J-Shuwa: A Large-Scale Web-Collected Japanese Sign Language-Japanese Parallel Corpus

Junwen Mo
The University of Tokyo
mo@nlab.ci.i.u-tokyo.ac.jp

Duc Minh Vo*
SB Intuitions
minh.duc.vo@sbintuitions.co.jp

Noriki Nishida
RIKEN
noriki.nishida@riken.jp

Shin'ichi Satoh
National Institute of Informatics
satoh@nii.ac.jp

Hideki Nakayama†
The University of Tokyo
nakayama@ci.i.u-tokyo.ac.jp

Abstract

Japanese Sign Language (JSL) is a low-resource sign language that has received limited attention in the AI research community, primarily due to the lack of large-scale, publicly available parallel corpora. In this work, we introduce J-Shuwa, a large-scale JSL-Japanese parallel corpus constructed from YouTube videos with hard-coded subtitles and closed captions. The corpus contains 197K parallel JSL-Japanese sentence pairs, totaling approximately 300 hours of video, making it the largest publicly available JSL dataset to date. We conduct sign language translation (SLT) experiments by training models on J-Shuwa and evaluating them on the JSL Dialogue Corpus under both zero-shot and fine-tuned settings. Our results demonstrate that J-Shuwa is effective for training SLT models. Beyond SLT, we believe that J-Shuwa can also serve as a valuable resource for future JSL research across a wide range of tasks. The dataset and code are publicly available at: https://github.com/SpaJune/J-Shuwa.

Figure 1: A synthetic example of right-to-left vertical subtitles with diverse watermarks and text on clothing, including Furigana annotations. Text is recognized using PaddleOCR (Cui et al., 2025). The background image is generated using ChatGPT (OpenAI) (OpenAI, 2025).

1 Introduction

Sign languages are visual languages with unique grammar that utilize spatial features to convey information, making them fundamentally distinct from spoken languages. To bridge the communication gap between hearing and deaf communities, research on sign language understanding has attracted increasing attention in the AI community (Bragg et al., 2019), including tasks such as Sign Language Recognition (SLR) (Koller et al., 2015; Li et al., 2020) and Sign Language Translation (SLT) (Camgöz et al., 2018; Zhou et al., 2021).

However, progress in sign language research is still limited, mainly because of data scarcity (Tan et al., 2024; Coster et al., 2024). To address this issue, recent work has focused on constructing large-scale sign language corpora from public platforms, primarily YouTube. Representative examples include YouTube-ASL (Uthus et al., 2023) and YouTube-SL-25 (Tanzer and Zhang, 2025). Notably, YouTube-SL-25 covers more than 25 sign languages, substantially expanding available resources for multilingual sign language research.

Nevertheless, Japanese Sign Language (JSL) still suffers from a lack of sufficient data resources. For example, YouTube-SL-25 contains only 62 hours of JSL content, which is insufficient to support robust model training and further research. This scarcity highlights the need for data expansion. Notably, YouTube-SL-25 includes only videos with closed captions (CCs). In contrast, we observe that a large number of JSL videos suitable for parallel corpus construction employ hard-coded subtitles (Hard-

*This work was conducted during the author's affiliation with The University of Tokyo.

†Corresponding author.

Findings of the Association for Computational Linguistics: ACL 2026, pages 36559–36574

Subs), where subtitles are embedded directly into video frames, which can still be utilized.

Processing HardSub videos requires inferring subtitle timing and textual content directly from visual data, making them significantly more challenging to handle than CC videos. Moreover, HardSub videos exhibit considerable variability across uploaders (see Figure 1), with subtitles appearing at different spatial locations and featuring diverse fonts and visual styles. Japanese-specific characteristics further exacerbate these challenges, including right-to-left vertical text and the frequent use of Furigana (phonetic annotations). These factors substantially increase the difficulty of accurately identifying suitable videos for corpus construction and reliably extracting subtitle text. As a result of these technical challenges, HardSub videos have been largely overlooked in prior work.

Motivated by these observations, we mine publicly available videos to construct a large-scale JSL-Japanese parallel corpus by leveraging both CC and HardSub videos, namely **J-Shuwa** (*Shuwa (手話)* is the Japanese term for *sign language*).

Our data construction pipeline begins by collecting large-scale raw candidate videos from YouTube. We employ action recognition models to filter videos whose primary content involves sign language. Videos containing CCs are directly identified through metadata. To detect HardSub videos, we apply optical character recognition (OCR) tools to identify text in common subtitle regions within video frames. Once HardSub videos are detected, we use VideoSubFinder[1] to locate subtitle segments and extract corresponding subtitle screenshots. An OCR system is then applied to these screenshots to extract the subtitle text.

To further improve data quality, we apply a series of heuristic-based filters to remove obviously erroneous samples, such as segments with extremely short durations, unnatural recognized text, and severe mismatches between subtitle length and display duration. Videos containing such samples are subsequently manually inspected and either retained or discarded. Overall, the dataset is constructed through a combination of automated processing and selective manual verification.

Our contributions can be summarized as follows:

- We introduce **J-Shuwa**, a large-scale, publicly available JSL-Japanese parallel corpus. Our corpus contains approximately 197K video-text pairs, amounting to nearly 300 hours of data, enabling further research on JSL.

- We conduct a study of SLT on the JSL Dialogue Corpus (Bono et al., 2014; Bono and Osugi, 2026) under different experimental setups, including zero-shot evaluation and in-domain adaptation via fine-tuning, demonstrating both the effectiveness and current limitations of the proposed corpus.

The YouTube video IDs and segment timestamps, along with the source code for subtitle text extraction, are publicly available[2]. We hope this corpus will encourage the community to devote greater attention to JSL research.

2 Related Work

2.1 Sign Language Understanding for JSL

Sign language understanding for JSL has received relatively limited attention within the AI community. Existing studies have primarily focused on fingerspelling recognition or a small set of predefined gestures with fixed meanings (e.g., *good morning*), using both sensor-based approaches (Ji et al., 2022; Chu et al., 2021) and vision-based methods (Shin et al., 2024; Murai et al., 2025). While these works have contributed to low-level recognition, fingerspelling alone is insufficient for comprehensive semantic understanding of JSL.

Multiple corpora have been developed for JSL linguistic research, including the JSL Colloquial Corpus (Bono et al., 2014; Bono and Osugi, 2026) and the Kogakuin University Japanese Sign Language Multi-Dimensional Database (KoSign) (Nagashima, 2021). These corpora provide high-quality, fine-grained annotations of JSL utterances at both the sentence and word levels. However, their relatively limited scale makes them unsuitable as training datasets. This scarcity of sufficiently large and publicly available data has significantly hindered the development of sign language understanding models for JSL and partly explains the limited research activity in this area.

Our work aims to bridge this gap by constructing a larger-scale JSL-Japanese parallel corpus as a resource for sign language understanding research.

