SHuBERT: Self-Supervised Sign Language Representation Learning via Multi-Stream Cluster Prediction

Shester Gueuwou[1], **Xiaodan Du**[1], **Greg Shakhnarovich**[1], **Karen Livescu**[1], **Alexander H. Liu**[2]
[1]TTI-Chicago, [2]MIT CSAIL
{shesterg,xdu,greg,klivescu}@ttic.edu, alexhliu@mit.edu
http://shubert.pals.ttic.edu

Abstract

Sign language processing has traditionally relied on task-specific models, limiting the potential for transfer learning across tasks. Pre-training methods for sign language have typically focused on either supervised pre-training, which cannot take advantage of unlabeled data, or context-independent (frame or video segment) representations, which ignore the effects of relationships across time in sign language. We introduce SHuBERT (Sign Hidden-Unit BERT), a self-supervised contextual representation model learned from approximately 1,000 hours of American Sign Language video. SHuBERT adapts masked token prediction objectives to multi-stream visual sign language input, learning to predict multiple targets corresponding to clustered hand, face, and body pose streams. SHuBERT achieves state-of-the-art performance across multiple tasks including sign language translation, isolated sign language recognition, and fingerspelling detection.

1 Introduction

Sign language presents unique challenges compared to other language modalities, because of the relative scarcity of data and its multi-channel nature, combining manual, facial, and other body movements, which can be quick and highly coarticulated (Bellugi and Fischer, 1972). Existing approaches to sign language processing have typically relied on models designed and trained for specific tasks, such as sign language translation (SLT) from signed to written languages (Camgoz et al., 2018; Shi et al., 2022; Zhang et al., 2024), isolated sign language recognition (ISLR) (Kezar et al., 2023), and fingerspelling detection and recognition (Shi et al., 2019; Fayyazsanavi et al., 2024; Georg et al., 2024). Pre-training approaches allow for pooling data across tasks, and several pre-training methods have been successful for sign language tasks (Uthus et al., 2023; Rust et al., 2024). However, these have typically focused on either supervised pre-training, which cannot take advantage of unlabeled data, context-independent (frame or video segment) representations, which ignore the effects of relationships across time in sign language, or contextual representations of only some aspects of sign language (see Sec. 2). These limitations have historically constrained the performance and scalability of sign language processing systems.

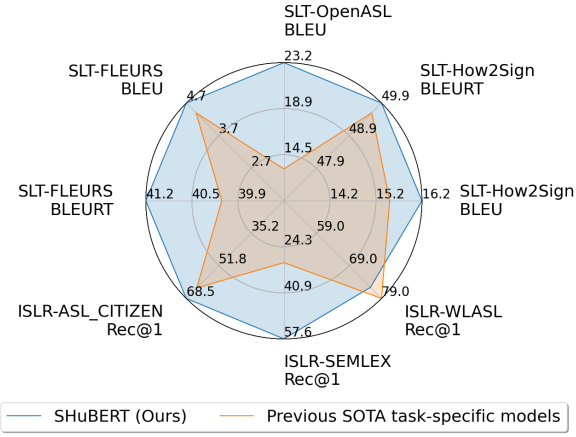

Figure 1: Comparison between our results using fine-tuned SHuBERT and results of the previous state-of-the-art task-specific models on a suite of tasks, datasets, and metrics. Note: The orange shade does not represent a single model but a collection of the previous SOTA results for models trained on public data. The SHuBERT-based results improve on all but one task-specific SOTA model. See Sec. 4 for details.

The success of self-supervised representations for written and spoken language, such as BERT for written language (Devlin et al., 2019) and HuBERT for speech (Hsu et al., 2021), has yet to be realized for sign language. Self-supervised learning seems particularly relevant for sign language, for which annotated datasets are scarce. But the unique multi-channel and other visual properties of sign languages suggest a specialized approach.

In this work, we present SHuBERT (Sign Hidden-Unit BERT) (Fig. 2), a self-supervised rep-

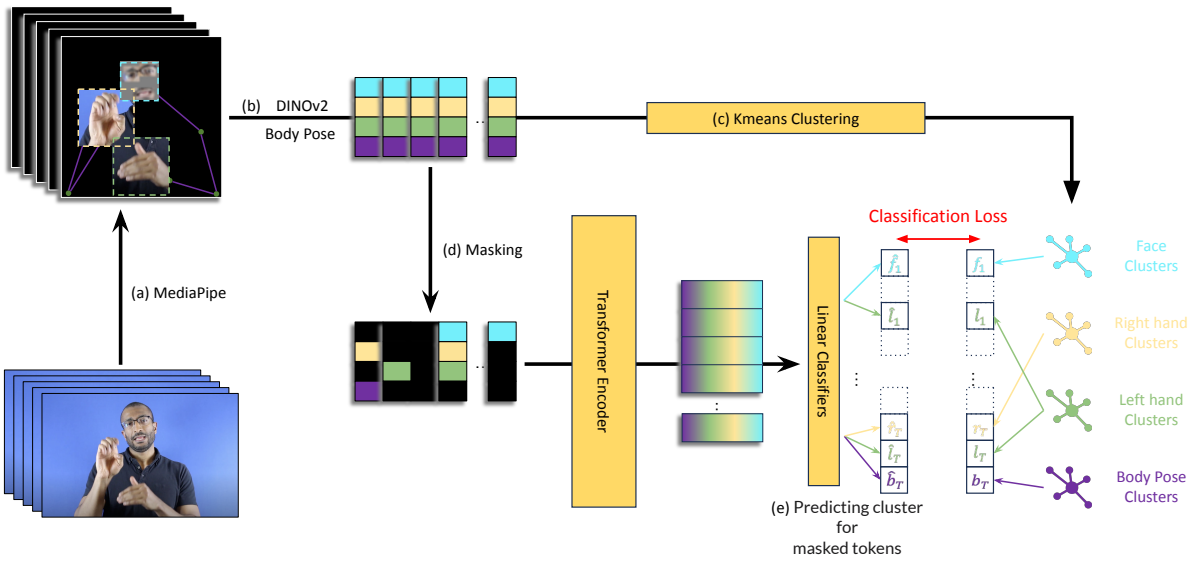

Figure 2: SHuBERT pre-training. (a) We locate a set of landmarks in each frame of the input video using MediaPipe (Lugaresi et al., 2019), with inter-frame interpolation to fill in missing landmarks. From these, we extract the upper body pose, crop the hand and face regions, and blur and partially mask the face crop for a measure of privacy and robustness. (b) We use DINOv2 (Oquab et al., 2023) to extract features for the hands and face, yielding a four-stream representation (two hands, face, body pose) for each frame. (c) We assign the feature vectors for frame t to cluster indices using pre-computed k-means clusters, yielding assignments $(f_t, l_t, r_t, b_t) \in [k]^4$ for face, left and right hand, and body pose, respectively. (d) We partially mask the features, and the masked features form the input to the transformer encoder. (e) We train SHuBERT to predict the cluster assignments for each masked input frame, $(\widehat{f}_t, \widehat{l}_t, \widehat{r}_t, \widehat{b}_t)$.

resentation learning approach that learns *contextual frame representations for all sign language channels jointly*. SHuBERT adapts the masked prediction paradigm of BERT and HuBERT to the characteristics of sign language video, and learns by predicting cluster assignments of multiple masked feature streams representing the hands, face, and body pose. The learned representations transfer effectively to multiple sign language understanding tasks, achieving state-of-the-art performance on several SLT benchmarks, multiple ISLR benchmarks, and fingerspelling detection, and improving over specialized models for each task (Fig. 1).

2 Related Work

Sign language understanding (recognition and translation) tasks have received increasing attention in the last few years (Camgoz et al., 2018; Shi et al., 2022; Lin et al., 2023; Kezar et al., 2023). For translation, early work mainly focused on gloss-based methods, which rely on (the rare and small) datasets with manually labeled glosses (Camgoz et al., 2018). More recent work has turned to larger and more naturalistic datasets without gloss labels. The most commonly used datasets are in American Sign Language (ASL) (Duarte et al., 2021; Shi et al., 2022; Uthus et al., 2023), German Sign Language (DGS) (Camgoz et al., 2018), British Sign Language (Albanie et al., 2021), and Chinese Sign Language (Zhou et al., 2021a). Of these, recent ASL datasets are the most naturalistic, and include large quantities of natively produced sign language (rather than translated from a spoken language, which has properties of "translationese" (Desai et al., 2024b)). For this reason we focus on ASL data and tasks, but our approach is applicable and extensible to any sign language.

2.1 Pre-Training for Text And Speech

Pre-training is a cornerstone of modern language processing across modalities. For written language, encoder models like BERT (Devlin et al., 2019) and its variants (e.g., (Liu, 2019; Lan et al., 2020)), based on masked language modeling, have served as dependable representations for language understanding tasks. In speech processing, self-supervised learning approaches (Mohamed et al., 2022) have taken inspiration from text encoder models while addressing the unique challenges

posed by continuous audio, which has no inherent segmentation into tokens nor a pre-defined token vocabulary. For example, Hidden-Unit BERT (HuBERT) (Hsu et al., 2021) adapts BERT by adding an offline clustering step to provide pseudo-labels for masked prediction. Such self-supervised representations, combined with task-specific fine-tuning, remain the state of the art for many speech tasks.

