BUTID: A Large-scale Sign Language Translation Dataset and Benchmarks for Turkish Sign Language

Karahan Şahin
CVSSP
University of Surrey, Guilford, UK
ks0085@surrey.ac.uk

Lale Akarun
Computer Engineering Department
Bogazici University, Istanbul, Turkey
akarun@boun.edu.tr

Kadir Gökgöz
Department of Linguistics
Bogazici University, Istanbul, Turkey
kadir.gokgoz@boun.edu.tr

Murat Saraçlar
Electrical and Electronic Engineering Department
Bogazici University, Istanbul, Turkey
murat.saraclar@boun.edu.tr

Abstract

With recent advances in deep learning, Sign Language Translation (SLT) technologies have shifted toward large-scale training with pretrained large language models (LLMs). However, most language models and datasets are provided for American Sign Language (ASL) and British Sign Language (BSL). To achieve similar advances for Turkish Sign Language (Türk İşaret Dili - TID) translation, we present the first large-scale open-source translation dataset, comprising over 500 hours of video footage and aligned Turkish translations by utilizing publicly available YouTube content. In this work, our dataset provides a scalable and reproducible framework for SLT. We provide SLT benchmarks using baseline models based on pretrained LLMs as well as the additional improvements with augmentations on text and sign modalities.

1. Introduction

Sign Language Translation (SLT) is a multi-disciplinary field, spanning various research domains, including Computer Vision (CV), Natural Language Processing (NLP) [4, 40, 60], and Linguistics [18, 24]. The core objective of SLT is to translate sign language input into spoken or written text. This task has been studied for over two decades to develop systems, aiming to overcome the communication barrier between sign language users and those who rely on spoken languages.

However, this is not an easy challenge due to the complexity of sign language. Unlike spoken languages, sign languages are multi-cue, meaning there are simultaneous articulators that include hand shapes, facial expressions, body movements to convey different linguistic information [5, 50]. Another significant challenge in sign language translation is the scarcity of large-scale datasets. Compared to spoken languages, which have shown significant improvements with recent developments in NLP [16, 52], based on vast amounts of text and speech data available, datasets in sign language have been limited to less than 100 hours until recently [10, 12, 56].

Figure 1. The BUTID dataset

Nevertheless, SLT research has improved significantly over the past two decades. Starting with the earlier works on Sign Language Recognition (SLR) [6, 49] and Continuous Sign Language Recognition (CSLR) [17, 27, 59] systems, they are utilized in the prediction of glosses to extract the intermediate discrete representation of a continuous sequence. With the advancements in deep learning [3, 55], the literature on SLT has shifted towards the end-to-end translation works [9, 11, 26]. These techniques have evolved the usage of the pre-trained transformer-based architectures [14, 61]. These techniques have shown that leveraging a pretrained model in textual domains results in significant improvements over the translation capabilities over sign language data.

Despite these advancements, SLT still has several persistent challenges. As we have discussed, one major limitation is the scarcity of large, diverse datasets necessary for training machine learning models. However, this challenge has been tackled, and exponential improvements have been obtained within the last 3-4 years with the development of open-access data collection from YouTube or by archives of TV channels [2, 28, 46, 51, 54]. Yet sign languages vary widely across different regions and communities, and collecting comprehensive datasets that capture this diversity is resource-intensive, and researchers working on less-studied sign languages may lack access to large-scale resources similar to the aforementioned studies.

This is the case in Turkish Sign Language (TID) as well. For TID, the first continuous sign language translation dataset was introduced in late 2024, namely the E-TSL dataset [63], which consists of 24 hours of sign language videos. However, the dataset is limited to the domain of language use, being restricted to the educational domain. To improve upon this, we propose the first large-scale open-domain dataset for TID.

In this work, we introduce the BUTID dataset, comprising over 500 hours of video footage from open-access YouTube content, capturing daily language use. The dataset provides a representative lexical and structural alignment between Turkish Sign Language and Turkish, addressing the scarcity of large-scale sign language corpora. To follow recent advancements in SLT, we establish baselines aligned with contemporary research [25, 54], emphasizing the role of models trained on the target language rather than the multilingual model.

2. Related Work

In this section, we will explore existing literature on sign language translation datasets and methodologies.

2.1. Datasets

The earlier datasets on sign language processing emerged in the early 2000s, primarily focusing on Sign Language Recognition (SLR) [8, 47] and Continuous Sign Language Recognition (CSLR) tasks [37, 56]. Most of these initial datasets were collected under highly controlled environments, such as laboratories or studio settings, and rely on labour-intensive annotations.

However, with the advancement in computer vision areas, the focus of data has shifted toward the collection of real-world data [46]. One of the foundational datasets at this is the RWTH-Phoenix Weather Dataset for German Sign Language (DGS) [19]. Initially developed from video footage of German weather forecasts, this dataset established an early benchmark for SLT research. The original version included only 3.25 hours of video footage with 1,980 sentence pairs; the later version, namely **Phoenix-2014T**, expanded to 11 hours of footage and approximately 9,000 sign-sentence pairs as the first benchmark on SLT [10].

With the recent advancement of deep learning, the usage of large-scale data in training has shown significant performance improvement within model performance [15, 38]. To incorporate this, one of the first large-scale SLT datasets, the BSL-1K dataset [1] represents a major leap forward, containing 1 million sentence pairs and covering 1,060 hours of footage. The BOBSL [2] dataset is further extended over the BSL-1K dataset, featuring 1.2 million sentence pairs and 1,467 hours of video content sourced from BBC broadcasts. In this dataset, automatic sentence alignment was employed to map spoken language content to sign language videos, which significantly reduced manual annotation efforts while maintaining large-scale coverage [7].

