Computers, Materials & Continua

Tech Science Press

https://doi.org/10.32604/cmc.2026.074016

ARTICLE

Korean Sign Language Recognition and Sentence Generation through Data Augmentation

Soo-Yeon Jeong[1], Ho-Yeon Jeong[2] and Sun-Young Ihm[3,*]

[1]Division of Software Engineering, Pai Chai University, Daejeon, Republic of Korea
[2]Department of Artificial Intelligence, Kyung Hee University, Yongin, Republic of Korea
[3]Department of Computer Engineering, Pai Chai University, Daejeon, Republic of Korea

*Corresponding Author: Sun-Young Ihm. Email: sunnyihm@pcu.ac.kr

Received: 30 September 2025; Accepted: 15 January 2026; Published: 12 March 2026

ABSTRACT: Sign language is a primary mode of communication for individuals with hearing impairments, conveying meaning through hand shapes and hand movements. Contrary to spoken or written languages, sign language relies on the recognition and interpretation of hand gestures captured in video data. However, sign language datasets remain relatively limited compared to those of other languages, which hinders the training and performance of deep learning models. Additionally, the distinct word order of sign language, unlike that of spoken language, requires context-aware and natural sentence generation. To address these challenges, this study applies data augmentation techniques to build a Korean Sign Language dataset and train recognition models. Recognized words are then reconstructed into complete sentences. The sign recognition process uses OpenCV and MediaPipe to extract hand landmarks from sign language videos and analyzes hand position, orientation, and motion. The extracted features are converted into time-series data and fed into a Long Short-Term Memory (LSTM) model. The proposed recognition framework achieved an accuracy of up to 81.25%, while the sentence generation achieved an accuracy of up to 95%. The proposed approach is expected to be applicable not only to Korean Sign Language but also to other low-resource sign languages for recognition and translation tasks.

KEYWORDS: Korean sign language recognition; LSTM; data augmentation; sentence completion

1 Introduction

Korean Sign Language (KSL) is the primary language used by individuals who are deaf or hard of hearing, and features a distinct grammatical structure. However, effective communication remains challenging due to significant grammatical differences between KSL and spoken Korean, as well as the lack of integration with widely-used speech-based translation technologies. To address these challenges, various sign language recognition technologies have been explored, with deep learning-based approaches gaining particular attention [1,2]. These methods analyze sign language gestures using cameras and translate them through machine learning models. However, most existing studies focus only on word-level recognition, with relatively little research on generating context-aware and natural sentences by combining recognized words. In KSL, meaning is conveyed not only through hand gestures but also through movements, facial expressions, and hand shapes. Therefore, recognizing individual words alone is often insufficient for effective communication. Additional research is needed to generate coherent sentences following recognition.

Previous research on sign language recognition has the following limitations. First, most existing studies focus on word-level recognition using video data, which imposes inherent recognition constraints. These studies are mainly centered on word classification and do not consider sentence-level conversion [3]. However, if recognition is limited to individual words, users must manually combine the words and infer context on their own, which can hinder real-time communication. For example, if the recognized signs are "we", "want", "deaf", and "help", the intended sentence might be "We want to help the deaf". However, a simple word concatenation approach might yield "We want deaf help", which distorts the meaning. Such misinterpretations often result in distorted or inaccurate meanings compared to natural Korean sentences. In optimization terms, generating natural sentences requires evaluating different word arrangements to find the most accurate and fluent expression. This process is analogous to biological optimization algorithms, where systems explore various solutions while converging toward optimal outcomes. Second, a critical lack of data makes data augmentation essential. Deep learning models are highly sensitive to both the quality and quantity of training data, yet, sign language datasets are far more difficult to construct and typically smaller in scale than text-based datasets. To address this, we apply data augmentation to existing sign language video data, expanding the dataset and enabling the model to learn more general patterns.

Accordingly, this study proposes a method to improve the performance of KSL recognition using data augmentation and generate natural sentences from recognized sign words using GPT-based natural language processing. By applying data augmentation techniques to existing KSL videos, we expand the training data and enhance the model's generalization capabilities. As sign language data is hard to acquire and varies significantly in expression, augmentation plays a vital role in boosting recognition performance. Furthermore, contrary to previous approaches that rely on simple parallel matching, we implement a context-aware sentence generation approach to produce more natural translations. Using GPT, individually recognized sign words are connected into fluent, semantically accurate sentences, thereby improving the overall quality of translation. Finally, we evaluate the effectiveness of this approach by comparing the recognition accuracy before and after data augmentation, and by assessing the GPT-based sentence generation results.

The remainder of this paper is organized as follows: Section 2 reviews prior work on sign language recognition, with a focus on KSL and data augmentation techniques. Section 3 details the proposed methodology, including the sign recognition model enhanced by data augmentation and the GPT-based sentence generation method. Section 4 presents experimental results and compares the performance of the proposed system with existing approaches. Section 5 concludes the paper and discusses future research directions.