Sign languages share similar challenges to speech, with no pre-existing token lexicon and variable-length units (gestures) with no explicit boundaries, and our approach takes inspiration from HuBERT. However, sign language video has its own unique challenges: the many sources of variation (signer appearance, background, lighting, camera angles), the multiple streams of gestures (hands, face, body), the high dimensionality of video, and the relative dearth of data. These challenges are addressed in SHuBERT by focusing on the relevant streams (via pose tracking) combined with multi-stream masking and clustering.

2.2 Pre-Training for Sign Language

Supervised pre-training. For translation of sign language video, the supervised pre-training approach has focused on (pre-)training a translation model on a large (but often noisy) out-of-domain dataset, followed by fine-tuning on a smaller in-domain dataset. This type of pre-training leverages large collections of annotated data, with some systems (Uthus et al., 2023; Tanzer and Zhang, 2024; Tanzer, 2024a; Zhang et al., 2024) trained on up to 6,600 hours of sign language content (Uthus et al., 2023) to achieve state-of-the-art performance. However, much of this data remains private. In addition, these approaches often involve substantial computational resources: The supervised model of Zhang et al. (2024), for example, was trained on 128 TPU-v3 chips for 20 days. Jiao et al. (2024) propose an alternative pre-training approach that greatly improves efficiency by using pose information only (rather than image pixels); however, this approach pre-trains and fine-tunes on the same training data. Unlike these approaches, Uni-Sign (Li et al., 2025) is a supervised pre-training approach, based on mT5 (Xue et al., 2021), that has been applied to multiple tasks including both translation and ISLR. Like all supervised methods, these approaches can not take advantage of available unlabelled sign language data.

Self-supervised pre-training. Previous work has compared multiple context-independent self-supervised techniques for ISLR, finding masked autoencoders (MAE) particularly effective (Sandoval-Castaneda et al., 2023). SSVP-SLT (Rust et al., 2024) adapts MAEs for large-scale sign language pre-training, achieving competitive performance on ASL-to-English translation. This approach is computationally demanding, using 64 A100 GPUs for 14 days, and takes a maximum of 128 input frames ($\sim$8 seconds) at a time so is unable to model longer-term dependencies. Other lines of work on SLT (e.g., Chen et al. (2022a)) have used a pre-trained S3D model (Xie et al., 2018), which requires the video sequence to be segmented into chunks, with each chunk treated as independent. All of these approaches learn context-independent representations of individual frames or video segments, whereas SHuBERT learns contextual frame representations and can operate on long video directly.

The only previous self-supervised approach of which we are aware for *contextual* sign representation learning is SignBERT+ (Hu et al., 2023), which extends the earlier SignBERT (Hu et al., 2021). This approach learns a representation specifically for hand poses, via masked reconstruction of hand joints, and has strong results on ISLR, continuous (gloss-based) sign recognition, and sign translation on the RWTH-PhoenixT German Sign Language dataset (Camgoz et al., 2018). However, this approach is inherently limited by not modeling the face and global body pose, and the results are obtained by combining SignBERT+ with a dataset-specific image pixel (RGB) representation. In addition, SignBERT+ is pre-trained on the union of datasets on which it is tested; that is, it is exposed to the fine-tuning data during pre-training. In contrast, SHuBERT models all components of sign language jointly, and is pre-trained on data that is disjoint from the fine-tuning data for the downstream tasks.

2.3 Multi-Stream Models of Sign Language

Several previous methods have taken advantage of the observation that sign language naturally decomposes into multiple streams of hand, face, and body motions.[1] For example, prior work incudes multi-stream models for fingerspelling recog-

[1]The term "multi-stream" has been used in different senses in prior work. For example, DSTA-SLR (Hu et al., 2024) creates multiple streams consisting of different geometric representations of the same skeleton data, while we are concerned with streams that correspond to different body parts.

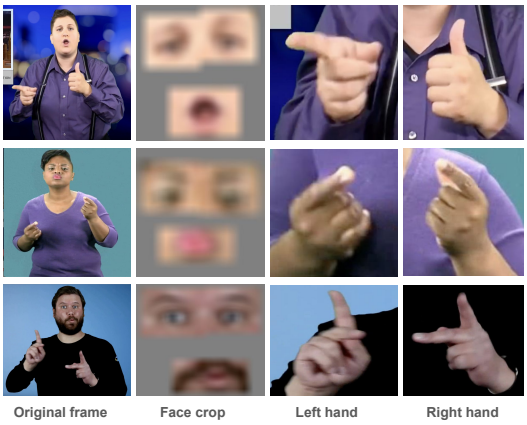

Figure 3: Sample frames of several signers and the corresponding input channels used for SHuBERT: blurred face crop, left hand crop, and right hand crop. In addition to the face and left/right hand features extracted from these crops, each frame is represented by an additional feature vector corresponding to the upper body pose extracted from MediaPipe (see Fig. 2).

nition (which combines hand and mouthing gestures) (Shi, 2023), SLT (Camgoz et al., 2020; Zhou et al., 2021b; Chen et al., 2022b; Shi et al., 2022; Gueuwou et al., 2025), and ISLR (Pu et al., 2016; Jiang et al., 2021). This factorization into multiple streams can enable dramatic improvements in data and compute efficiency over single-stream models that use the full image (Gueuwou et al., 2025).

Like this prior work, SHuBERT also adopts the idea of multiple streams. However, unlike prior approaches, SHuBERT learns a *self-supervised* representation from the multiple streams jointly that performs well on multiple tasks.

3 Sign Hidden-Unit BERT (SHuBERT)

SHuBERT is a transformer encoder (Vaswani et al., 2017) that learns contextualized representations of sign language video frames through self-supervised learning. The pre-training approach is outlined in Fig. 2. In the following sections we describe the video features used in SHuBERT (Sec. 3.1) and the self-supervised training approach (Sec. 3.2).

3.1 Multi-Stream Feature Pre-Processing

Fig. 3 provides examples of SHuBERT's input video features, described in detail below.

Handshapes We use the MediaPipe Hand Landmarker[2] model, which has hand detection accuracy ~95% on OpenASL.[3] Upon inspection, we find that the majority of the remaining 5% of "failed" detections occur when the hands are outside the frame. For these cases, we interpolate from the nearest frames with successful detections. Dilated bounding boxes for the detected hand landmarks (for both left and right hand) are cropped and resized to 224×224.

Facial Features The signer's face contains important non-manual markers for sign languages (Bragg et al., 2019). Previous approaches either use the full face, compromising privacy (Gueuwou et al., 2025), or blur the whole face in an attempt to protect privacy, potentially losing essential non-manual markers (Rust et al., 2024). Our design attempts to balance the need to preserve linguistic information with the goal of enhancing privacy. We identify the whole face, mouth and eye regions in the frame from the relevant MediaPipe facial landmarks. The face pixels are greyed out *except for the pixels in the mouth and eye regions*. We then apply Gaussian blur to the entire face region and resize it to 224×224.

Image Feature Extractor for Hands and Face We use DINOv2 (Oquab et al., 2023), which has proven successful in previous sign language work (Wong et al., 2024; Gueuwou et al., 2025), as the feature extractor for face and hand image crops. An additional benefit of DINOv2 representations is that they yield meaningful clusters after quantization (Zheng et al., 2024), which is an important property since SHuBERT training targets are clustered input features. While many other prior approaches use keypoint estimation tools, these have some weaknesses in capturing handshapes (Moryossef et al., 2021) and facial expressions (Kuznetsova and Kimmelman, 2024).

To adapt the general image feature extractor to a face feature extractor for sign language, we randomly sample 5 million face crops from videos in the YouTube-ASL (Uthus et al., 2023) dataset and use them for continued pre-training of DINOv2 for 1 epoch. We do the same for the hand feature extractor, using 5 million randomly sampled hands (mix of left and right hand crops) from YouTube-ASL. For both the face and hand streams, crop regions of interest (ROIs) are processed through the face or hand fine-tuned DINOv2 models, yielding

[2] https://ai.google.dev/edge/mediapipe/solutions/vision/hand_landmarker

[3] Estimated from a small-scale experiment on 100 OpenASL videos, with manual verification of detections.

