Phonological Representation Learning for Isolated Signs Improves Out-of-Vocabulary Generalization

Anonymous ACL submission

Abstract

Sign language datasets are often not representative in terms of vocabulary, underscoring the need for models that generalize to unseen signs. Vector quantization is a promising approach for learning discrete, token-like representations, but it has not been evaluated whether the learned units capture spurious correlations that hinder out-of-vocabulary performance. This work investigates two phonological inductive biases: Parameter Disentanglement, an architectural bias, and Phonological Semi-Supervision, a regularization technique, to improve isolated sign recognition of known signs and reconstruction quality of unseen signs with a vector-quantized autoencoder. The primary finding is that the learned representations from the proposed model are more effective for one-shot reconstruction of unseen signs and more discriminative for sign identification compared to a controlled baseline. This work provides a quantitative analysis of how explicit, linguistically-motivated biases can improve the generalization of learned representations of sign language.

1 Introduction

The development of robust sign language models is often constrained by the limited scale and vocabulary of available datasets (Bragg et al., 2019; De Sisto et al., 2022). This data scarcity makes the ability to generalize to out-of-vocabulary (OOV) signs a significant challenge for the field. Sign languages are highly productive systems, where a finite set of phonological features, such as handshape, location, and movement patterns can be combined to form a vast lexicon (Stokoe, 1960; Brentari, 1998). A model that learns to represent these underlying phonological components of signs can improve sign recognition (Kezar et al., 2023b), and we hypothesize such phonological learning bias could improve reconstruction of novel combinations of those components in unseen signs.

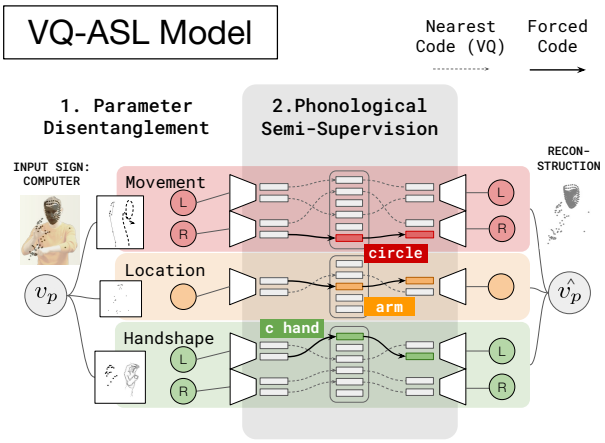

Figure 1: We show that out-of-vocabulary signs are more accurately reconstructed when trained under two phonological inductive biases: first, *disentangling* the input stream by ASL parameter; then, encoding each stream as a sequence of learned components, some of which are aligned with expert labels.

Vector-Quantized Variational Autoencoders (VQ-VAEs) have emerged as a powerful method for learning discrete latent representations of data, which are attractive for their potential use as tokens in sequence models (Van Den Oord et al., 2017; Razavi et al., 2019). Prior work has explored VQ for creating data-driven representations of signs, often as an alternative to linguistic glosses (Abzaliev and Mihalcea, 2024; Tasyurek et al., 2025). However, standard VQ models are trained with a compression objective that may encourage the learning of "tangled" representations, where the learned codes capture spurious, dataset-specific correlations that do not generalize to unseen signs (Higgins et al., 2017). This reflects a theoretical limitation formally proven by Locatello et al. (2019), who showed that the unsupervised learning of disentangled representations is "fundamentally impossible" without inductive biases on both the models and the data. This finding shows that a search for

principled sources for such biases is not just motivated, but necessary.

This paper therefore investigates whether phonologically-motivated inductive biases can improve a VQ-VAE's ability to reconstruct and recognize out-of-vocabulary (OOV) signs. We draw these biases from the Prosodic Model (Brentari, 1998) and the ASL-LEX 2.0 database (Sehyr et al., 2021), two lexicon-wide descriptions of ASL structure, and implement them through two mechanisms (Figure 1). The first is *Parameter Disentanglement* (PD), an architectural method that uses a multi-stream VQ-VAE to learn separate codebooks for distinct articulators and movement parameters. The second is P*honological Semi-Supervision* (PSS), a supervisory method that uses an auxiliary classification loss with expert phonological labels to regularize the latent codebooks and align them with established linguistic features.

This paper proceeds as follows: Section 2 reviews relevant background literature. Section 3 details the VQ-ASL framework. Section 4 describes the experimental setup for a controlled ablation study. Section 5 presents the results, and Section 6 discusses their implications and the limitations of the study. The main finding is that the proposed interventions offer complementary benefits, improving OOV generalization for both sign reconstruction and recognition, and revealing a trade-off between reconstruction fidelity and the discriminative quality of the learned representations.

2 Background

We review prior work in sign language representation learning, the theoretical foundations of disentangled representation learning, and the linguistic models that motivate our proposed inductive biases.