2 Related Works

2.1 Trends in Sign Language Recognition Research

Sign Language Recognition (SLR) is an actively researched field aimed at facilitating smooth communication between deaf or hard-of-hearing individuals and those who hear. With recent advances in deep learning, video-based SLR models have been applied to various national sign languages. For example, in studies on American Sign Language (ASL), a large-scale word-level video dataset called WLASL has been constructed, enabling systematic comparisons of deep learning models such as I3D, 3D-ResNet, and CNN-LSTM for word-level recognition tasks [4]. In the case of British Sign Language (BSL), 3D CNNs combined with Bi-LSTM networks have been employed to effectively analyze spatiotemporal features of sign movements, thereby enhancing recognition performance for co-articulated and complex sentence structures [5]. Meanwhile, in Chinese Sign Language (CSL), research has focused on sign-to-text translation using large-scale video datasets and Transformer-based models, which have shown promising results in improving context-aware sentence-level translation performance [6]. These research efforts across different

countries have generally benefited from substantial data availability, enabling the development of high-precision models.

In Korea, deep learning approaches are also being applied to KSL. A representative dataset used in KSL research is the KETI KSL dataset, which has supported various related studies. Ko et al. (2020) proposed a model that uses OpenPose to extract keypoints from the face, hands, and body, which are then fed into a sequence-to-sequence neural network to translate sign language videos into natural language sentences [7]. This study contributed to improving recognition accuracy using the KETI dataset. Sim et al. (2022) built a parallel sign-sentence dataset based on COVID-19 government briefings and explored sentence translation using a Transformer-based TSPNet model, achieving a BLEU-4 score of 3.63, which set a benchmark for sign language translation performance [8]. Kim et al. (2022) developed an online platform for sign language education using deep learning-based SLR technologies. Their platform allowed learners to study sign language in a web-based environment and receive real-time feedback on recognition results. The model was trained using sign language videos from the National Institute of the Korean Language and AI HUB [9].

However, the accessibility of sign language data varies significantly by country. Nations, such as the U.S., U.K., and China—where SLR research is more active—have large-scale datasets that provide a favorable environment for training deep learning models. In particular, the U.S. has multiple large sign language datasets (e.g., RWTH-PHOENIX-Weather, ASLLVD), which have enabled a wide range of research efforts [10,11]. In contrast, countries, such as Korea, where SLR research is still developing, face limitations in both the quantity and quality of available data, making it difficult to train high-performance models. Contrary to text data, sign language data must be collected in video format and account for variables, such as camera angles, lighting conditions, and individual variation in signing, making dataset construction more complex. Therefore, in data-scarce environments, data augmentation techniques are essential to compensate for limited resources.

In addition, KSL research faces structural and linguistic limitations. Korean and KSL differ significantly in grammatical structure and expression. Korean uses a wide range of grammatical particles to indicate syntactic relationships—such as subject, object, conjunctive, and interjection particles. In contrast, KSL tends to omit many of these particles, expressing syntactic relationships through spatial arrangement, context, and facial expressions. Only several particles, such as "부터" ("from") and "와" ("and") are explicitly used as independent signs [12]. For instance, the Korean sentence "우리는 농인을 돕고 싶다" ("We want to help the deaf") would be expressed in KSL as a series of key concepts: "우리, 농인, 돕다, 원하다" ("we, deaf person, help, want"), omitting subject and object markers. Consequently, simply listing recognized signs often fails to convey the complete intended meaning. This grammatical discrepancy highlights the importance of transforming recognized sign words into grammatically and semantically accurate Korean sentences. In Korean, particles and verb endings are essential for both grammatical structure and meaning. Therefore, it is crucial to generate sentences that include appropriate particles and verb endings by considering word context and semantic relationships. Most existing SLR studies focus on improving word-level recognition accuracy, with relatively little research on converting recognized sign sequences into natural Korean sentences. As the word order and grammatical markers in KSL differ significantly from spoken Korean, simple word concatenation often fails to produce natural translations. Prior studies have mainly used word-to-word matching methods, which often result in awkward or unnatural translations due to a lack of contextual consideration. Recent state-of-the-art SLR models such as CRAM-SLR, Multilingual Skeleton-Based SLR, and Multi-Scale Spatio-Temporal Feature Enhancement primarily focus on improving recognition accuracy through skeleton-based feature extraction, temporal attention mechanisms, and multi-scale motion modeling. While these approaches advance the classification of sign units, they remain centered on the recognition stage. To overcome these limitations, this study proposes a sentence transformation

model that takes word-level KSL recognition outputs as input and generates natural, grammatically correct sentences. Furthermore, to address the problem of data scarcity, we apply data augmentation techniques to improve sentence-level recognition and translation performance.

