

# Cross-Lingual Transfer of Self-Supervised Representations for Sign Language Translation: A Case Study for Turkish Sign Language

Özge Can Özel<sup>1\*</sup>, Saliha Nur Şahin<sup>1\*</sup>, Murat Saraçlar<sup>1</sup>

<sup>1</sup>Boğaziçi University

ozge.ozel@std.bogazici.edu.tr saliha.sahin@std.bogazici.edu.tr murat.saracilar@bogazici.edu.tr

\*Equal contribution

## Abstract

Sign Language Translation (SLT) is especially difficult for low-resource languages such as Turkish Sign Language (TID), where aligned sign-text data are scarce. Self-supervised encoders pretrained on large sign corpora can reduce this dependence, but existing models are developed and evaluated almost exclusively on high-resource sign languages. One such model is SHuBERT (Sign Hidden-Unit BERT), a publicly released self-supervised contextual representation model trained on American Sign Language (ASL). We ask whether SHuBERT can be adapted to TID translation on the open-domain BUTID benchmark. Pretrained SHuBERT is incorporated into an encoder-decoder-based translation model and evaluated under both frozen and fine-tuned settings. This is compared against two references trained under the same recipe: a pose baseline, and a raw-feature setting using DINOv2-encoded hand/face crops and pose. All variants outperform the pose baseline, frozen SHuBERT performs on par with raw-feature setting, and fine-tuning yields the strongest result, reaching 7.15 BLEU-4 on the test set (+1.62 over the pose baseline, +0.77 over the raw-feature setting). Our results suggest that reusing a high-resource, ASL-pretrained self-supervised encoder can benefit low-resource SLT, but this benefit emerges through fine-tuning rather than frozen transfer alone.

## 1 Introduction

Sign languages convey meaning through complex visual patterns involving hand shapes, movements, facial expressions, and body posture. Sign Language Translation (SLT) remains challenging due to long-range spatio-temporal dependencies, and is especially difficult for low-resource and under-represented sign languages such as Turkish Sign Language (TID), where aligned sign-text data are scarce.

Self-supervised pretraining on large unlabeled sign corpora has emerged as a way to reduce annotation requirements (Rust et al., 2024; Gueuwou et al., 2025). In particular, SHuBERT (Gueuwou et al., 2025), a transformer-based contextual representation model pretrained on 1,000 hours of ASL data, attains state-of-the-art results on ASL translation.

We study cross-lingual transfer of SHuBERT’s ASL-pretrained representations to TID translation on the open-domain BUTID dataset (Şahin et al., 2025). We combine these representations across transformer layers via a learned weighted sum and train a linear projection together with an encoder-decoder-based translation model. The end-to-end translation model is evaluated under both a frozen-SHuBERT setting and a setting in which SHuBERT’s own weights are fine-tuned jointly with the rest of the model.

To isolate where the improvement comes from, we compare these two settings against two references trained under the same recipe: a pose baseline using MediaPipe (Lugaresi et al., 2019) landmarks, and a raw-feature setting using SHuBERT’s own four input streams (DINOv2-encoded (Oquab et al., 2024) hand/face crops and pose) without its transformer encoder.

Our contributions are:

- We evaluate cross-lingual transfer of SHuBERT, an ASL-pretrained self-supervised encoder, to TID translation.
- We compare frozen and fine-tuned configurations of pretrained SHuBERT against a raw-feature setting and a pose baseline.
- We analyze which SHuBERT transformer layers carry transferable information via their learned mixing weights.

## 2 Related Work

Early SLT systems relied on gloss annotations (Camgöz et al., 2018), but the high cost of labeling has motivated a shift toward gloss-free, end-to-end approaches (Zhang et al., 2024). These systems benefit from large open-domain corpora such as YouTube-ASL (Uthus et al., 2023), but remain largely restricted to high-resource languages.

Self-supervised pretraining has recently been adapted to sign language videos. One such approach is SHuBERT (Gueuwou et al., 2025), which extends HuBERT (Hsu et al., 2021) (originally developed for speech, where it learns representations by a masked prediction paradigm) to a multi-stream sign representation, achieving state-of-the-art results on several ASL benchmarks. Because SHuBERT inherits HuBERT’s transformer stack, it plausibly inherits a comparable layer-wise organization. In speech, layer-wise analyses of self-supervised models show that transformer layers encode qualitatively different information with depth, intermediate layers frequently carrying richer linguistic content than the final layers (Pasad et al., 2021). The standard downstream protocol aggregates all layers through learned weights rather than relying on any single one (Yang et al., 2021). We investigate whether analogous layer-specialization effects hold for SHuBERT when transferred to TID.

Prior TID resources focused largely on isolated sign recognition, the BosphorusSign corpus (Camgöz et al., 2016) and its curated BosphorusSign22k benchmark (Özdemir et al., 2020), alongside the large-scale AUTSL dataset (Sincan and Keles, 2020). Within TID, self-supervised features have so far been applied mainly to isolated recognition (Sero et al., 2025).