[1] https://sourceforge.net/projects/videosubfinder/

[2] https://github.com/SpaJune/J-Shuwa

Dataset	Language	Vocab.	Dur.	# Sentences	Sources
RWTH-PHOENIX-2014T (Camgöz et al., 2018)	DGS	3K	11	8.2K	TV
KETI (Ko et al., 2019)	KSL	10K	419	14.6K	Lab
BOBSL (Albanie et al., 2021)	BSL	77K	1,447	-	TV
CSL-Daily (Zhou et al., 2021)	CSL	2K	23	20.6K	Lab
How2Sign (Duarte et al., 2021)	ASL	16K	79	35K	Lab
SP-10 (Yin et al., 2022)	various	17K	14	11K	Web
OpenASL (Shi et al., 2022)	ASL	33K	288	98K	Web
Auslan-Daily (Shen et al., 2023)	Auslan	14K	45	25K	TV&Web
ISLTranslate (Joshi et al., 2023)	ISL	11K	-	31K	Web
AfriSign (Gueuwou et al., 2023b)	various	20K	152	-	Web
JWSign (Gueuwou et al., 2023a)	various	729K	2,530	-	Web
YouTube-ASL (Uthus et al., 2023)	ASL	60K	984	610K	Web
YouTube-SL-25 (Tanzer and Zhang, 2025)	various	-	3,207	-	Web
CSL-News (Li et al., 2025)	CSL	5K	1,985	-	TV
iLSU-T (Stassi et al., 2025)	LSU	38K	201	86K	TV
YouTube-SL-25 (JSL data)	JSL	13K*	62	40K*	Web
J-Shuwa (Ours)	JSL	31K	~300	197K	Web

Table 1: Statistics of existing SLT datasets and parallel corpora. Vocab. denotes vocabulary size, and Dur. denotes duration in hours. Information is taken from the original papers unless otherwise specified. * indicates statistics computed by the authors.

2.2 SLT Dataset

Various datasets have been proposed in recent years to facilitate SLT research. PHOENIX14T (Camgöz et al., 2018) and CSL-Daily (Zhou et al., 2021) are well-established datasets containing carefully annotated video-text-gloss triplets for German Sign Language (DGS) and Chinese Sign Language (CSL), respectively. How2Sign (Duarte et al., 2021) provides an ASL dataset derived from the original How2 dataset (Sanabria et al., 2018). While these datasets offer high annotation quality, they are difficult to scale due to the substantial involvement required from sign language experts and Deaf annotators.

To address this scalability challenge, recent efforts have explored leveraging publicly available online data. OpenASL (Shi et al., 2022) collects ASL videos from YouTube. YouTube-ASL (Uthus et al., 2023) further expands coverage to 984 hours of ASL content. For other sign languages, ISLTranslate (Joshi et al., 2023) introduces a corpus for Indian Sign Language (ISL) mined from a few YouTube channels. Auslan-Daily (Shen et al., 2023) constructs an Australian Sign Language (Auslan) dataset by curating publicly available TV programs and online resources, combined with extensive expert annotations. More broadly, YouTube-SL-25 (Tanzer and Zhang, 2025) aggregates 3,207 hours of sign language videos across over 25 sign languages. A more comprehensive comparison of existing datasets is provided in Table 1.

In contrast, publicly available resources for JSL remain extremely limited. For example, YouTube-SL-25 includes only 62 hours of JSL content. In this work, we aim to substantially expand the scale of JSL-Japanese parallel data by mining HardSub and CC videos from YouTube, enabling broader coverage and increased data diversity.

3 J-Shuwa Corpus

J-Shuwa is a JSL-Japanese parallel corpus constructed from YouTube videos containing JSL content with either CCs or HardSubs. In this section, we describe our corpus construction pipeline, including identifying sign language videos suitable for parallel corpus construction and the video processing procedure. We then present efforts to improve the scalability by creating a subtitle-sign segment detection model that integrates HardSub video identification and segmentation. In addition, we explore the use of Large Vision Language Models (LVLMs) to generate cleaner subtitle text.

3.1 Stage 1 - Video Collection and Coarse Filtering

We first identify YouTube channels that produce content relevant to JSL, resulting in approximately 464 channels. These channels span several categories, including content created by deaf individuals, sign language instructional videos, and local government broadcasts. We then collect all videos

from each identified channel for subsequent data filtering, totaling around 72K videos.

We remove low-quality videos using a procedure similar to (Uthus et al., 2023). Specifically, we retain videos that satisfy the following criteria: (1) duration $\geq$ 15 seconds, (2) frame rate $\geq$ 20 FPS, (3) height $\geq$ 360, and (4) width $\geq$ height. We also filter out videos or channels that cover specific topics, such as sign language songs. These videos are performance-oriented, typically lack grammatical structure, and mainly consist of signing individual words to represent Japanese sentences.

The remaining videos are further screened using an action recognition model. We employ the SlowFast model (Feichtenhofer et al., 2019) pretrained on the Kinetics-400 dataset (Kay et al., 2017). For each video, we sample multiple 10-second segments and predict action labels for each segment. The number of sampled segments is proportional to the video duration, with a maximum of 10 segments per video. A video is retained if at least 60% of its sampled segments are classified as *Sign Language Interpreting*. Moreover, if a channel exhibits a high proportion of videos identified as sign language content, we also retain the remaining videos from that channel for further processing.

3.2 Stage 2 - HardSub Videos Identification and Processing

The retained videos can be roughly divided into three categories: **videos without subtitles**, **HardSub videos**, and **CC videos**. To construct a parallel corpus, we need to identify the latter two categories. CC videos are straightforward to identify and process from metadata. Our primary effort is devoted to identifying HardSub videos that can be utilized for corpus construction.

Candidate HardSub videos identification. We apply OCR to detect the presence of text in typical subtitle regions, such as the bottom or right areas of the video frames. For each video, we sample a subset of frames and compute the proportion of frames in which text is detected. Videos with a low proportion of detected text are filtered out, and the remaining videos are retained as candidates for further processing.

Video segmentation and text extraction. To segment videos with HardSubs, we rely on VideoSubFinder, a subtitle detection tool that identifies subtitle timing. After specifying candidate subtitle regions of a video, VideoSubFinder analyzes these regions and returns timestamps corresponding to subtitle appearances. In addition, it saves screenshots of the detected subtitles. We then apply PaddleOCR (Cui et al., 2025) to extract textual content from these screenshots, resulting in a dataset with aligned video segments and detected texts.

Heuristic metadata checks. The data produced solely by this automatic pipeline is unavoidably noisy, since videos without subtitles may still contain detectable text in typical subtitle regions, like the bottom area in Figure 1. We further inspect the resulting metadata table and apply heuristic rules to identify anomalous entries. Specifically, we flag videos that exhibit characteristics including the following: (1) most subtitles persist for abnormally long or short durations; (2) the extracted text is linguistically unnatural; (3) the whole video is long but contains only a small number of detected subtitles; or (4) there is a mismatch between subtitle duration and the amount of information conveyed in the text. For flagged videos, we manually review the original content and decide whether to retain or discard them from the dataset.

3.3 Stage 3 - Improving Corpus Scalability and Quality of Subtitle Extraction

As the project progressed, we identified several issues that adversely affected both the scalability and the quality of the constructed corpus. In particular, two major challenges emerged: (1) the complexity of identifying and processing HardSub sign language videos, and (2) the limited quality of subtitle text extracted using traditional OCR systems.