2.1 Representation Learning for Sign Language

The popular input modality for modern sign language recognition is skeletal pose data, extracted from video using tools like MediaPipe or OpenPose (Lugaresi et al., 2019; Cao et al., 2019). This representation can be seen as a form of *phonetic bias*, because it abstracts away from signer-specific visual details like clothing or background, focusing instead on the underlying articulatory movements while enhancing privacy (Bragg et al., 2020). Various neural architectures have been employed to learn features from this data. 3D Convolutional Neural Networks (3D CNNs) can capture spatio-temporal features directly from video, but are computationally intensive and may learn spurious visual cues (Pu et al., 2021). Graph Convolutional Networks (GCNs) are naturally suited to skeletal data, as they explicitly model the topological structure of the human body and can learn the dynamic relationships between joints over time (Yan et al., 2018; Jiang et al., 2021). More recently, contrastive learning objectives have been used to align visual representations of signs with textual descriptions, improving the grounding of the learned features without supervision (Jiang et al., 2024; Hao et al., 2021).

The goal of obtaining discrete representations for signs includes manual and learned efforts. Formal symbolic systems like SignWriting provide a manual transcription system, analogous to written text (Sutton, 1990). Data-driven approaches, particularly those using vector quantization, have also been explored to learn discrete tokenizations of sign language, often as an intermediate step for sign language translation or production (Moryossef et al., 2021; Saunders et al., 2020). A recent preprint, "Disentangle and Regularize," also investigates articulator-based disentanglement for sign language production, though with a focus on continuous latent spaces and different regularization techniques (Tasyurek et al., 2025). Our work is distinct in its focus on vector quantization and the use of phonological semi-supervision to structure the discrete codebooks for OOV generalization in isolated sign recognition.

2.2 Disentangled Representation Learning

The goal of disentangled representation learning is to produce a latent space where each dimension, or group of dimensions, corresponds to a distinct, meaningful factor of variation in the data (Higgins et al., 2017). A model that successfully disentangles the underlying generative factors is hypothesized to exhibit better compositional generalization, data efficiency, and robustness. In the context of signing, these factors may be defined as the five phonological parameters: handshape, palm orientation, location, movement, and non-manual markers.

A foundational work in this area is the β-VAE, proposed by Higgins et al. (2017). This framework modifies the standard Variational Autoencoder (VAE) objective by introducing a hyperparameter, β, that increases the weight of the Kullback-Leibler (KL) divergence term in the loss

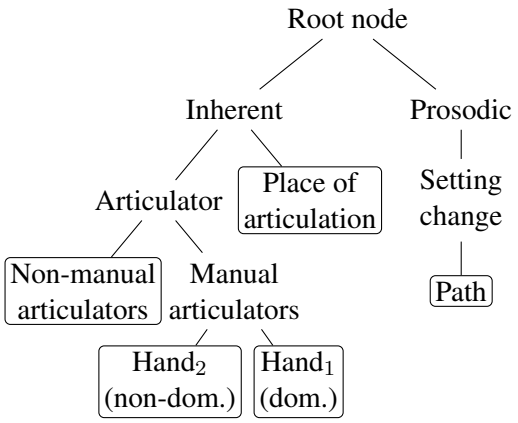

Figure 2: Brentari's Prosodic Model (Brentari, 1998) is hierarchical; the boxed variables represent disentangled parameters in this work.

function. A value of $\beta > 1$ imposes a stronger constraint on the information capacity of the latent bottleneck, forcing the model to learn a more efficient, and therefore more disentangled, representation by balancing reconstruction accuracy against the statistical independence of the latent variables.

However, the pursuit of unsupervised disentanglement faced a significant theoretical challenge from Locatello et al. (2019). They provided a formal proof demonstrating that for any dataset generated from disentangled latent factors, there exists an infinite family of transformations that can produce a perfectly entangled latent space while yielding the exact same data distribution. As an unsupervised model only has access to the observed data, it cannot distinguish between the true disentangled model and its entangled counterparts. Their conclusion is sobering: "the unsupervised learning of disentangled representations is fundamentally impossible without inductive biases on both the models and the data." This study provides the central motivation for this work, which seeks to identify and implement a principled source for such biases.

2.3 Linguistic Theory as a Stratified Inductive Bias

Our work responds to the challenge posed by Locatello et al. (2019) by using formal linguistic theory as a robust source of inductive biases, an approach aligned with Knowledge-Infused Learning (KiL) (Gaur et al., 2022). We apply this knowledge in a stratified manner for a deep infusion: a low-level architectural bias enforces parameter disentanglement on the articulators, while a higher-level phonological semi-supervision acts as a conceptual bottleneck that guides the model through linguistically-defined concepts (Koh et al., 2020). The specific knowledge for this work is the Prosodic Model of sign language phonology, developed by Brentari (1998).

