolution.

Both dataset videos have non-square spatial resolutions, so each frame is center-cropped to a square region with side length $\min(H, W)$, ensuring a consistent spatial format across samples. To reduce non-informative motion, videos are temporally trimmed to retain the central signing segment. Finally, MediaPipe Selfie Segmentation is applied to preserve primarily the hands and face regions.

4 Methodology

We propose a framework for learning motion-aware phonological embeddings from video sequences. The framework is designed to capture structured motion information and organize it into an embedding space that reflects phonological similarity in isolated sign language.

The framework consists of two components. The first component learns motion-aware spatio-temporal representations through a two-step process as shown in Figure 1. A spatio-temporal feature extractor captures local motion patterns, while a temporal attention module highlights the most informative temporal segments. The second component structures these representations in the embedding space using metric learning to align those similar to each other and split apart those different. The implementation of the model, which uses feature and temporal components, is detailed in the appendix A.

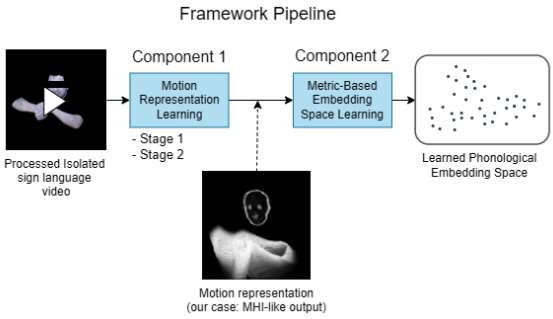

Figure 1: Pipeline of the framework, from input motion representations to the learned embedding space.

4.1 Motion Representation Learning

This component focuses on learning motion-aware spatio-temporal representations without relying on class-level supervision. To emphasize motion dynamics, the input is transformed into a motion-enhanced representation derived from temporal differences between frames, highlighting changes over time while reducing reliance on appearance cues such as color. The specific formulation of this representation is described in Appendix A.1.1.

The first stage focuses on learning local spatio-temporal motion patterns. This captures short-term motion dynamics and allows the encoder to model local changes and transitions in movement. This stage can be implemented using different spatio-temporal modeling approaches.

To capture the global organization of the sequence, the second stage performs temporal aggregation by assigning importance weights to different temporal segments. This emphasizes the most informative motion patterns while suppressing less relevant variations, and can be achieved using attention-based models that capture temporal dependencies across the sequence

The framework employs an auxiliary motion-based output as a learning objective to guide the encoder toward capturing motion dynamics from the input data. This objective can be instantiated using different motion representations. In this work, we adopt a Motion History Image (MHI). This objective provides structured guidance toward temporal variations and spatial motion patterns.

4.2 Metric-Based Embedding Space Learning

This component maps motion representations into a compact latent space in which geometric distances reflect motion-based similarity relationships. This

is achieved through metric learning objectives that learn relative distance constraints between samples, resulting in a structured space where similarity emerges as geometric proximity.

In this setting, class labels serve as a proxy for defining motion similarity relationships rather than as explicit semantic targets. The framework promotes an embedding space where samples with similar motion patterns are grouped together, while preserving meaningful inter-class relationships without enforcing global separation between all classes.

In this work, these relational constraints are instantiated using a semi-hard triplet loss objective, additional details are provided in Appendix A.2.2. To monitor the quality of the learned space, we computed several geometric metrics on the validation embeddings. These include intra-class compactness, inter-class separation, the inter/intra distance ratio, centroid margin statistics, nearest-centroid accuracy, and silhouette score using cosine distance. These metrics are not directly optimized during training, but are used to assess the geometric organization of the embedding space and guide model selection.

5 Evaluation of Phonological Embeddings

We evaluate the embedding space by analyzing its structure at the class level. The goal is to assess whether the learned representations capture meaningful motion characteristics between signs. We first obtain representative embeddings for each class and then analyze their relationships to identify groups of signs with similar motion patterns.

5.1 Preprocessing for Evaluation

After obtaining motion-aware embeddings, we applied a filtering step to remove atypical samples within each class. Specifically, samples whose individual pairwise distances exceed the 95th percentile of intra-class distances are excluded. This step reduces the influence of extreme instances, particularly in classes where such outliers are present, leading to more stable representations.

We compute a representative embedding for each class using a medoid-based approach. For each class, pairwise cosine distances are computed among all embeddings, and the sample that minimizes the average distance to the others is selected as the class medoid.

As medoids correspond to actual samples, they preserve realistic motion patterns and enable interpretable class-level representations. These medoids are then used for subsequent analysis by comparing distances between class representatives to identify groups of signs that share similar motion-based phonological patterns.

5.2 Phonological Clustering Modeling

To analyze the structure of the embedding space at the class level, we perform clustering over class medoids. Each medoid provides an embedding which represents each class, enabling analysis of inter-class relationships. We first apply Principal Component Analysis (PCA) to reduce the dimensionality of the L2-normalized embeddings to 50 components. To estimate the number of clusters, we use the Elbow Method over the range of $k \in \{2, \ldots, 49\}$. The elbow point was observed at k=11 for LSP and k=12 for ASL. Based on these values, we further evaluated K-means clustering for $k \in \{11, 12, 13\}$, analyzing the phonological patterns emerging within the resulting clusters for both LSP and ASL. Finally, we selected k=13, as it produced clusters with more coherent and distinguishable phonological patterns. For visualization purposes, we performed an additional 2D reduction using t-SNE (t-distributed stochastic neighbor embedding). Figure 2 provides the cluster visualizations using 2D t-SNE for LSP and ASL. The complete visualization for both can be found in the Appendix B.

