Disentangle and Regularize: Sign Language Production with Articulator-Based Disentanglement and Channel-Aware Regularization Supplementary Material

Sümeyye Meryem Taşyürek*
Hacettepe University

Tuğçe Kızıltepe†
Aselsan Inc.
Hacettepe University

Hacer Yalim Keles‡
Hacettepe University

1. Methodology

1.1. Pose Autoencoder

The total latent space is 80-dimensional in both dataset experiments, selected based on ablation studies comparing different configurations by back-translation performance. The face component is fixed at 16 dimensions, and the remaining 64 dimensions are proportionally divided based on joint counts: Upper body (8 joints): mapped from 8×3 to 8 dimensions, Right/Left hand (21 joints each): mapped from 21×3 to 28 dimensions per hand, Face (128 joints): mapped from 128×3 to 16 dimensions. To accommodate the characteristics of each dataset, we use slightly different encoder architectures. For **PHOENIX14T** [1], each articulator is encoded using a single linear projection layer, allowing direct mapping from raw 3D joint inputs to their respective latent representations. This simple and efficient structure is sufficient for PHOENIX14T, which has a more constrained linguistic domain and relatively lower articulatory variability. The decoder modules also consist of single linear layers per region, mirroring the encoder structure. For **CSL-Daily** [6], which involves broader linguistic context and higher articulatory variability, we introduce an additional projection layer with nonlinearity for Right/Left hand and face regions. Specifically, each region is encoded using a two-layer MLP comprising a linear projection, a PReLU activation, and a second linear layer. This increased representational capacity helps stabilize the latent space, particularly for hand and face regions where variability is highest. The intermediate hidden dimension is set to 40 for both hands and 96 for the face. Decoder modules follow the same structure, using symmetric two-layer MLPs with matching intermediate widths.

*meryemtasyurek@cs.hacettepe.edu.tr
†tkiziltepe@aselsan.com
‡hacerkeles@cs.hacettepe.edu.tr

1.2. Transformer Model

Input text is represented by 768-dimensional BERT-based word embeddings, which are linearly projected to 512-dimension to match the model's internal dimensions. The encoder consists of 3 layers with 4 attention heads and 1024-dimensional feed-forward networks, along with positional encoding to retain temporal ordering of the text sequence.

The decoder adopts a non-autoregressive structure to mitigate error accumulation. It includes 6 layers with 8 attention heads and the same 1024-dimensional feed-forward size. Temporal dynamics are initialized using learned time queries derived from a fixed stationary reference pose, where both hands rest downward as when signers are not performing a sign. This pose is projected into the decoder space and expanded across all time steps to provide a consistent initialization for sequence generation.

2. Implementation Details

Datasets. We evaluate our model on two continuous sign language datasets. PHOENIX-2014T [1] contains 8,247 German Sign Language (DGS) sentences aligned with gloss and spoken German. We use the 3D pose annotations from the CVPR 2025 SLRTP Challenge [5], where 2D keypoints extracted via MediaPipe Holistic [4] are uplifted to 3D using Ivashechkin *et al*. 's method [3], yielding $N \times 178 \times 3$ tensors. CSL-Daily [6] includes 20,654 Chinese Sign Language (CSL) sentences from 10 signers. We extract 3D poses using MediaPipe Holistic, standardizing each frame to 178 keypoints (upper body, hands, face). To ensure consistency across signers and datasets, poses are normalized by centering on the neck and scaling by shoulder width.

Autoencoder training settings. We assign region-specific weights to the reconstruction loss: $w_{\text{RH}} = w_{\text{LH}} = 1.5$, $w_{\text{F}} = 1.0$, $w_{\text{B}} = 0.5$. These weights prevent high-variance

upper body motions from overshadowing fine-grained hand and facial cues. L1 regularization with $\lambda = 1 \times 10^{-4}$ promotes encoder sparsity. Optimization is performed using the Adam optimizer, with a learning rate of 2×10^{-4} and beta parameters set to $(0.5, 0.9)$. The pose autoencoder is trained for 270 epochs on both PHOENIX14T and CSL-Daily datasets.