Brentari's model describes signs as being composed of a hierarchical and simultaneous arrangement of phonological parameters (Figure 2). This linguistic framework provides a blueprint for a *stratified* inductive bias, which we implement using mechanisms from Knowledge-Infused Learning (Gaur et al., 2022). The architectural bias in our model is informed by Brentari's factorization of signs into articulators (hands, face), place of articulation (location), and prosodic features (movement). We additionally use the discrete, contrastive features described by the theory, such as the handshapes cataloged in ASL-LEX 2.0 (Sehyr et al., 2021), to regularize the learning process through an auxiliary loss. For a complete description of the ASL-LEX features, see Appendix A. Despite the availability of several phonological models (e.g., H-M-H, Liddell & Johnson, 1989; Dependency Model, van der Kooij, 2002), we adopt Brentari's prosodic model as best suited for semi-supervised learning. It offers (1) a clear inventory of contrastive features aligned with linguistic theory, and (2) a hierarchical structure that maps well onto the discrete components produced by our model.

By leveraging both architectural and regularization-based KiL strategies, guided by different strata of Brentari's Prosodic Model, we aim to provide the structured priors necessary for learning a meaningful and generalizable disentangled representation of signs.

3 The VQ-ASL Framework

This section provides a detailed technical description of the proposed VQ-ASL framework. It begins by detailing the baseline architecture, a Transformer-based VQ-VAE (Section 3.1), and subsequently elaborates on the two novel, phonologically-motivated interventions: Parameter Disentanglement (PD) and Phonological Semi-Supervision (PSS) (Section 3.2).

3.1 Baseline Transformer VQ-VAE

The baseline model is a Vector-Quantized Variational Autoencoder (VQ-VAE), which learns a discrete latent representation of an input se-

quence (Van Den Oord et al., 2017). The encoder and decoder components are implemented using Transformer architectures (Vaswani et al., 2017) to effectively model the sequential nature of sign language pose data.

The **encoder**, E, takes a sequence of pose vectors $X \in \mathbb{R}^{T \times D}$ as input, where T is the number of frames and D is the dimensionality of the pose vector for each frame. The Transformer encoder processes this sequence and outputs a set of N_p continuous latent vectors, $Z_e = E(X)$, where $Z_e \in \mathbb{R}^{N_p \times L_c}$ and L_c is the latent channel dimension. In the baseline model, these N_p vectors form a single, unstructured latent sequence.

The **vector quantization layer** serves as the information bottleneck. It contains a learnable codebook, $C \in \mathbb{R}^{K \times L_c}$, which consists of K discrete code vectors, $\{c_j\}_{j=1}^{K}$. For each continuous vector $z_e^{(i)} \in Z_e$, the quantizer identifies the nearest code vector in the codebook using L2 distance:

$$k_i = \arg\min_j \|z_e^{(i)} - c_j\|_2^2.$$

The output of the VQ layer is a sequence of quantized vectors $Z_q \in \mathbb{R}^{N_p \times L_c}$, where each vector $z_q^{(i)}$ is the selected codebook vector c_{k_i}. Because the $\arg\min$ operation is non-differentiable, a straight-through estimator is used to copy gradients from the decoder input Z_q directly to the encoder output Z_e during backpropagation.

The **decoder**, D, has a symmetric Transformer architecture to the encoder. It takes the sequence of quantized vectors Z_q as input and reconstructs the original pose sequence, $\hat{X} = D(Z_q)$.

The total loss function for the baseline model, $\mathcal{L}_{\text{Baseline}}$, is a sum of three components, following the formulation of Van Den Oord et al. (2017):

$$\mathcal{L}_{\text{Baseline}} = \mathcal{L}_{\text{recon}} + \mathcal{L}_{\text{codebook}} + \beta \mathcal{L}_{\text{commit}}.$$

The reconstruction loss, $\mathcal{L}_{\text{recon}}$, is the mean squared error (MSE) between the input and reconstructed poses, $\mathcal{L}_{\text{recon}} = \|X - \hat{X}\|_2^2$, which trains the encoder and decoder. The codebook loss, $\mathcal{L}_{\text{codebook}} = \|\text{sg}[Z_e] - Z_q\|_2^2$, updates the codebook vectors to move them closer to the encoder outputs, where 'sg' denotes the stop-gradient operator. The commitment loss, $\mathcal{L}_{\text{commit}} = \|Z_e - \text{sg}[Z_q]\|_2^2$, regularizes the encoder output to remain close to the chosen code vectors, with β as a weighting hyperparameter.

3.2 Phonological Inductive Biases

We propose two inductive biases rooted in ASL phonology: parameter disentanglement and phonological semi-supervision.

3.2.1 Parameter Disentanglement (PD)

Parameter Disentanglement enforces a factorization of the latent space through an architectural bias. Instead of a single monolithic encoder-decoder pair, the model is structured into multiple parallel streams, each dedicated to a phonetically independent parameter of the sign. The input pose space X and the latent parameter space N_p are partitioned into channels corresponding to these components. The operationalized parameters are derived from Brentari's Prosodic Model (Brentari, 1998): articulators (left/right hand, facial expressions), prosodic features (left/right hand path movement), and place of articulation (location relative to other parts of the body).

