iSign: A Benchmark for Indian Sign Language Processing

Abhinav Joshi[†] **Romit Mohanty**[†] **Mounika Kanakanti**[¶†]
Andesha Mangla[⋆] **Sudeep Choudhary**[◇] **Monali Barbate**[◇]
Ashutosh Modi[†]
[†]IIT Kanpur
[¶]Max Planck Institute for Psycholinguistics [⋆]ISLRTC [◇]Microsoft IDC India
{ajoshi, ashutoshm}@cse.iitk.ac.in

Abstract

Indian Sign Language has limited resources for developing machine learning and data-driven approaches for automated language processing. Though text/audio-based language processing techniques have shown colossal research interest and tremendous improvements in the last few years, Sign Languages still need to catch up due to the need for more resources. To bridge this gap, in this work, we propose **iSign**: a benchmark for Indian Sign Language (ISL) Processing. We make three primary contributions to this work. First, we release one of the largest ISL-English datasets with more than 118k video-sentence/phrase pairs. To the best of our knowledge, it is the largest sign language dataset available for ISL. Second, we propose multiple NLP-specific tasks (including SignVideo2Text, SignPose2Text, Text2Pose, Word Prediction, and Sign Semantics) and benchmark them with the baseline models for easier access to the research community. Third, we provide detailed insights into the proposed benchmarks with a few linguistic insights into the workings of ISL. We streamline the evaluation of Sign Language processing, addressing the gaps in the NLP research community for Sign Languages. We release the dataset, tasks, and models via the following website: https://exploration-lab.github.io/iSign/.

1 Introduction

As per the WHO estimate, about 63 million people belong to the Deaf and Hard of Hearing (DHH) community in India (WHO, 2016; Varshney, 2016). Consequently, Indian Sign Language (ISL) is widely used in the Indian subcontinent. Moreover, according to Ethnologue (2022) (a reference publication documenting information about living languages of the world), ISL is the world's most widely used sign language. However, there is a considerable deficit of sign language interpreters, e.g., according to the Government of India organization

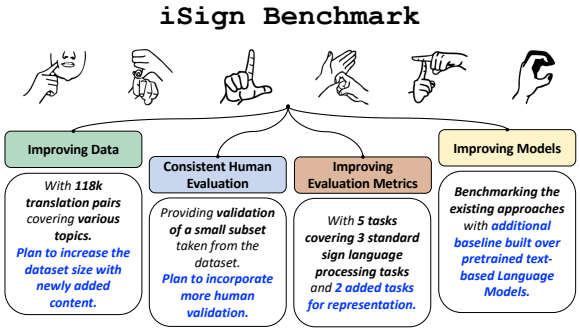

Figure 1: iSign Benchmark: The proposed benchmark for Indian Sign Language Processing.

Indian Sign Language Research and Training Center (ISLRTC) (https://islrtc.nic.in/), there are only 300 certified sign language interpreters in India. NLP technologies can help in this case.

Similar to spoken languages, sign languages are region-specific, for example, people in North America use American Sign Language (ASL), and people in Germany use Deutsche Gebärdensprache (DGS). Considerable efforts have been made to develop technologies for automatically processing sign language in other countries. However, when it comes to ISL, very limited technological advancements have been made; for example, there is a lack of standard benchmarks for ISL, resulting in low development and a lack of comparison of Machine Learning (ML) based solutions for ISL, e.g., word recognition, translation, generation, etc. In contrast, relatively speaking, other sign languages (e.g., American Sign Language (ASL), Deutsche Gebärdensprache (DGS)) have a sufficient number of annotated resources for data-driven approaches (Table 2). Natural Language Processing (NLP) has made rapid progress in the last few years (Min et al., 2021). Most of these approaches have targeted textual datasets. Easy access to textual datasets and leaderboards in languages like English has facilitated the development, reliabil-

Findings of the Association for Computational Linguistics ACL 2024, pages 10827–10844
August 11-16, 2024

Figure 2: An example showing the translation of the phrase “What, Where, How, and When" in Indian Sign Language. The text box length overlaps with the signs with a pause position in between.

ity, and standardization of experimentation on several tasks (Wang et al., 2019). However, there has been limited progress in visual modality-based languages, like sign languages (also referred to as signed languages: https://en.wikipedia.org/wiki/Sign_language), due to the limited availability of large-scale datasets. Moreover, from a modeling perspective, sign languages are data-hungry due to the complex relationship between different entities in visual modalities like signs, gestures, finger-spelling, and facial expressions. In this paper, to address the lack of a large-scale dataset for ISL processing and to promote the development of sign language processing techniques, we propose **iSign**. In a nutshell, we make the following contributions:

- We introduce **iSign**, a new benchmark for Indian Sign Language processing. Figure 1 provides an overview and design philosophy of the benchmark (inspired from (Gehrmann et al., 2021)).
- We create a dataset with 118, 228 ISL-English video-sentence/phrase pairs. To the best of our knowledge, this is the largest dataset for ISL.
- **iSign** includes 3 standard sign language processing tasks: *SignVideo2Text Translation, SignPose2Text translation, Text2Sign Translation, Sign/Gloss Recognition*. Two additional tasks of *Sign Presence Detection* and *Sign Semantic Similarity Prediction* are introduced. The two additional tasks are added to encourage representation learning and contextualized learning in Signed Languages. We develop baseline models and report results for each of the task. We release the data, tasks, and baseline models (https://exploration-lab.github.io/iSign/).
- We conduct a detailed analysis of the ISL dataset and analyze how it differs from spoken languages. We provide linguistic insights into the functioning of ISL, covering various aspects like structural differences, the significance of non-manual markers, the use of space, the use of fingerspelling and co-reference, and role shifts. We hope that the detailed analysis will open up a new set of computational challenges from the linguistic perspective of ISL and further encourage various research directions.

2 Related Work

Recently, the research community has been actively interested in developing tools and techniques for processing sign languages. Since sign languages contain both visual, gestural, and language modalities, both the vision (Li et al., 2020a) and natural language (Yin et al., 2021) research communities have developed techniques. Several tasks for sign language processing have been proposed, for example, sign language detection (Moryossef et al., 2020), identification (Monteiro et al., 2016), segmentation (Bull et al., 2020), recognition (gloss detection) (Imashev et al., 2020; Sincan and Keles, 2020), generation (Saunders et al., 2020c,b; Xiao et al., 2020; Rastgoo et al., 2022), and translation (Jiang et al., 2023; Müller et al., 2022; Muller et al., 2022; Moryossef et al., 2021; Yin and Read, 2020a,b; Camgoz et al., 2018b, 2020).

Isolated Sign Language Recognition/Gloss Recognition: Many benchmarks have been proposed for gloss recognition (§4) in sign languages other than ISL (Mesch and Wallin, 2012; Fenlon et al., 2015; Gutierrez-Sigut et al., 2016; Martinez et al., 2002b; Zahedi et al., 2005; Efthimiou and Fotinea, 2007; Tavella et al., 2022). There are very few datasets for ISL like Rekha et al. (2011); Nandy et al. (2010); Kishore and Kumar (2012); Selvaraj et al. (2022), INCLUDE dataset (Sridhar et al., 2020), ISL-CSLRT dataset (Elakkiya and Natarajan, 2021), CISLR (Joshi et al., 2022), and ISLTranslate (Joshi et al., 2023). Table 1) provides a comparison with other isolated sign language recognition datasets.

Sign Language Translation Datasets: Various datasets (Yin et al., 2021) for sign language translation have been proposed in recent years for dif-

Datasets	Sign-Language	Words	Videos	Avg. Videos/ Word	Signers	Modalities	Categories
Boston ASLLVD	American	2742	9794	3.6	6	RGB	-
DEVISIGN-L	Chinese	2000	24000	12	8	RGB, depth	-
DGS Kinnect	German	40	3000	75	15	RGB, depth	-
GSL	Greek	20	840	42	6	RGB	-
LAS64	Argentinian	64	3200	50	10	RGB	-
LSE-sign	Spanish	2400	2400	1	2	RGB	-
Perdue RVL-SLLL	American	39	546	14	14	RGB	-
PSL Kinnect 30	Polish	30	300	10	-	RGB, depth	-
RWTH-BOSTON-50	American	50	483	9.7	3	RGB	-
WLASL	American	2000	21,083	10.5	119	RGB	-
Nandy et al. (2010)	Indian	22	600	27.3	-	RGB	-
Kishore and Kumar (2012)	Indian	80	800	10	-	RGB	-
INCLUDE	Indian	263	4287	16.3	7	RGB	15
ISL-CSLRT	Indian	186	700	3.8	7	RGB	-
CISLR (Joshi et al., 2022)	Indian	4765	7050	1.5	71	RGB	57

Table 1: The Indian-Sign Language Dataset comparison with other Sign-Language datasets for Isolated Sign Language Recognition.