For ASL, the YouTubeASL dataset [54] stands out as a notable resource. This open-source dataset initially provided 900 hours of footage translated into English. Recent expansions of the dataset, namely Youtube-SL-25 [51], have increased the dataset size to 2,800 hours, covering more than 25 sign languages. The other similar large-scale datasets, including multilingual SLT, are the AfriSign dataset, [22], and JWSign [21], which has 2530 hours of Bible translations of 98 sign languages to demonstrate further the importance of the scale of pretraining data in SLT. You can see the collection of various datasets on SLT in Table 1.

In the context of TID, the development of SLT datasets has been more limited compared to other languages such as ASL, BSL, or DGS. The Educational Turkish Sign Language (E-TSL) dataset [63] presents the first continuous dataset for TID. Although not matching the scale of larger ASL datasets, E-TSL comprises approximately 24 hours of TID footage, totaling 1,418 video clips featuring 11 different signers. This dataset focuses on educational content, specifically Turkish language lessons for 5th, 6th, and 8th grades. Additionally, the YouTube-SL-25 dataset [51] includes 18 hours of TID videos, contributing to the available resources for Turkish Sign Language. While these datasets are smaller in scale compared to extensive ASL datasets, they represent significant steps toward advancing SLT research for TID.

2.2. Sign Language Translation

With the improvements in Neural Machine Translation (NMT) [3, 32, 55], Camgöz et al. [10] have shown NMT-based end-to-end SLT systems with the first public dataset, Phoenix2014T, using a 2D-CNN + RNN-based approach. Subsequent advancements in end-to-end translation systems, also known as sign-to-text models, introduced various novel methodology. Building upon this, Kim et al. [26] proposed a similar end-to-end system but employed

Table 1. Overview of recent large-scale sign language translation datasets, including our BUTID dataset

Dataset Name	Sign Lang.	Spoken Lang.	# Instances	# Hours	# Signers
YouTube-SL-25 [51]	25 SL	Multi-lingual	2.16M	3,207	$>$ 3,000
JWSign [21]	98 SL	Multi-lingual	N/A	2,530	$>$ 1,500
BOBSL [2]	BSL	English	1.2M	1,467	39
YouTube-ASL [54]	ASL	English	610K	984	$>$ 2519
BUTID	TID	Turkish	237K	536	27
OpenASL [46]	ASL	English	288	280	200
E-TSL [63]	TID	Turkish	1486	24	11
CSL-Daily	CSL	Chinese	21K	23	10
PHOENIX14T [10]	DGS	German	8,257	10	9

body key-point coordinates as input for their translation networks, evaluating their methods on a Korean Sign Language dataset.

With the emergence of large language models (LLMs), De Coster et al. [14] proposed a frozen pretrained transformer model by initializing a transformer translation model with pretrained BERT-based [15] and mBART-50 [13] models to develop SLT systems.

As LLMs have advanced, the importance of pretraining has become increasingly evident, especially when paired with large-scale resources such as BOBSL [2], YouTubeASL [51, 54], and JWSign [21]. Transformer-based pretraining strategies have become standard practice for SLT tasks. Uthus et al. [54] demonstrated the potential of using the YouTube-ASL dataset for large-scale ASL translation model training. They fine-tuned a T5 model [39] on YouTube-ASL and subsequently fine-tuned it on smaller benchmark datasets like How2Sign [12]. Rust et al. [42] further improved performance on How2Sign by pretraining a video encoder on YouTube-ASL, initialized from a self-supervised image encoder model, namely Hiera-based [43] architecture, trained with a masked autoencoding objective.

The most recent approaches employ pretraining strategies for SLT that go beyond translation-based objectives. Zhang et al. [61] proposed multiple training objectives, such as alignment, while Wong et al. [57] suggested a pseudo-gloss pretraining strategy that automatically extracts pseudo-glosses from sentences to pretrain the sign encoder. Similarly, Gong et al. [20] demonstrated the effectiveness of a vector-quantization-based pretraining strategy with representation alignment pretraining with pretrained LLMs.

3. The BUTID Dataset

In this section, we present our large-scale open-source dataset for Turkish Sign Language (TID) translation, the largest Turkish SLT dataset. This dataset includes 539 hours of sign language video, automatically aligned with corresponding to Turkish text, providing a large corpus for both linguistic and computer science research.

The dataset comprises 237,718 video clips from 1,700 YouTube videos, featuring manually annotated captions alongside signed translations specifically aimed at accessible content for individuals with disabilities in the "Engelsiz TRT" channel of Turkish Radio and Television (TRT) Corporation. The captions and signed content used in our paper were created by the Audio Description Association (Sesli Betimleme Derneği — SeBeDer).

The dataset includes 15 main categories of television series, consisting of various genres from science fiction to historical dramas, resulting in large lexical variation suitable for representative SLT training. A total of 27 unique signers, who are Children of Deaf Adults (CODA) translators, are featured throughout the dataset, with diversity in terms of signing styles and individuals with age and gender variations.

3.1. Data Selection

The absence of exact alignment between signed content and captions poses a significant challenge in developing sign language translation datasets, primarily due to structural and temporal differences between signed and spoken languages. These issues have been noted in the previous literature in the development of the SLT dataset [1, 2, 61]. Although these datasets have often included manual annotation for the development of a solution, we propose an automatic procedure in this paper. The conversational nature of the video content often includes natural pauses or gaps in dialogue, which can facilitate semi-automated alignment. We designed an alignment algorithm leveraging these natural dialogue gaps and pose estimation techniques, which we will explain in detail in Section 3.3.