3.3.1 Subtitle-Sign Segment Detection Model

A key bottleneck in scalability arises from the need to identify HardSub sign language videos, a process that involves multiple heterogeneous steps. In addition, the video segmentation relies heavily on VideoSubFinder. VideoSubFinder suffers from several limitations. First, it requires the manual specification of candidate subtitle regions for different videos. Especially, this requirement becomes problematic when subtitle orientation changes within a single video. Second, VideoSubFinder cannot determine whether a detected subtitle segment actually co-occurs with sign language, which is critical for processing videos containing mixed content. To address these limitations, we develop a unified model that can identify subtitle-sign co-occurring segments. We formulate this task as a sequence labeling problem over video frames using the **BIO** tagging scheme. Specifically, the label **O** denotes

frames without subtitles or sign language, **B** indicates the beginning of a segment in which subtitles and sign language co-occur, and **I** represents the continuation of such a segment.

Model Architecture. The architecture follows a straightforward vision-temporal encoder design. We adopt SigLIP2 (Tschannen et al., 2025) for frame-level feature extraction. Temporal dependencies are modeled using a bidirectional Mamba (Gu and Dao, 2023; Liang et al., 2024) backbone, which captures long-range context across video frames.

Training Detail. We use the segmentation results from our previously constructed corpus as the basis for model training. From this corpus, we manually correct a small subset of representative videos from each channel to create a clean finetuning set by removing irrelevant or misaligned segments and adding missing segments. The remaining samples, which may contain noise, are retained for pretraining. The model is first pretrained on the noisy dataset and subsequently adapted to the clean dataset.

After two-stage training, we obtain a subtitle-sign segment detection model. This model can assist in identifying HardSub sign language videos with diverse subtitle orientations and locations. Besides, it can segment videos based on the subtitles. Although the model is not perfect in practice, it substantially reduces manual effort and provides a scalable and convenient solution for ongoing dataset expansion, which is one of the important use cases of our corpus.

3.3.2 Subtitle Extraction

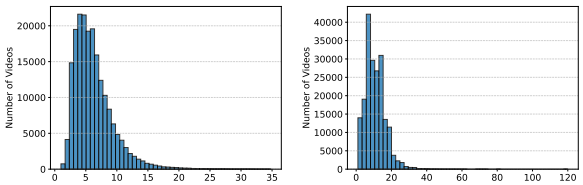
Traditional OCR systems, such as PaddleOCR (Cui et al., 2025), exhibit several limitations when extracting subtitle text from video frames. They often segment a single sentence into multiple fragments, and lack robustness to complex subtitle layouts (see Figure 1). As a result, extensive and intricate post-processing is required, yet the extracted subtitles still remain error-prone. This post-processing is also unstable and difficult to reproduce consistently.

Motivated by recent advances in the OCR capabilities of LVLMs, we adopt an instruction-based OCR approach. Specifically, we employ Qwen3-VL-32B-Thinking (Bai et al., 2025) to extract subtitle text from video frames. The model is instructed to extract only the visible subtitle text, ignore Furigana annotations, and output results strictly in JSON format to ease postprocessing. To improve robustness in identifying the correct subtitle content, we additionally provide two neighboring frames as contextual inputs. Further details of the prompting strategy are provided in Appendix B.

We utilize the trained subtitle-sign segment detection model and LVLM-based OCR to reprocess the entire corpus. Additionally, we construct a separate validation set from YouTube, which will be discussed in the following section.

Statistics	JSW		JSW valid.		JDC	
	Len.	Dur.	Len.	Dur.	Len.	Dur.
Min	1	1.00	1	1.53	1	0.15
Max	120	34.80	44	31.76	79	19.70
Mean	10.76	6.28	12.52	7.20	11.70	4.16
Median	10	5.60	12	6.33	8	2.86
Std.	5.66	3.24	6.05	3.65	10.63	3.73
90th percentile	18	10.21	20	11.87	26	9.34

Table 2: Statistics of subtitle lengths and video clip durations. JSW, JSW valid., and JDC denote the J-Shuwa corpus, the J-Shuwa validation set, and the JSL Dialogue Corpus, respectively. Dur. denotes video duration in seconds, and Len. denotes text length measured at the word level.



(a) Duration (b) Caption Length

Figure 2: Distribution of video durations and subtitle lengths in J-Shuwa.

3.4 Dataset Statistics

Finally, we construct a JSL-Japanese parallel corpus comprising approximately 197,742 sign language sentences, with a cumulative duration of around 300 hours, segmented from 6,322 videos. The statistics of sentence durations and caption lengths are summarized in Table 2, and their distributions are illustrated in Figure 2. Among these, 161,230 sentences are obtained by processing video segments with hard-coded subtitles using the procedure described above, substantially enriching the available SLT data for JSL. We apply SpaCy (Honnibal et al., 2020) for text preprocessing and vocabulary analysis, retaining only content words, which yields approximately 31K unique vocabulary items.

The YouTube-SL-25 dataset contains 1,075 JSL videos with closed captions. Our constructed dataset includes 1,008 such videos, of which 609 overlap with YouTube-SL-25. Based on a rough

inspection, the differences can be attributed to the following factors: (1) each dataset identifies some JSL channels that are not included in the other; (2) YouTube-SL-25 was refreshed in May 2024 to incorporate newer videos (Tanzer and Zhang, 2025), whereas our dataset is based on a snapshot collected in June 2023; and (3) we exclude videos featuring sign language songs, which are included in YouTube-SL-25.

4 Experiment

In this section, we conduct experiments on training SLT models with the J-Shuwa corpus to evaluate its utility.

4.1 Experiment Setup

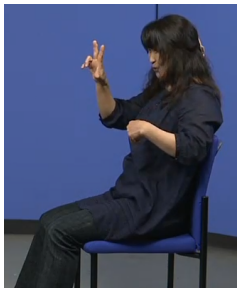
Benchmark. We construct two benchmarks for evaluation, the statistics of which are shown in Table 2:

- **J-Shuwa validation set.** A held-out validation set of 2,309 sentences, collected from additional videos published after June 2023 from the same channels as the training data, with no overlap with the training split.
- **JSL Dialogue Corpus.** A subset of the JSL Colloquial Corpus (Bono et al., 2014; Bono and Osugi, 2026), consisting of 1,372 dialogue segments (<20s) with high-quality Japanese translations. Each segment is recorded from semi-frontal and side views (Figure 3), yielding 2,744 segments in total.[3]

Baseline Models. Following (Uthus et al., 2023), we adopt a similar architecture to establish our baseline. The input modality is skeletal data extracted by Mediapipe Holistic (Lugaresi et al., 2019). We embed skeletal data by a two-layer MLP projection network to obtain video representations, which are then fed into the language models. Here, we fine-tune two types of backbone models: (1) an encoder-decoder model, mT5-base (Xue et al., 2021), and (2) decoder-only models, including Qwen3-1.7B and Qwen3-4B (Yang et al., 2025). For Qwen3-based models, we use Instruct variants. For the training sets, in addition to using J-Shuwa (JSW) alone, we also experiment with joint training on YouTube-ASL (YASL) (Uthus et al., 2023) to assess potential gains. More training details, including hyperparameter settings and data preprocessing procedures, are available in Appendix C.

Evaluation Metrics We use BLEU (Papineni et al., 2002), computed with SacreBLEU (Post, 2018), to measure surface-level similarity, and BLEURT (Sellam et al., 2020), using the BLEURT-20 checkpoint, to assess semantic similarity.