Dataset	Language	Sentences	Vocab. (corresponding Text)	Hours
Purdue RVL-SLLL (Martinez et al., 2002a)	ASL	2.5k	104	-
Boston 104 (Dreuw et al., 2007)	ASL	201	103	-
How2Sign (Duarte et al., 2021)	ASL	35k	16k	79
OpenASL (Shi et al., 2022)	ASL	98k	33k	288
YouTube-ASL (Shi et al., 2022)	ASL	610k	60k	984
AfriSign (Gueuwou et al., 2023)	KSL, ZSL, SASL GSL, NSL, ZISL	98k	20k	-
BOBSL (Albanie et al., 2021)	BSL	993k	72k	1447
CSL Daily (Zhou et al., 2021)	CSL	20.6k	2k	23
Phoenix-2014T (Camgoz et al., 2018a)	DGS	8.2k	3K	11
SWISSTXT-Weather (Camgöz et al., 2021)	DSGS	811	1k	-
SWISSTXT-News (Camgöz et al., 2021)	DSGS	6k	10k	-
KETI (Ko et al., 2018)	KSL	14.6k	419	28
VRT-News (Camgöz et al., 2021)	VGT	7.1k	7k	100
ISL-CSLRT (Elakkiya and Natarajan, 2021)	ISL	100	-	-
ISLTranslate (Joshi et al., 2023)	ISL	31k	11k	55
iSign (ours)	ISL	118k	40k	252

Table 2: Comparison of continuous sign language translation datasets. Please refer to the App. Table 7 for details.

ferent sign languages, e.g., ASL (Martinez et al., 2002a; Dreuw et al., 2007; Uthus et al., 2023), Chinese sign language (Zhou et al., 2021), Korean sign language (Ko et al., 2018), Swiss German Sign Language - Deutschschweizer Gebardensprache (DSGS) and Flemish Sign Language - Vlaamse Gebarentaal (VGT) (Camgöz et al., 2021). Table 2 provides a comparison with datasets for other sign languages.

Sign Language Generation: The field of sign language production predominantly revolves around the generation of hand movements, with the majority of existing approaches leveraging GAN architectures and sometimes in combination with transformer architecture (Stoll et al., 2020; Saunders et al., 2022, 2020d; Zelinka and Kanis, 2020; Saunders et al., 2020a).

3 iSign Benchmark

Dataset Creation: For creating **iSign**, we primarily use three publicly available and authentic resources on YouTube: ISLRTC videos[1] (a Government of Indian initiative), ISH News (News channel in ISL),[2] and phrases from DEF (Deaf Enabled Foundation, a non-profit working for the DHH community).[3] These YouTube channels provide permissions to scrape videos and use them for research. Each of these videos contains a single signer communicating information (educational content or news about current affairs) in ISL and a corresponding transcript in English. The videos are pre-processed and split to obtain video-sentence

[1] http://tinyurl.com/mr3v4ead
[2] http://tinyurl.com/4a9xe6rk
[3] http://tinyurl.com/3rzw7xff

Dataset Translations	ISL-Signer Translations (references)
Where are you going to, man? I said.	Where are you going to, man? I asked.
Where are you going this fine day? I said to the puppy.	Where are you going on this special day? I asked the puppy.
and name your dog.	name your dog.
You might begin. I am a little brown dog.	What do you begin? I am a little brown dog.
Not I.	Not, I did not come.
I said to the horse as he went by.	I asked the horse as he went by.
as he went by, up in the hills	I asked the puppy as he went by.
Autobiography	to tell your life story in your own language.
Page 117. It was impossible for me to...	Page 117. The two fences
climb because every step was 6 feet high.	were impossible for me to climb because they were 6 ft high.
60 feet above the ground.	I was high above the ground at
and blew my hair aside to get a better view of my face.	He blew my hair aside to get a better view of my face.
Each minister looked at the line and was puzzled.	Each minister looked at the line and was puzzled.
No one could think of any way to make it longer.	No one could think of any way to make it longer.
I turned back to join the crew.	Gulliver turned back to join his crew.
Dinner was brought for the farmer in a dish.	A farmer brought food in a dish
Land with no vegetation.	meaning is no vegetation.
and some had wells to supply water.	and some had wells to supply water.
My wife and my children.	wife and my children. had covered drains.

Table 3: The Table shows a sample of English translations present in the created dataset compared to sentences translated by ISL Signer. Blue and Red colored text highlight the difference between semi-automatically generated English sentences and gold sentences generated by the ISL instructor.

Metric	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	WER	ROUGE-L	ROUGE-1	ROUGE-2	ROUGE-L-SUM
Score	76.3	73.42	71.2	69.3	73.83	33.83	81.9	82.2	64.4	81.8

Table 4: The Table shows the Translation scores for a sample of 593 sentence pairs from the created dataset when compared to references translated by ISL Signer.

level (or phrase level) pairs. Fig. 2 shows an example from **iSign**. The exact splitting, the preprocessing process, and data statistics are described in the App. A. Since the YouTube channels keep getting populated with new content, we will continue to grow **iSign** with more video-sentence pairs. Note that in the current version of the benchmark, we ignore all the videos where two signers communicate with each other, as in sign language, a two-signer communication might use different sets of reference points for communication. Moreover, in such videos, the orientation of the signer might change completely, resulting in more variation from the visual perspective due to different camera angle placement. We speculate a key-point detection framework might be helpful here as the 3D key points can be normalized to keep the same viewing angle; however, the problem of different gesture/sign usage for change in reference point will remain. We leave the exploration of including multiple signer videos for future work.

Comparison with Existing Datasets: Table 1 shows word-level datasets and Table 2 compares **iSign** with other video-based sign language datasets. For ISL, there are two existing publicly available datasets: ISLTranslate (Joshi et al., 2023) (having $31k$ video-sentence pairs) and CISLR (Joshi et al., 2022) (having 7k video-word pairs) for Isolated Sign Language Recognition. These two datasets were the largest among the previous ISL datasets. We also include these ISL datasets in **iSign**, after due permissions from the authors. It results in $118,228$ video-English sentence/phrase/words pairs in **iSign**. In the past, other works like Uthus et al. (2023); Shi et al. (2022) followed a similar strategy of using YouTube videos with captions for generating translation datasets.

Validation: To verify the reliability of the video-sentence/phrase ISL-English pairs present in the dataset, we took the help of three certified ISL signers. The signers worked with us on a pro-bono basis (details in App. A.2). Due to the limited availability of certified ISL signers, we could only use a small randomly selected sign-text pairs sample (593 pairs) for human translation and validation. We asked ISL instructors to translate the videos. Each video is provided with one reference translation by the signers. Table 3 shows a sample of

sentences created by the ISL instructor. To quantitatively estimate the reliability of the translations in the dataset, we compare the English translation text present in the dataset with the ones provided by the ISL instructors. Table 4 shows the translation scores for 593 sentences in the created dataset. Overall, the BLEU-4 score is 69.3 (indicative of high reliability), ROUGE-L (Lin, 2004) is 81.9, and WER (Word Error Rate) is 33.83. To provide a reference for comparison, for text-to-text translations, the BLEU score of human translations ranges from 30-50 as reported by Papineni et al. (2002). Ideally, it would be better to have multiple reference translations available for the same signed sentence in a video; however, the high annotation effort and the lower availability of certified ISL signers make it a challenging task.

4 iSign Tasks

We propose various tasks in **iSign** to evaluate and compare different models developed for ISL processing. The tasks are described below.