Continuous, sentence-level translation resources are comparatively recent, and work on TID translation remains limited. E-TSL (Öztürk and Keles, 2024) introduces a 24-hour educational-domain dataset with pose-based baselines, and BUTID (Şahin et al., 2025) provides the largest open-domain TID translation benchmark to date, which we adopt in this work. Even so, TID resources remain small relative to high-resource sign languages, motivating approaches that reuse representations learned elsewhere rather than training from scratch.

Cross-lingual transfer offers one such route: a translation model pretrained on a high-resource language benefits a low-resource one—for example, Alhejab et al. (2026) pretrain a T5 model on

ASL and fine-tune on Saudi Sign Language, reporting a roughly threefold BLEU-4 gain over training from scratch. Closest in goal, Mo et al. (2025) pretrain a multi-stream skeleton masked autoencoder on YouTube-ASL and transfer it to several unseen sign languages for recognition and translation. Transfer has also been studied within TID: Kindiroglu et al. (2024) evaluate supervised cross-dataset transfer between two Turkish isolated SLR datasets (BosphorusSign22k and AUTSL) using a temporal graph-convolutional model.

Our work differs from all of these in two respects: (i) we transfer a publicly released self-supervised encoder (SHuBERT) *cross-lingually*, from high-resource ASL; and (ii) we target continuous TID translation, a setting underrepresented in prior work, rather than isolated recognition.

## 3 Methods

### 3.1 Dataset

We use the BUTID dataset (Şahin et al., 2025), a large-scale open-domain Turkish Sign Language corpus of broadcast television content (drama, comedy, and historical series) from the “Engelsiz TRT” platform, spanning 27 signers. The official splits are partitioned at the clip level rather than the series, episode, or signer level. Different clips from the same series, and potentially the same episode or signer, can appear across train, dev, and test. We use its *multi-caption* subset of long-form, multi-sentence clips (on average 4 sentences per clip): 106,665 / 5,925 / 5,925 train / dev / test clips (287.7 / 16.1 / 16.1 hours), averaging  $\sim 9.7$  s and  $\sim 22$  text tokens per clip.<sup>1</sup>

### 3.2 SHuBERT-Based Translation Model

The translation pipeline is illustrated in Figure 1. We run a combined translation model, following the method used in the SHuBERT paper (Gueuwou et al., 2025) but substituting ByT5 (Xue et al., 2022) with a multilingual T5 model (mT5-small) (Xue et al., 2021). We extract four visual streams (left hand, right hand, face, upper-body pose), via MediaPipe (Lugaresi et al., 2019) landmarks and

<sup>1</sup>These split sizes and durations differ from those in the original BUTID paper (e.g., 5,717 / 5,712 dev / test clips); we use the version available at the time of writing. We release our exact train/dev/test splits as CSV files, with permission from the dataset creators, at [https://drive.google.com/drive/folders/1XOAC0CKR3sS9vKnoFIfIuW\\_pj0m\\_yrk0?usp=sharing](https://drive.google.com/drive/folders/1XOAC0CKR3sS9vKnoFIfIuW_pj0m_yrk0?usp=sharing).

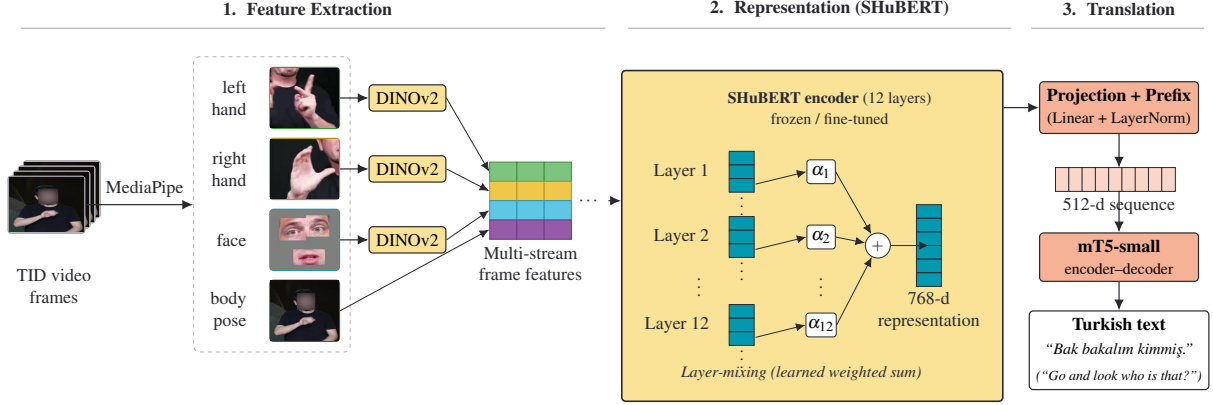


Figure 1: Overview of the SHuBERT-to-mT5 TID translation pipeline: four visual streams are encoded by SHuBERT, its per-layer 768-d states are combined by a learned weighted sum, projected into mT5-small, and decoded to Turkish.

fine-tuned DINOv2 models (Oquab et al., 2024)<sup>2</sup>, for every BUTID clip. The four streams are then projected and concatenated to form a feature vector for each frame. The resulting sequence passes through SHuBERT’s 12-layer transformer encoder (768-dimensional hidden states, 86M parameters), initialized from its public ASL-pretrained checkpoint<sup>2</sup>. We combine its per-layer hidden states with a learned weighted sum (Yang et al., 2021) rather than using a single layer’s output. The result is projected through one linear layer into mT5-small’s embedding space and prepended with an embedded task-prefix, then passed through mT5-small as a standard encoder-decoder, trained with token-level cross-entropy and a label smoothing factor of 0.2.