Each stream $X^{(s)} \subset X$ possesses its own encoder E_s, decoder D_s, and a dedicated codebook C_s. The input keypoints are filtered for each stream such that they contain minimal overlap between parameters:

- $X^{(\text{RH})}, X^{(\text{LH})}$: the 21 coordinates for the right or left hand, uniformly translated such that the wrist is at the origin. Uniform frame sample.
- $X^{(\text{MOVR})}, X^{(\text{MOVL})}$: the wrist keypoint for the right or left hand. All frames.
- $X^{(\text{NMM})}$: the face keypoints translated such that the nose is at the origin. Uniform frame sample.
- $X^{(\text{BODY})}$: all the keypoints except face and hands. Uniform frame sample.

The total latent space is the concatenation of the stream-specific latents, $N_p = \sum_s N_{p,s}$, with the total bottleneck size held matching that of the baseline architecture. The total reconstruction loss is the sum of the reconstruction losses from each individual stream. To capture linguistic constraints, the codebooks for the left and right hand articulators ($C^{(\text{LH})}$ and $C^{(\text{RH})}$) and movements ($C^{(\text{LMOV})}$ and $C^{(\text{RMOV})}$) are shared, enforcing a symmetry constraint on these inventories. Furthermore, each stream employs specific preprocessing and attention-masking strategies.

Hyperparameter	Description	Search Range	Final Value
Transformer Dim	Hidden dimension of the Transformer layers.	[64, 512]	256
Transformer Layers	Num. layers in encoder/decoder.	[1, 6]	5
Latent Dim (L_c)	Dimensionality of each code vector.	[4, 32]	32
Num. Latent Vecs (N_p)	Number of vectors in the bottleneck.	[10, 100]	30
Codebook Size (K)	Number of entries in each codebook.	[10, 500]	200
Commitment Cost (β)	Weight for the VQ commitment loss.	[1e-6, 0.1]	3e-6
Diversity Weight (γ)	Weight for codebook diversity loss.	[0.01, 3.0]	3.0
Learning Rate	Adam optimizer learning rate.	[1e-5, 0.005]	8.61e-5
Dropout	Dropout probability in Transformer layers.	[0.0, 0.5]	0.2

Table 1: Hyperparameter Search and Final Configuration. A search was conducted to find the best values for the baseline model. These values were then held constant across all model variants for a fair comparison.

3.2.2 Phonological Semi-Supervision (PSS)

Phonological Semi-Supervision introduces a regularization-based bias by using expert labels from the ASL-LEX 2.0 database to guide the organization of the codebooks (Sehyr et al., 2021). This regularization provides a weak supervisory signal that encourages the learned discrete codes to align with meaningful, contrastive phonological features.

For each phonological parameter in ASL-LEX (e.g., Handshape, which has over 60 distinct values), a subset of codes within the corresponding codebook (e.g., the shared hand articulator codebook, $C_{A_{LH/RH}}$) is arbitrarily pre-assigned to represent these expert-defined features. During training, if a sign has a known phonological label y_f, an auxiliary loss is applied. With a specified probability, the quantization step is forced to select the code vector c_{y_f} that corresponds to the ground-truth label. The codebook and commitment losses are then computed with respect to this forced code, encouraging the encoder to produce outputs that are semantically aligned with ASL-LEX 2.0 features.

4 Experimental Setup

This section details the experimental design, including the dataset (Section 4.1), model configurations (Section 4.2), implementation details, and evaluation metrics (Section 4.3) used to systematically assess the impact of the proposed phonological inductive biases.

4.1 Dataset and Splits

The experiments are conducted on the Sem-Lex Benchmark, a large-scale dataset for American Sign Language (ASL) modeling (Kezar et al., 2023a). It consists of over 84,000 videos of isolated signs produced by 41 deaf ASL signers, covering a vocabulary of 3,149 unique signs. Crucially, the dataset is cross-referenced with ASL-LEX 2.0, providing the expert-annotated phonological feature labels, $\mathcal{Y}$, required for the Phonological Semi-Supervision (PSS) intervention (Sehyr et al., 2021). The dataset can be formally represented as $\mathcal{D} = \{(X_i, \Phi_i, y_i)\}_{i=1}^{N}$, where X_i is the input pose sequence, Φ_i is the set of its phonological labels, and y_i is the sign's identifier.

To evaluate OOV generalization, this work utilizes the benchmark's "unseen gloss" split. The dataset is partitioned such that the vocabularies of the training, validation, and test sets are disjoint. Any sign evaluated in the test set has not been seen during training, providing a direct measure of the model's ability to generalize to novel signs rather than novel instances of familiar signs.

4.2 Models Compared

The study is designed as a controlled ablation to isolate the effects of Parameter Disentanglement (PD) and Phonological Semi-Supervision (PSS). Four model configurations are compared:

1. **Baseline:** The standard Transformer VQ-VAE described in Section 3.1.

2. **VQ-ASL-PD:** The baseline model augmented with the multi-stream Parameter Disentanglement architecture.

3. **VQ-ASL-PSS:** The baseline model trained with the Phonological Semi-Supervision auxiliary loss.

4. **VQ-ASL (Full):** The proposed model incorporating both PD and PSS.