Task 1) ISL-to-English Translation: It is a standard task of translating a sign (source) language to a spoken (target; in text form) language. As done in previous work, we use standard neural machine translation metrics to benchmark the baseline models on this task, including BLEU, METEOR, and ROUGE-L. Input (modality) to the translation system is a video in the form of a sequence of RGB images or pose-based features. There is a significant difference in performance based on the input modality. Hence, we add two subtasks under this task to facilitate the development and comparison of various approaches: **ISLVideo-to-English Translation** and **ISLPose-to-English Translation**. The former uses image-based features as input (Camgoz et al., 2020; Shi et al., 2022; Chen et al., 2022; Cheng et al., 2023), and the latter uses pose-based features in the form of body key points (Uthus et al., 2023; Selvaraj et al., 2022). Image-based approaches have shown promising results for sign language translation tasks. However, image-based architectures are compute-heavy and require more time for inference. Moreover, regarding large-scale application perspective, including images may result in signer-based biases creeping into the model learning. In contrast, extracting body pose features is fast and easy on edge devices; hence, pose-based approaches are more practical.

Task 2) English-to-ISLPose Generation: The goal of this task is to transform textual input into a sequence of body poses that correspond to the sign language representation of the input sentence (Saunders et al., 2020d). Hence, this task aims to generate a sign language video. The generated translations are evaluated using the Dynamic Time Warping (DTW) metric (Müller, 2007). The DTW algorithm measures the alignment between the generated pose sequence and the ground truth, allowing us to assess how well the generated poses match the expected sign language representation.

Task 3) Word/Gloss Recognition (Isolated Sign Recognition): Given a video of a signer (performing gestures and actions), the task is to predict the corresponding gloss label (word). We follow an existing work, CISLR (Joshi et al., 2022), which contains a low number of average videos per word (1.5 videos per word), and formulate the gloss recognition task as a one-shot learning task. Overall, the CISLR task contains 4765 sign video samples, which act as prototypes, and the task is to classify the remaining 2285 videos into one of the 4765 categories. We consider the standard metric of Top-1, Top-5, and Top-10 classification accuracy scores to evaluate this task.

Task 4) Word Presence Prediction: To capture the quality of sign representations learned by algorithms, we define a new task of Word Presence Prediction. Given a pair of a word (as a query) and a sentence (as a candidate) as two signed ISL videos, the task is to predict if the query word is present (used) in the signed sentence. For similarity comparison, we consider the cosine similarity of the representations obtained for the pair of ISL videos. We evaluate the performance over this task using the standard classification accuracy, i.e., if the learned representations are able to predict the word presence given a query and a candidate pair. Since this task can also be treated as a retrieval task, we also consider Top Rank (Avg.) by ranking the entire pool of candidates for a particular query. Note that the better the rank, the better the representations learned by the model. (more details in the App. C)

Task 5) Semantic Similarity Prediction: Given a pair of a word (as a query) and a sentence (as a candidate) as two signed ISL videos, we propose a new task of predicting the semantic similarity of videos. We select the ISL description videos corresponding to an ISL word as the candidate. For example, a sample for this task will contain an ISL video for the word "revision," the corresponding ISL sen-

Task ID	Task Name	# Samples
1.	ISL-to-English Translation	118, 228
2.	English-to-ISL Pose Generation	118, 228
3.	Word/Gloss Recognition	7,050
4.	Word Presence Prediction	1,523
5.	Semantic Similarity Prediction	593

Table 5: Number of samples present in the **iSign** benchmark for various tasks. Note that tasks 4 and 5 are only used for validating the quality of the learned representation and have no trainset.

tence will describe the same word, "a change that is made to something, or the process of doing this." For similarity comparison, we consider the cosine similarity of the representations obtained for the pair of ISL videos. To evaluate the representations learned by the models, we determine if the correct match is present within the top 5% of the cosine similarity scores of the query-candidate pairs. Additionally, we consider the rank metric, indicating the position of the true match within this ranked list. More details about the evaluation metrics can be found in the App. C

Overall, Task 1, Task 2, and Task 3 correspond to standard tasks proposed for processing other sign languages as well; nevertheless, given the scale of data, the tasks are new for ISL. Task 4 and Task 5 are newly introduced in this paper to promote representation learning in ISL. Table 5 summarizes the number of samples present for each task in the **iSign** benchmark.

5 Models, Experiments and Results

Baseline Models: We experimented with various models for the proposed task as described next (model training details provided in App. B).

1) ISL-to-English Translation: For this task, we follow Camgoz et al. (2020) and validate the performance for both the sub-tasks. For ISLVideo-to-English Translation, we use spatial embeddings extracted from pre-trained CNNs as input for our model. For ISLPose-to-English Translation, we follow Saunders et al. (2020d) and create a sequence of poses to act as input to the SLT (Sign Language Transformer) model (Camgoz et al., 2020). For pose key points, we use the Mediapipe pose estimation pipeline (MediaPipe, 2023).

2) English-to-ISLPose Generation: We utilize a transformer-based architecture as introduced in Saunders et al. (2020d) to generate body key points corresponding to the textual input.

3) Word/Gloss Recognition: For this task, we follow CISLR (Joshi et al., 2022) as the baseline. CISLR uses the state-of-the-art model (Inception3D (I3D) (Carreira and Zisserman, 2017)) on the WLASL dataset (Li et al., 2020b) and trains it on 2000 classes. Further, the penultimate layer of the trained model is used to generate features corresponding to the prototype videos in the dataset, and each test sample video is assigned the gloss corresponding to the nearest prototype using cosine similarity between the obtained features.

4) Word Presence Prediction: As the motivation of this task is to validate the representations learned by a neural architecture, we consider a pre-trained I3D, trained on the Human Kinetics Dataset (Kay et al., 2017), as a baseline and report the findings.

5) Semantic Similarity Prediction: We use average cosine similarity scores between the sign representations of the word and their corresponding description to measure the similarity. The representations are obtained using a pre-trained I3D network (Carreira and Zisserman, 2017).

Results: Table 6 shows the baseline results for all the tasks in **iSign**. The BLEU scores for Sign-to-Text translation (Task 1) are a bit low, though we follow the previous baselines (Camgoz et al., 2020) that perform well (with a BLEU score of $\sim$ 20) on the RWTH Phoenix 2014T dataset (Camgoz et al., 2018a). We speculate a high gap between the two datasets to be the primary reason for the observed difference. For Task 2, we obtained a DTW score of 22.69, pointing towards high variation in generated and ground truth poses. For Task 3, we found the I3D features used by Joshi et al. (2022) to be performing better than the representations extracted via the models trained for Task 1. For Task 4 and 5, we compute the Top 5% accuracy by performing a one vs all prediction, i.e., a query's feature representation will have the highest similarity with the corresponding candidate's feature representation. We also report the average rank assigned via similarity corresponding to all the candidates (i.e., the lower the rank, the better the learned representations). For Task 4, we found the T5base + I3D features getting Top-5% Acc. of 52% and average rank as 193 out of 1523 candidates (also check App. Table 9). Overall, all the baseline performance points towards a huge scope of future developments in ISL processing (also see Limitations section).

Poor Performance of Existing Baseline Models: The current baseline models for several tasks

Neural Machine Translation					
Task	**Source Language**	**Target Language**	**Model**	**BLEU-4**	**ROUGE-L**
SignVideo-to-Text	ISL	English	Camgoz et al. (2020)	0.56	**19.58**
SignVideo-to-Text	ISL	English	T5(small)+I3D(2000)	0.24	11.41
SignPose-to-Text	ISL	English	Camgoz et al. (2020)	0.77	9.52
SignPose-to-Text	ISL	English	T5(large)+Mediapipe(75)	0.09	19.11
SignPose-to-Text	ISL	English	T5(small)+Mediapipe(75)	0.8	16.46
SignPose-to-Text	ISL	English	T5(base)+Mediapipe(75)	**1.47**	16.67
SignPose-to-Text	ISL	English	SLT+Mediapipe(75)	0.36	7.60
Sign Language Generation					
Task	**Source Language**	**Target Language**	**Model**	**DTW**	
Text-to-SignPose	English	ISL	Saunders et al. (2020c)	22.69	
Word Level Translation/Gloss Prediction					
Task	**Source Language**	**Target Language**	**Model**	**Acc. (Top-1%)**	**Acc. (Top-5%)**
ISLR	ISL	English/Gloss	Joshi et al. (2022)	16.81	20.04
Sign Representation Learning					
Task	**Query**	**Candidate**	**Model**	**Top 5% Acc.**	**Rank (Avg.)**
Word Presence	word	example sentence	T5-base + I3D	52	193/1523
Semantic Similarity	word	description sentence	T5-base + I3D	67	44/593

Table 6: Results of various baseline models on proposed tasks.