2.2 Data Augmentation

Although deep learning has shown outstanding performance and rapid advancement across various artificial intelligence fields, it still heavily relies on the quantity and quality of training data. This dependency is especially prominent in video-based deep learning technologies; without sufficient data, models are prone to overfitting to specific datasets and may struggle to generalize effectively in real-world environments. To address this challenge, researchers have actively explored and applied data augmentation techniques. The most common video data augmentation methods include image rotation, scaling, horizontal or vertical flipping, brightness and contrast adjustments, and noise injection [13]. More recently, studies have utilized Generative Adversarial Networks (GANs) to generate synthetic data that closely resemble real-world data, effectively increasing the volume of training data.

In the context of sign language recognition, data augmentation has also been explored as a means to improve model performance. Walsh et al. showed that simple augmentation methods—such as photo-realistic variations in camera angle and background—can mitigate data scarcity issues, even in resource-rich contexts like ASL [14]. Perea-Trigo et al. developed the CALSE-1000 Spanish Sign Language dataset and applied a variety of augmentation techniques, including face swapping, affine transformations, and brightness and noise variations, significantly enhancing recognition accuracy [15]. Moryossef et al. proposed a text-based augmentation method that generates synthetic parallel data to overcome the shortage of training data in sign language translation tasks, improving overall translation performance [16]. However, existing approaches tend to emphasize spatial transformations of video data, often overlooking the temporal characteristics specific to sign language. In the case of KSL, a major limitation is the extreme scarcity of data—public platforms often provide only a single video per word, resulting in a critically small dataset. In such data-limited settings, models are highly susceptible to overfitting.

Furthermore, sign language is inherently dynamic, comprising not only static hand shapes but also time-series elements, such as hand movement, speed, and direction. Thus, basic visual transformations alone are insufficient to address the fundamental challenge of data scarcity. To overcome these limitations, this study applies four augmentation techniques—noise injection, quantization, time warping, and time shifting—based on existing KSL videos. These four methods are specifically designed to account for the time-series nature of sign language data, enabling more effective data expansion and improved model generalization.

3 Proposed Method

This study proposes a method to improve the performance of KSL recognition models through data augmentation, while also converting recognized sign words into grammatically and contextually appropriate Korean sentences. In particular, this approach addresses the challenges of limited data availability and unnatural sentence generation observed in existing SLR systems by constructing a dataset using data augmentation and applying a context-aware sentence generation model. Section 3.1 presents the overall system structure and process in which the proposed method is applied. Section 3.2 presents data augmentation techniques for efficiently expanding limited KSL datasets to enhance the performance of both recognition and translation models. Finally, Section 3.3 describes in detail how to generate natural Korean sentences from recognized sign words by considering contextual and grammatical relationships.

3.1 System Architecture

The proposed sign language recognition and sentence generation system proposed consists of three main modules: dataset construction based on data augmentation, sign language video recognition, and sentence generation with contextual awareness. Fig. 1 illustrates the architecture of the system comprising these three modules.

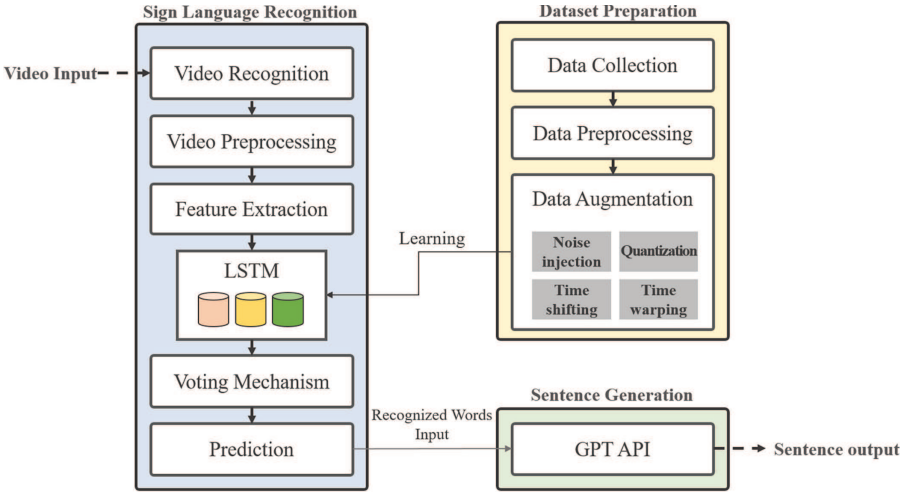

Figure 1: Overall architecture of the proposed KSL recognition–generation framework. The system comprises three main modules: (1) dataset preparation with data augmentation, (2) LSTM-based sign language recognition, and (3) GPT-based sentence generation.

The Dataset Preparation module collects training data, preprocesses it, and applies data augmentation techniques for sign language recognition. In this study, sign language videos were collected using the API provided by the Culture Public Data Portal [17]. By specifying parameters, such as service key, row count, and page number, video URLs for specific sign language words were obtained. This module also applies data augmentation techniques to the collected videos to expand the dataset. The techniques used include noise injection, quantization, time warping, and time shifting, which are described in detail in Section 3.2.