We collected our clips in two caption groups, namely "single-caption translation" and "multi-caption translation" groups. In the "single caption" group, we select the clips with appropriate padding at the beginning and end of captions. This enables us to locate where the signing starts and ends without any noise around translation. In the "multi-caption" translation, we merge the subsequent sequences

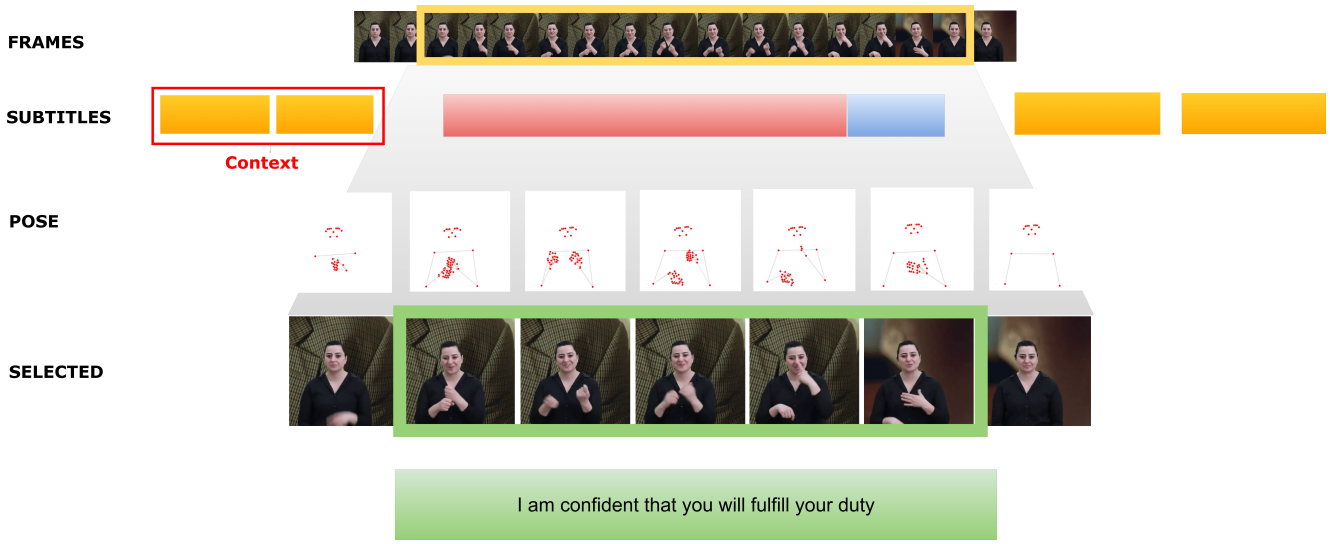

Figure 2. Visualization of sign language translation alignment. The yellow rectangle highlights the caption duration, followed by the padding duration in blue. The red section represents the original time duration of the subtitles. The green rectangle indicates the selected signing duration after filtering out inactive frames. The figure includes raw video frames, corresponding pose representations, and the final subtitle output, demonstrating the alignment between visual signing and textual translation.

with brief pauses of less than 2 seconds in between to locate sequences with no conversational turns between source video content.

3.2. Dataset Preprocessing

We employed three main preprocessing methods for both text and signed video content.

Signer Box Annotation. In all videos, signers appear in the bottom-right corner of the videos; the bounding box detection must be done for isolating signing activity. We have done bounding box annotations for each video semi-automatically.

Text Cleaning. There are various descriptive captions in the subtitle file, so we applied text cleaning methods, where we removed captions like songs, speaker-denoting descriptors, and conversational turn-denoting non-alphabetic characters.

Initializing Proper Nouns. The proper names of people are transformed into their initial letter, such as for the name "Ahmet", it will be transformed into "A.". This is the grammatical strategy utilized in TID as well as other sign languages [33].

3.3. Sign-Caption Alignment

Aligning captions with the corresponding sign sequences was one of the most challenging issues of dataset creation. Since sign sequences are the result of live interpretation, the boundaries of sign videos often do not overlap with caption start and end points, resulting in the extraction of mismatched and noisy content pairs affecting the translation performance. We designed an alignment algorithm, as illustrated in Figure 2. The algorithm detects sign boundaries using captions and pose estimation, identifying active frames and applying temporal pruning [34].

Caption Gap Detection: Identify captions with gaps of more than two seconds before the next caption sequence and captions themselves with durations exceeding two seconds.

Sequence Extraction: Extract the video sequence starting from the caption's timestamp and extend it by 1500 ms beyond the caption's end to account for potential interpretation length without overlapping with the next sequence.

Active Signing Detection: Apply pose estimation within the signer's bounding box to identify active frames where the signer's hands are positioned one third of the distance between hip and shoulder, indicating signing activity. Use the leftmost and rightmost active frames to determine the signing subsequence.