show poor performance, pointing toward developing more sign-language-specific neural architectures in the future. The machine translation baselines, which show SOTA performance on datasets like Phoenix-2014T and CSL-Daily, do not perform well for the created dataset. Most existing approaches for sign language translation (Camgoz et al., 2018a; De Coster and Dambre, 2022; De Coster et al., 2021; Chen et al., 2022) depend on intermediate gloss labels for translations. As glosses are aligned to video segments, they provide fine one-to-one mapping that facilitates supervised learning in learning effective video representations. Previous work has reported a drop of about 10.0 in BLEU-4 scores without gloss labels (Camgoz et al., 2018a). However, considering the annotation cost of gloss-level annotations, it becomes imperative to consider gloss-free sign language translation approaches. Moreover, gloss mapping in continuous sign language might remove the grammatical aspects of sign language. The presence of gloss labels for sign sentences in a dataset helps translation systems work at a granular level of sign translation. However, generating gloss annotations for a signed sentence is an additional challenge due to the scarcity of certified signers. The large number of samples in the created dataset makes the gloss-level annotation infeasible. There are a few recent works on Sign language translation Voskou et al. (2021); Yin and Read (2020c), which try to remove the requirement for a glossing sequence for training and propose a transformer-based architecture for end-to-end translations. Moreover, the noteworthy point is that the dataset size of **iSign** differs from Phoenix-2014T and CSL-Daily by a significant margin (118k vs. 8.2k, 20.6k), making **iSign** more diverse with high variance in gestures/signs, which might also be one of the reasons for poor performance. Furthermore, in a recent work, Müller et al. (2022, 2023) report the current performance of automatic translation systems to be very low compared to the text-based translation systems for spoken languages.

Evaluation Metric for Sign Generation: In the current version of **iSign**, we use the Dynamic Time Warping (DTW) algorithm to evaluate the quality of generated sign pose key points. The DTW algorithm has limitations in dealing with significant variations in motion between the generated and ground truth sequences. Moreover, the DTW may not be suitable for measuring the quality of generated sign language. While designing the evaluation metrics, it is essential to consider sign language's linguistic aspects and ensure they correlate strongly with human evaluation (more details in App. C).

6 ISL Linguistics and Computational Challenges

Sign language functioning differs from spoken languages by a significant margin. In this section,

we highlight some ISL-specific features that might be helpful in the development of dedicated sign-language-specific neural architecture and facilitate understanding of ISL in the NLP research community. We created this list of insights after discussing it with a Professor of Indian Sign Language linguistics, Dr. Andesha Mangla, who is also one of the co-authors of this paper.

Structural Differences: ISL is a form of visual language that consists of signs, gestures, fingerspelling, and facial expressions going in parallel to communicate a sentence, making it quite different from spoken language in the structural form (Sinha, 2017). At a rudimentary level, the building blocks of sign and spoken languages differ. In spoken languages, a combination of sounds results in the formation of words, while in sign languages, a mixture of manual and non-manual parameters forms words. Moreover, the usage of visual-spatial and manual modality in sign languages allows the production of various concepts in parallel. For example, using physical space for multiple purposes, using head, eyes, and body to represent different entities, actions, etc., and using non-manual expressions for various concepts. In terms of linguistics, iconicity plays a more significant role in the production and perception of sentences when compared to spoken languages (Zeshan, 2000; Sinha, 2017; Brentari, 2019).

Significance of Non-Manual Markers: In ISL, non-manual markers like facial expressions, body language, etc., play a vital role in giving semantics to the produced sentences, both at lexical as well as grammatical levels (Sinha, 2017). For example, the word "HAPPY" is signed with a smiling face, whereas the word "SAD" is signed with a sad facial expression. The order of non-manual markers goes in parallel with the manual markers, giving the sentence meaning. For example, the same sentence signed with a forward head tilt and wide-open eyes will transform the statement sentence into a yes-no question. Moreover, non-manual markers are also used in the production of complex sentences, such as conditional sentences. The various use cases and parallel nature of non-manual markers in sign language production make it more challenging for a sequential language-based model to adapt to ISL.

Use of Space in ISL: Physical signing space is crucial in making production and communication more efficient in ISL (Sinha, 2017). Physical singing space provides a medium for assigning various reference points required in a specific sentence, like referring to designated locations for people, places, or any topic/subject. While communicating a sentence, the referents in a narration are assigned various locations in the signing space, which are then referred to in the conversation using the pointing sign toward the same space. Space provides a medium for references and grounds the language to actual space. For example, the word "AEROPLANE" will be signed in the upper portion of the signing space, whereas the word "CAR" will be signed in the lower portion of the signing space. As the references are created specific to sentences for each conversation, the linguistic structure of the language becomes more complex, and the conversation becomes challenging to separate into independent sentences.

Fingerspelling and Co-reference: In ISL, names for characters, places, etc., are produced in various ways. For introducing a new name in the conversation, the name is fingerspelled and simultaneously assigned a short sign consisting of the initials, which refers to the same name in future sentences (Sinha, 2017). For example, a girl character named "Neha" is introduced in the conversation by signing "FEMALE+CHILD (= girl) NAME N-E-H-A (fingerspelled) SIGN SHORT FEMALE+N" at the start of the conversation. Later on, the name Neha is co-referenced with the assigned short sign "FEMALE+N." Note that assigning a short sign is not unique and varies. For example, the same name, Neha, can be given a short sign using visual features and physical characteristics. For example, if there is a picture available of the child Neha, in which she is wearing a pair of spectacles, the sign introduction can look like "FEMALE+CHILD (= girl) NAME N-E-H-A (fingerspelled) SIGN SHORT FEMALE+SPECS," where "FEMALE+SPECS" becomes the assigned short sign. Moreover, other variations can exist, combining the first and second examples to create a short sign "N+SPECS," N coming from the name, and SPECS for the visual feature of spectacles. For co-referencing, ISL also makes use of signing space (Sinha, 2017). For example, a character, place, etc., is given a location in the signing space during the introduction, and co-references are then made by pointing toward that location in space. For introducing new concepts, things, actions, etc., if there is no available sign, the signer describes the concept by fingerspelling it.

Role Shifts in ISL: The visual modality provides signed languages with multiple ways of speaking

the same sentence. This is similar to audio in spoken languages, where a speaker makes use of voice modulations to enact another person's role. In ISL, Signers use role shifts to indicate different entities (Sinha, 2017). In role shift, the signer takes on the roles of the different participants in a narrative and enacts their roles (Lillo-Martin, 2012). The respective roles are indicated by head and shoulder position, eye gaze, and non-manual expressions. This unique property of Role Shifts makes translation more challenging. As in written languages, the role shifts are generally indicated by a pretext like "X said in a soft tone," followed by the respective dialogue.

Demographics and Dialects: ISL, being used in a diverse country like India, incorporates numerous variations due to regional, cultural, and geographic diversities. Some of the anecdotal evidence indicates that eastern regions like West Bengal and southern regions like Tamil Nadu and Kerala have a higher degree of variation when compared with the Delhi (northern region) dialect of ISL (Jepson, 1991; Zeshan, 2003; Johnson and Johnson, 2008; Zeshan et al., 2023). Apart from the location/region of usage, socio-economic factors like age, gender, education, etc., have been shown to affect the dialects of ISL, leading to more variations. Anecdotal evidence indicates that around 75% vocabulary is found to be similar across India, and the variation in the remaining 25% is majorly found in categories like numbers, colors, months, weekdays, kinship terms, food, etc. (Jepson, 1991; Zeshan, 2003; Johnson and Johnson, 2008; Zeshan et al., 2023).

Extended Usage of Verbs and Nouns: Apart from the co-referencing feature discussed in the main paper, the usage of verbs and nouns is expressed in different ways depending on the actual physical characteristics of the objects. For example, ISL has a generic sign for OPEN. However, when talking about specific objects, such as opening a drawer, opening (starting) a computer, opening one's eyes, etc., the sign gets modified depending on the object.

Use of Classifiers: In Indian Sign Language, classifiers describe locations and movements. For example, a classifier for vehicles can be used to depict multiple things like their location, direction of motion, manner of movement, etc. The entity the classifier refers to depends on the context. For example, if the classifier is used after signing CAR, it will refer to CAR, but if it is signed after BUS, it will refer to BUS. The use of classifiers enables sign languages to describe visual details, manner of movement, etc., in greater detail than can be done in spoken languages.