When analyzing the clusters, we observed phonological patterns that differentiated clusters from one another. These patterns were primarily associated with the number of hands, movement symmetry, and major location, together with additional cluster-specific phonological characteristics. Regarding the number of hands, signs were categorized as one or two hands. Movement symmetry was analyzed only in two-handed signs. Two movement patterns were observed, including symmetrical movement, where both hands perform the same movement either simultaneously or alternately, and asymmetrical movement, where one hand acts as the base while the other performs the active movement. As for the major location, which is the location where the sign is made, we found neutral space, upper neutral space, mid neutral space, upper head space, lower face space, head space, and upper head space. Regarding the specific phonological characteristics for each cluster, different forms can be found, such as whether the sign is performed with multiple movements or a single movement,

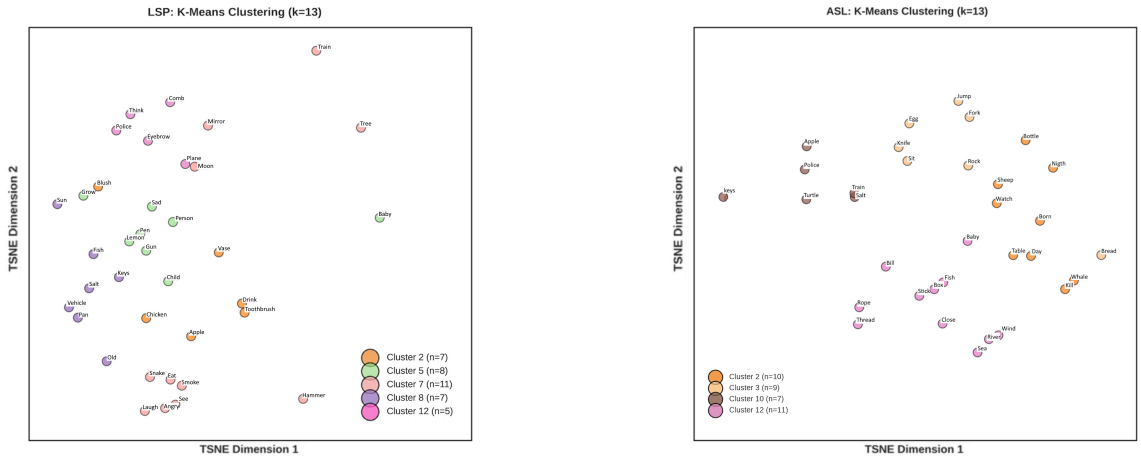

Figure 2: Clusters using the MHI-based embeddings for LSP (left) and ASL (right)

whether all fingers are selected or only some to perform the movement, whether it is an inward or outward path, or if the hands are intersecting, and others. Basically, we have focused on analyzing the movement, since embeddings are based on MHI, which focuses on capturing movement. Each cluster is defined by dominant phonological properties such as number of hands, symmetry, location, and movement characteristics.

Table 1 and Table 2 present the phonological characterization of the resulting clusters. Representative signs correspond to those matching the dominant cluster pattern, while outliers denote signs that deviate from one or more of these properties.

6 Discussion

The clustering results suggest that the learned embedding space organizes signs according to motion-based properties that align with the annotated phonological features. In both ASL and LSP, clusters tend to group signs based on phonological characteristics, at least with respect to the features we considered and annotated (number of hands, handshape symmetry, bimanual movement, movement symmetry, major location, and location symmetry). This indicates that the clusters capture, at least partially, the phonological structure of signs.

Importantly, similar structural patterns are observed across both datasets, despite differences in language and recording conditions. As shown in Figure 3, shared signs exhibit consistent clustering behavior across languages. This suggests that the learned representations capture motion regularities that are stable across linguistic contexts, although further analysis is required to determine the extent to which these patterns align with phonological structure.

6.1 Phonological Clustering Analysis LSP

For LSP, the thirteen clusters (C0–C12) reveal three major phonological groupings organized according to number of hands and movement symmetry. Group A (asymmetrical bimanual signs) comprises clusters C0 and C11; Group B (symmetrical bimanual signs) encompasses clusters C1, C3, C4, C6, C9, and C10; and Group C (one-handed signs) includes clusters C2, C5, C7, C8, and C12, as summarized in Table 1.

In Group A (asymmetrical bimanual signs), the signs in both clusters (C0 and C11) are performed in neutral space, but they differ in the specific phonological characteristics. C0 signs involve multiple movements such as TE, JUMP, FORK, BANANA, and others, while C11 signs involve a single movement such as DRAW, WORM, and CUP.

Group B includes symmetrical bimanual signs. Analyzing each cluster in this group reveals differences in major location and specific phonological characteristics. For example, C1, C3, C4, C6, and C10 are performed in neutral space, differing in the specific phonological characteristics. C1 signs have an inward or outward path such as CLOSE, DAY, BIG, NIGHT, and others. Clusters C3 and C6 also have all fingers selected, but they differ in that C3 (BOX, BILL, TABLE, EGG, and others) uses all fingers in multiple configurations, while C6 (PATH, COOK, SOCCER, and others) uses all fingers in a fist or open palm configuration. C4 signs are made with one or two fingers selected, includ-

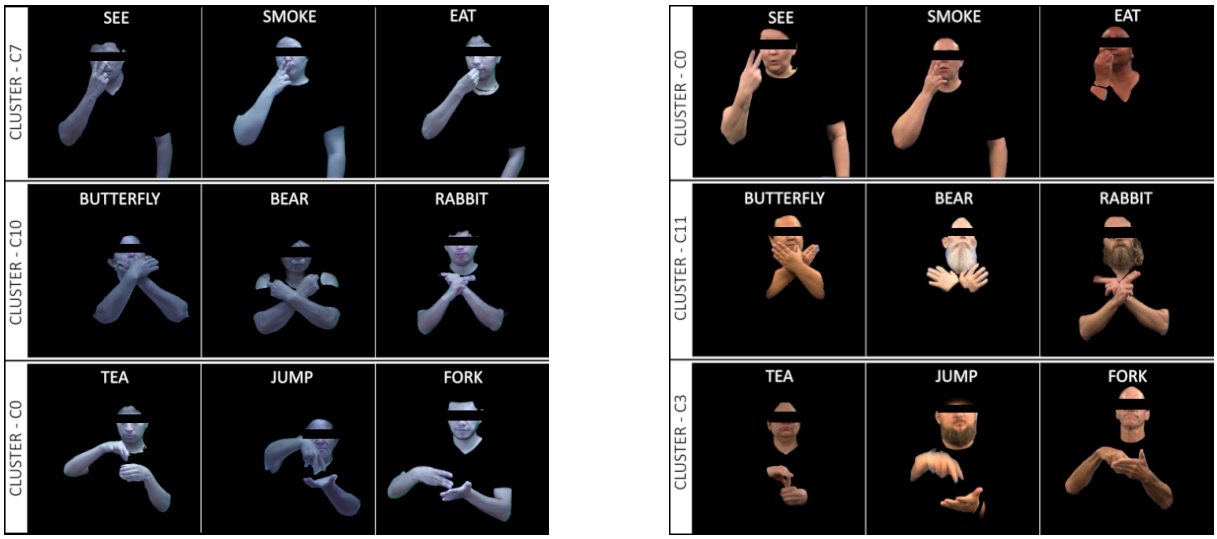

Figure 3: Examples of signs belonging to representative clusters C7, C10, and C0 in the LSP dataset (left) and C0, C11, and C3 in ASL dataset (right).