Table 2: Reconstruction Fidelity (MSE) on Seen ($\mathcal{D}_{\text{train}}$) vs. Unseen ($\mathcal{D}_{\text{test}}$) Data. Lower is better. The overall MSE columns show the generalization gap, while the channel-wise columns provide a detailed breakdown of test performance for the disentangled models.

	Overall MSE		**Channel-wise Test MSE**					
Model	$\mathcal{D}_{\text{train}}$	$\mathcal{D}_{\text{test}}$	$X^{(\text{RH})}$	$X^{(\text{LH})}$	$X^{(\text{NMM})}$	$X^{(\text{BODY})}$	$X^{(\text{MOVR})}$	$X^{(\text{MOVL})}$
Baseline	0.039	0.058						
VQ-ASL-PSS	0.037	0.055						
VQ-ASL-PD	0.032	0.046	0.051	0.019	0.117	0.126	0.023	0.009
VQ-ASL (Full)	**0.029**	**0.041**	**0.015**	**0.005**	**0.005**	**0.068**	**0.017**	**0.009**

A hyperparameter search was conducted exclusively on the Baseline model using the validation set, which contains unseen signs. The search was performed with Optuna (Akiba et al., 2019), a hyperparameter optimization framework. The optimization objective was to minimize reconstruction loss while simultaneously maximizing codebook utilization (measured by increase in perplexity over training) to prevent codebook collapse. The optimal hyperparameters found for the baseline were then fixed and used for all other model variants to ensure that any observed performance differences are attributable to the specific interventions and not to differences in tuning. Table 1 summarizes the key hyperparameters and their selected values. For further implementation details see Appendix B.

4.3 Evaluation Metrics

The models were evaluated using two primary analyses designed to test our hypotheses regarding OOV generalization.

Reconstruction Error. We measured the Mean Squared Error (MSE) between the ground-truth (X) and reconstructed ($\hat{X}$) pose sequences—the distance between each skeletal point in the original and reconstructed sequence across every frame—as an intrinsic evaluation of each models' reconstruction quality. We hypothesized that while OOV reconstruction would be worse than in-vocabulary (IV), this generalization gap would be lessened by the proposed inductive biases.

Phonological Alignment. After training each model, we froze their encoders and additionally trained two MLP probes on the quantized encodings Z_q to measure how well the learned codes align with labeled ASL phonology.

The first probe $M_{\text{ISR}}(Z_q, \theta) \approx p(y_g|v)$ is trained for the downstream task of isolated sign recognition. This probe measures the how *discriminative* the learned features are for signs that were not seen in training. If a model leverages spurious correlations in the train set, then these learned components may not reliably distinguish OOV signs.

The second probe $M_{\text{PFR}}(Z_q, \theta)$ is trained to recognize the phonological features in ASL-LEX 2.0: $f_{\text{PFR}} : Z_q \rightarrow y_\phi$. For the supervised models (VQ-ASL-PSS, VQ-ASL (Full)), this probe is an intrinsic evaluation of how successfully the PSS objective was learned, as well as confirmation that the ASL-LEX 2.0 features generalize to unseen signs. For the purely self-supervised models (Baseline, VQ-ASL-PD), M_{PFR} measures the extent to which ASL-LEX 2.0 features may emerge through induction from data alone.

We hypothesized that a positive correlation will exist between the two probes' performance, that is, successfully recognizing ASL-LEX 2.0 features will facilitate the ISR task on signs unseen during training. We also hypothesized that, for the supervised models (VQ-ASL-PSS and VQ-ASL (Full)) when the ISR probe misclassifies a sign, its errors with respect to PFR will be less severe if the disentanglement intervention (PD) is applied.

5 Results and Analysis

Our results suggest that out-of-vocabulary (OOV) generalization is improved by modeling different aspects of lexical structure. Separating the model by phonological parameter (*disentanglement*) improves the generation of sign form, creating a more productive system for novel combinations and reducing reconstruction MSE by 21%. In parallel, using phonological labels as a constraint (Phonological Semi-Supervision) stabilizes the model's understanding of sign identity, increasing sign recognition MRR by 14%.

Table 3: Phonological Alignment Results. We report Mean Reciprocal Rank (MRR) and Recall@10 (%) for two MLP probes trained on the frozen latent codes (Z_q) for both In-Vocabulary ($\mathcal{D}_{\text{train}}$) and Out-of-Vocabulary ($\mathcal{D}_{\text{test}}$) signs.