Simultaneous Articulation: The use of multiple articulators in parallel allows a signer to represent different entities simultaneously. For example, the left hand can depict a bridge, and the right hand can show a car moving under the bridge.

We believe the working and insights of sign-language linguistics would promote the incorporation of domain knowledge into computational methods for sign-language processing.

7 Conclusion and Future Directions

We propose **iSign**, a benchmark for Indian Sign Language Processing, to bridge the gap between developments in spoken languages and signed languages and provide a standardization medium for accelerating the improvement. We release one of the largest available datasets for ISL with 118k video-sentence/phrase pair sentences to facilitate research. We believe the released dataset will not only help improve translation models but also open up various ways for advancing natural language modeling techniques like contextualized representation learning, mask language modeling, capturing semantic similarities, etc., and encourage research in the NLP community. As a part of the benchmark, we incorporate various sign language tasks. In the future, we plan to grow the dataset by adding more samples. We also plan to add more ISL-based tasks (e.g., Sign Sentence Retrieval) to the benchmark. Additionally, we provide some linguistic insights into the functioning of ISL and discuss the open challenges in sign language processing. Accordingly, we plan to incorporate linguistic priors into the models.

Acknowledgements

We would like to thank anonymous reviewers for their insightful comments. We would like to thank the Indian Sign Language Research And Training Center (ILSRTC) team for helping us validate the quality of the curated translation dataset. We would also like to extend our immense gratitude towards the ISLRTC members for providing us full support in sharing the ISL content creation process. Finally, we thank the ISL content creators on YouTube (ISH-TV and DEF), who gave us permission to use the videos and thus made this work possible.

Limitations

The large number of ISL-English sentence pairs in the initial version of the dataset makes it challenging to validate. Though we provide a small-scale validation with an ISL instructor's help, the entire corpus's validation is tedious, and it is infeasible to look into minute aspects of ISL-English translations in an initial version. In the future, we plan to extend the validation task and add a gold translation set with multiple references provided for a set of 5K ISL sentences. For an initial version of the benchmark, it is imperative to consider some possible technical constraints in the dataset. We believe addressing these constraints is a long-term goal, and mentioning them in detail will open up various directions for future work on analyzing and improving the benchmark.

Alignment in ISL-English Pairs: As the **iSign** translation dataset consists of sentence-level videos clipped from a longer video using multiple strategies, pauses in the available audio signal for ISLRTC videos, frame pattern heuristics for DEF videos, and available English caption timestamps for ISH videos, there are multiple ways in which the alignment between ISL signing sentence and corresponding English Sentence is disrupted. For example, in ISLRTC videos, though the audio in the background is aligned with the corresponding signs in the video, it could happen in a few cases that the audio was fast compared to the corresponding sign representation and may miss a few words at the beginning or the end of the sentence.

Co-referencing in ISL: As mentioned in Section 6, ISL linguistics involves assigning various short signs for names, places, etc., with various spatial reference points to refer to places defined for a specific conversation, making the context an essential feature for a complete translation of the content. As the translation dataset in **iSign** was created by segmenting a longer video into shorter segments representing a sentence, there is a high possibility that the names in the sentences are introduced in different sentences and assigned a short sign, which may be difficult to refer to in the later sentences. We consider this a major limitation of the created ISL-English pair dataset, as translating the same names would be difficult without a given reference to the created short sign or allocated space, and the independently made translations would result in a lower score in evaluation metrics like BLEU. One way of addressing this challenge would be to perform a task similar to NER on the ISL videos. However, the unavailability of resources and the scarcity of certified signers make ISL video annotation challenging.

Presence of Role Shifts: As a major portion of the created dataset comes from educational content created by ISLRTC, a lot of material contains stories with fictional characters. In stories, there exists a high possibility of Role Shifts when producing sentences spoken by a character. A similar example in spoken storytelling would be the change in voice to articulate a character's voice, for example, the use of a squeaky voice for enacting the sentences of a small mouse. Since these Role Shifts are produced via non-manual markers, they would result in slight gesture variations for the same sentence, making the translated sentence more challenging to predict. Lastly, in this paper, we do not compare ISL with other sign languages (like ASL and DGS); to the best of our knowledge, no such previous study exists, and performing such a study would require a considerable amount of effort in terms of humans having sign language expertise. In the future, we plan to explore such a study, as this can help with cross-model transfer, i.e., adapt models for rich resource sign languages to low resource sign languages.

Ethical Considerations

We create a dataset from publicly available resources without violating copyright. We are not aware of any direct ethical concerns regarding our dataset. Moreover, the dataset involves people of Indian origin and is created mainly for Indian Sign Language translation. The ISL-specific insights are obtained from a professor working primarily on ISL linguistics. The annotations are done by the ISL professor and their team on a pro bono basis.

Please note we do not endorse the use of the benchmark data for non-research (commercial and real-life) applications, and the primary motivation for creating the iSign benchmark is to consolidate all the research happening in parallel for ISL. Moreover, we believe providing a platform by maintaining a common leaderboard for multiple tasks will advance the field with more transparency and reproducibility. Sign language datasets include visual modality with the inclusion of facial expression and non-manual markers being an integral part of language, which does pose privacy challenges. Though the videos in the dataset are in the public domain, the models trained on the dataset may contain signer-specific features and may not generalize to real-world usage.

References

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

Diane Brentari. 2019. *Sign Language Phonology*. Key Topics in Phonology. Cambridge University Press.

Hannah Bull, Michèle Gouiffès, and Annelies Braffort. 2020. Automatic segmentation of sign language into subtitle-units. In *European Conference on Computer Vision*, pages 186–198. Springer.

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

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

Necati Cihan Camgoz, Oscar Koller, Simon Hadfield, and Richard Bowden. 2020. Sign language transformers: Joint end-to-end sign language recognition and translation.

Necati Cihan Camgöz, Ben Saunders, Guillaume Rochette, Marco Giovanelli, Giacomo Inches, Robin Nachtrab-Ribback, and Richard Bowden. 2021. Content4all open research sign language translation datasets. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, page 1–5. IEEE Press.

Joao Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308.

Yutong Chen, Ronglai Zuo, Fangyun Wei, Yu Wu, Shujie Liu, and Brian Mak. 2022. Two-stream network for sign language recognition and translation. In *2022 Neural Information Processing Systems*.

Yiting Cheng, Fangyun Wei, Bao Jianmin, Dong Chen, and Wen Qiang Zhang. 2023. Cico: Domain-aware sign language retrieval via cross-lingual contrastive learning. In *CVPR*.

Mathieu De Coster and Joni Dambre. 2022. Leveraging frozen pretrained written language models for neural sign language translation. *Information*, 13(5).

Mathieu De Coster, Karel D'Oosterlinck, Marija Pizurica, Paloma Rabaey, Severine Verlinden, Mieke Van Herreweghe, and Joni Dambre. 2021. Frozen pretrained transformers for neural sign language translation. In *1st International Workshop on Automated Translation for Signed and Spoken Languages*.

Philippe Dreuw, David Rybach, Thomas Deselaers, Morteza Zahedi, and Hermann Ney. 2007. Speech recognition techniques for a sign language recognition system. In *Proc. Interspeech 2007*, pages 2513–2516.

Amanda Duarte, Shruti Palaskar, Lucas Ventura, Deepti Ghadiyaram, Kenneth DeHaan, Florian Metze, Jordi Torres, and Xavier Giro-i Nieto. 2021. How2Sign: A Large-scale Multimodal Dataset for Continuous American Sign Language. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Eleni Efthimiou and Stavroula-Evita Fotinea. 2007. Gslc: creation and annotation of a greek sign language corpus for hci. In *International Conference on Universal Access in Human-Computer Interaction*, pages 657–666. Springer.

R Elakkiya and B Natarajan. 2021. Isl-csltr: Indian sign language dataset for continuous sign language translation and recognition. *Mendeley Data*.

Ethnalogue. 2022. The Indian Sign Language.

Jordan B Fenlon, Kearsy Cormier, and Adam C. Schembri. 2015. Building bsl signbank: The lemma dilemma revisited. *International Journal of Lexicography*, 28:169–206.