3.4. Dataset Statistics

After filtering the subtitles, we selected sequences totaling approximately 536 hours of sign language footage. The final dataset comprises 237,718 video clips extracted from 1,734 unique videos. Clips vary in duration between 3 to 20 seconds, with our caption-based selection algorithm filtering out any segments exceeding 20 seconds. You can find the duration distribution across clips in Figure 3. The sentence alignment process was then applied, which helped create a more evenly distributed clip duration across the dataset. This alignment process yields a balanced representation of shorter and longer sequences, addressing the initial skewed distribution seen in the raw data. However, the word

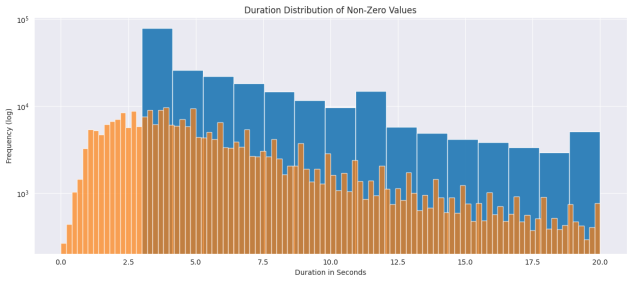

Figure 3. Comparison of raw and aligned duration distributions (in seconds) for video clips in the BUTID dataset, illustrating the effect of alignment on clip length frequency.

distribution on the clips, shown in Figure 4 is less affected by this realignment.

Overall, the dataset contains 121,463 unique words, encompassing a diverse range of morphological variations inherent to the Turkish language. Notably, 91,033 words appear fewer than five times, with 56,655 occurring only once (singletons), accounting for 74.95% and 46.64% of the vocabulary, respectively. This high proportion of rare and singleton words is challenging from an SLT perspective.

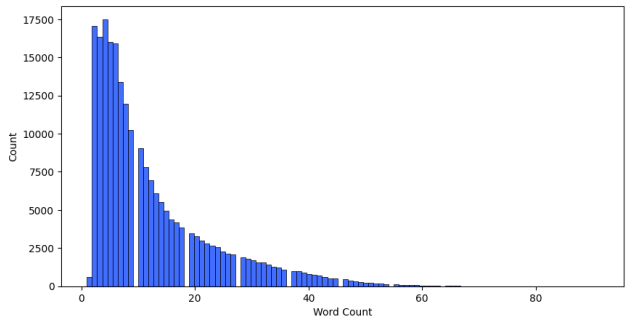

Figure 4. The word distribution per clip of BUTID

3.5. Categorical Distribution

Figure 5 illustrates the categories of the clips, which are extracted from TV content. While drama, action, and comedy are the most common categories, the dataset's diversity ensures a rich vocabulary. Additionally, the channel features substantial historical content, including Ottoman Turkish lexicons, adding language complexity. Therefore, we excluded these videos during the prior filtering process.

4. Translation Method

In this section, we introduce our sign language translation (SLT) method, which leverages LLM-based translation similar to translation methods proposed in recent studies [25, 42, 54] using frame-level pose representations as visual features. We detail the architectural framework and

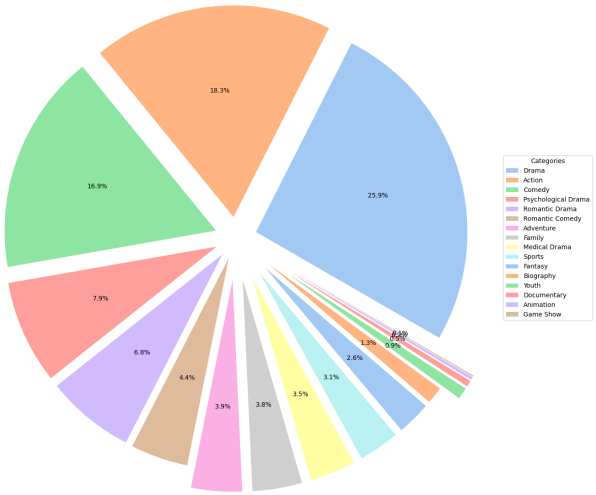

Figure 5. The categorical distribution across clips of BUTID

key parameters guiding our approach.

4.1. Visual Features

Pose Estimation. The estimations are done via the MediaPipe library's [31] Holistic module with model complexity set to 2, and minimum tracking confidence and minimum detection confidence values set to 0.3.

Keypoint Selection and Normalization. Frame-level pose normalization was applied separately for each cue, using min-max normalization on the x, y, and z coordinates individually. In our final keypoint selection, we utilized all available landmarks for the hands. For the body pose, we selected 20 points focused on the upper body. Additionally, for facial landmarks, we chose 35 points, following the approach of Uthus et al. [54]. In total, there are 255 visual features.

4.2. SLT Architecture

The models that utilize pretrained large language models in SLT implement a feature projection layer for mapping visual features extracted from sign language videos into the token space of the pre-trained language model using a linear layer before feeding into the network. This projection layer inputs a matrix of visual features, defined as $V \in \mathbf{R}^{T \times D_v}$, extracted from a sign language video, where T denotes the number of time steps (frames) and D_v is the dimensionality of the visual feature vectors.

Sign Encoder. For this experiment, we have utilized a projection layer which is two MLP layers with GeLU activation function [23].

Encoder-Decoder-based SLT. We employ the pretrained encoder-decoder T5 model [39] as proposed in the YouTubeASL dataset [51, 54, 61]. This approach uses nor-

Table 2. Dataset Splits of the BUTID Dataset

Metric	Single Caption			Multi-caption		
	Train	Dev	Test	Train	Dev	Test
# clips	108,530	5,540	5,554	106,665	5,717	5,712
# hours	138.47	6.54	6.57	346.96	18.86	18.91
Avg. clip duration (seconds)	4.59	4.32	4.33	11.71	11.77	11.79
Avg. text length (char)	29.92	27.15	27.22	114.23	114.54	115.20
Avg. text length (token)	6.03	5.61	5.64	22.27	22.37	22.49
Avg. sentences	1.12	1.12	1.14	3.99	4.01	4.03

malized pose data embedded through a linear projection layer. The model is pretrained on the YouTubeASL dataset and fine-tuned on the How2Sign benchmark [12] for enhanced SLT performance.