## 4 Experiments

We evaluate this pipeline under four conditions:

1. **Frozen SHuBERT:** SHuBERT is kept frozen; only the weighted-sum, projection, and mT5-small are trained.
2. **Fine-tuned SHuBERT:** SHuBERT is fine-tuned jointly with the rest of the model, warm-started from the converged frozen checkpoint in (1).
3. **Raw-feature:** To isolate SHuBERT’s impact on translation performance, SHuBERT is removed entirely. The four raw streams are

concatenated and projected directly into mT5-small’s embedding space by a single linear layer. Through this, we isolate how much of the translation signal comes from SHuBERT’s own representation as opposed to the raw visual features it is computed from.

4. **Pose baseline:** Pose-based baseline, trained under the same conditions as our other experiments, uses raw MediaPipe landmarks (8 body, 21 per hand, and 35 face points per frame). The landmarks are normalized relative to the hands and face, rescaled using a shared body-derived scale, flattened, and projected directly into mT5-small’s embedding space through a single linear layer.

### 4.1 Training Configuration

All experiments were conducted on a single NVIDIA RTX 4090 GPU. Models are optimized with Adafactor (Shazeer and Stern, 2018), weight decay 0.01, a linear learning-rate schedule, batch size 4 with gradient accumulation over 64 steps (effective batch size 256), and the same label smoothing (0.2) throughout. Early stopping is performed based on validation performance. For all experiments except SHuBERT fine-tuning, we use a peak learning rate of  $10^{-3}$  with a 2.4-epoch warmup and train for up to 100 epochs.

For fine-tuning, we use a discriminative learning rate: SHuBERT’s own parameters receive a peak learning rate of  $5 \times 10^{-5}$  in a separate optimizer parameter group from the projection, layer-mixing, and mT5-small parameters. Rather than opening SHuBERT’s weights from a random or

<sup>2</sup>We use the pretrained SHuBERT encoder and the face or hand fine-tuned DINOv2 feature extractors released in the official SHuBERT repository: <https://github.com/ShesterG/SHuBERT>.

purely frozen starting point, we warm-start fine-tuning from the best checkpoint of the corresponding frozen run, and open SHuBERT’s encoder and layer-mixing weights simultaneously for continued training, with a shorter, 1-epoch warmup and a peak rate of  $10^{-4}$  for the non-SHuBERT parameters.

Appendix A reports, for every system in Tables 1–3, the number of epochs run to the reported (best-dev) checkpoint and the corresponding optimizer steps.

## 4.2 Evaluation Metrics

To assess the quality of sign language translations, we employ standard evaluation metrics: BLEU (Papineni et al., 2002), ROUGE-L (Lin, 2004) and BERTScore (Zhang et al., 2020).

We report BLEU-1 through BLEU-4 to show how performance changes with the length of the matched  $n$ -gram. For a candidate translation and reference translation, BLEU combines  $n$ -gram precision with a brevity penalty.<sup>3</sup>

We additionally report ROUGE-L, which measures the longest common subsequence (LCS) between a hypothesis and its reference and converts the resulting precision and recall into an F-measure (Lin, 2004).<sup>4</sup>

Finally, we compute BERTScore (Zhang et al., 2020), which matches the most similar token in the other sentence using cosine similarity and aggregates these matches into precision, recall, and  $F_1$ .<sup>5</sup>

Several of the BLEU-4 margins we report are under 1 point, we therefore test the pairwise comparisons discussed in Section 5 for significance using paired bootstrap resampling (2000 resamples per test) on BLEU-4, via SacreBLEU’s --paired-bs implementation (Post, 2018; Koehn, 2004).

## 4.3 Decoding Configuration

At inference, we decode with beam search (4 beams) and a length penalty of 1.0, blocking repetition of any 3-gram, and allow up to 128 new tokens with early stopping at the end-of-sequence

token. We use the same decoding configuration for all systems, on both the development and test sets.

## 5 Results and Discussion

Table 1 reports results at each configuration’s best checkpoint by dev BLEU-4. Both ASL-pretrained SHuBERT-based variants outperform pose baseline on every metric. Within the SHuBERT-based variants, the frozen weighted-sum model performs comparably to the raw-feature setting. Fine-tuning SHuBERT’s own weights yields the largest gain observed, surpassing both the frozen and raw-feature configurations, with consistent trends between dev and test.