Sebastian Gehrmann, Tosin Adewumi, Karmanya Aggarwal, Pawan Sasanka Ammanamanchi, Anuoluwapo Aremu, Antoine Bosselut, Khyathi Raghavi Chandu, Miruna-Adriana Clinciu, Dipanjan Das, Kaustubh Dhole, Wanyu Du, Esin Durmus, Ondřej Dušek, Chris Chinenye Emezue, Varun Gangal, Cristina Garbacea, Tatsunori Hashimoto, Yufang Hou, Yacine Jernite, Harsh Jhamtani, Yangfeng Ji, Shailza Jolly, Mihir Kale, Dhruv Kumar, Faisal Ladhak, Aman Madaan, Mounica Maddela, Khyati Mahajan, Saad Mahamood, Bodhisattwa Prasad Majumder, Pedro Henrique Martins, Angelina McMillan-Major, Simon Mille, Emiel van Miltenburg, Moin Nadeem, Shashi Narayan, Vitaly Nikolaev, Andre Niyongabo Rubungo, Salomey Osei, Ankur Parikh, Laura Perez-Beltrachini, Niranjan Ramesh Rao, Vikas Raunak, Juan Diego Rodriguez, Sashank Santhanam, João Sedoc, Thibault Sellam, Samira Shaikh, Anastasia Shimorina, Marco Antonio Sobrevilla Cabezudo, Hendrik Strobelt, Nishant Subramani, Wei Xu, Diyi Yang, Akhila Yerukola, and Jiawei Zhou. 2021. The GEM benchmark: Natural language generation, its evaluation and metrics. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 96–120, Online. Association for Computational Linguistics.

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

Eva Gutierrez-Sigut, Brendan Costello, Cristina Baus, and Manuel Carreiras. 2016. Lse-sign: A lexical database for spanish sign language. *Behavior Research Methods*, 48(1):123–137.

Huggingface. 2022. T5.

Alfarabi Imashev, Medet Mukushev, Vadim Kimmelman, and Anara Sandygulova. 2020. A dataset for linguistic understanding, visual evaluation, and recognition of sign languages: The k-rsl. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 631–640.

Jill Jepson. 1991. Urban and rural sign language in india. *Language in Society*, 20(1):37–57.

Zifan Jiang, Amit Moryossef, Mathias Müller, and Sarah Ebling. 2023. Machine translation between spoken languages and signed languages represented in SignWriting. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1706–1724, Dubrovnik, Croatia. Association for Computational Linguistics.

Jane E Johnson and Russell J Johnson. 2008. Assessment of regional language varieties in indian sign language. *SIL Electronic Survey Report*, 6:2008.

Abhinav Joshi, Susmit Agrawal, and Ashutosh Modi. 2023. ISLTranslate: Dataset for Translating Indian Sign Language. In *Findings of the Association for Computational Linguistics (ACL)*. Association for Computational Linguistics.

Abhinav Joshi, Ashwani Bhat, Pradeep S, Priya Gole, Shashwat Gupta, Shreyansh Agarwal, and Ashutosh Modi. 2022. CISLR: Corpus for Indian Sign Language recognition. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. 2017. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*.

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

PVV Kishore and P Rajesh Kumar. 2012. A video based indian sign language recognition system (inslr) using wavelet transform and fuzzy logic. *International Journal of Engineering and Technology*, 4(5):537.

Sang-Ki Ko, Chang Jo Kim, Hyedong Jung, and Choong Sang Cho. 2018. Neural sign language translation based on human keypoint estimation. *ArXiv*, abs/1811.11436.

Dongxu Li, Cristian Rodriguez, Xin Yu, and Hongdong Li. 2020a. 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.

Dongxu Li, Cristian Rodriguez, Xin Yu, and Hongdong Li. 2020b. 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.

D Lillo-Martin. 2012. 17. utterance reports and constructed action. insign language: An international handbook, r. pfau, m. steinbach and b. woll (eds), 365–387.

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.

Aleix M. Martinez, Ronnie B. Wilbur, Robin Shay, and Avinash C. Kak. 2002a. Purdue rvl-slll asl database for automatic recognition of american sign language. *Proceedings. Fourth IEEE International Conference on Multimodal Interfaces*, pages 167–172.

Aleix M. Martinez, Ronnie B. Wilbur, Robin Shay, and Avinash C. Kak. 2002b. Purdue rvl-slll asl database for automatic recognition of american sign language. *Proceedings. Fourth IEEE International Conference on Multimodal Interfaces*, pages 167–172.

MediaPipe. 2023. MediaPipe Holistic.

Johanna Mesch and Lars Wallin. 2012. From meaning to signs and back:lexicography and the swedish sign language corpus. In *LREC 2012*.

Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heinz, and Dan Roth. 2021. Recent advances in natural language processing via large pre-trained language models: A survey. *arXiv preprint arXiv:2111.01243*.

Caio DD Monteiro, Christy Maria Mathew, Ricardo Gutierrez-Osuna, and Frank Shipman. 2016. Detecting and identifying sign languages through visual features. In *2016 IEEE International Symposium on Multimedia (ISM)*, pages 287–290. IEEE.

Amit Moryossef, Ioannis Tsochantaridis, Roee Aharoni, Sarah Ebling, and Srini Narayanan. 2020. Real-time sign language detection using human pose estimation. In *European Conference on Computer Vision*, pages 237–248. Springer.

Amit Moryossef, Kayo Yin, Graham Neubig, and Yoav Goldberg. 2021. Data augmentation for sign language gloss translation. *arXiv preprint arXiv:2105.07476*.

Mathias Müller, Sarah Ebling, Eleftherios Avramidis, Alessia Battisti, Michèle Berger, Richard Bowden, Annelies Braffort, Necati Cihan Camgöz, Cristina España-bonet, Roman Grundkiewicz, Zifan Jiang, Oscar Koller, Amit Moryossef, Regula Perrollaz, Sabine Reinhard, Annette Rios, Dimitar Shterionov, Sandra Sidler-miserez, and Katja Tissi. 2022. Findings of the first WMT shared task on sign language

translation (WMT-SLT22). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 744–772, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Mathias Muller, Zifan Jiang, Amit Moryossef, Annette Rios Gonzales, and Sarah Ebling. 2022. Considerations for meaningful sign language machine translation based on glosses. *ArXiv*, abs/2211.15464.

Mathias Müller, Zifan Jiang, Amit Moryossef, Annette Rios, and Sarah Ebling. 2023. Considerations for meaningful sign language machine translation based on glosses. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 682–693, Toronto, Canada. Association for Computational Linguistics.

Meinard Müller. 2007. Dynamic time warping. *Information retrieval for music and motion*, pages 69–84.

Anup Nandy, Jay Shankar Prasad, Soumik Mondal, Pavan Chakraborty, and Gora Chand Nandi. 2010. Recognition of isolated indian sign language gesture in real time. In *International conference on business administration and information processing*, pages 102–107. Springer.

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.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision.

Razieh Rastgoo, Kourosh Kiani, Sergio Escalera, Vassilis Athitsos, and Mohammad Sabokrou. 2022. All You Need In Sign Language Production. ArXiv:2201.01609 [cs].

J Rekha, J Bhattacharya, and S Majumder. 2011. Shape, texture and local movement hand gesture features for indian sign language recognition. In *3rd international conference on trendz in information sciences & computing (TISC2011)*, pages 30–35. IEEE.

B. Saunders, N. Camgoz, and R. Bowden. 2022. Signing at scale: Learning to co-articulate signs for large-scale photo-realistic sign language production. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5131–5141, Los Alamitos, CA, USA. IEEE Computer Society.

Ben Saunders, Necati Cihan Camgöz, and Richard Bowden. 2020a. Adversarial training for multi-channel sign language production. *CoRR*, abs/2008.12405.

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

Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. 2020c. Progressive transformers for end-to-end sign language production. In *European Conference on Computer Vision*, pages 687–705. Springer.

Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. 2020d. Progressive Transformers for End-to-End Sign Language Production. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Prem Selvaraj, Gokul Nc, Pratyush Kumar, and Mitesh Khapra. 2022. OpenHands: Making sign language recognition accessible with pose-based pretrained models across languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2114–2133, Dublin, Ireland. Association for Computational Linguistics.

Bowen Shi, Diane Brentari, Greg Shakhnarovich, and Karen Livescu. 2022. Open-domain sign language translation learned from online video.

Ozge 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.

Samar Sinha. 2017. *Indian Sign Language: An Analysis of Its Grammar*. Gallaudet University Press.

Advaith Sridhar, Rohith Gandhi Ganesan, Pratyush Kumar, and Mitesh Khapra. 2020. *INCLUDE: A Large Scale Dataset for Indian Sign Language Recognition*, page 1366–1375. Association for Computing Machinery, New York, NY, USA.