Transformer training settings. For L1 loss, we use; $w_{\text{RH}} = 14$, $w_{\text{LH}} = 10$, $w_{\text{F}} = 2$, decided based on ablation studies comparing different weighting schemes.

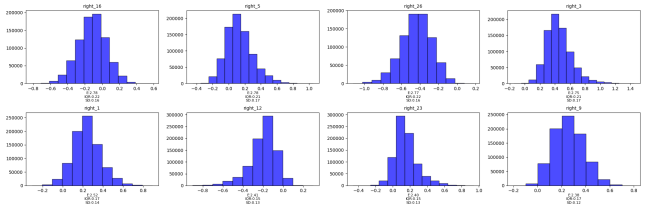
Hands carry the core lexical content in sign language, with the right hand as dominant in DGS and CSL, conveying most meaning, and the left hand serving a supportive role. The body offers coarse spatial context, while the face encodes grammatical and emotional cues but involves redundant motion across many keypoints. We apply similar region-based weighting in both the autoencoder and transformer training settings to reflect these functional roles. Regions are weighted accordingly, guided by ablation results.

We follow a two-phase training schedule: the transformer is first trained using only the weighted L1 loss to ensure stable latent reconstruction, and KL regularization is introduced only in the second phase to align predicted latent statistics with empirical channel priors without causing early collapse.

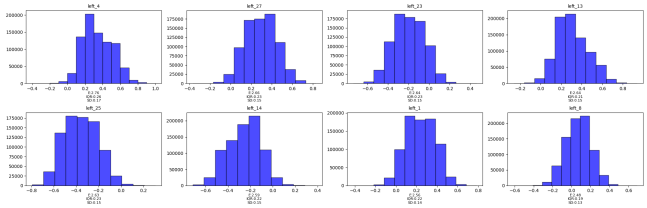
We employ the Adam optimizer with a learning rate of 2×10^{-4}, weight decay of 1×10^{-4}, and a ReduceLROnPlateau scheduler (factor 0.9, patience 40) in both phases of the training. Early stopping based on validation loss is applied to prevent overfitting. Training is performed on a 4×NVIDIA A100-SXM4-40GB setup using PyTorch Lightning [2].

3. Statistical Analysis of the Learned Latent Spaces

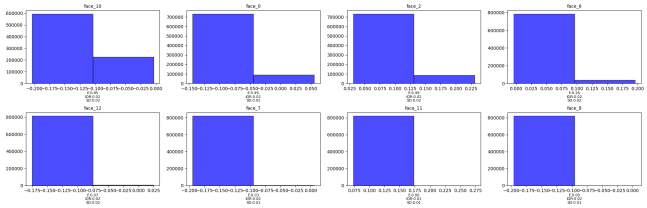
The statistical analysis of the learned latent representations reveals meaningful distinctions in the encoding behavior across body regions. In Figure 1, we present example histograms and corresponding entropy values for selected channel distributions from each structurally disentangled latent subspace (face, body, right hand, and left hand) learned by the autoencoder. As can be seen, the face channels have minimal variance. In contrast, the right and left hand channels show broader, more dispersed distributions with higher entropy and standard deviation across most dimensions, confirming the rich variability and critical role of manual articulators in sign expression. The body encodings fall somewhere in between, reflecting more stable but still semantically relevant movement, however body region spans a larger physical space and thus generates higher magnitude latent activations, even when the underlying motion is less semantically dense. To mitigate this effect and



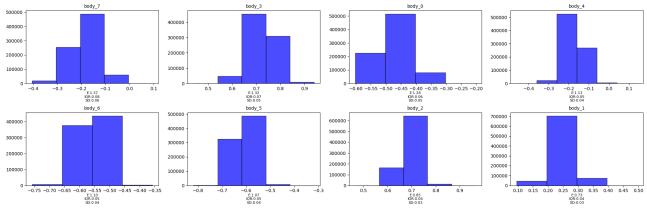
(a) Right hand channels (avg. entropy: 2.01).