Paired bootstrap resampling on the test set confirms that fine-tuned SHuBERT significantly outperforms both the raw-feature and frozen configurations ( $p < 0.001$  for both), while the difference between frozen SHuBERT and the raw-feature setting is not significant ( $p = 0.354$ ). The pose baseline is also significantly below the raw-feature, frozen, and fine-tuned SHuBERT configurations ( $p < 0.001$  for all three comparisons).

The absence of a significant difference between the frozen SHuBERT model and the raw-feature setting suggests that the pretrained representations may not transfer optimally to TID without adaptation.

One plausible explanation is that SHuBERT’s pretrained encoder compresses its input toward representations useful for ASL. Under frozen transfer to a typologically distant language such as TID, this compression may discard information relevant to the target language, offsetting the benefit of useful contextual structure. This would also explain why fine-tuning, which allows SHuBERT’s representations to adapt away from ASL-specific structure, recovers a clear advantage over both frozen configurations.

A translation baseline for the BUTID multi-caption subset is reported in the official dataset paper (Şahin et al., 2025): BLEU-1 37.61, BLEU-4 2.42, ROUGE-L 22.16, and BERTScore 50.96 on the test set. The two setups differ in several ways: our models use a single linear projection rather than a two-layer MLP sign encoder, the training configurations differ, and, most importantly, the evaluation metrics are computed using different procedures. We therefore treat the reported BUTID scores as context rather than a directly comparable baseline, and base our conclusions on the systems

<sup>3</sup>The implementation is based on SacreBLEU (Post, 2018) with signature BLEU|nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:1.4.2.

<sup>4</sup>ROUGE-L is computed using the Python rouge package: <https://github.com/pltrdy/rouge>.

<sup>5</sup>Computed with Hugging Face evaluate using dbmdz/bert-base-turkish-cased: <https://huggingface.co/dbmdz/bert-base-turkish-cased>. BERTScore is computed without baseline rescaling.

| Method                    | Development Set |              |              |             |              |              | Test Set     |              |              |             |              |              |
|---------------------------|-----------------|--------------|--------------|-------------|--------------|--------------|--------------|--------------|--------------|-------------|--------------|--------------|
|                           | B-1             | B-2          | B-3          | B-4         | R-L          | BERT         | B-1          | B-2          | B-3          | B-4         | R-L          | BERT         |
| Pose Baseline             | 28.85           | 15.30        | 9.13         | 5.75        | 17.25        | 58.81        | 28.73        | 15.11        | 8.94         | 5.53        | 16.95        | 58.63        |
| DINOv2+Pose (raw-feature) | 30.60           | 16.76        | 10.26        | 6.57        | 18.83        | 59.39        | 30.42        | 16.41        | 9.97         | 6.38        | 18.28        | 59.18        |
| Frozen SHuBERT            | 29.87           | 16.40        | 10.10        | 6.53        | 18.51        | 59.38        | 29.95        | 16.30        | 9.98         | 6.42        | 18.60        | 59.34        |
| Fine-tuned SHuBERT        | <b>31.71</b>    | <b>17.89</b> | <b>11.21</b> | <b>7.35</b> | <b>20.00</b> | <b>60.45</b> | <b>31.57</b> | <b>17.68</b> | <b>10.99</b> | <b>7.15</b> | <b>19.81</b> | <b>60.26</b> |

Table 1: SLT results on the BUTID development and test sets (Multi-Caption split). B-1 to B-4 denote BLEU-1 to BLEU-4, R-L is ROUGE-L, and BERT is the mean F1 of BERTScore.

| Init.      | Adapt.     | B-1          | B-4         | R-L          | BERT         |
|------------|------------|--------------|-------------|--------------|--------------|
| Random     | Frozen     | 25.87        | 3.81        | 12.62        | 54.93        |
| Random     | Fine-tuned | 26.93        | 4.40        | 14.04        | 55.92        |
| Pretrained | Frozen     | 29.95        | 6.42        | 18.60        | 59.34        |
| Pretrained | Fine-tuned | <b>31.57</b> | <b>7.15</b> | <b>19.81</b> | <b>60.26</b> |

Table 2: Impact of ASL pretraining on SHuBERT-based SLT (BUTID test set; checkpoints selected on dev BLEU-4, dev results follow the same trend). All variants share the same architecture; only the encoder initialization (*random* vs. pretrained SHuBERT) and adaptation differ.

we evaluate under a single fixed configuration.

### 5.1 Impact of ASL pretraining

In order to determine whether SHuBERT’s gains arise from ASL pretraining itself or merely from the added encoder capacity, we evaluate two additional variants. Instead of beginning with the ASL-pretrained SHuBERT encoder, weights are randomly initialized. The same warm-started fine-tuning protocol is applied to the randomly-initialized variant. All experiments share the same architecture and parameter count (Section 4.1).<sup>6</sup> Results are reported in Table 2.