Stephanie Stoll, Necati Cihan Camgöz, Simon Hadfield, and Richard Bowden. 2020. Text2sign: Towards sign language production using neural machine translation and generative adversarial networks. *Int. J. Comput. Vis.*, 128(4):891–908.

Federico Tavella, Viktor Schlegel, Marta Romeo, Aphrodite Galata, and Angelo Cangelosi. 2022. WLASL-LEX: a dataset for recognising phonological properties in American Sign Language. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 453–463, Dublin, Ireland. Association for Computational Linguistics.

David Uthus, Garrett Tanzer, and Manfred Georg. 2023. Youtube-ASL: A large-scale, open-domain american sign language-english parallel corpus. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Saurabh Varshney. 2016. Deafness in india. *Indian journal of otology*, 22(2).

Andreas Voskou, Konstantinos P. Panousis, Dimitrios I. Kosmopoulos, Dimitris N. Metaxas, and Sotirios P. Chatzis. 2021. Stochastic transformer networks with linear competing units: Application to end-to-end

sl translation. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11926–11935.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32.

World Health Organization WHO. 2016. World hearing day 2023.

Qinkun Xiao, Minying Qin, and Yuting Yin. 2020. Skeleton-based chinese sign language recognition and generation for bidirectional communication between deaf and hearing people. *Neural networks*, 125:41–55.

Kayo Yin, Amit Moryossef, Julie Hochgesang, Yoav Goldberg, and Malihe Alikhani. 2021. Including signed languages in natural language processing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7347–7360, Online. Association for Computational Linguistics.

Kayo Yin and Jesse Read. 2020a. Attention is all you sign: sign language translation with transformers. In *Sign Language Recognition, Translation and Production (SLRTP) Workshop-Extended Abstracts*, volume 4.

Kayo Yin and Jesse Read. 2020b. Better sign language translation with stmc-transformer. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5975–5989.

Kayo Yin and Jesse Read. 2020c. Better sign language translation with STMC-transformer. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5975–5989, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Morteza Zahedi, Daniel Keysers, Thomas Deselaers, and Hermann Ney. 2005. Combination of tangent distance and an image distortion model for appearance-based sign language recognition. In *Deutsche Arbeitsgemeinschaft für Mustererkennung Symposium*, volume 3663 of *Lecture Notes in Computer Science*, pages 401–408, Vienna, Austria.

Jan Zelinka and Jakub Kanis. 2020. Neural sign language synthesis: Words are our glosses. In *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 3384–3392.

Ulrike Zeshan. 2000. *Sign Language in Indo-Pakistan: A description of a signed language*. John Benjamins.

Ulrike Zeshan. 2003. Indo-pakistani sign language grammar: a typological outline. *Sign Language Studies*, pages 157–212.

Ulrike Zeshan, Nirav Pal, Deepu Manavalamamuni, Ankit Vishwakarma, Sibaji Panda, Jagdish Choudhary, and Inu Aggarwal. 2023. *Indian Sign Language*. National Institute of Open Schooling.

Hao Zhou, Wen gang Zhou, Weizhen Qi, Junfu Pu, and Houqiang Li. 2021. Improving sign language translation with monolingual data by sign back-translation. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1316–1325.

Appendix

Table of Contents

List of Figures

A iSign Dataset Details

We plan to release the dataset along with a benchmark webpage to maintain the leaderboard of existing approaches to the proposed tasks.

A.1 Dataset Creation Details

To create the translation dataset with ISL-English video-sentence/phrase pairs, we primarily use three publicly available resources: ISLRTC videos,[4] ISH News (News channel in ISL),[5] and phrases from DEF (Deaf Enabled Foundation).[6]. These YouTube channels provide permission to scrape videos and use them for research. The scraped videos contained a signer communicating a sentence to the spectators. We obtain a list of 114, 3373, and 651 videos from ISLRTC, ISH, and DEF, respectively. These videos generally range from 15-20 minutes and contain about 20-25 number of sentences on average from various sources. To create a video-sentence/phrase level pair dataset, we further split the long videos at the sentence level. For ISH videos, the source videos provide an English caption aligned with the video where the same content is present in the English language as subtitles. We make use of these timestamps to split the videos into sign sentences and combine them with respective captions for English translation. For ISLRTC videos, since the subtitles were not present, we used the available English audio to generate the respective translations. At the end of each sentence, there is a pause in audio and the signer's gesture; we clip the videos using an audio heuristic to create sentence-level segmentation of the signed video. For generating the respective transcripts, we use a speech-to-text model (Whisper (Radford et al., 2022)) for generating the text. Note that the speech-to-text model might not be 100% accurate, resulting in some noisy texts for a few audio clips. Hence, we further clean the generated text via manual inspection by listening to the audio where the sentences are noisy and make less contextual sense. DEF videos provide a word-of-the-day format where a word is communicated at the start of the video with its explanation in the middle and some example sentences where the word is being used at the end. We clip all these sections to gener-

[4] https://www.youtube.com/@islrtcnewdelhi4069/playlists

[5] https://www.youtube.com/channel/UC99w_Bzj8ikOz8Gpv0prbNg

[6] https://www.youtube.com/channel/UCM7U7CyJGRIbu4qsmSh3UZg

ate 3 sets of clippings from a video, namely words, word descriptions, and examples. Note that a few videos have multiple example sentences available for the same word. We segment each example sentence as a different entry in the created dataset.

Data Cleaning and Preprocessing: The videos (e.g., Fig. 4) contain the pictures corresponding book pages. We crop the signer out of the video by considering the face location of the first frame as the reference point and removing the remaining background in the videos. Further, the cropped videos are used to extract pose key points (Figure 5 and Figure 6).

Data Statistics: We had a total of 118, 228 pairs of video-translation. We used T5-tokenizer (Huggingface, 2022) for tokenizing the sentences, there were 15 tokens on an average for each translation text. For the words, we used " "(space) as the separating character. More details about the distribution can be found in Figure 3 and Table 8.

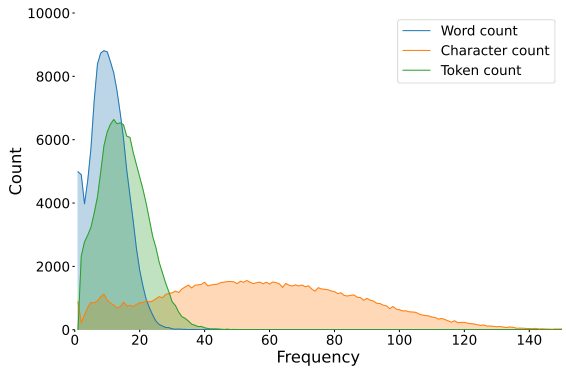

Figure 3: The figure shows the distribution of the number of words in the target translation text

Abbreviation	Language Name
DGS	Deutsche Gebärdensprache
CSL	Chinese Sign Language
BSL	British Sign Language
ASL	American Sign Language
GSL	Ghanaian Sign Language
NSL	Nigerian Sign Language
KSL	Kenyan Sign Language
ZSL	Zambian Sign Language
ZISL	Zimbabwean Sign Language
SASL	South African Sign Language
ISL	Indian Sign Language
DSGS	Deutschschweizer Gebärdensprache (Swiss German Sign Language)
VGT	Vlaamse Gebarentaal (Flemish Sign Language)

Table 7: Sign Language Abbreviations and full forms corresponding to Table 2.

Tokens	Average	15
	Minimum	2
	Maximum	146
	Median	14
	90^{th} percentile	25
Words	Average	10
	Minimum	1
	Maximum	120
	Median	10
	90^{th} percentile	17
Characters	Average	57
	Minimum	1
	Maximum	622
	Median	56
	90^{th} percentile	98
Frames	Average	215
	90^{th} percentile	371

Table 8: Statistics about the dataset: for the 118, 228 Video-Sentence/Phrase Pairs. The first four rows highlights the stats obtained from english phrase text sentence and the last row represents the stats for the corresponding Video-Sentence

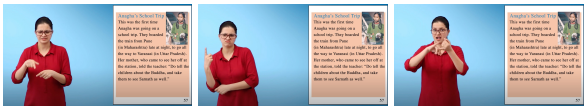

Figure 4: The figure shows an example of the educational content video where the signer signs for the corresponding textbook.