ing BED, BAG, KNIFE, SIT, and others. C10 is distinguished by intersecting hands (BUTTERFLY, BEAR, RABBIT, FLOWER, and others). This cluster has only one outlier: FIRE. Finally, C9 signs are articulated in the upper head space and involve forehead or overhead movements, including CROWN, CEILING, MOUNTAIN, and RAIN, among others. No outliers were identified in this cluster.

In group C (one-handed signs), the clusters are formed by signs made with a single hand, differentiated by their major location and specific phonological characteristics. C2 includes one-handed signs articulated in the lower face space, such as APPLE, TOOTHBRUSH, DRINK, CHICKEN. C5 and C8 are made in neutral space. The difference is that C5 is made in upper neutral space with a down or bottom path, such as GUN, LEMON, SAD, PERSON, GROW, and others. This cluster has only one outlier: BABY. C8 (VEHICLE, PAN, FISH, OLD, and others) is made in mid-neutral space with movement against the signer. Clusters C7 and C12 include signs articulated at head level, but they differ in location and finger configuration. C7 signs are produced in the upper or lower face region with two selected fingers, including SEE, SMOKE, SNAKE, among others. In contrast, C12 signs are articulated in the upper head space around the temple location, including POLICE, COMB, EYEBROW, PLANE and THINK. No outliers were identified in C12.

6.2 Phonological Clustering Analysis ASL

For ASL, the thirteen clusters (C0–C12) reveal four broad phonological groups organized according to number of hands and symmetry. Group A (asymmetrical bimanual signs) comprises clusters C2, C3, and C4; Group B (symmetrical bimanual signs) encompasses clusters C1, C6, C7, C9, C11, and C12; and Group C (one-handed signs) includes clusters C0, C5, and C8. Finally, Group D (mixed manual sings) consists of the single cluster C10, as summarized in Table 2.

In Group A (asymmetrical bimanual signs), we find signs where one hand is active and the other serves as a base for performing the movement. All clusters in this group are performed in neutral space, but each one is distinguished by a specific characteristic. For example, C2 signs are articulated relative to the non-dominant hand, including DAY, WHALE, BORN, NIGHT, and others. C3 is characterized by having movements articulated over the dominant hand, such as ROCK, KNIFE, JUMP, FORK, and others. Finally, C4 signs have the non-dominant hand with a palm shape, such as PEN, DRAW, READ, and COOK.

Group B includes symmetrical bimanual signs. Analyzing each cluster in this group reveals differences in major location and specific phonological characteristics. For example, C1, C6, C7, C11, and C12 are performed in neutral space, differing in their specific phonological characteristics. The C1 signs are characterized by movements that end near the lower face space, such as FIRE, CROCODILE, VASE, and DOOR. C6 and C7 also involve all fingers being selected, but differ in that C6 (HEALTH, CRAB, ANGRY, BREAK) has movements mostly outward to the sides, while

Cluster	Hands	Symmetry	Location	Feature	Signs	Outliers
Group A: Asymmetrical bimanual signs						
C0	2	Asymmetrical	Neutral space	Multiple movements	TEA, JUMP, BANANA, SEA, SHEEP, CROCODILE, HORSE, FORK, ROCK	STAR, COLUMN, BOTTLE
C11	2	Asymmetrical	Neutral space	Just one movement	DRAW, WORM, CUP	WATCH, DOOR
Group B: Symmetrical bimanual signs						
C1	2	Symmetrical	Neutral space	Inward or outward path	CLOSE, DAY, BIG, BIRD-CLAW, BREAK, WIND, NIGHT, THREAD	KILL, BOOK
C3	2	Symmetrical	Neutral space	Multiple configurations	HEALTH, BOX, BILL, TABLE, EGG, CRAB, BREAD	HOOK, RECTANGULAR-PAPER, RIVER
C4	2	Symmetrical	Neutral space	1–2 fingers selected	BED, BAG, KNIFE, SIT, READ, FINISH	HOUSE, GLASSES, CAT
C6	2	Symmetrical	Neutral space	All fingers selected in fist or open palm configuration	PATH, COOK, SOCCER, STICK, BORN, WOOD, ROPE	LARGE-VEHICLE
C9	2	Symmetrical	Upper head space	Forehead or overhead	SHAMPOO, CEILING, RAIN, SICK, MOUNTAIN, MONKEY, CROWN	—
C10	2	Symmetrical	Neutral space	Intersecting hands (Hands touching or crossed)	BUTTERFLY, BEAR, RABBIT, FLOWER, TURTLE	FIRE
Group C: One-handed signs						
C2	1	N/A	Lower face space	—	APPLE, TOOTHBRUSH, DRINK, BLUSH, CHICKEN	VASE, WHALE
C5	1	N/A	Upper neutral space	With top down or bottom up path	GUN, LEMON, SAD, PEN, PERSON, GROW, CHILD	BABY
C7	1	N/A	Head space	Two selected fingers	SEE, SMOKE, SNAKE	TREE, TRAIN
C8	1	N/A	Mid neutral space	Movement against the signer	VEHICLE, PAN, FISH, OLD, SUN	KEYS, SALT
C12	1	N/A	Upper head space	Temple location	THINK, POLICE, EYEBROW, COMB, PLANE	—

Table 1: Phonological characterization of the resulting clusters for Peruvian Sign Language (LSP) with $K = 13$.

C7 (LARGE-VEHICLE, VEHICLE, MONKEY, PATH) has movements mostly outward to the front. Cluster C11 is characterized by all signs involving hands crossing or contralateral movements, such as RABBIT, BUTTERFLY, WOOD, and BEAR. This cluster has no outliers. Cluster C12 is characterized by movements in central space, such as STICK, WIND, SEA, and BOX, among others. Finally, the C9 signs are performed in the upper head space with forward or overhead movements, including HOUSE, CROWN, SHAMPOO, BLUSH, and others.