In the frozen setting, ASL pretraining raises BLEU-1 from 25.87 to 29.95 and ROUGE-L from 12.62 to 18.60 without updating any encoder parameters, indicating that the pretrained features themselves encode transferable signing structure. All pairwise differences in Table 2 are statistically significant ( $p < 0.001$ , paired bootstrap resampling). This gap persists after fine-tuning: the ASL-pretrained encoder leads by 4.6 BLEU-1 and 5.8 ROUGE-L over its randomly initialized counterpart, and the frozen ASL-pretrained encoder alone

<sup>6</sup>We also attempted training both the pretrained and the randomly initialized encoder jointly with mT5-small end-to-end under a standard  $10^{-3}$  schedule, without warm-starting from a frozen checkpoint; in both cases the training resulted in models with poor performance (BLEU-4 remained below 1.0), motivating our staged protocol.

already surpasses the fully fine-tuned random one. While suggestive of a genuine pretraining benefit beyond added capacity, this comparison is confounded by the training protocol, and we cannot fully rule out that the randomly initialized model’s fine-tuned performance is limited by optimization difficulty rather than the absence of ASL pretraining.

### 5.2 Layer Importance Analysis

The layer-weighted-sum mechanism learns a softmax-normalized importance weight for each SHuBERT layer. We inspect these weights at the best checkpoint of all four weighted-sum configurations. Figures 2 and 3 compare the learned layer-weight distributions for randomly initialized and ASL-pretrained SHuBERT in the frozen and fine-tuned settings, respectively.

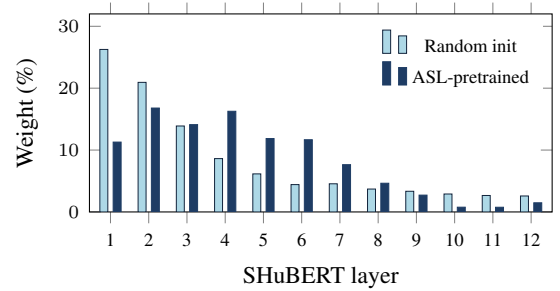


Figure 2: Layer-mixing weights, frozen setting: random initialization vs. ASL-pretrained SHuBERT.

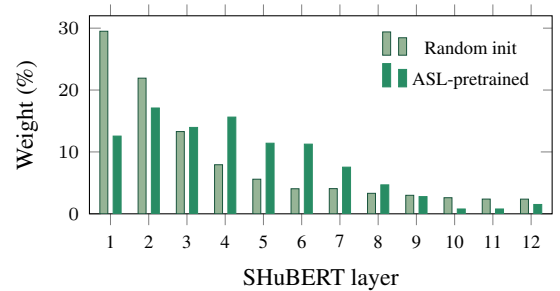


Figure 3: Layer-mixing weights, fine-tuned setting: random initialization vs. ASL-pretrained SHuBERT.

| Method             | B-1   | B-4  | R-L   | BERT  |
|--------------------|-------|------|-------|-------|
| Frozen (12 layers) | 29.95 | 6.42 | 18.60 | 59.34 |
| Frozen (Layer 4)   | 29.38 | 6.12 | 17.69 | 58.72 |

Table 3: Comparison of frozen SHuBERT using all layers and Layer 4 only (BUTID test set; checkpoints selected on dev BLEU-4, dev results follow the same trend).

ASL pretraining changes which layers the model relies on. Without pretraining, the weights concentrate sharply on the earliest layers (layers 1–2 alone carry 47–51% of the mass) and decrease steadily with depth. With ASL pretraining, the distribution is flatter: the model draws on the middle layers (roughly 1–5) fairly evenly before tailing off after layer 6.

The distribution remains nearly unchanged after fine-tuning for both settings, suggesting the frozen model already identifies the transferable layers, with fine-tuning refining representations rather than reweighting them. We note, however, that our fine-tuning runs are warm-started from a converged frozen checkpoint and trained for relatively few additional epochs (Section 4.1), so we cannot rule out that longer or non-warm-started fine-tuning would eventually shift the mixing weights further.

This concentration of weight in the early-to-intermediate layers raises the question of whether a single highly weighted layer can capture most of the benefit of the full weighted-sum representation. To examine this, we additionally evaluate a frozen model using only Layer 4, which individually receives one of the highest mixing weights ( $\sim 16\%$ ) in the frozen and pretrained SHuBERT setting. For this experiment, the projection and the decoder are retrained from scratch under the same frozen protocol as the full weighted-sum model; only the layer-mixing is replaced by a fixed selection of Layer 4. As seen in Table 3, the Layer 4-only model achieves performance comparable to the full 12-layer frozen representation: the difference is statistically significant ( $p = 0.004$ , paired bootstrap resampling) but the absolute margin is small (0.30 BLEU-4). Isolating Layer 4 recovers most of the weighted-sum’s benefit, indicating that the transferable signal is concentrated in a few layers rather than spread across the encoder.