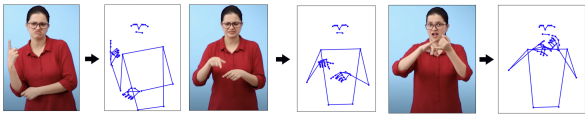

Figure 5: The figure shows an example of a frame from the iSign dataset video with the corresponding extracted keypoints.

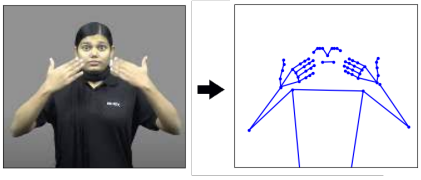

Figure 6: The figure shows an example of a frame from the CISLR dataset video with the corresponding extracted keypoints.

A.2 Annotation Details

We asked three certified ISL instructors to translate and validate a random subset from the dataset (discussed in Section 3). One of the instructors is an assistant professor of sign language linguistics. All the instructors are employed with ISLRTC, the organization involved in creating the sign language content; however, the instructors did not participate in videos present in the translation dataset. The instructors performed the validation voluntarily. It

Word Level Translation/Gloss Prediction					
Task	**Source Language**	**Target Language**	**Model**	**Acc. (Top-1%)**	**Acc. (Top-5%)**
ISLR	ISL	English/Gloss	I3D	10	32.75
ISLR	ISL	English/Gloss	T5-small+I3D	15	36.98

Sign Representation Learning					
Task	**Query**	**Candidate**	**Model**	**Top 5% Acc.**	**Rank (Avg.)**
Word Presence	word	Ex. Sent.	I3D	45	233/1523
			T5-small+I3D	48	219/1523
			T5-small+Mediapipe(75)	42	244/1523
			T5-base + Mediapipe(75)	52	198/1523
			T5-large + Mediapipe(75)	43	237/1523
Semantic Similarity	word	Descrip. Sent.	I3D	63	59/593
			T5-small+I3D	46	169/593
			T5-small+Mediapipe(75)	47	137/593
			T5-base+Mediapipe(75)	67	44/593
			T5-large+Mediapipe(75)	53	96/593

Table 9: Results of various baseline models on Task 3, 4 and 5. Ex. Sent. refers to Example Sentence and Descrip. Sent. refers to Description Sentence.

took the instructor about 3 hours to validate 100 sentences. They generated the English translations by looking at the video.

B Models and Experiment Details

Pose Keypoint Extraction Pipeline: We use the Mediapipe pose estimation pipeline.[7] For the choice of holistic key points, we follow Selvaraj et al. (2022), which returns the 3D coordinates of 75 key points (excluding the face mesh). Figure 5 and Figure 6 shows an example of the obtained 75 keypoints from the mediapipe pipeline. Further, we normalize every frame's key points by placing the midpoint of shoulder key points to the center and scaling the key points using the distance between the nose key point and the shoulders midpoint.

Data Splits: For Task-1 and Task-2, we use a split of 80%, 10%, and 10% for train, validation, and test set, respectively. For Task-3, we follow CISLR (Joshi et al., 2022) and use 4765 samples as prototypes and the remaining 2285 videos as test sets. For Task-4 and Task-5, we take 594 and 1525 positive pairs, respectively to report the results.

Hyperparameters and Training: We follow the code base of SLT (Camgoz et al., 2020) to train and develop the proposed SLT-based pose-to-text architecture by modifying the input features to be sign-pose sequences generated by the mediapipe. The model architecture is a transformer-based encoder-decoder consisting of 3 transformer layers each for both encoder and decoder. We use the Adam optimizer (Kingma and Ba, 2014) with a learning rate of 0.0001, $\beta = (0.9, 0.999)$ and weight decay of 0.0001 for training the proposed baseline with a batch size of 32. The architecture has 14,337,264 trainable parameters . For the generation task, we follow the code base released by (Saunders et al., 2020d). The model architecture is a transformer-based encoder-decoder consisting of 2 transformer layers for both the encoder and decoder. We use the Adam optimizer (Kingma and Ba, 2014) with a learning rate of 0.001 for training the proposed baseline with a batch size of 8. The architecture contains 3,67,68,768 trainable parameters. We perform all the experiments using the NVIDIA A40 GPU machine.

We follow the T5 based model for sign language translation. We use the AdamW optimiser with a learning rate of 0.0001, and weight_decay=$1e-6$. The batch size is 8 for a model with t5-large as backbone and 32 for a model with t5-small and base as backbones. We did transfer learning by freezing only the layers of T5 and finetune the embedding only upto 40 epochs, after that we train the whole model with reduced learning rate.

C Evaluation Metrics for Translation Tasks

In this section, we discuss some of the limitations of the current evaluation criteria for translation tasks and point toward the scope of building sign-language-specific evaluation methods in the future.

[7] https://ai.googleblog.com/2020/12/mediapipe-holistic-simultaneous-face.html

We further provide the evaluation metric details of the additional tasks of Word Presence Prediction and Semantic Similarity Predictions.

Sign Language to Spoken/Texual Language Translation: In the current version of the benchmark, we use the standard translation metrics, which have shown effective usage and interpretation when working with written textual languages to textual language translations. However, in sign languages, the high use of co-referencing (explained in Section 6) makes the translations more challenging to divide into independent sentences. Current evaluation metrics like BLEU and ROUGE depend on the textual translation match between the reference-candidate pair translations. By taking a geometrical average over multiple values of n, BLEU scores give a measure to compute textual similarities between the reference-candidate pair translations. However, as the translated sentences depend on the co-reference symbol/gesture assigned in the past sentences of conversation, the exact independent generation for a sentence in a conversation becomes impossible. One way to overcome this issue is to perform a NER (Named-entity recognition) over the text translations to remove all the references to names and assign the same token for training the neural architectures. However, the co-reference usage for various things and concepts still needs to be solved. In the future, we plan to study and understand the co-reference feature of sign language in more detail, examining if a clear distinction or pattern exists between the usage of fingerspelling, new concepts, and location assigning for later use and how various tags can be introduced in the respective textual translations to make the sentences independent, facilitating the current neural translation pipelines working on language sentence pairs.

Spoken/Texual Language to Sign Language Translation: Sign language generation is a more challenging task in terms of evaluation. In the current version of the benchmark, we use DTW scores to evaluate the quality of generated poses. Dynamic Time Warping, or DTW, tries to capture the similarity between two temporal sequences, takes care of the varying speed of the temporal sequences, and has shown practical usage in speech processing. However, for Indian Sign Language, the significant use of non-manual markers like facial expressions, body language, etc., makes the DTW a less effective metric for capturing the quality of translations. One way of dealing with these would be to develop weighing criteria for various body keypoints associated with non-manual markers or detect specific action units for various gestures produced via non-manual markers. Due to a wide variety of gestures and sign production styles, designing a metric for judging the quality of translations becomes more challenging. In the past, Saunders et al. (2020c) have proposed using back translation scores for judging the quality of generated keypoint poses. However, considering the low performance of machine translation systems for sign languages, the back-translation scores become less reliable. Moreover, a large-scale human evaluation study is needed to judge the effectiveness of various evaluation metrics, which currently needs to be added to the ISL literature.

Word Presence Prediction: In this task, we predict whether a specific query word is present in a candidate sign video. The primary evaluation metric for this task is the standard classification accuracy, which measures the model's ability to correctly predict the presence or absence of the query word within the candidate sentence. Additionally, since this task can be viewed as a retrieval task where the goal is to retrieve the correct candidate from a pool of candidates for a given query, another evaluation metric called Top Rank (Avg.) is considered. Top Rank (Avg.) involves ranking the entire pool of candidate sentences for a particular query and computing the average rank of the true candidate. A lower average rank signifies better performance, indicating that the model can retrieve the correct candidate more accurately from the pool.

Semantic Similarity Prediction: To evaluate the representations learned by the models for semantic similarity prediction, we use the top 5% accuracy and rank metrics. The top 5% accuracy determines if the true semantic match (the candidate video describing the query word) is ranked within the top 5% of the similarity scores obtained for all candidate videos. Achieving a high Top 5% accuracy indicates the model's effectiveness in accurately identifying the most semantically similar candidate videos for a given query. The rank specifies the position of the true semantic match within the ranked list of candidate videos based on the similarity scores. A lower rank indicates that the true semantic match is positioned higher in the list, signifying a better performance of the model in identifying highly similar candidate videos for the query.