(a) Semi-frontal view



(b) Side view

Figure 3: Screenshots from different angles of a video in the JSL Dialogue Corpus.

4.2 Experiment on J-Shuwa Validation Set

Setup. In our initial experiments, we evaluate within-distribution performance on the J-Shuwa validation set to compare performance under different configurations. We report the checkpoints achieving the highest BLEU-4 score during training. The results are shown in Table 3.

Model Comparison. Qwen3 models trained with LoRA consistently outperform mT5 models with full fine-tuning. Notably, Qwen3 models trained only on JSW achieve performance comparable to mT5 models jointly trained on JSW and YASL. We attribute this to the stronger inherent text generation capabilities of Qwen3, whereas mT5, being primarily a pretrained model, lacks generation ability without sufficient fine-tuning. The best performance is obtained by Qwen3-1.7B, while the difference with Qwen3-4B is relatively small.

Effect of Joint Training. Furthermore, joint training on JSW and YASL leads to consistent improvements across all models, yielding gains of approximately 2 to 3 BLEU-4 points compared to training on JSW alone. This demonstrates the effectiveness of cross-lingual transfer.

Effect of LoRA Configuration for Qwen3-Based Models. We observe that adapting only the attention modules results in lower BLEU and BLEURT scores than adapting all MLP modules within both the feed-forward and attention layers

[3]The corpus itself is publicly available, but the Japanese translations used in this study are based on internal experimental annotations by the Bono Lab of National Institute of Informatics, which are not included in the public release. They were used with permission from the principal investigator, Mayumi Bono.

Model	Training Schedule	LoRA Module	B1	B2	B3	B4	BLEURT
mT5-base	JSW		23.47	13.62	9.20	6.64	25.05
	JSW + YASL		29.61	18.83	13.16	9.68	35.10
Qwen3-4B	JSW	QV	24.04	14.06	9.38	6.72	30.27
	JSW	QKVO	25.39	15.16	10.36	7.59	31.23
	JSW	ALL	30.64	19.51	13.89	10.37	35.09
	JSW + YASL	ALL	33.31	22.11	15.96	12.02	39.91
Qwen3-1.7B	JSW	QV	22.35	12.54	8.32	5.98	27.92
	JSW	QKVO	24.97	14.71	9.97	7.23	30.63
	JSW	ALL	29.14	18.64	13.29	9.97	35.84
	JSW + YASL	ALL	34.41	22.97	16.64	12.58	42.08

Table 3: Results on the J-Shuwa validation set. JSW denotes J-Shuwa, and YASL denotes YouTube-ASL. In the LoRA module column, Q/K/V/O correspond to the query, key, value, and output projection layers, respectively; "ALL" includes Q/K/V/O as well as the up, down, and gate projection layers. B1/2/3/4 denote BLEU-1/2/3/4 scores.

	Front ∪ Side			Front View			Side View		
Training Set	B1	B2	BLEURT	B1	B2	BLEURT	B1	B2	BLEURT
JSW	10.23	2.80	16.13	9.78	3.11	17.19	8.94	2.05	15.07
→ FT	18.66±0.23	8.05±0.33	22.63±0.35	19.88±0.40	8.96±0.42	24.74±0.57	17.49±0.58	7.15±0.51	20.51±0.62
JSW + YASL	7.64	2.14	19.05	9.23	2.96	19.02	6.34	1.43	19.08
→ FT	21.62±0.63	9.99±0.60	27.22±0.58	23.63±0.89	11.34±0.69	29.66±0.86	19.71±0.94	8.66±0.70	24.77±0.58
Ablation study									
YSL25-JSL	8.69	1.99	12.23	9.19	2.09	11.72	8.19	1.89	12.74
→ FT	16.15±0.41	6.55±0.21	19.54±0.33	16.24±0.46	6.83±0.26	20.52±0.56	16.06±0.74	6.24±0.51	18.56±0.75
JSW-CC	9.72	2.10	11.66	10.02	2.27	12.56	9.42	1.93	10.76
→ FT	15.91±0.38	5.85±0.39	19.38±0.40	16.16±0.39	5.85±0.36	20.87±0.61	15.67±0.68	5.81±0.81	17.90±0.50
JSW-HardSub	8.46	2.28	12.88	9.82	2.86	14.12	7.06	1.67	11.64
→ FT	18.00±0.56	7.58±0.32	22.16±0.51	19.28±0.64	8.21±0.57	24.11±0.78	16.78±0.60	6.97±0.13	20.20±0.69
YASL → FT	17.88±0.47	7.93±0.49	22.15±0.53	18.71±0.74	8.61±0.57	23.50±0.49	17.07±0.49	7.26±0.44	20.80±0.89

Table 4: Performance on the JSL Dialogue Corpus test set under zero-shot inference (without → FT) and after fine-tuning (with → FT), including the ablation study with models trained on different subsets. Results for Front ∪ Side are computed over the aggregation of frontal-view and side-view samples.

(denoted as ALL in Table 3). This suggests that sign language is novel knowledge to the models. Updating only the attention components is insufficient for effective model adaptation.

We adopt the Qwen3-1.7B model with the ALL LoRA configuration as the default setup in subsequent experiments.

4.3 Experiment on JSL Dialogue Corpus

We evaluate the Qwen3-1.7B checkpoints that achieved the best performance in previous experiments on the more challenging, well-annotated JSL Dialogue Corpus, under both zero-shot and fine-tuning settings.

The dataset is divided into training, validation, and test sets, comprising 2058/136/550 segments (75%/5%/20%). Since the target translations are relatively short, with a median length of eight words (see Table 2), we report BLEU-1, BLEU-2, and BLEURT.

In the zero-shot setting, we directly evaluate the pretrained checkpoints on the test set without further adaptation. For fine-tuning, we further train the checkpoints on the training split of the JSL Dialogue Corpus. For each checkpoint, we conduct five runs and report the mean and standard deviation on the test set. The test set results are summarized in Table 4.

4.3.1 Zero-shot Performance

For zero-shot evaluation, BLEU-1 scores of 10.23 and 7.64 on pretrained models suggest that they capture key lexical items and coarse thematic elements. Although joint training on JSW and YASL results in lower BLEU scores, it achieves slightly higher BLEURT scores, indicating that cross-lingual transfer helps better capture semantic meaning.

Nevertheless, the overall ability to recover full semantic content remains limited, as reflected by

Reference:	ねえねえ、ねえ違うわよ、 長期間 は ダメ だけど数日なら 大丈夫 *No, no, that's not right.* *Long-term* *is* *not allowed* *, but a few days is* *fine* *.*
JSW:	覚えていても 長い間 覚え ていません *Even if (I) remember it, (I)* *don't* *remember it for a* *long time* *.*
JSW → FT:	今は固形ルーになってるけど前の方が美味しかった *Now it has become a solid roux, but the previous one tasted better.*
JSW + YASL:	長く 寝ては いけません *(You) must* *not* *sleep for a* *long time* *.*
JSW + YASL→ FT:	長く は ダメ 、一晩ぐらいなら 大丈夫 。 *Long-term* *is* *not allowed* *, but one night is* *fine* *.*

Table 5: An example from the JSL Dialogue Corpus, along with outputs from Qwen3-1.7B-based models before and after fine-tuning. Key semantic components in the reference are highlighted with distinct colors, and their aligned realizations in model outputs are marked using the same color.

the low BLEURT scores, indicating that the current training scale is insufficient to cover the diversity of JSL scenarios.