In Group C (One-handed signs), the clusters are formed by signs performed with a single hand. These clusters are differentiated by their major location and specific phonological characteristics. The C0 signs are performed in the lower face space with all or just two fingers selected, such as EAT, DRINK, BED, and FLOWER, among others. The C5 signs are performed in the upper head space, such as MOON, GROW, HORSE, and SUN. Finally, C8 includes signs articulated in the head space with minimal internal finger movement, including COMB, CAT, THINK, and HOOK, among others.

Ultimately, Group D includes only one cluster (C10) consisting of signs performed in neutral space using one or two fingers. It has been categorized as mixed because half of the signs use one hand (POLICE, APPLE, KEYS) and the other half use two (TRAIN, SALT, WORM).

7 Conclusions

We presented a framework for learning motion-aware representations from sign language videos using auxiliary MHI-based supervision and metric learning. The resulting embedding space organizes signs according to motion-based properties that are consistent with expert knowledge of phonological features across ASL and LSP datasets. The observed clustering structure suggests that motion representations capture meaningful patterns consistent with phonological organization, although some variability indicates limitations in capturing fine-grained articulatory details. These findings highlight the potential of motion-based representations for modeling parts of sign language phonology, while also pointing to the need for more diverse data and finer-grained annotations to better understand how different phonological features are reflected in the learned space. As future work, this ability to capture motion patterns could support

Cluster	Hands	Symmetry	Location	Feature	Signs	Outliers
Group A: Asymmetrical bimanual signs						
C2	2	Asymmetrical	Neutral space	Non-dominant hand as location	DAY, WHALE, BORN, NIGHT, KILL WATCH, SHEEP, BOTTLE, TABLE	SICK
C3	2	Asymmetrical	Neutral space	Movement over dominant hand	ROCK, KNIFE, JUMP, FORK TEA, SOCCER, BREAD, SIT	EGG
C4	2	Asymmetrical	Neutral space	Non-dominant palm shape	RECTANGULAR-PAPER, PEN DRAW, COOK, READ, CUP	BOOK, STAR SNAKE, BANANA
Group B: Symmetrical bimanual signs						
C1	2	Symmetrical	Neutral space	All movements end in front of the lower face space	FIRE, CROCODILE VASE, DOOR	LAUGH
C6	2	Symmetrical	Neutral space	All fingers selected, movement mostly outward to sides	HEALTH, CRAB ANGRY, BREAK	GUN HAMMER
C7	2	Symmetrical	Neutral space	All fingers selected, movement mostly outward to front	LARGE-VEHICLE, VEHICLE MONKEY, PATH, BIRD-CLAW	BAG GLASSES, PAN
C9	2	Symmetrical	Upper head space	Forehead or overhead movements	HOUSE, CROWN SHAMPOO, BLUSH	PERSON
C11	2	Symmetrical	Neutral space	Hands crossing or contralateral	RABBIT, BUTTERFLY WOOD, BEAR, MOUNTAIN	—
C12	2	Symmetrical	Neutral space	Movement in central space	STICK, WIND, SEA, BOX CLOSE, RIVER, ROPE, BILL	FISH
Group C: One-handed signs						
C0	1	N/A	Lower face space	All or just two finger selected	EAT, DRINK, BED FLOWER, LEMON, CHICKEN	SEE
C5	1	N/A	Upper space	Upper neutral space and head space	MOON, GROW, HORSE SUN, CHILD	BIG, FINISH RAIN, CEILING
C8	1	N/A	Head space	No internal movement (static fingers)	COMB, CAT, THINK TOOTHBRUSH, HOOK, PLANE	EYEBROW TREE
Group D: Mixed manual signs						
C10	1-2	N/A	Neutral space	1-2 selected fingers	TRAIN, SALT, WORM POLICE, APPLE, KEYS	—

Table 2: Phonological characterization of the resulting clusters for American Sign Language (ASL) with $K = 13$.

the development of tools that automatically detect and annotate motion-related characteristics of signs from video.

8 Limitations

Although the clustering analysis for LSP and ASL (Table 1 and Table 2) groups clusters into relatively broad categories, certain clusters reveal substructures based on movement onset and endpoint. For example, C1 (LSP) and C12 (ASL) group signs encoding movement but not path direction. Likewise, in C5 (LSP) the movement (top-down or bottom up) from the resting position to the signing location has been recognized as relevant even if it is not phonological. Another case is C11 (LSP) is defined by the final point of the movement, namely, the non-dominant hand, rather than path or handshape.

Additionally, the use of MHI representations introduces a limitation by flattening three-dimensional motion into a two-dimensional representation. Signs, which are inherently produced in 3D space, are therefore reduced to planar motion without depth information. This effect can lead to misinterpretations of spatial relationships. For example, the LSP sign VASE and the ASL sign HOOK are grouped with signs articulated in the lower face region, despite being produced in front of that region rather than on it. This is particularly relevant for body anchoring, a key phonological distinction in sign languages.

Our results strongly suggest that clusters capture aspects of phonological structure, particularly those related to motion and location. For handshape, results are less clear; for instance, some clusters group signs where all fingers are selected, while in others, only one or two of them are. Since handshape distribution is unequal, and given that finger selection determines handshape, we have not been able to fully assess whether these features are also captured by all clusters or contribute to their organization.

As we can see, assessing the relevance of each phonological feature requires a more diverse dataset, allowing us to determine how different components of phonological structure are reflected in clustering and whether certain features are more prominent than others.

References

Samit Biswas and Dac-Nhuong Le. 2024. Recognition of wh-question sign gestures in video streams using an attention driven c3d-bilstm network. *Procedia Computer Science*, 235:2920–2931.

Y. Chung and H. Lee. 2023. Joint triplet loss with semi-hard constraint for data augmentation and disease prediction using gene expression data. *Scientific Reports*, 13:18178.

Kearsy A. Cormier, Onno Crasborn, and Richard Bank. 2016. Digging into signs: Emerging annotation standards for sign language corpora. In *Proceedings of the 7th Workshop on the Representation and Processing of Sign Languages: Corpus Mining*, pages 35–40. ELRA.