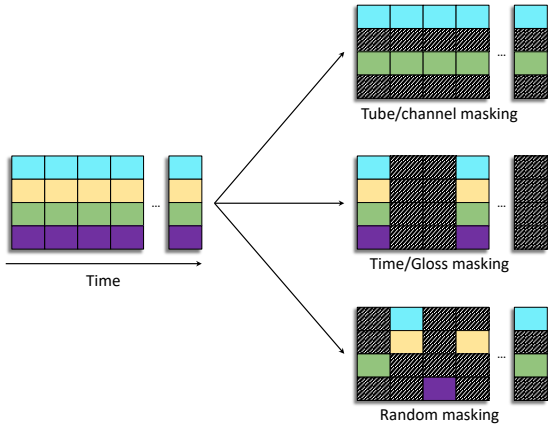

Figure 4: Three strategies for sequence masking. Each color-coded row corresponds to one of the four channels (face, right hand, left hand, body).

a 384-dimensional feature vector per crop, which we denote $\mathbf{x}_t^f, \mathbf{x}_t^l, \mathbf{x}_t^r \in \mathbb{R}^{384}$ for the face, left hand, and right hand features respectively.

Body Pose For (coarse) body pose, we extract seven key upper body landmarks (nose, shoulders, elbows, and wrists) and normalize their coordinates relative to the signing space, resulting in a compact 14-dimensional pose vector, $\mathbf{x}_t^b \in \mathbb{R}^{14}$.

3.2 Self-Supervised Training of SHuBERT

For each of the four channels (left/right hand, face, and pose), we layer-normalize the extracted features (producing zero mean and unit variance across all dimensions for each feature vector) and linearly project them to 256 dimensions, producing a 1024-dimensional input feature vector per frame. This joint representation of the four streams is masked (see below) and input to the transformer encoder. The output of the transformer is one representation vector per input frame, y_t, which may take into account information from the entire length of the input video, and from all four input channels.

Sequence Masking SHuBERT learns by predicting masked elements of the input feature sequence, given the observed (unmasked) data. We use a masking strategy designed for *multi-channel* sign language input. We consider three types of masking, illustrated in Fig. 4: channel masking, which masks entire channels (e.g., all face and left hand features in a video) to learn cross-channel dependencies; time masking, which masks all channels at selected temporal positions (e.g., face, hand, and body pose in frames 20-40 in a given video); and random masking, which independently masks random small frame spans in each channel. Based on our experiments comparing these strategies (Appendix A), we ultimately chose random masking.

Learning Objective We use offline k-means clustering (separately for each of the four channels) to create discrete target pseudo-labels (f_t, l_t, r_t, b_t) for the face, left hand, right hand, and body pose respectively. The transformer output vector for each frame is fed to four linear classifiers (one per channel) to predict the cluster assignments for masked channels for that frame.

As an example, suppose that face and body pose channel features for frame t, respectively $\mathbf{x}_t^f \in \mathbb{R}^{384}$ and $\mathbf{x}_t^b \in \mathbb{R}^{14}$, are masked. The k-means cluster assignments for these masked feature vectors are, respectively, f_t and b_t, each a number between 1 and k. The classifier predicts, from the output vector y_t for frame t, labels $\widehat{f}_t$ and $\widehat{b}_t$. The training objective is a cross-entropy loss between the target and predicted cluster assignments for the masked positions. The unmaksed positions are not included in the loss (but of course influence the predictions for the masked ones).

This self-supervised training produces the pre-trained SHuBERT model, which can then be fine-tuned for downstream sign language tasks using appropriate task-specific prediction layers and losses.

4 Experiments and Results

In this section, we describe the experimental setup for self-supervised training of SHuBERT, followed by its adaptation as a foundation model for multiple sign language processing tasks: sign language translation (Sec. 4.2), isolated sign language recognition (Sec. 4.3) and fingerspelling detection (Sec. 4.4). We also apply SHuBERT to phonological feature recognition, as a baseline for future work (see Appendix B).

4.1 Pre-Training SHuBERT

Data and Pre-Processing For pre-training SHuBERT, we use the YouTube-ASL dataset (Uthus et al., 2023). Note that *we excluded the clips that intersect with the OpenASL dataset* (Shi et al., 2022) to evaluate SHuBERT on OpenASL. To maintain the same training set size as the original YouTube-ASL dataset, we replaced the removed content with ASL videos from YouTube-SL-25 (Tanzer and Zhang, 2024) that are not present in YouTube-ASL.

Our final pre-training dataset comprises approximately 984 hours of ASL content.[4]

To reduce computation, we downsample the videos by removing every other frame (Uthus et al., 2023). The average frame rate of the post-processed videos is 14.89 fps.

Model Configuration and Training We train a base model consisting of 12 transformer blocks, with each block having an embedding dimensionality of 768, feed-forward dimensionality of 3072, and 12 attention heads. The complete model contains 86M parameters. For each channel (face, left hand, right hand, and body pose), we use k-means on 10% of the data to create 256 discrete clusters that serve as prediction targets. Figs. 5 to 8 in the Appendix provide examples of images in several clusters. Our masking strategy uses a span length of 3 frames (approximately 200ms), which is roughly the average duration of fingerspelling a single letter in ASL and is therefore roughly the smallest gesture length (Hanson, 1982). We train the model for 400K steps using 8 NVIDIA A6000 (48GB) GPUs, with a total training time of approximately 7 days.

We optimize the model using Adam with a peak learning rate of 5×10^{-4}, warming up for the first 8% of updates followed by linear decay. We batch videos to maintain efficiency while not exceeding 1,500 frames per GPU.

4.2 Sign Language Translation

For sign language translation, where the input sign language video is mapped to text in English, we use ByT5-Base (Xue et al., 2022), pre-trained on a large corpus of unlabeled multilingual text data, as a translation model to map from SHuBERT representations to written English, following prior work showing its strong performance on this task (Tanzer and Zhang, 2024; Zhang et al., 2024). We first extract video representations from SHuBERT and project them to ByT5's input space through a linear layer. We train the combined model (SHuBERT+projection layer+ByT5) with a cross-entropy loss and label smoothing factor of 0.2.

Similarly to Uthus et al. (2023) and Rust et al. (2024), we use a two-phase training strategy. In the first phase, we train the translation system on the weakly labeled YouTube-ASL dataset (with OpenASL removed, as described above) for 250K steps. During this phase, we use the AdamW optimizer with a peak learning rate of 5×10^{-4} for ByT5 and a reduced learning rate of 5×10^{-5} for SHuBERT parameters (when fine-tuned). The learning rate follows a cosine schedule with 10K warmup steps. We use a batch size of 2 utterances per GPU with gradient accumulation over 8 steps and use weight decay of 0.1.

In the second phase, we fine-tune on two target benchmark datasets (How2Sign (Duarte et al., 2021) and OpenASL (Shi et al., 2022)) for 50K steps, using a lower learning rate of 10^{-4} and 5K warmup steps. We also evaluate in a zero-shot setting, without any additional fine-tuning, on a third dataset (for which no training data exists), FLEURS-ASL (Tanzer, 2024b). We use a learned weighted sum of features from all SHuBERT layers rather than using a single layer's output, as is commonly done when using speech representations such as HuBERT (Yang et al., 2021). During decoding, we use beam search with a beam width of 5 and a maximum sequence length of 384 tokens.

We evaluate the final model on the three benchmark test sets of How2Sign, OpenASL, and FLEURS-ASL, using the standard BLEU (Papineni et al., 2002; Post, 2018)[5] and BLEURT (Sellam et al., 2020) translation metrics, as shown in Tab. 1 (see also example translations in Tabs. 11 to 13).

In all cases, our results using SHuBERT improve over the best prior published results using publicly available training data. On How2Sign, SHuBERT improves by +0.7 BLEU/+0.3 BLEURT over the prior state-of-the-art result (using public data) of SSV-SLT (Rust et al., 2024), which used slighty more pre-training data (1,054 vs. 984 hours) that included the training data of How2Sign. Better published results exist (as shown in Tab. 1), but they rely on private fine-tuning datasets so we cannot reproduce their settings nor compare to them meaningfully.

In the case of OpenASL, SHuBERT's improvement over the best prior result is +2.0 BLEU.[6] This larger improvement may be attributable to pre-training on similar-domain (generally native, natu-

[4]Note that YouTube-ASL encompasses parts of several datasets used in other work, including 72.4% of the OpenASL test set and 38.2% of the MSASL (Joze and Koller, 2018) test set. We do not compare to other work on MSASL for this reason.

[5]`SacreBLEUversionsignature:BLEU+c.mixed+#.1+s.exp+tok.13a+v.1.4.1.`

[6]Uni-Sign (Li et al., 2025) reports a similar BLEU score of 23.1 on OpenASL, but Uni-Sign is pre-trained on YouTube-ASL. Therefore, most of the test set is included in Uni-Sign's pre-training data, so we do not consider the results directly comparable.