4.3.2 Fine-tuning Performance

Fine-tuning Gains. The model pretrained solely on JSW achieves a BLEURT improvement of 6.50 after fine-tuning. In comparison, the jointly trained model exhibits a larger gain of 8.17, indicating greater benefit from fine-tuning. Despite these improvements, the overall scores remain relatively low, suggesting that the benchmark remains challenging.

Viewpoint Analysis. In the zero-shot setup, the performance is not significantly different between front-view and side-view camera angles. However, after fine-tuning, the difference becomes obvious: Performance gains are larger for front-view samples than for side-view ones.

This difference may be attributed to two factors. First, most sign language videos available on YouTube are recorded from a frontal perspective, leading to a strong viewpoint bias during pretraining. Second, side-view recordings may obscure important visual cues, making pose estimation more challenging and consequently degrading recognition performance.

These findings highlight the limitations introduced by **viewpoint bias** in YouTube sign language videos. While incorporating RGB modalities may alleviate this issue, it comes with increased computational cost. An alternative direction is to develop pose encoders that are robust to viewpoint variation and missing keypoints.

4.3.3 Ablation Study

In the ablation study, we decompose J-Shuwa into its CC and HardSub subsets (denoted as **JSW-CC** and **JSW-HardSub**) and conduct both zero-shot and post-adaptation experiments on each subset. In addition, we include JSL data from YouTube-SL-25 (denoted as **YSL25-JSL**) as a comparison baseline. For the fine-tuning setup, we further evaluate a strategy in which the model is pretrained solely on YouTube-ASL. The results are summarized in Table 4.

Effect of Data Subsets. In the zero-shot setting, JSW-HardSub achieves higher BLEURT scores than JSW-CC and YSL25-JSL, particularly for front-view samples. Moreover, JSW-HardSub benefits more from fine-tuning, achieving performance close to that of JSW, underscoring the contribution of large-scale parallel samples derived from HardSub videos. Both JSW-CC and JSW-HardSub yield lower BLEURT scores than the full JSW, suggesting that the two subsets are complementary and that their combination provides broader coverage.

Effect of ASL Data. Interestingly, the model pretrained solely on YASL also exhibits strong adaptation capability on this benchmark, achieving similar performance to models pretrained on JSW. This result can be partially attributed to the larger scale and high quality of ASL data in YASL, which employs skilled human annotators for data selection.

4.3.4 Qualitative Analysis

We present qualitative outputs from Qwen3-1.7B models trained under different schedules in Table 5. We can find that even without in-domain fine-tuning on the JSL Dialogue Corpus, the models are able to capture some key lexical items in certain cases. After in-domain fine-tuning, models produce outputs that are semantically closer to the reference. However, some models exhibit hallucinations and

generate degraded outputs, indicating overfitting.

More model outputs can be found in Appendix F.

5 Discussion & Conclusion

In conclusion, we construct a large-scale JSL-Japanese parallel corpus, named **J-Shuwa**, by leveraging both closed-captioned and hard-subtitled JSL videos from YouTube. We present our corpus construction pipeline, including the procedures for identifying suitable videos for corpus construction and processing HardSub sign language videos. We further conduct SLT experiments by training models on J-Shuwa and evaluating them on a benchmark derived from the JSL Dialogue Corpus. The experimental results demonstrate the effectiveness of J-Shuwa. Beyond SLT, J-Shuwa can serve as a foundation for a wide range of tasks for JSL, including cross-modal retrieval and sign language production.

Despite the substantially increased data scale, our experiments show that the models still struggle to produce faithful translations, underscoring that JSL translation remains a challenging problem. Given the relatively small size of the JSL user community, substantially expanding the dataset volume using open Internet resources alone may be less feasible. Future progress is therefore likely to rely on methodological advances, such as more powerful video encoders and improved cross-lingual transfer leveraging data from multiple sign languages.

We hope that J-Shuwa will foster broader engagement and closer collaboration within the research community, contributing to sustained progress in JSL research.

Limitations

Bias. Since the annotators are not experts in sign language, their judgments rely primarily on subjective visual perception. Consequently, the filtered data may be biased and may not accurately reflect the sign languages used within the Deaf community. For example, on YouTube, there exist at least two visually similar yet linguistically distinct systems that are frequently conflated (Chonan, 2001): **Japanese Sign Language (JSL)**, which naturally emerged within the Deaf community, and **Manually Coded Japanese (MCJ)**, also known as **Signed Japanese (SJ)**, which borrows lexical signs from JSL to represent spoken Japanese. A key limitation of this work is that, without the involvement of Deaf experts, we cannot reliably distinguish between these systems, resulting in a dataset that contains mixed content. Although JSL and SJ share similar lexical-level signing, they differ substantially in grammatical structure. While jointly training on data from both systems may act as a form of data augmentation and benefit lexical-level recognition, the lack of explicit separation may confuse models and hinder their ability to accurately capture sentence-level structure and semantics. Addressing this limitation will require closer collaboration with sign language experts and is left for future work.

Noisy Data. Due to limited annotation resources, the dataset inevitably contains noise, as it is not feasible to manually verify all samples. Consequently, some irrelevant videos are unavoidably included. Random inspections reveal several types of problematic segments, including poor subtitle-video alignment, clips without signing, and other quality issues. In future work, we plan to develop more effective data cleaning strategies and develop additional tools, such as JSL-Japanese quality estimation models, to more reliably filter the data and improve alignment quality.

Ethical Considerations

Similar to other sign language datasets derived from YouTube, such as YouTube-ASL (Uthus et al., 2023) and YouTube-SL-25 (Tanzer and Zhang, 2025), this work involves comparable ethical considerations. Facial expression is essential for conveying meaning in sign languages; consequently, most sign language datasets inevitably raise concerns about privacy. Our corpus is constructed from publicly accessible YouTube videos. Moreover, we release only the corresponding YouTube video IDs and segment-level timestamps, without redistributing any original video or subtitle content. This design ensures that content removal or access restrictions imposed by the original uploaders are respected and automatically reflected in the dataset.

We acknowledge that the dataset was not annotated by expert Deaf signers, which may introduce linguistic bias or misrepresentation. As a result, models trained on this dataset may not fully reflect natural sign language usage or the linguistic diversity of the Deaf community. This dataset should therefore not be considered representative of the full range of JSL or Deaf signing practices.

Acknowledgments

We thank the anonymous reviewers for their valuable feedback. We are also grateful to Mayumi

Bono and Tomohiro Okada at National Institute of Informatics for their valuable insights into Deaf culture and their expertise in JSL.

In this paper, we utilize the A Colloquial Corpus of Sign Languages in Japan (JSL) provided by the National Institute of Informatics and Tsukuba University of Technology through the IDR Dataset Service of the National Institute of Informatics. We sincerely thank the Deaf participants and sign language experts for their time and effort in producing high-quality data and annotations for the JSL Dialogue Corpus, which serves as a valuable benchmark for validating our constructed corpus.

This work was supported by JSPS/MEXT KAKENHI Grant Numbers JP22H05015 and JP23K28139, and JST ASPIRE Program Japan Grant Number JPMJAP25B3.