Onno A Crasborn, Harry Van der Hulst, and Els Van der Kooij. 2001. Signphon: A phonological database for sign languages. *Sign language & linguistics*, 4(1-2):215–228.

Ardo Damarjati, Ni D, Riskya Sembiring, and Alexander Gunawan. 2025. Optimizing few-shot learning for indonesian sign language video recognition using batch hard triplet loss. pages 106–112.

Philippe Dreuw and Hermann Ney. 2008. Towards automatic sign language annotation for the elan tool. In *Workshop Programme*, volume 50.

Ismail Elezi, Jenny Seidenschwarz, Laurin Wagner, Sebastiano Vascon, Alessandro Torcinovich, Marcello Pelillo, and Laura Leal-Taixé. 2023. The group loss++: A deeper look into group loss for deep metric learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2):2505–2518.

Karen Emmorey. 2020. Sign language: How the brain represents phonology without sound. *Current Biology*, 30(22):R1361–R1363.

Nancy Frishberg. 1975. Arbitrariness and iconicity: historical change in american sign language. *Language*, pages 696–719.

Andrea Frome, Yoram Singer, Fei Sha, and Jitendra Malik. 2007. Learning globally-consistent local distance functions for shape-based image retrieval and classification. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8.

Weifeng Ge. 2018. Deep metric learning with hierarchical triplet loss. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Jia Gong, Lin Geng Foo, Yixuan He, Hossein Rahmani, and Jun Liu. 2024. Llms are good sign language translators. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18362–18372.

Lianyu Hu, Liqing Gao, Zekang Liu, and Wei Feng. 2023. Continuous sign language recognition with correlation network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2529–2539.

Longtao Jiang, Min Wang, Zecheng Li, Yao Fang, Wengang Zhou, and Houqiang Li. 2024. Seds: Semantically enhanced dual-stream encoder for sign language retrieval. In *Proceedings of the 32nd ACM International Conference on Multimedia*, MM '24, page 5141–5150, New York, NY, USA. Association for Computing Machinery.

Edward S Klima and Ursula Bellugi. 1979. *The signs of language*. Harvard University Press.

Dongxu Li, Chenchen Xu, Xin Yu, Kaihao Zhang, Ben Swift, Hanna Suominen, and Hongdong Li. 2020. Tspnet: hierarchical feature learning via temporal semantic pyramid for sign language translation. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Zeyu Liang, Huailing Li, and Jianping Chai. 2023. Sign language translation: A survey of approaches and techniques. *Electronics*, 12(12).

Diane Lillo-Martin. 2009. Sign language acquisition studies. *The Cambridge handbook of child language*, pages 399–416.

P. Liu, J. Wang, Y. Shen, S. Mo, C. Qi, and J. Ma. 2026. Multimotion: Multi subject video motion transfer via video diffusion transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 7233–7241.

Jian He Low, Ozge Mercanoglu Sincan, and Richard Bowden. 2025. Sage: Segment-aware gloss-free encoding for token-efficient sign language translation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 5011–5020.

Yue Ma, Yulong Liu, Qiyuan Zhu, Ayden Yang, Kunyu Feng, Xinhua Zhang, Zhifeng Li, Sirui Han, Chenyang Qi, and Qifeng Chen. 2025. Follow-your-motion: Video motion transfer via efficient spatial-temporal decoupled finetuning. *arXiv preprint arXiv:2506.05207*.

Aurora Martinez del Río, Casey Ferrara, Sanghee J. Kim, Emre Hakgüder, and Diane Brentari. 2022. Identifying the Correlations Between the Semantics and the Phonology of American Sign Language and British Sign Language: A Vector Space Approach. *Frontiers in Psychology*, 13:806471.

Tim Meinhardt, Alexander Kirillov, Laura Leal-Taixe, and Christoph Feichtenhofer. 2022. Trackformer: Multi-object tracking with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Mara Moita, Ana Maria Abreu, and Ana Mineiro. 2023. Iconicity in the emergence of a phonological system? *Journal of Language Evolution*, 8(1):1–17.

S. Mopidevi, M. V. D. Prasad, and P. V. V. Kishore. 2023. Multiview meta-metric learning for sign language recognition using triplet loss embeddings. *Pattern Analysis and Applications*, 26:1125–1141.

Luis Naranjo-Zeledón, Mario Chacón-Rivas, Jesús Peral, and Antonio Ferrández. 2020. Phonological proximity in costa rican sign language. *Electronics*, 9(8):1302.

Carol Neidle. 2002. Signstream™ annotation: Conventions used for the american sign language linguistic research project. *American Sign Language Linguistic Research Project, Boston University*.

ElissaL Newport, Elizabeth F Ashbrook, and Richard P Meier. 1985. The acquisition of american sign language. the crosslinguistic study of language acquisition, vol. 1: The data, ed. by dan i. slobin, 881-938.

Felix Ott, David Rügamer, Lucas Heublein, Bernd Bischl, and Christopher Mutschler. 2023. Auxiliary cross-modal representation learning with triplet loss functions for online handwriting recognition. *IEEE Access*, 11:94148–94172.

Alexander Pondaven, Aliaksandr Siarohin, Sergey Tulyakov, Philip Torr, and Fabio Pizzati. 2025. Video motion transfer with diffusion transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Arezoo Sadeghzadeh and Md Baharul Islam. 2022. Triplet loss-based convolutional neural network for static sign language recognition. In *2022 Innovations in Intelligent Systems and Applications Conference (ASYU)*, pages 1–6.

Karahan Şahin and Kadir Gökgöz. 2024. Decoding sign languages: The sl-fe framework for phonological analysis and automated annotation. In *Proceedings of the LREC-COLING 2024 11th workshop on the representation and processing of sign languages: Evaluation of sign language resources*, pages 335–342.

Wendy Sandler and Diane Carolyn Lillo-Martin. 2006. *Sign language and linguistic universals*. Cambridge University Press.

Adam Schembri and Onno Crasborn. 2010. Issues in creating annotation standards for sign language description. In *sign-lang@ LREC 2010*, pages 212–216. European Language Resources Association (ELRA).