Method	SSL	PT data (hrs)	How2Sign		OpenASL		FLEURS-ASL	
			BLEU↑	BLEURT↑	BLEU↑	BLEURT↑	BLEU↑	BLEURT↑
Private data								
Tanzer (2024a)	×	∼2,800	18.1	50.8	-	-	**5.8**	**45.4**
Zhang et al. (2024)	×	∼6,600	**21.1**	**55.7**	-	-	-	-
Publicly available data								
SSVP (Rust et al., 2024)	✓	1,054	15.5	49.6	-	-	-	-
Uthus et al. (2023)	×	984	12.4	46.6	-	-	-	-
Tanzer and Zhang (2024)	×	3,207	15.4	47.9	-	-	4.4	40.1
SM (Gueuwou et al., 2025)	✓	984	14.3	-	-	-	-	-
VAP (Jiao et al., 2024)	×	-	12.9	-	21.2	-	-	-
Uni-Sign (Li et al., 2025)	×	984	14.9	49.4	23.1*	60.4*	-	-
OpenASL (Shi et al., 2022)	×	-	-	-	6.7	31.1	-	-
GloFE-VN (Lin et al., 2023)	×	-	-	-	7.1	36.7	-	-
C2RL (Chen et al., 2025)	×	-	-	-	13.2	-	-	-
Tanzer (2024b)	×	984	-	-	-	-	3.9	38.3
Ours	✓	984	**16.2**	**49.9**	**23.2**	**60.6**	**4.7**	**41.2**

Table 1: Translation results on the How2Sign, OpenASL, and FLEURS-ASL test sets. SSL: self-supervised learning (yes/no); PT: pre-training. * Uni-Sign is pre-trained on YouTube-ASL, which contains >72% of the OpenASL test samples, so we do not consider the results on OpenASL to be directly comparable to ours.

ral rather than interpreted signing) data in YouTube-ASL (but, as previously mentioned, none of the same data). The distinction between interpreted signing in a constrained visual environment (as in How2Sign and FLEURS-ASL) and natural signing in a less constrained environment (as in OpenASL and much of YouTube-ASL) is an important one that has not received sufficient attention, and is worth exploring further in future work. The properties of the visual environment affect the difficulty of the task, and natural signing has different characteristics from those of interpreted sign language (Desai et al., 2024c).

On FLEURS-ASL, a dataset designed for testing only, SHuBERT demonstrates strong performance in a *zero-shot* setting, surpassing both prior methods, which used over 3x as much pre-training data (3,207 hours) as ours.

We also note that some previous approaches use additional techniques such as auxiliary losses that contribute to their final results, such as additional contrastive learning with labelled data (Rust et al., 2024), joint training with text machine translation (Zhang et al., 2024), and multi-tasking with random dynamic clips from an original video (Tanzer, 2024b). It is possible that incorporating such techniques into our framework will further improve performance, but we leave this for future work.

Ablations We conduct several ablation studies to validate our design choices and analyze SHuBERT's behavior. These studies examine: (1) the impact of different masking strategies during pre-training, where random masking proves most effective based on BLEURT scores; (2) the importance of pre-training data scale, showing the clear benefit of using the full pre-training dataset; (3) the contribution of different architectural components, demonstrating that a weighted combination of layers significantly improves translation performance; and (4) the effects of fine-tuning versus keeping SHuBERT frozen during translation training, with fine-tuning providing moderate gains. The strong performance of the frozen, layer-weighted SHuBERT suggests that it is a promising approach for low-resource settings where parallel data may be limited. Detailed results and analysis of these ablations can be found in Appendix A.

4.3 Isolated Sign Language Recognition

For isolated sign language recognition (ISLR), the task of classifying a short video of a single sign, we include results for SHuBERT adapted with LoRA adapters (Hu et al., 2022). Unless otherwise specified, for all experiments, we train for 125 epochs with a batch size of 128 and perform early stopping according to validation results (R@1/P-I). We use an Adam optimizer (Kingma, 2014) with a learning rate of 10^{-4} and weight decay of 10^{-4}. For classification tasks, we first average SHuBERT rep-

Method	#Params	ASL Citizen			Sem-Lex			WLASL2000	
		Rec@1↑	Rec@5↑	Rec@10↑	Rec@1↑	Rec@5↑	Rec@10↑	P-I↑	P-C↑
ST-GCN (Desai et al., 2024a)	0.45M	0.60	0.82	0.88	-	-	-	-	-
SignCLIP (Jiang et al., 2024)	217M	0.60	0.84	0.89	0.30	0.48	0.55	-	-
I3D (Desai et al., 2024a)	25M	0.63	0.86	**0.91**	-	-	-	-	-
Sem-Lex (Kezar et al., 2023)	0.45M	-	-	-	0.69*	-	-	-	-
SignBERT (Hu et al., 2021)	-	-	-	-	-	-	-	39.40	36.74
SignBERT+ (Hu et al., 2023)	-	-	-	-	-	-	-	48.85	46.37
MSLU (Zhou et al., 2024)	-	-	-	-	-	-	-	56.29	53.29
NLA-SLR (Zuo et al., 2023)	-	-	-	-	-	-	-	61.05	58.05
Uni-Sign (Li et al., 2025)	580M	-	-	-	-	-	-	**63.52**	**61.32**
Ours (rank=1 LoRA)	0.17M	**0.65**	**0.87**	**0.91**	**0.54**	**0.74**	**0.80**	60.90	58.01

Table 2: ISLR results on the ASL Citizen, Sem-Lex, and WLASL2000 test sets. Note: For Sem-Lex, the result marked with an asterisk (*) is not directly comparable to Ours as it is for a reduced (and easier) test set, as mentioned in (Jiang et al., 2024). Additionally, the dataset version released by (Kezar et al., 2023) has a significant fraction of videos missing. For WLASL2000, evaluation metrics are per-instance (P-I) and per-class (P-C) Top-1 accuracy.

Method	SSL	Mean IoU ↑
Contrastive Learning (Yin et al., 2024)	×	0.28
SHuBERT (Ours)	✓	**0.40**

Table 3: Fingerspelling detection on ASL-Stem Wiki (Yin et al., 2024).

resentations across the time dimension and add a batch-norm layer followed by a linear layer as the classification head.

For LoRA training, we learn a rank-1 LoRA module for each linear layer in SHuBERT while keeping all the other weights frozen, resulting in training only 0.2% of the number of parameters of the original model. In addition to the aforementioned hyperparameters, we reduce the learning rate of the LoRA modules to 1/10 of the classification head's and also use 0.1 label smoothing.

Our ISLR results on ASL Citizen, WLASL2000 (original), and Sem-Lex are shown in Tab. 2. We note that we do not report on MSASL, as done in some of the prior work, because of the aforementioned overlap between its test set and YouTube-ASL. Following prior work, we report Recall at 1, 5 and 10, unless stated otherwise.[7] We achieve state-of-the-art performance on all datasets except WLASL2000 where Uni-Sign has a better result. We note that Uni-Sign fine-tunes 3,000 times more paramters than ours.[8]

4.4 Fingerspelling detection

Finally, we evaluate SHuBERT on the task of fingerspelling detection on the ASL-Stem-Wiki dataset (Yin et al., 2024). Given a sign language video input v, which consists of an ordered sequence of frames $\{v_1, v_2, \ldots, v_n\}$, the task is to identify all segments containing fingerspelling. The output is represented as a set F of frame intervals: $F = \{[s_1, e_1], [s_2, e_2], \ldots, [s_k, e_k]\}$, where each interval $[s_i, e_i]$ represents the start and end frames of a fingerspelling sequence, such that frames v_{s_i} through v_{e_i} contain fingerspelling. We follow the original ASL-Stem-Wiki evaluation pipeline (cross-validation) and evaluation metric (intersection over union, or IoU). The results are shown in Tab. 3. We see that by simply fine-tuning SHuBERT for this task, we increase the IoU by 42% compared to the previous state of the art method, which pre-trains and fine-tunes on the same dataset.

5 Conclusion

SHuBERT, our proposed self-supervised approach for learning sign language video representations, yields a transformer encoder that maps from multiple feature streams (face and hand appearance and upper body pose) to a stream of per-frame contextual representations. A single base SHuBERT model, when adapted to a range of sign language processing benchmarks including both translation and isolated sign recognition, achieves strong performance on all of them and almost always improves over the prior state of the art. SHuBERT is trained on public data and is publicly available.[9] Based on its strong performance on the tasks studied here, we expect that SHuBERT can serve as a base model for a broad range of sign language processing tasks.

[7] Note that some prior work reports ISLR results in terms of accuracy, which is equivalent to Recall at 1.

[8] Note: As described in the caption of Tab. 2, the Sem-Lex results in (Kezar et al., 2023) are not comparable with other methods, including ours.

[9] http://shubert.pals.ttic.edu

Limitations Our work has several limitations. First, although the results are competitive with or outperform prior work, the absolute performance is still quite poor. Neither our model nor others can replace human interpreters for broad-domain sign language translation. Second, our training data volume is significantly smaller than that of typical self-supervised speech and text models. We cannot say with certainty how our model would scale up to much larger datasets. Third, we have not carefully studied potential sources of bias in the model. From our qualitative visual inspection of images and their corresponding clusters (Figs. 5 to 8 in the Appendix), we observe that semantic properties appear to take precedence over attributes like skin color, gender, or eyewear. While these preliminary observations are encouraging, a more thorough investigation of potential biases would be valuable future work. Finally, the current scope of our work is limited to American Sign Language. Although sign languages use the same channels and share many elements, we do not know how well our model would generalize to other languages. Future work could address these limitations by expanding the training dataset and training on data from other sign languages.