The SLR module performs the core recognition task by processing either prerecorded or real-time sign language videos. The system extracts frames from the video, identifies hand movements and pose information, and conducts preprocessing tasks, such as frame normalization, background removal, and hand landmark extraction to derive meaningful features. For feature extraction, OpenCV and MediaPipe are used to obtain detailed hand and posture information from the video. MediaPipe Pose provides 33 landmark coordinates (x, y, z) for the user's body. From these, vectors between the shoulders and wrists are calculated, then normalized and analyzed using inner product-based angle computation to extract pose features. MediaPipe Hands analyzes joint data from each finger by calculating vectors between adjacent joints, thereby quantifying finger structure and movement. Combining these pose and hand landmark-based features, a total of 257 features per frame are extracted. To train the LSTM model, input data is formed by grouping 30 consecutive frames into sequence of shape [30, 257]. After preprocessing and feature extraction, an LSTM model is used to generate word-level predictions. The proposed model architecture follows our previous study [18], consisting of two LSTM layers with 256 hidden units each and a Dropout layer with a rate of 0.2. The LSTM output is passed through a Dense layer (128-dimensional) and a Softmax layer for final classification. The model was trained using the Adam optimizer (learning rate = 0.001) and the Sparse Categorical Cross-Entropy loss function. The LSTM processes the sequential data to recognize sign language

words expressed in the video frames, effectively capturing the temporal characteristics and continuity of the gestures.

Once all intended words from the user have been recognized, a Voting Mechanism is applied to determine the final prediction. The predicted word probabilities from each frame are aggregated over a defined temporal segment, and the word with the highest cumulative probability within that segment is selected as the final recognition result. Additionally, if no hand is detected for more than 10 consecutive frames, the system interprets this as the end of the current word and initiates processing for a new word. This majority-voting approach mitigates transient recognition errors and enhances the overall stability of word-level predictions. When a word-end signal is detected, the accumulated list of recognized words is passed to the Sentence Generation module. This final module develops a grammatically consistent and contextually appropriate Korean sentence, which is then provided to the user as the final output.

3.2 Dataset Construction through Data Augmentation

As previously noted, countries with relatively fewer sign language research initiatives, such as Korea, face challenges in acquiring sufficiently large datasets for training deep learning models. In this study, the sign language videos used from the Culture Public Data Portal provide only a single video per word. To enhance recognition performance, it is essential to consider the sequential dynamics of sign language, as each sign consists of a series of temporally dependent movements rather than isolated frames. Therefore, this study applies data augmentation techniques tailored to the characteristics of KSL to address challenges of data scarcity and overfitting. Four augmentation techniques were applied: noise injection, quantization, time warping, and time shifting. These methods were designed to align with the structural and temporal nature of sign language sequences.

First, noise injection was applied to the input sequence data. The noise injection method is described in Algorithm 1. Random noise was added to the first 78 channels representing hand movement features. For the hand-related features in channels 78 to 155, noise was added only when valid values were present, considering cases where certain features, such as finger joints might be missing or inaccurately detected. For the final 101 channels, which include hand position and body pose information, noise was added with a different intensity to simulate variation due to posture changes. This method allows the model to better adapt to minor variations in input features.

Algorithm 1: Noise injection

Input: Feature vector $F = \{x_1, x_2, \ldots, x_n\}$, noise factor σ
Output: Augmented vector F'
for each x_i in F **do**
    $x_i' = x_i + \text{Normal}(0, \sigma)$
end for
return F'

Second, quantization was used to convert continuous values into discrete ones, reducing the model's sensitivity to subtle numerical changes and promoting generalization. The quantization method is described in Algorithm 2. Using floating-point values as-is can cause the model to respond excessively to insignificant fluctuations, potentially hindering training. By quantizing the data to a fixed range, the model can focus on essential features. This integer transformation improves learning stability and enhances generalization performance. Third, time warping reflects individual differences in signing speed by slightly shifting the sequence along the time axis, either forward or backward, simulating variations in gesture start times.

During this process, certain frames are duplicated to create a naturally distorted sequence. The sequence length varies within a predefined maximum distortion ratio, and the final output sequence is adjusted to meet a fixed length requirement. This method reflects the fact that sign language expressions vary in speed across individuals, thereby improving the model's adaptability to various signing tempos.

Algorithm 2: Time warping

1. **Input**: Feature sequence $F = \{x_1, x_2, \ldots, x_n\}$, warp ratio α
2. **Output**: Warped sequence F'
3. Generate random warping factor $r \in [1, 1 + \alpha]$
4. Compute new sequence length $n' = \text{round}(r \times n)$
5. Resample F to new length n' using interpolation
6. **return** $\boldsymbol{F}$