J. Shin, A. S. Musa Miah, M. A. M. Hasan, K. Hirooka, K. Suzuki, H.-S. Lee, and S.-W. Jang. 2023. Korean sign language recognition using transformer-based deep neural network. *Applied Sciences*, 13:3029.

Markus Steinbach, Nivedita Mani, Anne Wienholz, Derya Nuhbalaoglu, and Annika Herrmann. 2019. Phonological priming in German Sign Language: An Eye Tracking Study Using the Visual World Paradigm. PsyArXiv preprint. Available online: https://psyarxiv.com/x5pts/ (accessed on 21 May 2020).

William C. Stokoe. 1960. Sign language structure: An outline of the visual communication system of the american deaf. *Studies in Linguistics: Occasional Papers*, 8.

Sieu Tran, Duc Nguyen Minh, Truong Nguyen Thanh, Hao Vo, Thanh Duc Ngo, Tien Do, Khiem Le, and Duy-Dinh Le. 2025. Generalizable sign language recognition via local temporal convolutions and region-aware pose encoding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 5030–5037.

Aamir Wali, Roha Shariq, Sajdah Shoaib, Sukhan Amir, and Asma Ahmad Farhan. 2023. Recent progress in sign language recognition: a review. *Machine Vision and Applications*, 34(6):127.

Zhiling Wang. 2025. Transformer-based motion predictor for multi-dancer tracking in non-linear movements of dancesport performance. *IEEE Access*, 13:100647–100666.

Chaocan Xue, Bineng Zhong, Qihua Liang, Haiying Xia, and Shuxiang Song. 2025. Unifying motion and appearance cues for visual tracking via shared queries. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(3):1987–1998.

Jinhui Ye, Xing Wang, Wenxiang Jiao, Junwei Liang, and Hui Xiong. 2024. Improving gloss-free sign language translation by reducing representation density. In *Advances in Neural Information Processing Systems*, volume 37, pages 107379–107402. Curran Associates, Inc.

Aoxiong Yin, Tianyun Zhong, Li Tang, Weike Jin, Tao Jin, and Zhou Zhao. 2023. Gloss attention for gloss-free sign language translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2551–2562.

Biao Zhang, Garrett Tanzer, and Orhan Firat. 2024. Scaling sign language translation. In *Advances in Neural Information Processing Systems*, volume 37, pages 114018–114047. Curran Associates, Inc.

Hengbo Zhang, Daming Liu, and Nana Fu. 2023. Continuous sign language recognition based on 3dcnn and blstm. In *Proceedings of the Fifth International Conference on Computer Information Science and Artificial Intelligence (CISAI 2022)*, volume 12566 of *Proceedings of SPIE*, page 125661Z.

A Appendix A: Detailed Architecture and Training Setup

The proposed model learns motion-aware representations from video sequences using a two-stage training procedure, where each stage corresponds to a specific component of the model. The first component is a motion-aware encoder that extracts spatio-temporal features from the input video and includes a mapping head module that decodes these

features into a motion representation used for auxiliary supervision. The second component is a projection module that maps these features into a compact embedding space for metric learning.

In the first stage, the motion-aware encoder is optimized using an auxiliary supervision signal to capture meaningful temporal dynamics. In the second stage, the learned representations are treated as fixed inputs, and the projection module is trained to structure the embedding space through metric learning. Training is performed independently for each dataset.

A.1 Stage 1: Motion Representation Learning with Auxiliary MHI Projection

In the first stage, the model learns spatio-temporal motion representations from video sequences guided by an auxiliary MHI-like projection. Rather than reconstructing an MHI as a final objective, this auxiliary task acts as a structural regularizer that encourages the model to capture meaningful motion dynamics.

A.1.1 Luminance-based preprocessing

Given an RGB video sequence of shape $B \times T \times 3 \times H \times W$, where B denotes the batch size, T the number of frames, 3 the RGB channels, and H and W the spatial dimensions, the input is transformed into a luminance-based motion representation. Each frame is converted into a four-channel representation consisting of a normalized luminance channel, a duplicated temporal luminance difference channel, and an edge-based luminance difference derived from temporal differences. The resulting representation is arranged as a tensor of shape $B \times 4 \times (T-1) \times H \times W$ and used as input to the model.

A.1.2 Motion 3D Backbone

The motion 3D backbone consists of a 3D convolutional network that processes the motion input and progressively reduces spatial resolution while increasing the number of feature channels. This produces a spatio-temporal feature tensor that preserves motion structure across frames while capturing both local spatial patterns and temporal dynamics.

A.1.3 Temporal Attention

A Transformer-based temporal attention module is used to aggregate temporal information. First, global spatial pooling is applied to each temporal slice to obtain frame-level descriptors, which are augmented with positional encodings and processed by a Transformer encoder to model temporal dependencies. The encoder produces an attention weight for each frame, which is used to compute a weighted aggregation of the spatio-temporal feature maps. Specifically, given feature maps $\mathbf{f}_t$ and attention weights w_t, the final representation is obtained as:

$$\mathbf{E} = \sum_{t=1}^{T} w_t \cdot \mathbf{f}_t \tag{1}$$

A.1.4 Motion Projection Head

The aggregated feature map is passed through a motion projection head that maps it into a single-channel motion representation. This module acts as a lightweight convolutional decoder that imposes a structural constraint on the learned features. Specifically, it applies upsampling followed by a sequence of convolutional layers with Group Normalization and ReLU activations, producing a dense 2D output. The final output is a single-channel map that captures motion intensity over space, analogous to a learned Motion History Image (MHI).

A.1.5 Training objective

The model is trained on 70% of the data. The encoder and motion projection head are jointly optimized using a Smooth L1 loss between the predicted and ground-truth MHI:

$$\mathcal{L}_{MHI} = \mathrm{SmoothL1}(MHI_{pred}, MHI_{target}) \tag{2}$$

Training is performed using Adam with a learning rate of 1×10^{-5} and weight decay of 1×10^{-7}, along with a StepLR scheduler (step size 1000, decay 0.5). After training, the motion projection head is discarded, and only the learned spatio-temporal feature representation is retained for the next stage.

A.2 Stage 2: Metric Learning in the Embedding Space

In the second stage, the model learns a compact embedding space from the spatio-temporal representations obtained in Stage 1. The feature maps produced by the motion encoder are precomputed and used as fixed inputs, enabling the embedding space to be structured independently of the feature extraction process.