Acknowledgment We are grateful to Shiry Ginosar, Anand Bhattad, Ju-Chieh Chou, and Chung-Ming Chien for their valuable suggestions throughout this project.

References

Samuel Albanie, Gül Varol, Liliane Momeni, Hannah Bull, Triantafyllos Afouras, Himel Chowdhury, Neil Fox, Bencie Woll, Rob Cooper, Andrew McParland, et al. 2021. BBC-Oxford British Sign Language dataset. *arXiv preprint arXiv:2111.03635*.

Ursula Bellugi and Susan Fischer. 1972. A comparison of sign language and spoken language. *Cognition*.

Danielle Bragg, Oscar Koller, Mary Bellard, Larwan Berke, Patrick Boudreault, Annelies Braffort, Naomi Caselli, Matt Huenerfauth, Hernisa Kacorri, Tessa Verhoef, et al. 2019. Sign language recognition, generation, and translation: An interdisciplinary perspective. In *Proc. SIGACCESS*.

Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. 2018. Neural sign language translation. In *Proc. CVPR*.

Necati Cihan Camgoz, Oscar Koller, Simon Hadfield, and Richard Bowden. 2020. Multi-channel transformers for multi-articulatory sign language translation. In *ECCV Workshops*.

Yutong Chen, Fangyun Wei, Xiao Sun, Zhirong Wu, and Stephen Lin. 2022a. A simple multi-modality transfer learning baseline for sign language translation. In *Proc. CVPR*.

Yutong Chen, Ronglai Zuo, Fangyun Wei, Yu Wu, Shujie Liu, and Brian Mak. 2022b. Two-stream network for sign language recognition and translation. In *Proc. SIGACCESS*.

Zhigang Chen, Benjia Zhou, Yiqing Huang, Jun Wan, Yibo Hu, Hailin Shi, Yanyan Liang, Zhen Lei, and Du Zhang. 2025. C^2RL: Content and context representation learning for gloss-free sign language translation and retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*.

Aashaka Desai, Lauren Berger, Fyodor Minakov, Nessa Milano, Chinmay Singh, Kriston Pumphrey, Richard Ladner, Hal Daumé III, Alex X Lu, Naomi Caselli, et al. 2024a. ASL Citizen: A community-sourced dataset for advancing isolated sign language recognition. In *Proc. NeurIPS*.

Aashaka Desai, Maartje De Meulder, Julie A. Hochgesang, Annemarie Kocab, and Alex X. Lu. 2024b. Systemic biases in sign language AI research: A Deaf-led call to reevaluate research agendas. In *LREC-COLING 2024 11th Workshop on the Representation and Processing of Sign Languages: Evaluation of Sign Language Resources*.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc. NAACL-HLT*.

Amanda Duarte, Shruti Palaskar, Lucas Ventura, Deepti Ghadiyaram, Kenneth DeHaan, Florian Metze, Jordi Torres, and Xavier Giro-i Nieto. 2021. How2Sign: A large-scale multimodal dataset for continuous American Sign Language. In *Proc. CVPR*.

Pooya Fayyazsanavi, Negar Nejatishahidin, and Jana Košecká. 2024. Fingerspelling PoseNet: Enhancing fingerspelling translation with pose-based transformer models. In *Proc. Winter Conference on Applications of Computer Vision (WACV)*.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André FT Martins. 2022. Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust. In *Proc. Conference on Machine Translation (WMT)*.

Manfred Georg, Garrett Tanzer, Saad Hassan, Maximus Shengelia, Esha Uboweja, Sam Sepah, Sean Forbes, and Thad Starner. 2024. FSboard: Over 3 million

characters of ASL fingerspelling collected via smartphones. *arXiv preprint arXiv:2407.15806*.

Shester Gueuwou, Xiaodan Du, Greg Shakhnarovich, and Karen Livescu. 2025. SignMusketeers: An efficient multi-stream approach for sign language translation at scale. In *Findings of the Association for Computational Linguistics: ACL*.

Shester Gueuwou, Sophie Siake, Colin Leong, and Mathias Müller. 2023. JWSign: A highly multilingual corpus of bible translations for more diversity in sign language processing. In *Findings of the Association for Computational Linguistics: EMNLP*.

VL Hanson. 1982. Use of orthographic structure by deaf adults: Recognition of fingerspelled words. In *Applied Psycholinguistics*.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE Trans. Audio, Speech, Lang. Process.*

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *Proc. ICLR*.

Hezhen Hu, Weichao Zhao, Wengang Zhou, and Houqiang Li. 2023. SignBERT+: Hand-model-aware self-supervised pre-training for sign language understanding. *IEEE Trans. Pattern Anal. Mach. Intell.*

Hezhen Hu, Weichao Zhao, Wengang Zhou, Yuechen Wang, and Houqiang Li. 2021. SignBERT: Pre-training of hand-model-aware representation for sign language recognition. In *Proc. CVPR*.

Lianyu Hu, Liqing Gao, Zekang Liu, and Wei Feng. 2024. Dynamic spatial-temporal aggregation for skeleton-aware sign language recognition. In *Proc. LREC-COLING*.

Songyao Jiang, Bin Sun, Lichen Wang, Yue Bai, Kunpeng Li, and Yun Fu. 2021. Skeleton-aware multi-modal sign language recognition. In *Proc. CVPR*.

Zifan Jiang, Gerard Sant, Amit Moryossef, Mathias Müller, Rico Sennrich, and Sarah Ebling. 2024. SignCLIP: Connecting text and sign language by contrastive learning. In *Proc. EMNLP*.

Peiqi Jiao, Yuecong Min, and Xilin Chen. 2024. Visual alignment pre-training for sign language translation. In *Proc. ECCV*.

Hamid Reza Vaezi Joze and Oscar Koller. 2018. MS-ASL: A large-scale data set and benchmark for understanding American Sign Language. *arXiv preprint arXiv:1812.01053*.

Lee Kezar, Jesse Thomason, Naomi Caselli, Zed Sehyr, and Elana Pontecorvo. 2023. The Sem-Lex benchmark: Modeling ASL signs and their phonemes. In *Proc. SIGACCESS*.

Diederik P Kingma. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Anna Kuznetsova and Vadim Kimmelman. 2024. Testing MediaPipe Holistic for linguistic analysis of non-manual markers in sign languages. *arXiv preprint arXiv:2403.10367*.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A lite BERT for self-supervised learning of language representations. In *Proc. ICML*.

Zecheng Li, Wengang Zhou, Weichao Zhao, Kepeng Wu, Hezhen Hu, and Houqiang Li. 2025. Uni-Sign: Toward unified sign language understanding at scale. In *Proc. ICLR*.

Kezhou Lin, Xiaohan Wang, Linchao Zhu, Ke Sun, Bang Zhang, and Yi Yang. 2023. Gloss-free end-to-end sign language translation. In *Proc. ACL*.

Yinhan Liu. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, et al. 2019. Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*.

Abdelrahman Mohamed, Hung-Yi Lee, Lasse Borgholt, Jakob D Havtorn, Joakim Edin, Christian Igel, Katrin Kirchhoff, Shang-Wen Li, Karen Livescu, Lars Maaløe, et al. 2022. Self-supervised speech representation learning: A review. *IEEE J. Sel. Top. Signal Process.*, 16(6):1179–1210.

Amit Moryossef, Ioannis Tsochantaridis, Joe Dinn, Necati Cihan Camgoz, Richard Bowden, Tao Jiang, Annette Rios, Mathias Müller, and Sarah Ebling. 2021. Evaluating the immediate applicability of pose estimation for sign language recognition. In *Proc. CVPR*.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. 2023. DINOv2: Learning robust visual features without supervision. *Trans. Machine Learning Research*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proc. ACL*.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proc. Conference on Machine Translation (WMT)*.

Junfu Pu, Wengang Zhou, and Houqiang Li. 2016. Sign language recognition with multi-modal features. In *Advances in Multimedia Information Processing–PCM 2016: 17th Pacific-Rim Conference on Multimedia*.

Phillip Rust, Bowen Shi, Skyler Wang, Necati Cihan Camgoz, and Jean Maillard. 2024. Towards privacy-aware sign language translation at scale. In *Proc. ACL*.

Marcelo Sandoval-Castaneda, Yanhong Li, Diane Brentari, Karen Livescu, and Gregory Shakhnarovich. 2023. Self-supervised video transformers for isolated sign language recognition. *arXiv preprint arXiv:2309.02450*.

Thibault Sellam, Dipanjan Das, and Ankur P Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proc. ACL*.

Zed Sevcikova 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. In *The Journal of Deaf Studies and Deaf Education*.

Bowen Shi. 2023. *Toward American Sign Language Processing in the Real World: Data, Tasks, and Methods*. Ph.D. thesis, TTI-Chicago.

Bowen Shi, Diane Brentari, Gregory Shakhnarovich, and Karen Livescu. 2022. Open-domain sign language translation learned from online video. In *Proc. EMNLP*.