4.3. LLM Backbone

In these experiments, we will be using the multilingual T5 model to provide a baseline LLM architecture for our dataset.

mt5-small. [58] The model builds upon the T5 architecture [38], expanding its parameter count to approximately 800 million. Each model is trained on the mC4 corpus, where 1.1% of the data is in Turkish, differing in the number of parameters in the architecture. We utilize T5 for a direct comparison between model parameter size with the same training strategies as well as the replicability purposes of [54, 61].

4.4. Augmentations

Random Drop. Random drop reduces the length of input sequences by removing a percentage of frames. This operation mimics scenarios where frames are missing or skipped during video recording or preprocessing. A random percentage of frames, between 10% and 30%, is selected for removal.

Skip Frames. Skip frame augmentation reduces the frame rate by selecting every other frame in the input sequence, similar to [51], we utilize this method for replicating undetected poses from the frames with blur or high-velocity movement.

Cue Masking. Cue masking applies artificial occlusion to specific cues, such as hand, pose, or face landmarks, similar to Jang et al. [25]. A random percentage of cues, ranging from 10% to 30%, is selected for masking, simulating scenarios where parts of the signer's body are obscured by motion blur, poor lighting, or camera angles.

4.5. Paraphrasing

Turkish exhibits significant word-order variation [35], which complicates BLEU Score evaluation due to its reliance on exact n-gram sequences. This often results in artificially low scores despite semantically correct translations. To address this, we employed the TURNA model [53], fine-tuned on the Turkish subset of OpenSubtitles2018 [30]. This dataset aligns well with our dataset, helping TURNA better handle variations in the Turkish language and improve evaluation reliability in caption setting. We use paraphrasing inferences both for augmentation and as target translations.

5. Experimental Setup

In this section, we detail our experimental setup, covering dataset creation, training procedures, and evaluation protocols.

5.1. Dataset Splits

To ensure a fair evaluation and account for translation length differences, we developed two subsets: "Single-Caption" and "Multi-Caption", stemming from our data collection strategies. The Single-Caption subset comprises short captions averaging 3–5 seconds, offering limited context. In contrast, the Multi-Caption subset includes sequences averaging 9–13 seconds, providing richer contextual information. This distinction is done as translation performance can vary with input length and may lead to bias in the decoding process.

We created three distinct dataset splits for each subset as train, development (dev), and test. These splits were carefully stratified based on the original video clips, ensuring each split maintained a consistent and uniform distribution of topics. As seen in Table 2, the average clip duration and text length are also matching across our subsets.

5.2. Training Details

The training was done for 20,000 steps on 4 AMD MI250x GPUs with a per-device batch size of 8 for all models. Gradients were accumulated over 8 steps in a total of 128 batch sizes per iteration. The training is initialized with warmup training for 1000 steps with a weight decay of 0.001. The learning rate of 0.0001 was used, and optimization was performed using Adafactor [45].

Table 3. Baseline results for SLT on the BUTID **Development** and **Test** datasets. Metrics: B-1 to B-4 denote BLEU-1 to BLEU-4 scores, R-L represents the ROUGE-L metric, and BERT indicates the Mean F1 score of BERTScore.

Subset	Aug	Development Set						Test Set					
		BLEU Scores				R-L	BERT	BLEU Scores				R-L	BERT
		B-1	B-2	B-3	B-4			B-1	B-2	B-3	B-4		
Single-Caption	✗	26.26	5.89	2.98	1.66	17.27	49.34	25.88	5.61	2.89	1.58	16.66	49.16
	✓	24.69	4.48	2.19	1.07	14.19	47.91	24.13	4.20	2.27	1.40	12.69	47.41
Multi-Caption	✗	37.34	11.75	5.04	2.40	22.27	51.06	37.61	11.79	5.19	2.42	22.16	50.96
	✓	39.05	13.04	5.86	2.84	24.11	52.33	39.31	13.20	6.02	2.88	23.98	52.31

5.3. Evaluation Metrics

BLEU. [36] (**Bi**lingual **E**valuation **U**nderstudy) is a metric widely used to objectively assess machine translation quality by comparing candidate translations to one or more reference translations. It calculates modified n-gram precision by measuring the overlap of n-grams, from 1 to 4 grams, between candidates and references while limiting over-counting

ROUGE. [29] (**R**ecall-**O**riented **U**nderstudy for **G**isting **E**valuation) is a widely-used metric for assessing machine-generated summaries or translations by measuring textual overlap with human reference texts. We will be using ROUGE-L metric, which assesses the longest common subsequence capturing sentence-level similarity.

BERT Score. [62] evaluates translation quality by computing cosine similarity between contextualized token embeddings from pre-trained BERT models, generating precision, recall, and F1 metrics. [1]

5.4. Generation Configuration

During inference on both development and test sets for both architectures, beam search with 4 beams was employed, applying a length penalty of 1.0 and preventing the repetition of any 3-gram sequences in the output. The decoder was allowed to generate up to 128 new tokens, with early stopping enabled to terminate generation upon reaching the end-of-sequence token.

6. Experimental Results

Table 3 summarizes the baseline results on the BUTID development and test datasets.

The Single-Caption setting yields the worst performance across all metrics, primarily due to the shorter sentence duration providing less contextual information for the model. This is particularly evident in the BLEU-3 and BLEU-4 scores, which heavily rely on longer n-grams and thus suffer when the available content is limited, namely an average of 3 seconds or 75 frames. Additionally, the ROUGE-L and BERT Score metrics also indicate weaker performance compared to the Multi-Caption setting. We hypothesize that this occurs due to the limited contextual information within the signing sequence.