Finally, the time shifting technique augments the input sequence by arbitrarily shifting it left or right along the time axis. The newly created sections from the shift are padded with zeros to indicate the initial state. This reflects real-world conditions in which the starting point of a sign in a video may not be consistent. Time shifting compensates for temporal variability, especially in sentence-level sign language recognition, where the boundaries between expression segments are often unclear. Each augmentation technique was implemented directly within the preprocessing pipeline developed in this study, and was designed to reflect the structural characteristics of the features extracted from MediaPipe: hand-related features (dimensions 1–155) and body posture features (dimensions 156–257). Specifically, Noise Injection applied varying degrees of random perturbation to the hand joint coordinates (dimensions 1–78) and posture features (dimensions 156–257) to simulate sensor noise and reduce recognition errors. Time Warping adjusted the sequence length by a random ratio within the range [1, 1 + α] to emulate variations in signing speed, while Time Shifting randomly shifted the sequence within a range of ±5 frames to account for irregular starting points of gestures. These techniques were carefully designed to expand the spatiotemporal diversity of the training data without altering its semantic meaning, thereby improving the model's robustness and generalization performance.

This study designed each augmentation method with attention to the temporal and visual characteristics of sign language data. By applying these diverse augmentation techniques, the model was trained on a wider variety of transformed input cases derived from a limited original dataset. Data augmentation is crucial for preventing overfitting and building robust recognition systems that can maintain high performance across varied environments and user conditions. These techniques also have broad applicability not only to sign language but also to other types of time-series data.

3.3 Context-Aware Sentence Generation

Sign language recognition systems output at the word level, which differs from conventional speech or text-based natural language processing systems. In particular, due to the omission of grammatical elements, non-standard word order, and reliance on non-verbal cues, sign language often expresses meaning in a simplified or condensed manner, even when conveying the same idea. Consequently, it can be difficult to generate natural-sounding sentences directly from the raw word outputs of sign language recognition systems. To address this challenge, this study proposes a method for generating contextually appropriate and natural sentences from lists of recognized sign language words. Rather than simply stringing words together, this process analyzes semantic relationships and automatically fills in omitted grammatical components—such as particles and verb endings—to form complete, fluent sentences.

To overcome the limitations of naïve word concatenation, a GPT-based sentence generation module was developed. This module takes the list of individually recognized words from the sign language input and reconstructs them into a full sentence. Using the GPT API, the module analyzes the grammatical and contextual relationships between words and attaches appropriate particles and verb endings to create coherent and natural Korean sentences. The conditions used in this sentence generation process is outlined in Table 1.

Table 1: Conditions for the sentence generation module.

Condition Item	Description
Synonym Group Selection	Only one word is selected from each synonym group for sentence generation.
Word Usage Restriction	Words not included in the input list are not used.
Particle and Ending Handling	Appropriate particles and verb endings are automatically added to form grammatically natural sentences.

When the sign language recognition module recognizes a sign, it returns a list of synonyms representing the meaning of the gesture. For sentence generation, only one representative word from each synonym group is used. For example, if the input is given in the form of synonym groups, such as [[“criminal”, “offender”], [“action”, “behavior”, “movement”], [“reason”, “meaning”, “intent”]], the sentence generation model selects one word from each group—say, criminal, action, and meaning—to generate a sentence, such as There is meaning in the criminal’s action. Furthermore, to ensure semantic accuracy, the model is constrained from using any words not included in the input. It also applies appropriate Korean particles and verb endings to maintain grammatical coherence between words, resulting in more natural and linguistically accurate sentences. This approach addresses the awkwardness often found in previous systems that relied solely on parallel word matching, and more conveys the signer’s intended meaning.

4 Experiments and Results

4.1 Data Sets and Evaluation Measures

To evaluate the proposed method, we used Korean Sign Language video materials obtained from the Culture Portal, a publicly accessible cultural data platform operated by the Ministry of Culture, Sports and Tourism of Korea and the Korea Culture Information Service Agency (KCISA) [17]. As this platform provides officially curated and managed cultural content, it offers a reliable basis for constructing the dataset used in this study. The platform provides one video per word; therefore, we applied the data augmentation techniques described in Section 3.2 to generate a total of 3552 word-based sequence data samples. The composition of this dataset is presented in Table 2.

Table 2: Summary of the dataset.

Dataset Component	Value
Number of Sequences	6,129,258
Sequence Length	30 frames
Features per Frame	257 features

All experiments were conducted on a system equipped with an Intel Core i9-10980XE CPU (3.00 GHz), 256 GB of RAM, and four NVIDIA GeForce RTX 3090 GPUs.

4.2 Results of the Experiments

The focus of this study was not on large-scale performance optimization but on exploring the feasibility of integrating KSL recognition with GPT-based sentence generation. As a proof of concept, we tested recognition accuracy for 10 selected sentences comprising a total of 32 unique words. Recognition accuracy was measured based on whether each word was successfully recognized, depending on the number of repetitions (n = 1, 3, 5 times). The test sentences consisted of expressions commonly used in daily life and communication contexts. Table 3 presents the test sentences and associated words required to express each sentence (see Table A1 for Korean version).

Table 3: Sentences used in the experiment.