Bowen Shi, Aurora Martinez Del Rio, Jonathan Keane, Diane Brentari, Greg Shakhnarovich, and Karen Livescu. 2019. Fingerspelling recognition in the wild with iterative visual attention. In *Proc. CVPR*.

Garrett Tanzer. 2024a. Fingerspelling within sign language translation. *arXiv preprint arXiv:2408.07065*.

Garrett Tanzer. 2024b. FLEURS-ASL: Including American Sign Language in massively multilingual multitask evaluation. *arXiv preprint arXiv:2408.13585*.

Garrett Tanzer and Biao Zhang. 2024. YouTube-SL-25: A large-scale, open-domain multilingual sign language parallel corpus. *arXiv preprint arXiv:2407.11144*.

Dave Uthus, Garrett Tanzer, and Manfred Georg. 2023. YouTube-ASL: A large-scale, open-domain American Sign Language-English parallel corpus. In *Proc. NeurIPS*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proc. NeurIPS*.

Ryan Wong, Necati Cihan Camgoz, and Richard Bowden. 2024. Sign2GPT: Leveraging large language models for gloss-free sign language translation. In *Proc. ICLR*.

Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. 2018. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *Proc. ECCV*.

Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. 2022. ByT5: Towards a token-free future with pre-trained byte-to-byte models. *Trans. ACL*.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proc. NAACL-HLT*.

Shu-wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y Lin, Andy T Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, et al. 2021. SUPERB: Speech processing universal performance benchmark. In *Proc. Interspeech*.

Kayo Yin, Chinmay Singh, Fyodor O Minakov, Vanessa Milan, Hal Daumé III, Cyril Zhang, Alex Xijie Lu, and Danielle Bragg. 2024. ASL STEM Wiki: Dataset and benchmark for interpreting STEM articles. In *Proc. EMNLP*.

Biao Zhang, Garrett Tanzer, and Orhan Firat. 2024. Scaling sign language translation. In *Proc. NeurIPS*.

Matthew Zheng, Enis Simsar, Hidir Yesiltepe, Federico Tombari, Joel Simon, and Pinar Yanardag. 2024. Stylebreeder: Exploring and democratizing artistic styles through text-to-image models. In *Proc. NeurIPS*.

Hao Zhou, Wengang Zhou, Weizhen Qi, Junfu Pu, and Houqiang Li. 2021a. Improving sign language translation with monolingual data by sign back-translation. In *Proc. CVPR*.

Hao Zhou, Wengang Zhou, Yun Zhou, and Houqiang Li. 2021b. Spatial-temporal multi-cue network for sign language recognition and translation. *IEEE Trans. Multimedia*, 24:768–779.

Wengang Zhou, Weichao Zhao, Hezhen Hu, Zecheng Li, and Houqiang Li. 2024. Scaling up multimodal pre-training for sign language understanding. *arXiv preprint arXiv:2408.08544*.

Ronglai Zuo, Fangyun Wei, and Brian Mak. 2023. Natural language-assisted sign language recognition. In *Proc. CVPR*.

Masking strategy	BLEU-1	BLEU	BLEURT
Channel masking	14.5	2.6	29.9
Time masking	15.1	2.3	31.2
Random masking	15.4	2.2	31.4

Table 4: Comparison of masking strategies in SHuBERT pre-training. SHuBERT is frozen and stopped after 100K steps, and fine-tuned and evaluated on How2Sign only.

Hours of pre-train data	BLEU-1	BLEU	BLEURT
984	15.4	2.2	31.4
98	12.7	0.7	29.1

Table 5: Impact of pre-training data size on SHuBERT's performance.

Layer of SHuBERT	BLEU-1	BLEU	BLEURT
None	15.3	2.5	31.6
Last Layer	21.4	4.7	35.0
Weighted Sum	29.3	7.1	39.5

Table 6: Contribution of several components of SHuBERT: direct use of input video features (None) vs. using SHuBERT's last layer vs. weighted combination of all layers. In all cases SHuBERT is frozen.

Fine-tune SHuBERT	BLEU-1	BLEU	BLEURT
×	21.4	4.7	35.0
✓	30.0	7.5	39.9

Table 7: Effect of fine-tuning on translation performance: Fine-tuning SHuBERT along with the translation model (✓) vs. using frozen SHuBERT representations (×).

A Ablations on SHuBERT

We conduct ablations, using the How2Sign translation task, to measure the importance of different factors in the SHuBERT pre-training and adaptation to the downstream task. Unless stated otherwise, in the ablation experiments we pre-train SHuBERT for 100K steps (instead of the full 400K) for a faster turnaround and freeze SHuBERT for the downstream task.

Masking Strategies Tab. 4 compares translation performance when using each of the three masking strategies (Fig. 4). We observe different signals from different metrics. While the BLEU scores suggest channel masking to be the best, random masking produces better BLEURT scores. Contradicting signals between different evaluation metrics for sign language translation has also been observed in prior work (Zhang et al., 2024). We chose to priotize BLEURT, as it generally has better alignment with human judgements of translation quality (Freitag et al., 2022), and therefore use random masking in all of our other experiments.

Data Scaling Behavior of SHuBERT In addition to the full pre-training dataset, we also train SHuBERT from scratch on a randomly selected 10% of the full data. In Tab. 5 we see that there is a noticeable drop in performance in BLEU and BLEURT, suggesting that data size is important. With multiple larger datasets now available—BOBSL (Albanie et al., 2021), JWSign (Gueuwou et al., 2023) and YouTube-SL-25 (Tanzer and Zhang, 2024) contain approximately 1500, 2500, and 3200 hours of data respectively—we expect that expanding the SHuBERT pre-training data may further improve performance. In addition, these larger datasets also include more language diversity, which may also improve performance and/or applicability to additional languages.

Isolating SHuBERT's Impact on Performance To understand SHuBERT's impact on translation performance, we conduct three experiments, shown in Tab. 6. In our baseline experiment ("None"), we directly feed the projected 4-channel features (face, left hand, right hand, body pose) to the ByT5 translation model, bypassing SHuBERT entirely, resulting in fairly poor performance. When we instead pass these features through a frozen pre-trained SHuBERT (400k steps, random masking) and use its final layer's output ("Last Layer"), we see significant improvement. Finally, computing a learned weighted combination of all SHuBERT layers ("Weighted Sum") further improves performance. These results demonstrate that each component of SHuBERT contribute to translation performance.

Frozen vs. Fine-Tuned SHuBERT We also investigate the impact of fine-tuning during translation training, as shown in Tab. 7. We compare two scenarios: using a frozen SHuBERT (400k steps, random masking) and only fine-tuning ByT5, versus fine-tuning both SHuBERT (from the same base model) and ByT5 together. Both scenarios use features from SHuBERT's last layer. Fine-tuning SHuBERT leads to substantial improvements compared to keeping it frozen, when using the final layer. However, referring back to Tab. 6, the relatively *small* difference between the fine-tuned performance and that of the frozen and layer-weighted SHuBERT is noteworthy. This observa-

Streams Used	BLEU ↑
Face only	0.6
Hands only	0.2
Upper body only	0.8
Hands + Upper Body	2.1
All streams	**2.4**

Table 8: Stream contributions in ASL-to-English translation.

Phonological Feature	Rec@1↑	
	Sem-Lex	ASL Citizen
Major Location	0.8477	0.9022
Minor Location	0.7130	0.8000
Second Minor Location	0.7328	0.8118
Contact	0.8684	0.9157
Thumb Contact	0.8474	0.8752
Sign Type	0.8464	0.9154
Repeated Movement	0.8265	0.8993
Path Movement	0.7275	0.7942
Wrist Twist	0.9058	0.9300
Selected Fingers	0.7953	0.8344
Thumb Position	0.8604	0.8819
Flexion	0.7264	0.7773
Spread	0.7942	0.8480
Spread Change	0.8160	0.8658
Nondominant Handshape	0.7632	0.8432
Handshape	0.6293	0.7080
Average	0.7938	0.8502

Table 9: Phonological feature recognition accuracy on two datasets, Sem-Lex and ASL Citizen.

tion is promising for low-resource sign languages, where we may have plentiful unlabeled video data for pre-training, but very limited parallel data for translation training.

Stream contributions in ASL-to-English translation To quantify the contribution of each input stream, we conduct a translation experiment with the How2Sign dataset (without pre-training), feeding the concatenation of the multiple streams directly to a language model (T5). The resulting BLEU scores are shown in Tab. 8, showing that all of the streams contribute to translation. These results should be interpreted as assessing the true relative importance of each stream, however, since this is a very small-scale experiment.

B Phonological Feature Recognition

We also conduct experiments on phonological feature recognition, that is the classification of linguistic features of signs, for two of the ISLR datasets. We report the Recall at 1 (prediction accuracy) for 16 commonly used phonological features (from ASL-LEX 2.0 (Sevcikova Sehyr et al., 2021)) in Tab. 9. The training setups and hyperparameters are identical to those of the full fine-tuning method in Sec. 4.3, except that we now train 16 classification heads simultaneously. We also find that removing weight decay gives a slight performance boost.