A.2.1 Embedding Projection Head

The spatio-temporal feature maps are processed by a lightweight multi-scale 2D convolutional pro-

jection head. The model consists of residual convolutional blocks with progressive downsampling, producing feature maps at multiple spatial scales. Each scale is aggregated using generalized mean pooling (GeM) or global average pooling and projected into a shared embedding space.

The resulting descriptors are combined through content-dependent scale attention, where weights are computed via a softmax over scale-specific features. The concatenated multi-scale representation is then passed through a bottleneck MLP with normalization and non-linear activation, followed by Layer Normalization. The final output is L2-normalized to obtain a 256-dimensional embedding.

A.2.2 Semi-hard Triplet Loss

The projection head is trained using a metric learning objective based on semi-hard triplet loss. Given an anchor a, a positive p from the same class, and a negative n from a different class, the objective enforces:

$$\mathcal{L}_{triplet} = \max\left(d(a,p) - d(a,n) + m,\, 0\right), \quad (3)$$

where $d(\cdot,\cdot)$ denotes the cosine distance between L2-normalized embeddings. Semi-hard negative mining selects negatives that satisfy:

$$d(a,p) < d(a,n) < d(a,p) + m,$$

providing informative training signals while avoiding the instability of fully hard negatives.

A.2.3 Training procedure

Mini-batches are constructed using PK sampling, with P classes and K samples per class, providing multiple positive and negative candidates per anchor. When no semi-hard negatives are available within a mini-batch, the hardest negative is selected as a fallback.

Training is performed using AdamW (learning rate 7×10^{-5}, weight decay 1×10^{-5}) with margin $m = 0.2$, for up to 10,000 epochs, using $P = 16$ and $K = 8$. All experiments were performed using two NVIDIA A100 80GB PCIe GPUs with CUDA 12.8.

A.2.4 Model selection

Model selection is performed using a composite validation score that reflects the geometric quality of the embedding space. This score combines complementary properties, including the inter/intra-class distance ratio to capture global separation and compactness, the average margin to reflect local class boundaries, and nearest-centroid accuracy using cosine distance as a proxy for embedding consistency.

Although the triplet loss promotes class separation, the objective is not strict separability but meaningful geometric organization, allowing samples with similar motion patterns to remain close even across different classes.

B Appendix B: Detailed Phonological Clustering LSP and ASL

For visualization purposes only, we performed an additional 2D reduction using t-SNE (t-distributed stochastic neighbor embedding) for both LSP and ASL.

B.1 Peruvian Sign Language

Based on the MHI results, 13 clusters (C0-C12) were obtained. Figures 5 provide complete cluster visualizations for LSP using 2D t-SNE.

B.2 American Sign Language

Based on the MHI results, 13 clusters (C0-C12) were obtained. Figures 6 provide complete cluster visualizations for ASL using 2D t-SNE.