Sentence	Words Used
We want to help the deaf.	we, want, deaf person, help
I ate bread for lunch.	I, lunch, bread, eat
I don't know the suspect's behavioral motive.	suspect, behavior, motive, don't know
Please tell me where the restroom is.	restroom, where, tell
How much is this? Please calculate it.	this, how much, calculate
How old are you?	you, age, what
I'm already tired and exhausted.	already, tired, exhausted
Please come to the staff meeting.	staff, meeting, come
Leaving work on time.	on time, home, go
Please approve this.	stamp, request

We measured the recognition success for each of the 32 words across the 10 sentences at repetition levels of 1, 3, and 5. The recognition accuracy by repetition count is shown in Table 4, and the recognized words at each repetition level are presented in Table 5.

Table 4: Word recognition accuracy results.

Repetition Count	Accuracy
1	0.468
3	0.625
5	0.812

Table 5: Recognized and unrecognized words by repetition count.

Repetition Count	Recognized Words	Unrecognized Words
1	we, want, deaf person, bread, eat, behavior, motive, tell, how much, age, already, exhausted, meeting, home, request	help, I, lunch, suspect, don't know, restroom, where, this, calculate, you, what, tired, staff, come, on time, go, stamp
3	we, deaf person, help, I, bread, eat, behavior, motive, don't know, where, this, calculate, you, age, already, tired, exhausted, meeting, home, stamp	want, lunch, suspect, restroom, tell, how much, what, staff, come, on time, go, request
5	we, want, deaf person, help, I, lunch, bread, eat, suspect, behavior, motive, don't know, where, tell, this, calculate, you, age, already, tired, staff, meeting, come, home, stamp, request	restroom, how much, what, exhausted, on time, go

In the one-time repetition test, an accuracy of approximately 46.8% was achieved. Relatively simple words, such as "우리" (we), "농인" (deaf person), "빵" (bread), and "먹다" (eat) were successfully recognized, while more complex or brief gestures, such as "돕다" (help), "나" (I), "화장실" (restroom), and "어디" (where) showed lower recognition rates. In the three-time repetition test, the accuracy improved to approximately 62.5%. Words that were initially misrecognized, such as "돕다", "나", "모르다" (don't know), "계산하다" (calculate), "당신" (you), and "피곤" (tired), showed improved recognition through repeated input. In the five-time repetition test, a significant improvement was observed, reaching an accuracy of approximately 81.25%. Numerous words that failed in earlier tests—such as "범인" (criminal), "점심" (lunch), "직원" (staff), and "오세요" (come)—were successfully recognized, demonstrating the effectiveness of repeated input in improving recognition accuracy. Overall, word recognition improved with increased repetition.

However, some words—such as "화장실" (restroom), "무엇" (what), "칼" (on time), and "가다" (go)—showed low recognition rates even after multiple repetitions. These words were associated with shorter motions, which are more difficult to recognize within the model's 30-frame input structure. In contrast, words that aligned better with the frame structure—such as "우리" (we), "농인" (deaf person), "빵" (bread), "먹다" (eat), "행동" (behavior), "동기" (motive), "나이" (age), "벌써" (already), "회의" (meeting), and "집" (house)—achieved better recognition. Well-recognized signs involved larger, more dynamic movements, while poorly recognized ones were smaller and performed with one hand, making them harder to distinguish from similar gestures. These results highlight the impact of motion size and duration on recognition performance, suggesting that improved recognition of short or subtle signs through model and dataset refinement could significantly enhance overall accuracy.

In the second experiment, we evaluated the model's ability to generate natural and semantically accurate sentences based on recognized word lists. We provided input word lists and assessed whether the generated sentences effectively conveyed the intended meaning. Multiple sentences generated from each word list were analyzed. Table 6 presents the sentence generation results, excluding duplicates.

Table 6: Sentence generation experiment results.

Given Word List	Target Sentence	Experimental Results
[we (casual), we (formal)], [want, crave, long for], [deaf person], [help]	We want to help the deaf.	We wish to help our deaf. We (humbly) wish to help the deaf. We would like to help the deaf. We (respectfully) long to help the deaf. We want to help the deaf (spoken/polite). We (humbly) wish to help the deaf. We wanna help the deaf. (casual, conversational) We deeply long to help the deaf. We wish to help our deaf. We (respectfully) long to help the deaf. We (humbly) wish to help the deaf. We (humbly) long to help the deaf. We deeply long to help the deaf. We long to help our deaf. **We (humbly) want to help the deaf who are longing.**
[oneself, I, I (formal)], [lunch], [bread], [eat]	I ate bread at lunch.	I had bread for lunch. It was me who ate bread at lunch. (emphasis on subject) I eat bread at lunch. (polite, present tense; habitual) I ate bread at lunch. (polite, past tense) I eat bread for lunch. (casual, descriptive; present or neutral tense) I eat bread at lunch. (polite, present tense; habitual) I had bread for lunch. (polite but casual tone) I eat bread at lunch. (casual, narrative style) Me, eat bread at lunch. I had bread for lunch. (emphasis on "I") I ate bread at lunch. (polite, past tense) I had bread for lunch. (formal, slightly more specific on the role of bread in the meal)
[perpetrator, criminal], [action, behavior, movement], [reason, meaning, pretext, significance], [don't know]	I don't know the criminal's motive.	I can't figure out why the perpetrator acted the way they did. I can't understand the perpetrator's behavior. I don't get the perpetrator's behavior. The reason behind the perpetrator's behavior is beyond understanding. I don't understand the perpetrator's movements. The perpetrator's behavior doesn't make logical sense. The reason for the perpetrator's behavior is unknown. I can't figure out the cause of the perpetrator's behavior. The reason behind the perpetrator's actions is unclear. I can't tell what made the perpetrator act that way. The reason for the perpetrator's behavior remains unknown. The perpetrator's actions cannot be explained. I can't understand the perpetrator's behavior. **The perpetrator's behavior cannot be known.** The perpetrator's behavior is incomprehensible. The rationale behind the perpetrator's actions cannot be grasped.