(b) Left hand channels (avg. entropy: 2.22).



(c) Face channels (avg. entropy: 0.14).



(d) Body channels (avg. entropy: 1.11).

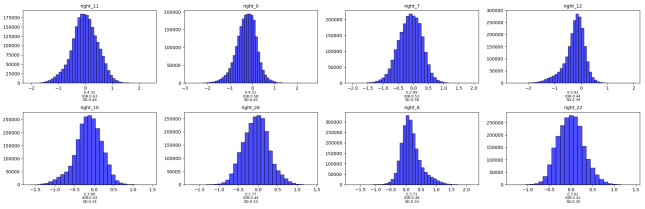
Figure 1. Histograms of top 8 latent channels with the highest entropy for each region of PHOENIX14T Dataset. Each subplot shows the distribution of a channel along with its entropy (E), interquartile range (IQR), and standard deviation (SD). (Zoom in for better visibility.)

maintain balanced learning, we increase the contribution of semantically richer regions by weighting.

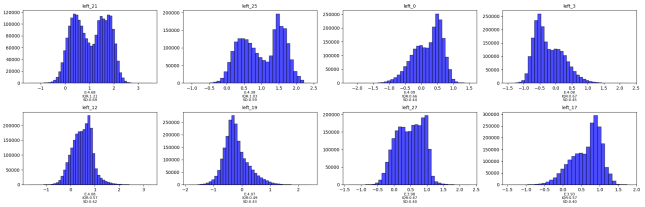
These distributional differences confirm that the latent space is semantically partitioned. They also highlight the need for targeted regularization and proper loss weighting. This encourages the model to focus more on dense, information-rich regions while still preserving reconstruction quality across all articulators.

In Figure 3, with separate PCA analysis for each articulator, we further analyze how the latent representations produced by our Transformer-based generator compare to those

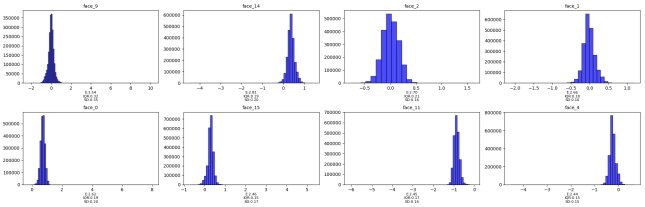
of the AE and how KL regularization influences these representations on PHOENIX14T dataset.



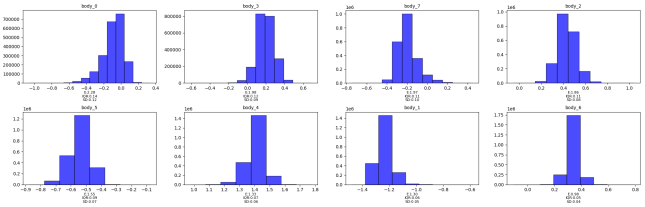
(a) Right hand channels (avg. entropy: 2.98).



(b) Left hand channels (avg. entropy: 3.19).



(c) Face channels (avg. entropy: 2.21).



(d) Body channels (avg. entropy: 1.65).

Figure 2. Histograms of top 8 latent channels with the highest entropy for each region of CSL-Daily Dataset. Each subplot shows the distribution of a channel along with its entropy (E), interquartile range (IQR), and standard deviation (SD). (Zoom in for better visibility.)

The latent space structure observed in CSL-Daily mirrors that of PHOENIX14T, with clear separation between articulator groups and similar relative entropy patterns, manual articulators exhibit the highest variability, followed by face and body (Figure 2). CSL-Daily encodings exhibit broader distributions across all articulator groups, as reflected by higher entropy values, particularly in the manual articulators. The left and right hand channels display the widest spread, with average entropies around 3.0, indicating increased variability and expressiveness in CSL's daily conversational context. Even face channels, which previously showed limited variance, now present more active and information-rich distributions. This broader activation aligns with CSL-Daily's more diverse signer pool and open-domain language, reinforcing the need for careful latent space design.