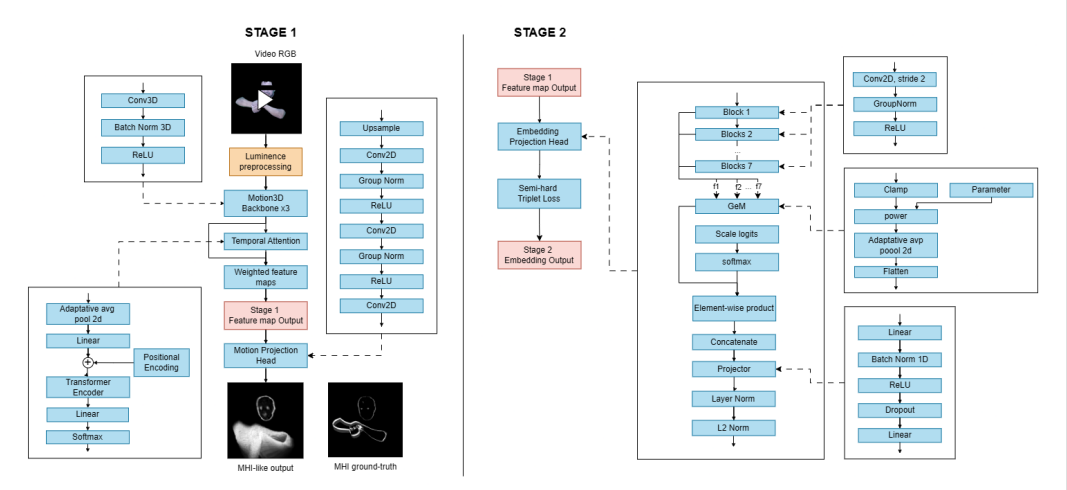

Figure 4: Stage 1 and Stage 2 of the model

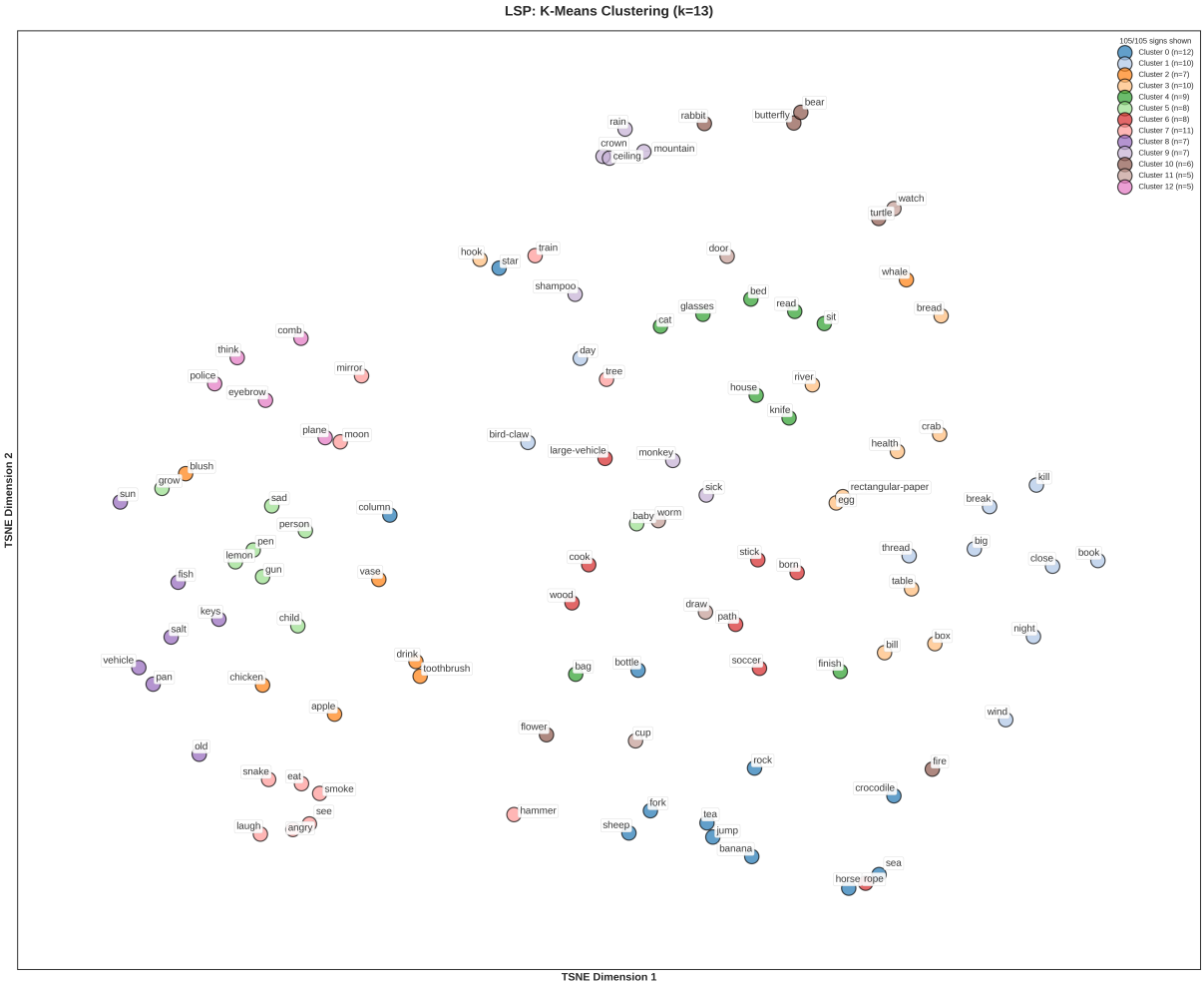

Figure 5: LSP Clustering using the MHI-based embeddings

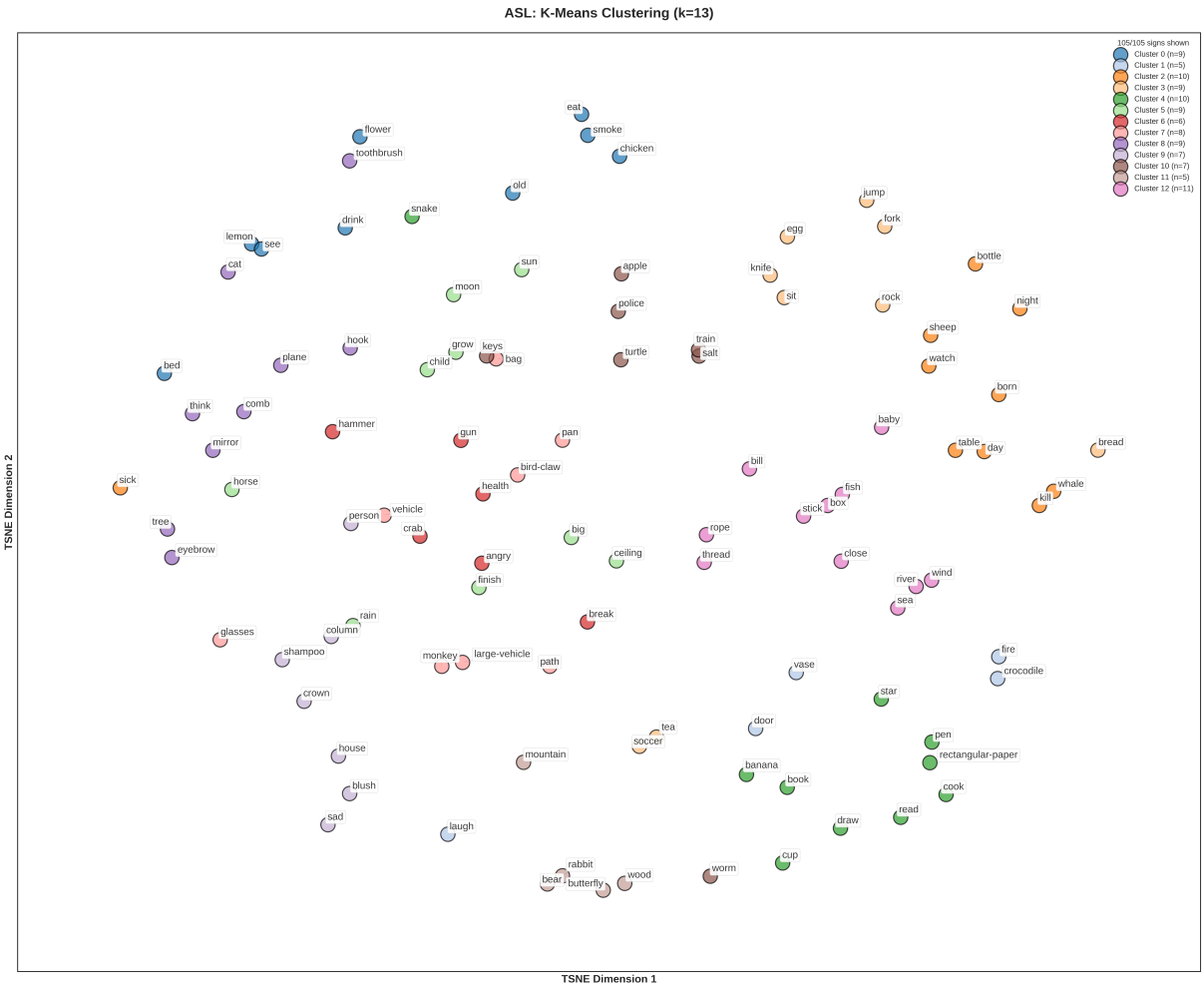

Figure 6: ASL Clustering using the MHI-based embeddings