Note: Bold indicates semantically mismatched outputs compared to the reference translation.

In the first word list, the generated sentences generally conveyed the intended meaning accurately. The model appropriately used particles and verb endings to produce natural sentences. However, some sentences lacked a subject particle for "우리" (we), resulting in unclear sentence subjects. Additionally, some expressions (e.g., "저희는 갈망하는 농인을 돕고 싶습니다", meaning "We want to help the deaf who are longing") could be misinterpreted. In this case, the sentence implies that the deaf person is the one who

longs for something, rather than conveying our intent to help, which differs from the original intent of "We want to help the deaf". For the second word list, the generated sentences were mostly natural, correctly using tense and grammatical particles, and showed minimal semantic deviation from the intended meaning. For the third word list, the model-generated sentences generally aligned with the intended meaning overall, though some combinations felt slightly awkward due to the large number of input words. In particular, certain sentences (e.g., "현행범이란 행동의 까닭을 알 수 없다", meaning "The reason behind the action known as the perpetrator is unclear") included grammatically unnatural expressions.

The sentence generation module achieved 95% semantic accuracy, calculated based on Table 6, where generated sentences were considered correct only when their meaning fully aligned with the reference translation. Mismatched cases are indicated in bold. To further clarify the nature of these semantic errors, Table 7 presents representative cases illustrating typical mismatch patterns observed during sentence generation. Each generated sentence was evaluated against a reference translation and considered correct if it accurately conveyed the intended meaning. This result demonstrates the model's capability to produce natural and semantically appropriate Korean sentences from recognized sign sequences. However, observed challenges, such as grammatical errors and unnatural expressions suggest that further improvement is needed, particularly in the preprocessing of word selection and postprocessing of sentence outputs.

Table 7: Representative semantic error cases in sentence generation.

Target Sentence	Generated Sentence	Error Analysis
We want to help the deaf.	We want to help the deaf who are longing.	An ambiguous verb candidate (*long for*) was incorrectly attached as a modifier of the noun phrase, resulting in unintended semantic modification.
I don't know the criminal's motive.	The perpetrator's behavior cannot be known.	The target sentence expresses uncertainty about the criminal's motive, but the generated sentence incorrectly shifts the meaning to express uncertainty about the behavior itself, fundamentally altering the intended semantics.

5 Conclusion and Future Works

This study proposes a data augmentation strategy and a GPT-based sentence generation approach to enhance KSL recognition accuracy and sentence translation quality. By applying data augmentation techniques—namely noise injection, quantization, time warping, and time shifting—based on hand and pose features, we mitigated overfitting in a limited dataset and enhanced the generalization performance of the model.

To validate the performance of the sign language recognition model, we conducted recognition accuracy experiments based on varying repetition counts. The results showed clear performance improvements: approximately 46.8% accuracy with one repetition, 62.5% with three repetitions, and 81.25% with five repetitions. Signs with large and distinct movements, such as "우리" (we), "농인" (deaf person), and "빵" (bread) were recognized with high accuracy, whereas short and static gestures, such as "화장실" (restroom), "무엇" (what), "칼" (on time), and "가다" (go) remained difficult to recognize even with repeated attempts. These results underscore the importance of motion length, gesture size, and hand movement complexity for

recognition performance. Future work should focus on improving recognition accuracy for short and static signs through data enhancement and model architecture optimization.

Additionally, the GPT-based sentence generation experiment, which aimed to convert recognized word lists into coherent sentences, achieved approximately 95% semantic accuracy. The model generated grammatically correct and natural Korean sentences from the provided word combinations. However, some instances exhibited awkward syntax or semantic ambiguity due to constraints in word selection during the input process. These issues suggest that further improvements could be made by refining the preprocessing step, incorporating syntactic analysis based on sentence components, and applying postprocessing techniques.

The proposed data augmentation and GPT-based natural language generation techniques contributed meaningfully to KSL translation. By producing sentence-level outputs, the study demonstrated the practical viability of the system. In future research, we plan to build a large-scale learning environment using the KETI KSL dataset and additional sign language video corpora, and to perform comparative experiments with various models such as CNN-LSTM and Transformer-based SLR architectures.