	ISR Probe				PFR Probe			
	$\mathcal{D}_{\text{train}}$		$\mathcal{D}_{\text{test}}$ (OOV)		$\mathcal{D}_{\text{train}}$		$\mathcal{D}_{\text{test}}$ (OOV)	
Model	**MRR**	**R@10**	**MRR**	**R@10**	**MRR**	**R@10**	**MRR**	**R@10**
Baseline	.452	52.3	.381	45.1	.515	58.3	.488	55.5
VQ-ASL-PD	.441	51.5	.372	44.2	.502	57.1	.475	54.1
VQ-ASL-PSS	.503	58.2	.435	50.4	.593	61.2	.565	58.4
VQ-ASL (Full)	**.528**	**61.3**	**.452**	**52.8**	**.618**	**62.3**	**.583**	**59.8**

5.1 Reconstruction Error

Table 2 presents the reconstruction performance for each model. The full VQ-ASL model achieves the lowest overall test MSE (0.041), indicating the most accurate OOV reconstruction. The VQ-ASL-PD model provides a substantial individual improvement, reducing the test MSE to 0.046 from the baseline's 0.058. In contrast, the VQ-ASL-PSS model offers a smaller gain in reconstruction fidelity, with a test MSE of 0.055. The combination of both interventions in the full model demonstrates a constructive interaction effect, yielding the best overall reconstruction quality.

5.2 Phonological Alignment

Table 3 presents the results from the two phonological alignment probes. The ISR probe results show a complementary narrative to the reconstruction findings. Here, PSS is the primary driver of performance, with the VQ-ASL-PSS model improving the OOV MRR to .435 over the baseline's .381. The VQ-ASL-PD model shows a slight degradation in OOV ISR performance, with an MRR of .372, highlighting a trade-off between the interventions. The PFR probe results for the unsupervised models show the Baseline achieves an OOV MRR of .488, indicating some emergent phonological structure. For the supervised models, the high OOV PFR scores for VQ-ASL-PSS (.565) and VQ-ASL (Full) (.583) confirm the PSS objective was learned successfully and generalizes. Across all models, a positive correlation exists between the MRR performance of the two probes. The full VQ-ASL model achieves the best performance on both probes, demonstrating the complementary nature of the two biases.

6 Discussion

The empirical results demonstrate that the proposed inductive biases improve OOV generalization in qualitatively different and complementary ways. This section interprets these findings, discusses their broader implications for future research, and concludes with the primary takeaway of this work.

6.1 Interpreting the Reconstruction-Recognition Trade-off

The results reveal a trade-off between reconstruction fidelity and representation discriminability. The architectural bias of Parameter Disentanglement (PD) is the primary driver of improved reconstruction, yet it slightly harms sign recognition accuracy when used alone. Conversely, Phonological Semi-Supervision (PSS) is the primary driver of recognition accuracy but offers only a minor benefit to reconstruction.

This finding provides empirical support for the arguments of Locatello et al. (2019). Our results demonstrate that a purely structural inductive bias (PD), while effective for a generative task like reconstruction, is insufficient to guarantee the emergence of a semantically meaningful latent space for a discriminative task. The model, guided only by architectural separation and reconstruction loss, learns to represent fine-grained motion details that are not necessarily contrastive for sign identification, thus slightly harming classification. It is only with the addition of a semantic inductive bias (PSS), which forces the model's representations to align with expert-defined, contrastive features, that the latent space becomes well-structured for recognition.

6.2 Implications and Future Work

The framework and findings presented here open several avenues for future research. The work is positioned as a step toward more complex sign language understanding tasks. The phonologically structured codebooks learned by VQ-ASL could serve as a powerful pre-trained tokenizer for models targeting continuous sign language recognition and translation, where generalizable decomposition strategies are necessary to mitigate the limited vocabularies in existing datasets.

Furthermore, this framework can be adapted to serve as a computational tool for exploring and validating linguistic hypotheses (Appendix C).

6.3 Limitations and Ethical Considerations

It is important to acknowledge the limitations of this study. First, the scope is restricted to isolated signs. This work does not address the significant challenges of co-articulation, prosody, and grammatical non-manual markers that are present in continuous, conversational signing. The proposed model is best viewed as a pre-training strategy for these more complex tasks.

Second, the evaluation of reconstruction quality relies on automated metrics (MSE on pose data), which are known to be imperfect proxies for human perception of motion quality and naturalness. Stronger claims about reconstruction would require human evaluation studies.

Third, the benchmarking in this paper is to isolate the effects of the proposed interventions. While providing a rigorous assessment of intervention effects, the models are not compared against published state-of-the-art results on different benchmarks. We do not intend to make claims about state-of-the-art performance in, for example, isolated sign recognition, but rather contribute novel architectural biases and carefully demonstrate their value in representation learning for sign languages.

Finally, any model trained on existing datasets is susceptible to inheriting their biases. The Sem-Lex dataset, while large, has a demographic composition that is not fully representative of the broader signing community, with a majority of signers being white and female (Kezar et al., 2023a). The performance of the VQ-ASL model may not generalize equally well across all demographic groups, and further work is needed to assess and mitigate these potential biases.

6.4 Conclusion

This work suggests that while vector quantization is a promising method for learning discrete sign representations, its success may be contingent on the inclusion of phonological priors. Our findings indicate that purely data-driven VQ models struggle to learn representations that are simultaneously generative and discriminative. However, by infusing the model with explicit linguistic knowledge—both through architectural constraints and weak supervision—we can guide the learning process toward a more structured and generalizable latent space. This provides a conservative but principled path forward for developing more robust sign language models.

References

Artem Abzaliev and Rada Mihalcea. 2024. Unsupervised discrete representations of american sign language. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19786–19793.