References

Samuel Albanie, Gül Varol, Liliane Momeni, Hannah Bull, Triantafyllos Afouras, Himel Chowdhury, Neil Fox, Bencie Woll, Rob Cooper, Andrew McParland, and Andrew Zisserman. 2021. Bbc-oxford british sign language dataset. *CoRR*, abs/2111.03635.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025. Qwen3-vl technical report. *Preprint*, arXiv:2511.21631.

Mayumi Bono, Kouhei Kikuchi, Paul Cibulka, and Yutaka Osugi. 2014. A colloquial corpus of Japanese Sign Language: Linguistic resources for observing sign language conversations. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1898–1904, Reykjavik, Iceland. European Language Resources Association (ELRA).

Mayumi Bono and Yutaka Osugi. 2026. A Colloquial Corpus of Sign Languages in Japan (JSL). Informatics Research Data Repository, National Institute of Informatics (dataset).

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. 2019. Sign language recognition, generation, and translation: An interdisciplinary perspective. In *The 21st International ACM SIGACCESS Conference on Computers and Accessibility, ASSETS 2019, Pittsburgh, PA, USA, October 28-30, 2019*, pages 16–31. ACM.

Necati Cihan Camgöz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. 2018. Neural sign language translation. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 7784–7793. Computer Vision Foundation / IEEE Computer Society.

Hirohito Chonan. 2001. Grammatical differences between japanese sign language, pidgin sign japanese, and manually coded japanese: Effects on comprehension. *Japanese Journal of Educational Psychology*, 49(4):417–426.

Xianzhi Chu, Jiang Liu, and Shigeru Shimamoto. 2021. A sensor-based hand gesture recognition system for japanese sign language. In *3rd IEEE Global Conference on Life Sciences and Technologies, LifeTech 2021, Nara, Japan, March 9-11, 2021*, pages 311–312. IEEE.

MMAction2 Contributors. 2020. Openmmlab's next generation video understanding toolbox and benchmark. https://github.com/open-mmlab/mmaction2.

Mathieu De Coster, Dimitar Shterionov, Mieke Van Herreweghe, and Joni Dambre. 2024. Machine translation from signed to spoken languages: state of the art and challenges. *Univers. Access Inf. Soc.*, 23(3):1305–1331.

Cheng Cui, Ting Sun, Manhui Lin, Tingquan Gao, Yubo Zhang, Jiaxuan Liu, Xueqing Wang, Zelun Zhang, Changda Zhou, Hongen Liu, Yue Zhang, Wenyu Lv, Kui Huang, Yichao Zhang, Jing Zhang, Jun Zhang, Yi Liu, Dianhai Yu, and Yanjun Ma. 2025. Paddleocr 3.0 technical report. *Preprint*, arXiv:2507.05595.

Amanda Cardoso Duarte, Shruti Palaskar, Lucas Ventura, Deepti Ghadiyaram, Kenneth DeHaan, Florian Metze, Jordi Torres, and Xavier Giró-i-Nieto. 2021. How2sign: A large-scale multimodal dataset for continuous american sign language. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 2735–2744. Computer Vision Foundation / IEEE.

Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. 2019. Slowfast networks for video recognition. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 6201–6210. IEEE.

Albert Gu and Tri Dao. 2023. Mamba: Linear-time sequence modeling with selective state spaces. *CoRR*, abs/2312.00752.

Shester Gueuwou, Sophie Siake, Colin Leong, and Mathias Müller. 2023a. 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. Association for Computational Linguistics.

Shester Gueuwou, Kate Takyi, Mathias Müller, Marco Stanley Nyarko, Richard Adade, and Rose-Mary Owusuaa Mensah Gyening. 2023b. Afrisign: Machine translation for african sign languages. In *Proceedings of the 4th Workshop on African Natural Language Processing, AfricaNLP@ICLR 2023, Kigali, Rwanda, May 1, 2023*.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Li Ji, Jiang Liu, and Shigeru Shimamoto. 2022. Recognition of japanese sign language by sensor-based data glove employing machine learning. In *4th IEEE Global Conference on Life Sciences and Technologies, LifeTech 2022, Osaka, Japan, March 7-9, 2022*, pages 256–258. IEEE.

Abhinav Joshi, Susmit Agrawal, and Ashutosh Modi. 2023. ISLTranslate: Dataset for translating Indian Sign Language. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10466–10475, Toronto, Canada. Association for Computational Linguistics.

Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, and 1 others. 2017. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*.

Sang-Ki Ko, Chang Jo Kim, Hyedong Jung, and Choongsang Cho. 2019. Neural sign language translation based on human keypoint estimation. *Applied sciences*, 9(13):2683.

Oscar Koller, Jens Forster, and Hermann Ney. 2015. Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers. *Comput. Vis. Image Underst.*, 141:108–125.

Dongxu Li, Cristian Rodriguez Opazo, Xin Yu, and Hongdong Li. 2020. Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. In *IEEE Winter Conference on Applications of Computer Vision, WACV 2020, Snowmass Village, CO, USA, March 1-5, 2020*, pages 1448–1458. IEEE.

Kunchang Li, Yali Wang, Yizhuo Li, Yi Wang, Yinan He, Limin Wang, and Yu Qiao. 2023. Unmasked teacher: Towards training-efficient video foundation models. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 19891–19903. IEEE.

Zecheng Li, Wengang Zhou, Weichao Zhao, Kepeng Wu, Hezhen Hu, and Houqiang Li. 2025. Uni-sign: Toward unified sign language understanding at scale. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Aobo Liang, Xingguo Jiang, Yan Sun, Xiaohou Shi, and Ke Li. 2024. Bi-mamba+: Bidirectional mamba for time series forecasting. *arXiv preprint arXiv:2404.15772*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg, and Matthias Grundmann. 2019. Mediapipe: A framework for building perception pipelines. *CoRR*, abs/1906.08172.

Ryota Murai, Naoto Tsuta, Duk Shin, and Yousun Kang. 2025. Point-supervised japanese fingerspelling localization via hr-pro and contrastive learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*.

Yuji Nagashima. 2021. Kogakuin university japanese sign language multi-dimensional database (kosign). `https://doi.org/10.32130/rdata.5.1`. Dataset.

OpenAI. 2025. Chatgpt.

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.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Joseph Redmon, Santosh Kumar Divvala, Ross B. Girshick, and Ali Farhadi. 2016. You only look once: Unified, real-time object detection. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 779–788. IEEE Computer Society.

Ramon Sanabria, Ozan Caglayan, Shruti Palaskar, Desmond Elliott, Loïc Barrault, Lucia Specia, and Florian Metze. 2018. How2: A large-scale dataset for multimodal language understanding. *CoRR*, abs/1811.00347.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Xin Shen, Shaozu Yuan, Hongwei Sheng, Heming Du, and Xin Yu. 2023. Auslan-daily: Australian sign language translation for daily communication and news. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Bowen Shi, Diane Brentari, Gregory Shakhnarovich, and Karen Livescu. 2022. Open-domain sign language translation learned from online video. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 6365–6379. Association for Computational Linguistics.

Jungpil Shin, Md Al, Abu Saleh, Kota Suzuki, and Koki Hirooka. 2024. Japanese sign language recognition by combining joint skeleton-based handcrafted and pixel-based deep learning features with machine learning classification. *Computer Modeling in Engineering & Sciences*, 139(3):2605.