### 5.3 Length-Stratified Analysis

Aggregate BLEU averages over utterances of very different lengths, which can mask systematic differ-

ences. To examine this, we stratify the development set clips into five bins by reference word count and compute SacreBLEU corpus-level BLEU-4 separately for each bin (Table 4).

| Length | # Sent. | $n$  | Pose  | Raw   | Frozen | FT           |
|--------|---------|------|-------|-------|--------|--------------|
| 2–8    | 2.23    | 1185 | 11.15 | 11.32 | 10.74  | <b>12.24</b> |
| 8–12   | 2.58    | 1184 | 7.31  | 8.00  | 7.47   | <b>8.46</b>  |
| 12–17  | 3.29    | 1185 | 6.61  | 7.19  | 7.27   | <b>7.77</b>  |
| 17–26  | 4.63    | 1184 | 5.44  | 6.10  | 6.42   | <b>7.10</b>  |
| 26–71  | 7.44    | 1185 | 4.30  | 5.64  | 5.50   | <b>6.27</b>  |

Table 4: Dev BLEU-4 by reference length, using five approximately equal-sized bins. # Sent.: average number of sentences per clip;  $n$ : references per bin; Raw: raw-feature; FT: fine-tuned SHuBERT.

Fine-tuned SHuBERT outperforms the pose baseline in every bin. Its margin remains similar across the first three bins ( $+1.09$ – $+1.16$  BLEU-4) and increases for longer clips, reaching  $+1.97$  in the longest bin. The raw-feature and frozen settings also outperform the pose baseline in most bins, although the frozen model trails pose in the shortest bin. BLEU-4 declines with length for every system, consistent with the greater difficulty of translating longer, multi-sentence utterances. In summary, the results suggest that fine-tuned SHuBERT becomes advantageous over the pose baseline as utterance length grows.

### 5.4 Qualitative Analysis

Table 5 shows four illustrative test-set examples comparing the fine-tuned SHuBERT model against the pose baseline.

In Example 1, SHuBERT reproduces the reference almost word-for-word while the pose baseline shifts tense and drops a clause.

In Example 2, SHuBERT matches the reference exactly, whereas the pose baseline loses the temporal marker “*bugün*” (today) and flattens the follow-up question into a statement.

In Example 3, both models correctly preserve the proper noun “Elizabeth” but generate an incorrect name for the addressee. However, SHuBERT accurately reproduces both the capture event and the successful completion of the operation, whereas the pose baseline omits the capture event. The higher sentence-level BLEU-4 (40.46 vs. 8.18) is consistent with the stronger performance of fine-tuned SHuBERT on longer, multi-clause examples discussed in Section 5.3.

In Example 4, the pose baseline reproduces the

reference’s opening exchange verbatim, including the numeral “*elli kere*” (fifty times), but neither system renders the idiom “*gözleri dolu dolu bakıyor*” (eyes welling up), each instead substituting an unrelated closing clause, suggesting this kind of non-literal, idiomatic expression remains difficult for both architectures.

## 6 Conclusion

We studied cross-lingual transfer of SHuBERT’s ASL-pretrained representations to low-resource, underrepresented Turkish Sign Language translation on BUTID. ASL-pretrained SHuBERT consistently outperforms random initialization, indicating that its pretrained representations contain transferable information. However, frozen SHuBERT performs similarly to the raw-feature setting, while fine-tuning the SHuBERT encoder yields the strongest results, reaching 7.15 BLEU-4 on the test set.

Our analyses further show that the transferable information is concentrated primarily in early-to-intermediate SHuBERT transformer layers, with a single highly weighted intermediate layer recovering most of the performance of the full weighted-sum representation. Fine-tuned SHuBERT also maintains an advantage over the pose baseline across utterance lengths, with a larger margin on longer inputs. Together, these results suggest that reusing a high-resource, ASL-pretrained self-supervised encoder can benefit low-resource sign language translation, though this benefit emerges through fine-tuning rather than frozen transfer alone.

## Limitations

Our study is limited to a single target sign language and dataset (TID/BUTID), so it remains unclear whether the observed transfer patterns generalize to other low-resource sign languages or domains. Computational constraints also limit our translation model to mT5-small and restrict the extent of SHuBERT fine-tuning; larger translation models may interact differently with the transferred representations. We rely on automatic metrics, which under-reflect semantic adequacy for morphologically rich, word-order-flexible Turkish and which we do not complement with human evaluation. Despite the improvements, absolute translation quality remains low on this open-domain benchmark for all systems.

All experiments are run with a single random seed; we do not report variance across training seeds, though we assess the significance of our main pairwise comparisons via bootstrap resampling on the test set.

We also do not evaluate transfer under more severely data-scarce conditions; such an experiment would more directly test whether SHuBERT’s benefit grows as target-language supervision becomes scarcer, and we leave this to future work.

## Acknowledgments

This work was supported in part by the Turkish Directorate of Strategy and Budget through the ROYAL Project under Grant 2019K12-149250. We also thank Prof. Dr. Lale Akarun and Karahan Şahin for their support and guidance, and the anonymous reviewers for their constructive feedback.