Takuya Akiba, Sota Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*.

Danielle Bragg, Naomi Caselli, Matt Huenerfauth, Hernisa Kacorri, and Meredith Ringel Morris. 2020. Exploring collection of sign language datasets: Privacy, participation, and model performance. In *The 22nd International ACM SIGACCESS Conference on Computers and Accessibility*, pages 1–14.

Danielle Bragg, Oscar Koller, Mary Bellard, Larwan Berke, Patrick Boudrealt, Annelies Braffort, Naomi Caselli, Matt Huenerfauth, Hernisa Kacorri, Tessa Verhoef, Christian Vogler, and Meredith Ringel Morris. 2019. Sign language recognition, generation, and translation: An interdisciplinary perspective. In *The 21st International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '19)*. ACM.

Diane Brentari. 1998. *A Prosodic Model of Sign Language Phonology*. MIT Press.

Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2019. OpenPose: Realtime multi-person 2D pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(1):172–186.

Mirella De Sisto, Vincent Vandeghinste, Santiago Egea Gómez, Mathieu De Coster, Dimitar Shterionov, and Horacio Saggion. 2022. Challenges with sign language datasets for sign language recognition and

translation. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, pages 2711–2718.

Manas Gaur, Kaushik Roy, Ugur Kursuncu, Ziad Al-Halah, Krishnaprasad Thirunarayan, and Amit Sheth. 2022. Advancing neuro-symbolic AI with knowledge-infused learning. *IEEE Intelligent Systems*, 37(5):103–110.

Yutong Hao, Zunshu Liu, Minsi Li, Jiahui Wu, Chen Li, Shidong Zhang, Yidong Ge, and Ying Shan. 2021. Self-supervised learning on pose data for sign language recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. 2017. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations (ICLR)*.

Eric Jang, Shixiang Gu, and Ben Poole. 2017. Categorical reparameterization with gumbel-softmax. In *International Conference on Learning Representations (ICLR)*.

Songyao Jiang, Bin Sun, Lichen Wang, Yue Bai, Kunpeng Li, and Yun Fu. 2021. Skeleton aware multi-modal sign language recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 3432–3441.

Zifan Jiang, Gerard Sant, Amit Moryossef, Mathias Müller, Rico Sennrich, and Sarah Ebling. 2024. SignCLIP: Connecting text and sign language by contrastive learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9171–9193, Miami, Florida, USA. Association for Computational Linguistics.

Lee Kezar, Elana Pontecorvo, Adele Daniels, Connor Baer, Ruth Ferster, Lauren Berger, Jesse Thomason, Zed Sevcikova Sehyr, and Naomi Caselli. 2023a. The Sem-Lex benchmark: Modeling ASL signs and their phonemes. In *The 25th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '23)*, pages 1–15.

Lee Kezar, Jesse Thomason, and Zed Sevcikova Sehyr. 2023b. Improving sign recognition with phonology. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. 2020. Concept bottleneck models. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 5338–5348. PMLR.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. 2019. Challenging common assumptions in the unsupervised learning of disentangled representations. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*.

Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg, and Matthias Grundmann. 2019. MediaPipe: A framework for building perception pipelines. arXiv preprint arXiv:1906.08172.

Chris J. Maddison, Andriy Mnih, and Yee Whye Teh. 2017. The concrete distribution: A continuous relaxation of discrete random variables. In *International Conference on Learning Representations (ICLR)*.

Amit Moryossef, Artem Abzaliev, Rada Mihalcea, and Yoav Goldberg. 2021. Unsupervised discrete representations of american sign language. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4531–4537.

J. Pu, W. Zhou, and H. Li. 2021. Learning to sign by watching tv. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ali Razavi, Aaron van den Oord, and Oriol Vinyals. 2019. Generating diverse high-fidelity images with VQ-VAE-2. In *Advances in Neural Information Processing Systems*.

Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. 2020. Everybody sign now: Translating spoken language to photo-realistic sign language video. *arXiv preprint arXiv:2011.09846*.

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

William C. Stokoe, Jr. 1960. *Sign Language Structure: An Outline of the Visual Communication Systems of the American Deaf*, volume 8 of *Studies in Linguistics: Occasional Papers*. Linstok Press.

Valerie Sutton. 1990. *Sign writing for everyday use*. Deaf Action Committee for Sign Writing.

Sumeyye Meryem Tasyurek, Tugce Kiziltepe, and Hacer Yalim Keles. 2025. Disentangle and regularize: Sign language production with articulator-based disentanglement and channel-aware regularization. arXiv preprint arXiv:2504.06610.

Aaron Van Den Oord, Oriol Vinyals, and Koray Kavukcuoglu. 2017. Neural discrete representation learning. In *Advances in Neural Information Processing Systems (NIPS)*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Sijie Yan, Yuanjun Xiong, and Dahua Lin. 2018. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

A Phonological Feature Assignment

The Parameter Disentanglement (PD) architecture and the Phonological Semi-Supervision (PSS) loss require a mapping from the phonological features described in ASL-LEX 2.0 to the specific model streams. Table 4 details this assignment for the 16 features used in this work, based on the descriptions provided by Kezar et al. (2023a).