Ariel E. Stassi, Yanina Boria, J. Matías Di Martino, and Gregory Randall. 2025. ilsu-t: an open dataset for uruguayan sign language translation. In *19th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2025, Tampa/Clearwater, FL, USA, May 26-30, 2025*, pages 1–10. IEEE.

Sihan Tan, Nabeela Khan, Zhaoyi An, Yoshitaka Ando, Rei Kawakami, and Kazuhiro Nakadai. 2024. A review of deep learning-based approaches to sign language processing. *Adv. Robotics*, 38(23):1649–1667.

Garrett Tanzer and Biao Zhang. 2025. Youtube-sl-25: A large-scale, open-domain multilingual sign language parallel corpus. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Michael Tschannen, Alexey A. Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier J. Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. 2025. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *CoRR*, abs/2502.14786.

David Uthus, Garrett Tanzer, and Manfred Georg. 2023. Youtube-asl: A large-scale, open-domain american sign language-english parallel corpus. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 40 others. 2025. Qwen3 technical report. *CoRR*, abs/2505.09388.

Aoxiong Yin, Zhou Zhao, Weike Jin, Meng Zhang, Xingshan Zeng, and Xiaofei He. 2022. MLSLT: towards multilingual sign language translation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 5099–5109. IEEE.

Hao Zhou, Wengang Zhou, Weizhen Qi, Junfu Pu, and Houqiang Li. 2021. Improving sign language translation with monolingual data by sign back-translation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 1316–1325. Computer Vision Foundation / IEEE.

A Dataset Detail

A.1 Pretraining Data

YouTube-ASL (Uthus et al., 2023) reports a total of 610K captions. However, due to download issues, changes in video availability or uploader permissions over time, as well as processing failures, we obtained 549,125 valid captions for training. For the JSL data from YouTube-SL-25, we successfully processed 39,101 items from 1,059 videos, which is comparable to the number of videos reported in their paper (Tanzer and Zhang, 2025) (1,075).

Topic	AniN	Cur	Pro
Count	1514	1196	34
Avg. Trans. Length	13.16	9.93	9.35
Std. Trans. Length	11.42	9.29	8.80
Vocab. Size	1009	907	70
Avg. Vid. Duration (s)	4.46	3.84	2.65
Std. Vid. Duration (s)	3.96	3.42	1.97

Table 6: Topic-wise statistics of the JSL Dialogue Corpus

A.2 JSL Dialogue Corpus

Currently available translation annotations in the JSL Dialogue Corpus cover several dialogue topics, namely AniN (animation), Cur (curry), and Pro

(proud of your country). The topic-wise statistics are shown in Table 6.

Split	Metric	AniN	Cur	Pro
Train	Count	1138	898	22
	Avg. Length	13.58	10.39	10.09
	Std. Length	11.92	9.70	9.47
	Vocab. Size	890	812	59
	Avg. Duration (s)	4.63	3.94	2.24
	Std. Duration (s)	4.11	3.50	1.28
Validation	Count	88	48	–
	Avg. Length	11.55	7.83	–
	Std. Length	9.30	6.55	–
	Vocab. Size	191	103	–
	Avg. Duration (s)	3.51	3.80	–
	Std. Duration (s)	2.72	3.59	–
Test	Count	288	250	12
	Avg. Length	11.99	8.67	8.00
	Std. Length	9.77	7.95	7.63
	Vocab. Size	421	352	29
	Avg. Duration (s)	4.08	3.46	3.40
	Std. Duration (s)	3.59	3.05	2.74

Table 7: Dataset Statistics by Split and Topic.

A.3 Dataset Split in Fine-tuning Experiments

The split is performed randomly while ensuring that semi-frontal and side-view recordings of the same instance are assigned to the same subset. The validation set is used to select suitable hyperparameters that stabilize the training process. The topic-wise statistic of each split is shown in Table 7.

B Subtitle Extraction

The instruction used for subtitle text extraction in Qwen3-VL-32B-Thinking is shown in Table 8. We choose frames from at most two neighboring segments as context.

C Training Detail

C.1 Data Preprocessing

For each video, we extract 160 keypoints using MediaPipe Holistic (Lugaresi et al., 2019), including 6 upper-body keypoints, 21 keypoints for each hand, and 112 facial keypoints, and discard the z-axis information. The resulting keypoint sequences are represented as 320-dimensional vectors (160 keypoints $\times 2$ coordinates) per frame.

MediaPipe Holistic extracts poses for only a single person. To reduce the influence of distractors in J-Shuwa and YouTube-SL-25, we first apply YOLO (Redmon et al., 2016) for human tracking and crop the signer region prior to pose extraction. If a video clip contains only one detected person, that individual is treated as the signer by default. When multiple people are detected, we apply an action recognition model[4] and select the person with the highest predicted probability for the Sign Language Interpreting class.

In contrast, YouTube-ASL primarily consists of single-person videos and has been manually verified by domain experts. Given its substantially larger scale and computational constraints, we directly apply pose extraction without additional preprocessing.

We normalize each pose skeleton by translating it to the midpoint of the shoulders and scaling it by the inter-shoulder distance. To prevent instability caused by erroneous pose estimates, we enforce a minimum shoulder distance of 0.1 during normalization. Missing keypoint coordinates are filled with a constant value of -5.

For each video, we sample up to 256 frames at a sampling rate of 2, corresponding to approximately 21 seconds at 24 frames per second. The maximum length of the target translation sequence is limited to 128 tokens.

C.2 Training Strategy

All models are trained using an instruction-based format by prefixing each sample with an instruction. For ASL samples, we use the prefix: “Translate the American Sign Language video to English:”. For JSL samples, we use: “Translate the Japanese Sign Language video to Japanese:”. These prefixes are applied consistently across all experimental settings. For Qwen3 models, we use the Instruct variant and therefore also apply the corresponding chat template.

C.3 Hyperparameters at Varying Stage

C.3.1 Training on JSW or JSW + YASL

We train all models using the AdamW (Loshchilov and Hutter, 2019) optimizer with a global batch size of 64 (8 samples per GPU across 8 GPUs) and employ a cosine decay learning rate scheduler with warmup. Training is conducted using BF16 mixed precision.

For different backbone models, we use different training strategies and hyperparameters. In experiments of fine-tuning on JSW or JSW + YASL, we fully fine-tune mT5 for up to 100 epochs, validating after each epoch, and apply early stopping when

[4]We use the UnMasked Teacher action recognition model (Li et al., 2023), implemented with the MMAction2 framework (Contributors, 2020).

```
Here are the surrounding frames to provide context for the scene:
[Context Image 1]
[Context Image 2]
...
[Context Image N]

Now analyze the following image and extract any subtitle text visible in it.
Ignore the furigana or small annotations above the main subtitle text if present.

Output only JSON in the following format:

{
  "subtitle_text": "Extracted subtitle string",
  "language": "Detected language code (e.g. en, ja, fr)"
}

If no subtitle is found, return:
{ "subtitle_text": null, "language": null }

[Target Image]
```

Table 8: Prompt used for subtitle extraction with contextual frames.

	mT5	Qwen3
Learning Rate	1e-3	
Min. Learning Rate	1e-6	
Warmup Ratio	10%	
AdamW betas	(0.9,0.999)	
Training Strategy	Full	LoRA
weight decay	1e-2	0
Max Training Epochs	100	20
LoRA - Rank	-	32
LoRA - Alpha	-	32
LoRA - Dropout	-	0.05
Early Stopping	5	✗

Table 9: Fine-tuning hyperparameters for mT5 (full fine-tuning) and Qwen3 (LoRA) on JSW and JSW + YASL.

the BLEU-4 score on the validation set does not improve for five consecutive epochs. For Qwen3-based models, we adopt Low-Rank Adaptation (LoRA) (Hu et al., 2022) for parameter-efficient fine-tuning. Due to computational resource constraints, Qwen3-based models are trained for a fixed number of epochs and validated every two epochs. Other hyperparameter settings can be found in Table 9.