## References

- Ali Alhejab, Tomáš Železný, Lamya Alkanhal, Ivan Gruber, Yazeed Alharbi, Jakub Straka, Václav Javorek, Marek Hruš, Badriah Alkalifah, and Ahmed Ali. 2026. [Saudi sign language translation using T5](#). In *Speech and Computer*, pages 331–343, Cham. Springer Nature Switzerland.
- Necati Cihan Camgöz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. 2018. [Neural sign language translation](#). In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7784–7793.
- Necati Cihan Camgöz, Ahmet Alp Kindiroğlu, Serpil Karabüklü, Meltem Keleşir, A. Sumru Özsoy, and Lale Akarun. 2016. [BosphorusSign: A Turkish Sign Language recognition corpus in health and finance domains](#). In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)*, pages 1383–1388, Portorož, Slovenia. European Language Resources Association (ELRA).
- Shester Gueuwou, Xiaodan Du, Greg Shakhnarovich, Karen Livescu, and Alexander H. Liu. 2025. [SHuBERT: Self-supervised sign language representation learning via multi-stream cluster prediction](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 28792–28810, Vienna, Austria. Association for Computational Linguistics.
- 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/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460.

| Example | System             | Output (Turkish)                                                                                                                                 | English gloss                                                                                                                                             |
|---------|--------------------|--------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------|
| 1       | Reference          | Seni bir daha hayal kırıklığına uğratmayacağım. En azından kendi adıma söz veriyorum. Ben elimden geleni yapacağım.                              | I won't let you down again. At least I promise on my own behalf. I will do everything I can.                                                              |
|         | Fine-tuned SHuBERT | Seni bir daha hayal kırıklığına uğratmayacağım. En azından benim adıma söz veriyorum. Elimden geleni yapacağım.                                  | I won't let you down again. At least I promise on my behalf. I will do everything I can.                                                                  |
|         | Pose base-line     | Seni hayal kırıklığına uğratamam. En azından ben kendim adıma söz verdim. Elimden gelen her şeyi yaptım.                                         | I can't let you down. At least I promised on my own behalf. I did everything I could.                                                                     |
| 2       | Reference          | Bugün çok garip bir şey oldu biliyor musun? Ne oldu?                                                                                             | Do you know something really strange happened today? What happened?                                                                                       |
|         | Fine-tuned SHuBERT | Bugün çok garip bir şey oldu biliyor musun? Ne oldu?                                                                                             | Do you know something really strange happened today? What happened?                                                                                       |
|         | Pose base-line     | Bir şey mi oldu biliyor musun? Oldu.                                                                                                             | Do you know if something happened? It did.                                                                                                                |
| 3       | Reference          | Komutanım, Elizabeth'i ele geçirdik. Operasyon başarıyla tamamlandı. Bu harika bir haber Emre. Peki konuştu mu? Şimdilik hayır ama konuşacaktır. | Commander, we've captured Elizabeth. The operation was completed successfully. That's great news, Emre. Did she talk? Not yet, but she will.              |
|         | Fine-tuned SHuBERT | Komutanım, Elizabeth'i ele geçirdik. Operasyon başarıyla tamamlandı. Harika haber Elif. Bunu mu söylüyorsun? Burada değil ama söyleyeceğim.      | Commander, we've captured Elizabeth. The operation was completed successfully. Great news, Elif. Is that what you're saying? Not here, but I'll tell you. |
|         | Pose base-line     | Komutanım, Elizabeth'in operasyonu başarıyla tamamlandı. Harika haber Lilith. Sen ne dediğinin farkında mısın? Hayır, konuşmayı tercih ederim.   | Commander, Elizabeth's operation was completed successfully. Great news, Lilith. Are you aware of what you're saying? No, I'd rather talk.                |
| 4       | Reference          | Bir şey mi oldu? Bilmiyorum abla. Elli kere sordum. Bir tuhaf, sadece gözleri dolu dolu bakıyor.                                                 | Did something happen? I don't know, sis. I've asked fifty times. It's strange, she just stares with teary eyes.                                           |
|         | Fine-tuned SHuBERT | Bir şey mi oldu? Bilmiyorum abla en başından beri soruyorum. Bozma moralini. Sanki bütün gün susuyorum.                                          | Did something happen? I don't know, sis, I've been asking since the very beginning. Don't get upset. It's like I've been quiet all day.                   |
|         | Pose base-line     | Bir şey mi oldu? Bilmiyorum abla. Elli kere sordum. Sanki biraz tatlı kokuyor.                                                                   | Did something happen? I don't know, sis. I've asked fifty times. It's like it smells a little sweet.                                                      |

Table 5: Qualitative examples from the BUTID test set, comparing the fine-tuned SHuBERT model against the pose baseline. English glosses are our own translations, provided for readability.