B Implementation Details

All model variants were designed with comparable total parameter counts and total bottleneck sizes ($N_p \times L_c$) to ensure that performance differences are attributable to architectural and training strategies, not model capacity. For the PD models, the total bottleneck size and the aggregate number of codebook entries were kept consistent with the baseline. The allocation of latent vectors and codebook entries to each phonological stream was proportional to the descriptive complexity of that parameter as defined in ASL-LEX 2.0 (Sehyr et al., 2021). For instance, the handshape articulators, which are described by multiple sub-features (e.g., selected fingers, flexion), received a larger portion of the latent space compared to simpler binary features like wrist twist. Sufficient capacity was allocated to each codebook to allow for the learning of features not explicitly annotated in ASL-LEX, such as palm orientation and complex path movements.

B.1 Mitigating Codebook Collapse

A common challenge in training VQ-VAEs is "codebook collapse," also known as the "dead code" problem (Razavi et al., 2019). This occurs when a large portion of the code vectors in the codebook are never selected as the nearest neighbor to any encoder output during training. As a result, these "dead" codes receive no gradients from the codebook loss term and are never updated, leading to an inefficient use of the model's representational capacity. In addition to the commitment loss term, which helps stabilize training, two techniques were employed to ensure high codebook utilization.

First, the Gumbel-Softmax distribution is used as a differentiable approximation to the discrete categorical sampling of codes (Jang et al., 2017; Maddison et al., 2017). Using a temperature-annealing schedule, this encourages exploration in the early stages of training, making it less likely for codes to become permanently unused.

Second, a dead code re-initialization strategy is implemented. Periodically during training (e.g., every 1000 steps), the usage of each code vector is tallied. Any code vector whose usage count falls below a predefined threshold is considered "dead" and is re-initialized. The re-initialization is performed by setting the dead code vector to be the average of a small random sample of encoder output vectors from the current mini-batch. This ensures that all parts of the codebook remain in a high-density region of the encoder's output space, making them likely to be selected and updated in subsequent training steps.

C Linguistic Hypotheses for Future Work

Several directions are of particular interest:

- **Hierarchical Hypothesis:** The hierarchical nature of phonological features in Brentari's model could be explored more explicitly. Techniques like residual vector quantization, where a second codebook quantizes the error of the first, could be used to model the relationship between a high-level feature like "handshape" and its constituent sub-features like "selected fingers" and "flexion."
- **Gradient Hypothesis:** Not all phonological parameters may be equally suited to discrete representation. By analyzing gradients within the codebook space or selectively disabling quantization for certain channels (e.g., movement path), one could experimentally measure which aspects of signing are more continuous in nature.
- **Phonotactic Hypothesis:** The model implicitly learns the rules of valid phoneme combinations (phonotactics). By feeding the decoder random permutations of learned codes, the reconstruction error could serve as a proxy for phonotactic legality. This could be used to predict which novel combinations of features would form "plausible" new signs in ASL.

Table 4: Mapping of ASL-LEX 2.0 Phonological Features to VQ-ASL Streams

Phonological Feature	**Description**	**Assigned Stream**
Major Location	Broad location of the sign (e.g., neutral, head)	$X^{(\text{BODY})}$
Minor Location	Specific location of the sign (e.g., forehead, cheek)	$X^{(\text{BODY})}$
Selected Fingers	Which fingers are active in the handshape	$X^{(\text{LH})}, X^{(\text{RH})}$
Flexion	The joint configuration of the selected fingers	$X^{(\text{LH})}, X^{(\text{RH})}$
Flexion Change	Whether the flexion of fingers changes	$X^{(\text{MOVL})}, X^{(\text{MOVR})}$
Spread	Whether selected fingers touch one another	$X^{(\text{LH})}, X^{(\text{RH})}$
Spread Change	Whether the spread of fingers changes	$X^{(\text{MOVL})}, X^{(\text{MOVR})}$
Thumb Position	Position of the thumb relative to fingers	$X^{(\text{LH})}, X^{(\text{RH})}$
Thumb Contact	Whether the thumb makes contact with fingers	$X^{(\text{LH})}, X^{(\text{RH})}$
Sign Type	Number of hands and symmetry	$X^{(\text{LH})}, X^{(\text{RH})}$
Movement	The primary path movement of the hand(s)	$X^{(\text{MOVL})}, X^{(\text{MOVR})}$
Repeated Movement	Whether the movement is repeated	$X^{(\text{MOVL})}, X^{(\text{MOVR})}$
Wrist Twist	Whether the hand rotates about the wrist	$X^{(\text{MOVL})}, X^{(\text{MOVR})}$
Non-Manual Signal	Presence of a required facial expression	$X^{(\text{NMM})}$
Mouth Morpheme	Presence of a required mouth gesture	$X^{(\text{NMM})}$
Head Movement	Presence of a required head movement	$X^{(\text{NMM})}$