References

[1] Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. Neural sign language translation. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7784–7793, 2018. 1

[2] William Falcon and The PyTorch Lightning team. PyTorch Lightning, 2019. 2

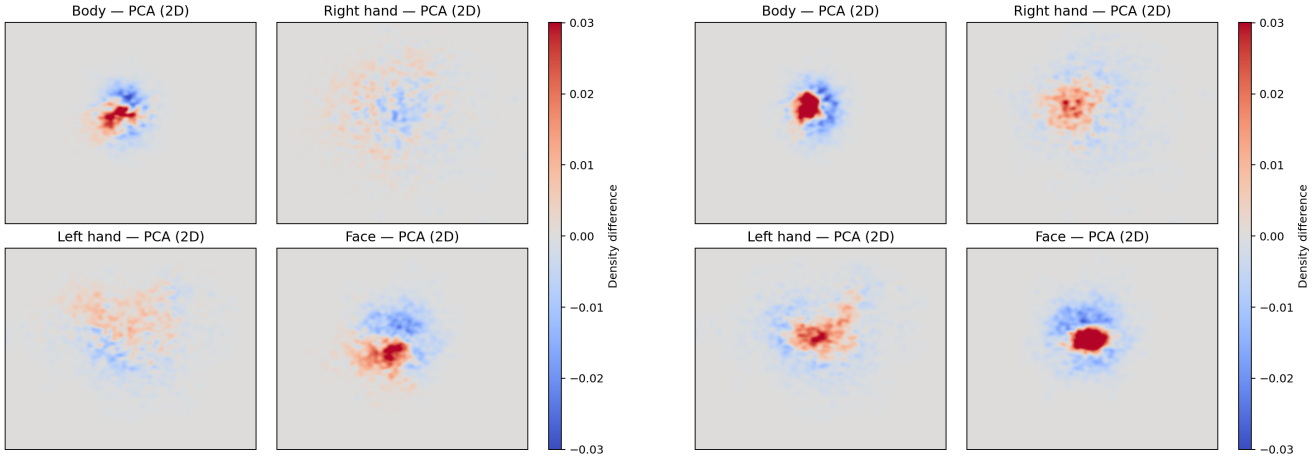
[3] Maksym Ivashechkin, Oscar Mendez, and Richard Bowden. Improving 3d pose estimation for sign language, 2023. 1

[4] 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. Mediapipe: A framework for building perception pipelines, 2019. 1

[5] Harry Walsh, Ed Fish, Ozge Mercanoglu Sincan, Mohamed Ilyes Lakhal, Richard Bowden, Neil Fox, Kearsy Cormier, Bencie Woll, Kepeng Wu, Zecheng Li, Weichao Zhao, Haodong Wang, Wengang Zhou, Houqiang Li, Shengeng Tang, Jiayi He, Xu Wang, Ruobei Zhang, Yaxiong Wang, Lechao Cheng, Meryem Tasyurek, Tugce Kiziltepe, and Hacer Yalim Keles. Slrtp2025 sign language production challenge: Methodology, results, and future work. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2025. 1

[6] Hao Zhou, Wengang Zhou, Weizhen Qi, Junfu Pu, and Houqiang Li. Improving sign language translation with monolingual data by sign back-translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 1

(a) GT regional PCA **(b)** Transformer w/o KL **(c)** Transformer w/ KL



(d) GT - Transformer w/o KL **(e)** GT - Transformer w/ KL

Figure 3. (a–c) **Regional PCA projections.** (d-e) **Density-difference maps (GT – Transformer)**: red indicates areas with higher density than the AE, blue indicates lower.