Further, augmentation impacts performance differently depending on the caption type. In the Single-Caption setting, augmentation slightly decreases performance, as seen in the decrease of BLEU-4 from 1.66 to 1.07 in the Development Set and from 1.58 to 1.40 in the Test Set. This further suggests that the lack of contextual continuity in single-caption inputs restricts the benefits of augmentation. In contrast, in the Multi-Caption setting, augmentation leads to a more significant improvement across all evaluation metrics. For example, BLEU-4 improves from 2.40 to 2.84 in the Development Set and from 2.42 to 2.88 in the Test Set. This demonstrates that when more contextual information is available, sign language augmentations have increasing performance effects for our dataset.

6.1. Effect of Sign Augmentation and Paraphrasing

We have also applied augmentation to both signing sequences and text translations to analyze their impact on SLT performance. However, due to the constraints of the paraphrase generation model, which was trained on a caption-based dataset, we limited textual augmentation to single-caption sequences.

Table 4. Effect of sign augmentation and paraphrasing on SLT performance for the BUTID Test dataset.

Paraphrase	Augment	B-4	R-L	BERT
✗	✓	1.40	12.69	47.41
✓	✗	**3.42**	**22.23**	**53.22**
✓	✓	2.72	17.99	51.14

As shown in Table 4, paraphrases generated by TURNA model [53] significantly improves translation quality. When paraphrasing is applied without sign augmentation, BLEU-4 improves from 1.40 to 3.40, ROUGE-L increases from 12.69 to 22.23, and BERTScore rises from 47.41 to 53.22. This highlights the positive impact of paraphrasing on trans-

[1] More recently, metrics like BLEURT [44] or COMET [41] have been introduced to address these limitations by incorporating semantic representations on sentence similarity; however these metrics are not available in Turkish.

Figure 6. Examples of predicted and ground truth (GT) sentences for Turkish Sign Language translation. Each example consists of sign language video frames, the predicted sentence (Pred), the ground truth sentence (GT), and their corresponding translations in English.

lation fluency and semantic consistency. However, combining paraphrasing with sign augmentation leads to a decreasing effect in BLEU-4 (3.40 $\rightarrow$ 2.72) which aligns with our previous findings on the effect of sign augmentation.

6.2. Qualitative Analysis

As we have discussed, translation into the Turkish language poses challenges for surface-level matching evaluation metrics like BLEU. The extensive morphological variations often result in mismatches, even when the root words remain the same. Additionally, the flexible word order affects n-gram-based evaluations, particularly for BLEU-3 and BLEU-4, as changes in word arrangement disrupt exact phrase matches.

Figure 6 illustrates the challenges of using BLEU scores for evaluating Turkish Sign Language (TID) translations, particularly due to the agglutinative nature of Turkish and flexible word order. In each example, the predicted sentence (Pred) is compared against the ground truth (GT), yet all translations receive a BLEU-4 score of 0, despite their semantic similarity. This demonstrates how BLEU, which relies on exact word matches and fixed n-gram sequences, struggles to capture meaning when variations in word order and morphology occur.

7. Conclusion

We introduced a large-scale open-domain dataset for Turkish Sign Language (TID) Translation (SLT), comprising 536 hours of sign video, segmented into 237,718 clips from 1,734 videos. The videos consist of two subsets, namely "single-caption" and "multi-caption" datasets, where the captions are loosely aligned with the videos. As a benchmark, we utilized LLM-based architecture for sign language translation, establishing the language-specific requirements of SLT systems. Additionally, we integrated sign and text augmentation, adapting to Turkish's agglutinative morphology and flexible word order to enhance translation performance. In summary, we developed a Turkish SLT benchmark informed by recent literature, leveraging modern architectures and multimodal augmentations. We aim to implement a decoder-based LLM that leverages contextual information from other textual sources. Given the importance of context in recent studies [25, 48] and our findings on the single-caption subset's limitations, embedding previous contextual cues will be a key focus for improving translation quality. One limitation of the dataset is loose alignment between captions and video, especially in the multi-caption setting. Our future work will involve providing better alignment. We aim to further work on this data to provide a multi-disciplinary research ground for TID.

Acknowledgement

The computations in this work are partially performed at TUBITAK ULAKBIM, High Performance and Grid Computing Center (TRUBA resources), and EuroHPC LUMI, whose funding and resources were instrumental in completing this research. We acknowledge EuroHPC Joint Undertaking for awarding project ID EHPC-AI-2024A03-075 access to LUMI at CSC, Finland. We express our gratitude to the Sesli Betimleme Derneği (SeBeDer) for their encouragement and for providing consent for this work.