To the best of our knowledge, no previous work has reported phonological feature recognition accuracies on the ASL Citizen dataset. Similarly, though the Sem-Lex authors (Kezar et al., 2023) report phonological feature prediction accuracies, they are not comparable to ours, which are computed on the entire test set available to the public. Thus, we hope that our results in Tab. 9 can serve as a benchmark for future work.

C ASL Phonological Feature Classification Details

American Sign Language (ASL) can be described through a set of phonological features, similarly to the description of spoken languages via features. These features capture the essential components of sign formation, including hand configuration, movement patterns, and spatial relationships. Tab. 10 presents the number of classes for each of the phonological features we use in our phonological feature prediction analysis for Sem-lex and ASL-Citizen, and below we list the values of each feature. This commonly used feature set is from ASL-LEX 2.0 (Sevcikova Sehyr et al., 2021)).

Handshape v, 5, y, h, open_b, c, baby_o, flat_h, o, l, 1, a, open_8, w, curved_5, d, flatspread_5, i, f, s, p, flat_b, curved_4, flat_o, g, open_e, 4, closed_b, bent_1, 3, flat_horns, goody_goody, flat_m, bent_v, flat_1, r, 8, curved_v, open_h, curved_1, horns, flat_ily, flat_n, bent_l, stacked_5, ily, e, flat_v, curved_l, spread_open_e, curved_h, 7, closed_e, t, flat_4, open_f, k, and spread_e.

Nondominant Handshape v, 5, y, none, open_b, Dominance Condition Violation, B, 1, a, open_8, C, s, h, o, flat_b, curved_5, p, c, S, closed_b, 4, flat_m, bent_v, flat_1, flat_h, baby_o, curved_v, i, f, bent_1, Symmetry Violation, flatspread_5, flat_o, curved_1, open_h, stacked_5, g, l, bent_l, 3, 8, spread_open_e, e, horns, w, r, Lax, curved_l, open_e, flat_4, O, curved_b, A, ily, flat_v, and flat_horns.

Minor Location Neutral, Head Away, Body Away, Hand Away, Palm, Finger Tip, Forehead,

Phonological Feature	Number of Classes	
	Sem-Lex	ASL Citizen
Major Location	5	6
Minor Location	37	37
Second Minor Location	37	38
Contact	2	2
Thumb Contact	3	3
Sign Type	6	6
Repeated Movement	2	2
Path Movement	8	8
Wrist Twist	2	2
Selected Fingers	12	12
Thumb Position	2	2
Flexion	8	8
Spread	3	3
Spread Change	3	3
Nondominant Handshape	56	57
Handshape	58	58

Table 10: Number of classes for each phonological feature represented in two ASL datasets, Sem-Lex and ASL Citizen. Most features have the same number of classes across datasets, while a few features have values that don't appear in one of the datasets (for example, Second Minor Location has 37 classes that appear in Sem-Lex and 38 classes in ASL Citizen).

Finger Front, Mouth, Chin, Other, Upper Arm, Torso Top, Forearm Back, Cheek Nose, Wrist Front, Palm Back, Finger Back, Finger Radial, Under Chin, Finger Ulnar, Wrist Back, Shoulder, Arm Away, Forearm Ulnar, Torso Mid, Heel, Clavicle, Eye, Forearm Front, Neck, Torso Bottom, Upper Lip, Head Top, Elbow Back, Hips, and Waist.

Second Minor Location Neutral, Head Away, Torso Bottom, Finger Tip, Hand Away, none, Palm, Forearm Back, Finger Back, Body Away, Torso Top, Finger Front, Chin, Arm Away, Upper Arm, Finger Ulnar, Eye, Hips, Neck, Palm Back, Forearm Front, Finger Radial, Mouth, Heel, Torso Mid, Other, Waist, Cheek Nose, Forehead, Elbow Back, Under Chin, Clavicle, Shoulder, Forearm Ulnar, Head Top, Upper Lip, and Forearm Radial.

Sign Type Symmetrical Or Alternating, One Handed, Dominance Violation, Asymmetrical Different Handshape, Asymmetrical Same Handshape, and Symmetry Violation.

Path Movement Curved, Back And Forth, Straight, Circular, None, Z-shaped, Other, and X-shaped.

Flexion Fully Open, Curved, Bent, Flat, none, Fully Closed, Stacked, and Crossed.

Selected Fingers im, imrp, p, i, t, m, ip, imp, mr, imr, r, and mrp.

Major Location Neutral, Head, Body, Hand, and Arm.

Flexion Change 1.0, 0.0, and none.

Spread Change 1.0, 0.0, and none.

Thumb Contact 1.0, 0.0, and none.

Spread 1.0, 0.0, and none.

Thumb Position Closed and Open.

Repeated Movement 1.0 and 0.0.

Contact 1.0 and 0.0.

Wrist Twist 0.0 and 1.0.

D Cluster Samples

We visualize clustering results for the face, left hand, right hand, and upper body pose in Figs. 5 to 8. All cluster samples were randomly selected (i.e., without manual curation or cherry-picking). Each row represents a cluster. For each channel (Face, Left hand, Right hand, Upper Body), we include 10 random examples for 5 random clusters. While there is variability within each cluster, and some clusters contain a large mix of poses, we can also see a great deal of systematic behavior, where the images in a cluster tend to correspond to similar gestures regardless of signer appearance or other visual properties. The caption for each figure provides our interpretations of some of the clusters.

E Translation examples

We provide example translations produced by our model given inputs from three ASL datasets: How2Sign (instructional content, Tab. 11), FLEURS-ASL (zero-shot setting, Tab. 12), and OpenASL (general domain with native signers and varying background, Tab. 13).

(1) Reference	And that's a great vital point technique for women's self defense.
(Uthus et al., 2023)	It's really great for women's self defense.
(Rust et al., 2024)	This is a really great point for women's self defense.
Ours	If you're a bigger person we're talking about really self defense here.
(2) Reference	In this clip I'm going to show you how to tape your cables down.
(Uthus et al., 2023)	In this clip we're going to show you how to cut a piece of clay.
(Rust et al., 2024)	In this clip I'm going to show you how to clip the cable, the cable.
Ours	In this clip I'm going to show you how to brand out the cable strings.
(3) Reference	In this segment we're going to talk about how to load your still for distillation of lavender essential oil.
(Uthus et al., 2023)	In this clip we're going to talk about how to feed a set of baiting lizards for a lava field oil.
(Rust et al., 2024)	In this clip we're going to talk about how to feed the trail for draining clean for laborer oil.
Ours	In this clip we're going to talk about how to take our stick for disinfectant oil.
(4) Reference	You are dancing, and now you are going to need the veil and you are going to just grab the veil as far as possible.
(Uthus et al., 2023)	Their hopping and dancing is now, they're going to need their squat and squat and they're going to be able to move independently.
(Rust et al., 2024)	So that she's going to get her hips up as far as she can, and now she's going to lift her head up as far as possible.
Ours	Her dancing and now she needs her feather to grab it with her foot as far as possible.
(5) Reference	But if you have to setup a new campfire, there's two ways to do it in a very low impact; one is with a mound fire, which we should in the campfire segment earlier and the other way to setup a low impact campfire is to have a fire pan, which is just a steel pan like the top of a trash can.
(Uthus et al., 2023)	But if you have to set up a new campfire, there are two ways to do a low impact fire, one is a cone fire, which we have to do in the tent earlier, and the other one is to set up a campfire in a fire pan.
(Rust et al., 2024)	But if you have to set up a new campfire, this is one way to do it in a low impact. One is a monk fire. One is a campfire. The other one is to set a campfire in a campfire. That's just a post like the top of the post.
Ours	But if you have to set a new campfire, there are two ways to do a low impact one is a bond fire, which we should do in your campfire, another one is to set a campfire in a fire pan that is just just set a pan like the top of it pan.
(6) Reference	So, this is a very important part of the process.
(Uthus et al., 2023)	Alright, let's get started.
(Rust et al., 2024)	It's an important part of the process.
Ours	This is a very important part of the process.

Table 11: Qualitative translation examples from the How2Sign dataset, comparing SHuBERT-based translations to previous models.