During inference, we use beam search with a beam width of 5.

C.3.2 In-domain Fine-tuning on the JSL Dialogue Corpus

In the in-domain fine-tuning experiments on the JSL Dialogue Corpus, we adopt the same hyperparameter settings as the Qwen3-1.7B configuration described above, with the exception of the learning rate. The number of training epochs, optimizer, and learning rate schedule remain unchanged, while the total number of training steps varies according to the dataset size. We tune only the learning rate on the validation set and ultimately set it to 3e-4 for fine-tuning. We report results from the final training checkpoints rather than those with the best validation performance, as the small validation set leads to unstable evaluation metrics during training.

C.4 Computational Cost

Most training runs were conducted on a single node equipped with 8 NVIDIA A100 GPUs (80 GB). A small number of runs were conducted on nodes with 8 NVIDIA H200 GPUs.

The approximate wall-clock training times on A100 nodes for different Qwen3-based models and training datasets are reported in Table 11. The reported times include the validation procedure, during which inference was performed and contributed to the overall runtime. Actual training time may vary depending on system conditions such as I/O performance and resource contention; therefore, these results should be regarded as indicative rather than exact measurements.

In the testing stages, the inference is conducted on a single NVIDIA RTX 6000 Ada.

D Zero-shot Performance on How2Sign

We report the zero-shot performance of checkpoints jointly trained on J-Shuwa and YouTube-ASL on the How2Sign dataset in Table 10 for reference.

Model	Training Schedule	B1	B2	B3	B4	BLEURT
T5 (Uthus et al., 2023)	YASL	20.93	10.35	6.14	3.95	34.98
mT5	JSW + YASL	22.51	10.96	6.34	3.88	35.88
Qwen3-4B	JSW + YASL	23.84	12.37	7.35	4.67	38.44
Qwen3-1.7B	YASL	24.91	13.09	7.92	5.04	39.73
Qwen3-1.7B	JSW + YASL	25.55	13.50	8.23	5.31	39.79

Table 10: Zero-shot performance on How2Sign.

Model	Training Set	Training Time (Approx.)
Qwen3-4B	JSW	9h30min
	JSW + YASL	32h
Qwen3-1.7B	JSW	6h
	JSW + YASL	20h4min

Table 11: Approximate wall-clock training time for Qwen3-based models and training sets.

The mT5 model achieves similar performance to the T5 baseline reported in the original study, despite being jointly trained on J-Shuwa and YouTube-ASL. One possible factor contributing to the performance gap is the difference in training data (See Appendix A). In addition, unlike T5, which is English-centric, mT5 is a multilingual model with a larger parameter footprint and broader language coverage, which may dilute its performance on English.

Qwen3-based models achieve better performance than mT5. Notably, Qwen3-1.7B achieves the best performance in our experiments, aligning with the trends observed in JSL translation.

E Annotation Cost

All annotation jobs described in the main text, including data selection and video segmentation, were performed by the authors. The annotation process was semi-automatic and conducted intermittently throughout the research, without a fixed schedule or a dedicated annotation phase. As a result, no precise estimate of the annotation time or cost can be provided.

F Case Study

We provide more model outputs in Table 13 and 12.

G Usage of Generative AI

The authors used generative AI to assist with language editing and improving readability.

Reference:	ひよこはひよこで望遠鏡をのぞいている と、猫と目が合うんだ。猫とひよこはたまたま視線が合ってしまった。 *While the chick is peering through binoculars, it ends up making eye contact with a cat. The cat and the chick just happen to lock eyes.*
JSW:	20代の頃は双眼鏡で見ていましたが *When (I) was in my twenties, (I) was looking through binoculars, but...*
JSW → FT:	猫は双眼鏡で鳥を見つけて、双眼鏡で見ていたら、向こうのビルと視線が合う の。 *The cat finds a bird with binoculars, and while looking through the binoculars, its gaze meets a building over there.*
JSW + YASL:	目を閉じて視線を外して視線を外すようにします。 *They close their eyes and try to avert their gaze.*
JSW + YASL→ FT:	ねずみは、双眼鏡でぐるっと見ていた二匹が、ちょうど目が合ってお互いに気づくんだ。 *The two mice who were looking around with binoculars just happened to make eye contact and notice each other.*

Table 12: Another example from the JSL Dialogue Corpus, along with outputs from Qwen3-1.7B-based models before and after fine-tuning. Key semantic components in the reference are highlighted with distinct colors, and their aligned realizations in model outputs are marked using the same color.

Reference: ねえねえ、ねえ違うわよ、長期間はダメだけど数日なら大丈夫
No, no, that's not right. Long-term is not allowed, but a few days is fine.

Unfine-tuned	**Fine-tuned**
Qwen3-1.7B (JSW) 覚えていても 長い間覚えていません *Even if (I) remember it, (I) don't remember it for a long time.*	**Qwen3-1.7B (JSW)** 今は固形ルーになってるけど前の方が美味しかった *Now it has become a solid roux, but the previous one was tastier.*
Qwen3-1.7B (JSW + YASL) 長く寝てはいけません *(You) must not sleep for a long time.*	**Qwen3-1.7B (JSW + YASL)** 長くはダメ、一晩ぐらいなら大丈夫。 *Long-term is not good, but one night is okay.*
Qwen3-4B (JSW) ねえちょっと長い間 簡単に覚えられました *Hey, for a little long while, I was able to remember it easily.*	**Qwen3-4B (JSW)** ねえ。… *Hey...*
Qwen3-4B (JSW + YASL) 「頭髪は長すぎよ」 *"Your hair is too long."*	**Qwen3-4B (JSW + YASL)** あー、忘れたよ。でも、おーおー、忘れたよ。でも、いいよ。 *Ah. I forgot. But, oh, oh, I forgot. But it's fine.*
mT5 (JSW) ○○○と申します どうぞよろしくお願い致します *My name is XXX. Nice to meet you.*	**mT5 (JSW)** 長くはダメ、一晩ぐらいなら大丈夫って言われたらいいって思う *I think it would be fine if I were told that long-term is no good, but one night is okay.*
mT5 (JSW + YASL) 油断してはいけません。 *You must not let your guard down.*	**mT5 (JSW + YASL)** 長くはダメ、一晩ぐらいなら大丈夫って言われた、残ったらダメって *I was told that long-term is no good and one night is okay, and that it's no good if it remains.*

Table 13: Qualitative examples — left: unfine-tuned; right: fine-tuned.