References

[1] Samuel Albanie, Gül Varol, Liliane Momeni, Triantafyllos Afouras, Joon Son Chung, Neil Fox, and Andrew Zisserman. Bsl-1k: Scaling up co-articulated sign language recognition using mouthing cues. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pages 35–53. Springer, 2020. 2, 3

[2] Samuel Albanie, Gül Varol, Liliane Momeni, Hannah Bull, Triantafyllos Afouras, Himel Chowdhury, Neil Fox, Bencie Woll, Rob Cooper, Andrew McParland, and Andrew Zisserman. BOBSL: BBC-Oxford British Sign Language Dataset. *arXiv*, 2021. 2, 3

[3] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate, 2016. 1, 2

[4] Danielle Bragg, Oscar Koller, Mary Bellard, Larwan Berke, Patrick Boudreault, Annelies Braffort, Naomi Caselli, Matt Huenerfauth, Hernisa Kacorri, Tessa Verhoef, Christian Vogler, and Meredith Ringel Morris. Sign language recognition, generation, and translation: An interdisciplinary perspective. In *Proceedings of the 21st International ACM SIGACCESS Conference on Computers and Accessibility*, page 16–31, New York, NY, USA, 2019. Association for Computing Machinery. 1

[5] Diane Brentari. A prosodic model of sign language phonology. *A Bradford Book*, 1998. 1

[6] Patrick Buehler, Andrew Zisserman, and Mark Everingham. Learning sign language by watching tv (using weakly aligned subtitles). In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2961–2968. IEEE, 2009. 1

[7] Hannah Bull, Triantafyllos Afouras, Gül Varol, Samuel Albanie, Liliane Momeni, and Andrew Zisserman. Aligning subtitles in sign language videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11552–11561, 2021. 2

[8] Necati Cihan Camgöz, Ahmet Alp Kındıroğlu, Serpil Karabüklü, Meltem Kelepir, Ayşe Sumru Özsoy, and Lale Akarun. 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, 2016. 2

[9] Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. Neural sign language translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7784–7793, 2018. 1

[10] Necati Cihan Camgöz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. Rwth-phoenix-weather 2014 t: Parallel corpus of sign language video, gloss and translation. *CVPR, Salt Lake City, UT*, 3:6, 2018. 1, 2, 3

[11] Necati Cihan Camgoz, Oscar Koller, Simon Hadfield, and Richard Bowden. Sign language transformers: Joint end-to-end sign language recognition and translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1

[12] Amanda Cardoso Duarte, Shruti Palaskar, Lucas Ventura Ripol, Deepti Ghadiyaram, Kenneth DeHaan, Florian Metze, Jordi Torres Viñals, and Xavier Giró Nieto. How2sign: A large-scale multimodal dataset for continuous american sign language. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition: Virtual, 19-25 June 2021: proceedings*, pages 2734–2743. Institute of Electrical and Electronics Engineers (IEEE), 2021. 1, 3, 6

[13] Hugh A Chipman, Edward I George, Robert E McCulloch, and Thomas S Shively. mbart: multidimensional monotone bart. *Bayesian Analysis*, 17(2):515–544, 2022. 3

[14] Mathieu De Coster, Karel D'Oosterlinck, Marija Pizurica, Paloma Rabaey, Severine Verlinden, Mieke Van Herreweghe, and Joni Dambre. Frozen pretrained transformers for neural sign language translation. In *18th Biennial Machine Translation Summit (MT Summit 2021)*, pages 88–97. Association for Machine Translation in the Americas, 2021. 1, 3

[15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, 2019. Association for Computational Linguistics. 2, 3

[16] OpenAI et al. Gpt-4 technical report, 2024. 1

[17] Gaolin Fang and Wen Gao. A srn/hmm system for signer-independent continuous sign language recognition. In *Proceedings of Fifth IEEE International Conference on Automatic Face Gesture Recognition*, pages 312–317. IEEE, 2002. 1

[18] JJ Fenlon. *Seeing sentence boundaries: the production and perception of visual markers signalling boundaries in signed languages*. PhD thesis, UCL (University College London), 2010. 1

[19] Jens Forster, Christoph Schmidt, Thomas Hoyoux, Oscar Koller, Uwe Zelle, Justus H Piater, and Hermann Ney. Rwth-phoenix-weather: A large vocabulary sign language recognition and translation corpus. In *LREC*, pages 3785–3789, 2012. 2

[20] Jia Gong, Lin Geng Foo, Yixuan He, Hossein Rahmani, and Jun Liu. Llms are good sign language translators. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18362–18372, 2024. 3

[21] Shester Gueuwou, Sophie Siake, Colin Leong, and Mathias Müller. JWSign: A highly multilingual corpus of Bible translations for more diversity in sign language processing. In

Findings of the Association for Computational Linguistics: EMNLP 2023, pages 9907–9927, Singapore, 2023. Association for Computational Linguistics. 2, 3

[22] Shester Gueuwou, Kate Takyi, Mathias Müller, Marco Stanley Nyarko, Richard Adade, and Rose-Mary Owusuaa Mensah Gyening. Afrisign: Machine translation for african sign languages. In *4th Workshop on African Natural Language Processing*, 2023. 2

[23] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016. 5

[24] Matt Huenerfauth. A multi-path architecture for machine translation of english text into american sign language animation. In *Proceedings of the student research workshop at HLT-NAACL 2004*, pages 25–30, 2004. 1

[25] Youngjoon Jang, Haran Raajesh, Liliane Momeni, Gül Varol, and Andrew Zisserman. Lost in translation, found in context: Sign language translation with contextual cues, 2025. 2, 5, 6, 8

[26] San Kim, Chang Jo Kim, Han-Mu Park, Yoonyoung Jeong, Jin Yea Jang, and Hyedong Jung. Robust keypoint normalization method for korean sign language translation using transformer. In *2020 International Conference on Information and Communication Technology Convergence (ICTC)*, pages 1303–1305. IEEE, 2020. 1, 2

[27] Oscar Koller, Sepehr Zargaran, Hermann Ney, and Richard Bowden. Deep sign: Hybrid cnn-hmm for continuous sign language recognition. In *BMVC*, pages 136–1, 2016. 1

[28] Dongxu Li, Cristian Rodriguez, Xin Yu, and Hongdong Li. Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. In *The IEEE Winter Conference on Applications of Computer Vision*, pages 1459–1469, 2020. 2