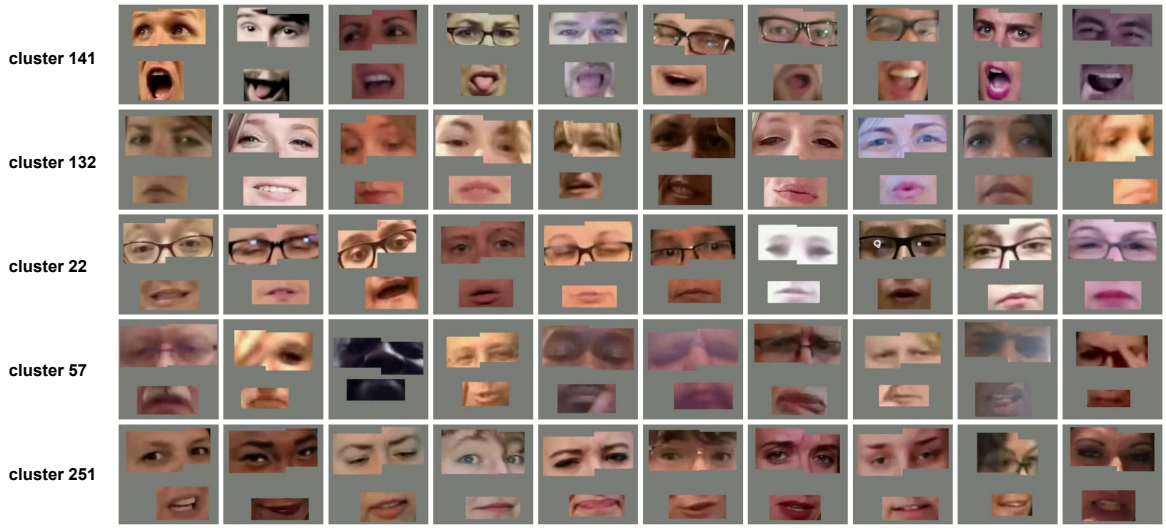

Figure 5: Sample face clusters. Each row represents a distinct cluster and 10 random examples from it. Cluster 141 includes mainly open-mouthed expressions with raised eyebrows, cluster 57 seems to capture closed or squinting eyes with neutral mouths, and cluster 251 corresponds to a slightly tilted head with direct gaze and little mouth opening. NOTE: For clarity, we show unblurred cropped faces here.

Figure 6: Sample left hand clusters. Each row represents a distinct cluster and 10 random examples from it. Cluster 125 shows pointing configurations with the index finger extended. Cluster 115 generally corresponds to closed fist formations oriented with the thumb on top. Cluster 51 seems to be a mix of poses without a consistent description. Cluster 98 seems to include mainly transitional hand movements around the chest.

Figure 7: Sample right hand clusters. Each row represents a distinct cluster and 10 random examples from it. Cluster 42 captures mainly multi-finger pointing gestures. Cluster 76 corresponds to clasped or overlapped hands in resting positions.

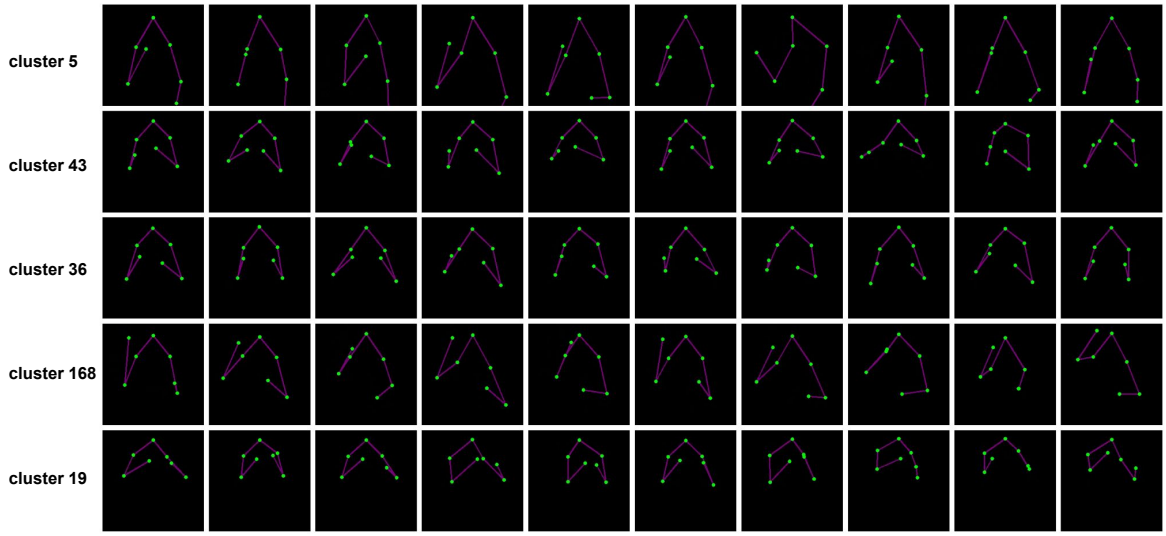

Figure 8: Sample upper body clusters. Each row represents a distinct cluster and 10 random examples from it. Cluster 5 seems to correspond to configurations of the upper body pose where the right hand is at shoulder level and the left hand is down. Cluster 43 seems to be a configuration where the two hands are raised and close to each other near the chest, and the signer is facing slightly to the right. This might correspond to signs being performed with both hands involved/active. Cluster 36 appears similar, but with the hands slightly farther apart. Cluster 19 is generally similar to cluster 43, except that the signer tends to be facing slightly to the left. Finally in cluster 168, the right hand is usually above the shoulder and close to the face/head.

(1)	Reference	During the 1980s he worked on shows such as Taxi, Cheers, and The Tracy Ullman Show.
	(Tanzer, 2024b)	In the 1980s, she worked in theaters like taxesi, cheesy, and tracy.
	Ours	In the 1980's, she worked for theaters like Taxi Chers, Tracy Ullman Shaw.
(2)	Reference	The rise of new technologies allows us to see and investigate brain structures and processes never seen before.
	(Tanzer, 2024b)	There is a new technique to detect brains and vision.
	Ours	Increasing new technology that allows people to consider investigating their brain structures and brain structures.
(3)	Reference	The Articles required unanimous consent from all the states before they could be amended and states took the central government so lightly that their representatives were often absent.
	(Tanzer, 2024b)	The law requires all states to agree on a standard and that it is a legal requirement.
	Ours	The state's agreement requires all standardized agreements to remove the standards of representatives from the state to represent the state's amendments.

Table 12: Qualitative translation examples from the FLEURS-ASL dataset, comparing SHuBERT-based translation (zero-shot) to a previous approach (Tanzer, 2024b).

(1)	Reference	thank you
	(Shi et al., 2022)	thank you
	Ours	thank you
(2)	Reference	come on
	(Shi et al., 2022)	come on
	Ours	maybe
(3)	Reference	now i've come this far and it 's a different team
	(Shi et al., 2022)	how do you feel about it
	Ours	it feels like a crash in the team
(4)	Reference	i was there from the beginning to the end and time went by fast
	(Shi et al., 2022)	the students were thrilled by this
	Ours	i just wanted to leave because it went ahead and started
(5)	Reference	i'm here at nad's 50th wow
	(Shi et al., 2022)	the nad has been <unk> for many years
	Ours	well the nad is shocked to have 50 years of dhh
(6)	Reference	i entered the yap 2018 competition and won
	(Shi et al., 2022)	the competition was started with ideas
	Ours	i enrolled in that competition in 2018 and then i won
(7)	Reference	you can check out their kickstarter in the link below
	(Shi et al., 2022)	you can watch the conversation at lake county
	Ours	you can check out their kickstarter link below
(8)	Reference	that is one thing i found interesting and wanted to share with you today
	(Shi et al., 2022)	i also am the president of the jr. nad conference here
	Ours	that's one interesting thing she wanted to share with you
(9)	Reference	those are the different types of bills
	(Shi et al., 2022)	schools have switched to teaching students
	Ours	i looked at several different types of interpreting services
(10)	Reference	dry january has picked up in popularity since it began in 2012
	(Shi et al., 2022)	krispy kreme is bringing back its original playstation in 2016
	Ours	the qury dry january started in 2012
(11)	Reference	we will be happy to respond give you support and listen to your concerns
	(Shi et al., 2022)	please review and submit your time passion and support this important issue
	Ours	the nad is willing to respond and support your concerns
(12)	Reference	there were videos posted on the internet that showed a person walking on the grass completely engulfed in flames
	(Shi et al., 2022)	a video shows the officer walking up to his shoulder and before he was shot
	Ours	videos posted on social media of him walking on a grass walking completely with fire
(13)	Reference	and people would become carpenters laborers mechanics plowers and farmers
	(Shi et al., 2022)	the next year 1880 the nad was established in the first operation 30 of the house in 2015
	Ours	and he was forced to wear a wearing a labover meganic and a financial warper
(14)	Reference	for nad youth programs related information please contact us via facebook at the nad youth programs or email us through
	(Shi et al., 2022)	you can contact us through our website where you can check our facebook page online at <unk>
	Ours	if you want to contact the nad youth program you can contact us through our facebook page at the nad youth program or youth program through our website
(15)	Reference	last week suspects gregory mcmichael and his son travis were arrested and charged with felony murder and aggravated assault
	(Shi et al., 2022)	last week a black man named <unk> <unk> was arrested and charged with felony murder and aggravated assault
	Ours	last week two suspects gregory mcmichael and his son travis were arrested and charged with felony murder and wounded by another gravated assaulter

Table 13: Qualitative translation examples from the OpenASL dataset, comparing SHuBERT-based translations to a previous model (Shi et al., 2022)

.