- Ahmet Alp Kindiroglu, Ozgur Kara, Ogulcan Ozdemir, and Lale Akarun. 2024. [Transfer learning for cross-dataset isolated sign language recognition in under-resourced datasets](#). In *2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG)*, pages 1–8.
- Philipp Koehn. 2004. [Statistical significance tests for machine translation evaluation](#). In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.
- Chin-Yew Lin. 2004. [ROUGE: A package for automatic evaluation of summaries](#). In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.
- Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg, and Matthias Grundmann. 2019. [MediaPipe: A framework for perceiving and processing reality](#). In *Third Workshop on Computer Vision for AR/VR at IEEE Computer Vision and Pattern Recognition (CVPR) 2019*.
- Junwen Mo, MinhDuc Vo, and Hideki Nakayama. 2025. [Improving sign language understanding with a multi-stream masked autoencoder trained on ASL videos](#). In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 1202–1218, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.
- Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, and 7 others. 2024. [DINOv2: Learning robust visual features without supervision](#). *Trans-*

- actions on Machine Learning Research. Featured Certification.
- Oğulcan Özdemir, Ahmet Alp Kındıroğlu, Necati Cihan Camgöz, and Lale Akarun. 2020. [Bosphorus-Sign22k sign language recognition dataset](#). In *Proceedings of the LREC2020 9th Workshop on the Representation and Processing of Sign Languages: Sign Language Resources in the Service of the Language Community, Technological Challenges and Application Perspectives*, pages 181–188, Marseille, France. European Language Resources Association (ELRA).
- Şükrü Öztürk and Hacer Yalim Keles. 2024. [E-TSL: A continuous educational Turkish sign language dataset with baseline methods](#). In *2024 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA)*, pages 1–7.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. [Bleu: a method for automatic evaluation of machine translation](#). In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
- Ankita Pasad, Ju-Chieh Chou, and Karen Livescu. 2021. [Layer-wise analysis of a self-supervised speech representation model](#). In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 914–921.
- Matt Post. 2018. [A call for clarity in reporting BLEU scores](#). In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.
- Phillip Rust, Bowen Shi, Skyler Wang, Necati Cihan Camgöz, and Jean Maillard. 2024. [Towards privacy-aware sign language translation at scale](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8624–8641, Bangkok, Thailand. Association for Computational Linguistics.
- Karahan Şahin, Lale Akarun, Kadir Gökgöz, and Murat Saraçlar. 2025. [BUTID: A large-scale sign language translation dataset and benchmarks for Turkish sign language](#). Presented at the 3rd Workshop on Sign Language Recognition, Translation and Production (SLRTP), CVPR 2025. Remote paper.
- Onur Şero, Oğulcan Özdemir, and Lale Akarun. 2025. [Turkish sign language recognition through self-supervised learning features and heatmap](#). In *2025 33rd Signal Processing and Communications Applications Conference (SIU)*, pages 1–4.
- Noam Shazeer and Mitchell Stern. 2018. [Adafactor: Adaptive learning rates with sublinear memory cost](#). In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 4596–4604. PMLR.
- Özge Mercanoglu Sincan and Hacer Yalim Keles. 2020. [AUTSL: A large scale multi-modal Turkish sign language dataset and baseline methods](#). *IEEE Access*, 8:181340–181355.
- Dave Uthus, Garrett Tanzer, and Manfred Georg. 2023. [YouTube-ASL: A large-scale, open-domain American sign language-English parallel corpus](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 29029–29047. Curran Associates, Inc.
- 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](#). *Transactions of the Association for Computational Linguistics*, 10:291–306.
- 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 *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.
- 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, Tzu-Hsien Huang, Wei-Cheng Tseng, Ko tik Lee, Da-Rong Liu, Zili Huang, Shuyan Dong, Shang-Wen Li, Shinji Watanabe, Abdelrahman Mohamed, and Hung yi Lee. 2021. [SUPERB: Speech Processing Universal PERFORMANCE Benchmark](#). In *Inter-speech 2021*, pages 1194–1198.
- Biao Zhang, Garrett Tanzer, and Orhan Firat. 2024. [Scaling sign language translation](#). In *Advances in Neural Information Processing Systems*, volume 37, pages 114018–114047. Curran Associates, Inc.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. [BERTScore: Evaluating text generation with BERT](#). In *International Conference on Learning Representations (ICLR)*.

## A Training Epochs and Compute Budget

All experiments use a single fixed random seed (42). With batch size 4 and 64-step gradient accumulation, one optimizer step aggregates 256 clips, so each epoch over the 106,665-clip train split (Section 3.1) corresponds to  $\approx 417$  optimizer steps.

Table 6 reports, for every system in Tables 1–3, the number of training epochs completed up to the checkpoint selected by best dev BLEU-4 (Section 4.2), and the corresponding number of optimizer steps (each aggregating 64 microbatches of batch size 4, per Section 4.1). The frozen and

fine-tuned SHuBERT variants (and their random-init counterparts) are two-stage: fine-tuning warm-starts from the corresponding frozen run’s own best checkpoint (Section 4.1), so we report the fine-tuning phase’s epochs/steps as an addition on top of its frozen phase’s own cost.

| System          | Epochs | Steps  |
|-----------------|--------|--------|
| Pose Baseline   | 66     | 27,500 |
| Raw-feature     | 33     | 13,750 |
| Frozen SHuBERT  | 38     | 15,833 |
| + fine-tuning   | +8     | +3,333 |
| Frozen, layer 4 | 51     | 21,250 |
| Random/Frozen   | 40     | 16,666 |
| + fine-tuning   | +8     | +3,333 |

Table 6: Epochs and optimizer steps to the reported (best-dev) checkpoint, per system. Rows marked “+ fine-tuning” are additional cost on top of the frozen row directly above them.