[29] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, 2004. Association for Computational Linguistics. 7

[30] Pierre Lison, Jörg Tiedemann, and Milen Kouylekov. OpenSubtitles2018: Statistical rescoring of sentence alignments in large, noisy parallel corpora. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, 2018. European Language Resources Association (ELRA). 6

[31] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, et al. Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*, 2019. 5

[32] Minh-Thang Luong. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*, 2015. 2

[33] Bahtiyar Makaroglu. Blend formation in turkish sign language: Are we missing the big picture? *Journal of Language and Linguistic Studies*, 17(1):139–157, 2021. 4

[34] Oğulcan Özdemir, İnci M Baytaş, and Lale Akarun. Multi-cue temporal modeling for skeleton-based sign language recognition. *Frontiers in Neuroscience*, 17:1148191, 2023. 4

[35] Balkız Öztürk. Null arguments and case-driven agree in turkish. *Minimalist essays*, pages 268–287, 2006. 6

[36] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 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, 2002. Association for Computational Linguistics. 7

[37] Siegmund Prillwitz, Thomas Hanke, Susanne König, Reiner Konrad, Gabriele Langer, and Arvid Schwarz. Dgs corpus project–development of a corpus based electronic dictionary german sign language/german. In *sign-lang@ LREC 2008*, pages 159–164. European Language Resources Association (ELRA), 2008. 2

[38] Alec Radford. Improving language understanding by generative pre-training. 2018. 2, 6

[39] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020. 3, 5

[40] Razieh Rastgoo, Kourosh Kiani, and Sergio Escalera. Sign language recognition: A deep survey. *Expert Systems with Applications*, 164:113794, 2021. 1

[41] Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online, 2020. Association for Computational Linguistics. 7

[42] Phillip Rust, Bowen Shi, Skyler Wang, Necati Cihan Camgoz, and Jean Maillard. 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, 2024. Association for Computational Linguistics. 3, 5

[43] Chaitanya Ryali, Yuan-Ting Hu, Daniel Bolya, Chen Wei, Haoqi Fan, Po-Yao Huang, Vaibhav Aggarwal, Arkabandhu Chowdhury, Omid Poursaeed, Judy Hoffman, et al. Hiera: A hierarchical vision transformer without the bells-and-whistles. In *International Conference on Machine Learning*, pages 29441–29454. PMLR, 2023. 3

[44] Thibault Sellam, Dipanjan Das, and Ankur P Parikh. Bleurt: Learning robust metrics for text generation. *arXiv preprint arXiv:2004.04696*, 2020. 7

[45] Noam Shazeer and Mitchell Stern. Adafactor: Adaptive learning rates with sublinear memory cost. In *International Conference on Machine Learning*, pages 4596–4604. PMLR, 2018. 6

[46] Bowen Shi, Diane Brentari, Greg Shakhnarovich, and Karen Livescu. Open-domain sign language translation learned from online video. *arXiv preprint arXiv:2205.12870*, 2022. 2, 3

[47] Ozge Mercanoglu Sincan and Hacer Yalim Keles. Autsl: A large scale multi-modal turkish sign language dataset and baseline methods. *IEEE access*, 8:181340–181355, 2020. 2

[48] Ozge Mercanoglu Sincan, Necati Cihan Camgoz, and Richard Bowden. Is context all you need? scaling neural sign language translation to large domains of discourse. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 1955–1965, 2023. 8

[49] T. Starner, J. Weaver, and A. Pentland. Real-time american sign language recognition using desk and wearable computer based video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(12):1371–1375, 1998. 1

[50] William C Stokoe. Sign language structure. *Annual review of anthropology*, pages 365–390, 1980. 1

[51] Garrett Tanzer and Biao Zhang. Youtube-sl-25: A large-scale, open-domain multilingual sign language parallel corpus, 2024. 2, 3, 5, 6

[52] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023. 1

[53] Gökçe Uludoğan, Zeynep Balal, Furkan Akkurt, Meliksah Turker, Onur Gungor, and Susan Üsküdarlı. TURNA: A Turkish encoder-decoder language model for enhanced understanding and generation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 10103–10117, Bangkok, Thailand, 2024. Association for Computational Linguistics. 6, 7

[54] David Uthus, Garrett Tanzer, and Manfred Georg. Youtube-asl: A large-scale, open-domain american sign language-english parallel corpus, 2023. 2, 3, 5, 6

[55] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2017. 1, 2

[56] Ulrich von Agris and Karl-Friedrich Kraiss. Signum database: Video corpus for signer-independent continuous sign language recognition. In *4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies*, pages 243–246, 2010. 1, 2

[57] Ryan Wong, Necati Cihan Camgoz, and Richard Bowden. Sign2gpt: Leveraging large language models for gloss-free sign language translation, 2024. 3

[58] Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 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, 2021. Association for Computational Linguistics. 6

[59] Hee-Deok Yang, Stan Sclaroff, and Seong-Whan Lee. Sign language spotting with a threshold model based on conditional random fields. *IEEE transactions on pattern analysis and machine intelligence*, 31(7):1264–1277, 2008. 1

[60] Kayo Yin, Amit Moryossef, Julie Hochgesang, Yoav Goldberg, and Malihe Alikhani. Including signed languages in natural language processing, 2021. 1

[61] Biao Zhang, Garrett Tanzer, and Orhan Firat. Scaling sign language translation, 2024. 1, 3, 5, 6

[62] Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*, 2020. 7

[63] Şükrü Öztürk and Hacer Yalim Keles. 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, 2024. 2, 3