Acknowledgement: Not applicable.

Funding Statement: This work was partly supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP)-Innovative Human Resource Development for Local Intellectualization Program grant funded by the Korea government (MSIT) (IITP-2026-RS-2022-00156334, 50%) and the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2021R1C1C2011105, 50%).

Author Contributions: Conceptualization, Sun-Young Ihm; methodology, Sun-Young Ihm; validation, Soo-Yeon Jeong, Ho-Yeon Jeong and Sun-Young Ihm; formal analysis, Soo-Yeon Jeong; investigation, Soo-Yeon Jeong and Sun-Young Ihm; writing—original draft preparation, Soo-Yeon Jeong and Sun-Young Ihm; writing—review and editing, Soo-Yeon Jeong and Sun-Young Ihm; visualization, Soo-Yeon Jeong; supervision, Sun-Young Ihm; project administration, Sun-Young Ihm; funding acquisition, Sun-Young Ihm. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The data that support the findings of this study are openly available in [Culture Open Data Portal] at [https://www.culture.go.kr].

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A

Table A1: Table 3 in Korean.

Sentence	Words Used
우리는 농인을 돕고 싶다.	우리, 원하다, 농인, 돕다
나는 점심에 빵을 먹었다.	나, 점심, 빵, 먹다
범인의 행동동기를 모르겠다.	범인, 행동, 동기, 모르다
화장실이 어딘지 알려주세요.	화장실, 어디, 가르치다
이건 얼마에요 계산해주세요.	이것, 얼마, 계산하다
몇살이세요?	당신, 나이, 무엇
벌써 피곤해서 힘들다.	벌써, 피곤, 힘들다
직원 회의에 오세요	직원, 회의, 오세요

(Continued)

Table A1 (continued)

Sentence	Words Used
칼퇴근	칼, 집, 가다
결재 부탁드립니다.	도장, 부탁

References

1. Pu J, Zhou W, Li H. Dilated convolutional network with iterative optimization for continuous sign language recognition. In: Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence; 2018 Jul 13–19; Stockholm, Sweden.
2. Camgoz NC, Hadfield S, Koller O, Ney H, Bowden R. Neural sign language translation. In: Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2018 Jun 18–23; Salt Lake City, UT, USA. p. 7784–93.
3. Tanzer G, Shengelia M, Harrenstien K, Uthus D. Reconsidering sentence-level sign language translation. arXiv:2406.11049. 2024.
4. Li D, Opazo CR, Yu X, Li H. Word-level deep sign language recognition from video: a new large-scale dataset and methods comparison. In: Proceedings of the 2020 IEEE Winter Conference on Applications of Computer Vision (WACV); 2020 Mar 1–5; Snowmass Village, CO, USA. p. 1448–58.
5. Albanie S, Varol G, Momeni L, Afouras T, Chung JS, Fox N, et al. BSL-1K: scaling up co-articulated sign language recognition using mouthing cues. In: Proceedings of the Computer Vision—ECCV 2020; 2020 Aug 23–28; Glasgow, UK. Cham, Switzerland: Springer International Publishing; 2020. p. 35–53.
6. Zhou H, Zhou W, Zhou Y, Li H. Spatial-temporal multi-cue network for continuous sign language recognition. Proc AAAI Conf Artif Intell. 2020;34(7):13009–16. doi:10.1609/aaai.v34i07.7001.
7. Ko SK, Kim CJ, Jung H, Cho C. Neural sign language translation based on human keypoint estimation. Appl Sci. 2019;9(13):2683. doi:10.3390/app9132683.
8. Sim H, Sung H, Lee S, Cho H. Sign language dataset built from S. Korean government briefing on COVID-19. KIPS Trans Softw Data Eng. 2022;11(8):325–30. doi:10.3745/KTSDE.2022.11.8.325.
9. Kim MS, Yoo MH, Jang JH, Choi JY, Na GE, Kim MS, et al. Implementation of deep learning-based sign language acquisition online platform. J Digit Contents Soc. 2022;23(11):2147–57. doi:10.9728/dcs.2022.23.11.2147.
10. Zhang X, Duh K. Handshape-aware sign language recognition: extended datasets and exploration of handshape-inclusive methods. In: Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2023; 2023 Dec 6–10; Singapore. p. 2993–3002.
11. Zhou Y, Xia Z, Chen Y, Neidle C, Metaxas DN. A multimodal spatio-temporal GCN model with enhancements for isolated sign recognition. In: Proceedings of the LREC-COLING 2024 11th Workshop on the Representation and Processing of Sign Languages: Evaluation of Sign Language Resources; 2024 May 25; Torino, Italia. p. 408–19.
12. National Institute of Korean Language, Korean Association of the Deaf, Korean sign language studies. Vol. 2. Seoul, Republic of Korea: National Institute of Korean Language; 2010.
13. Shorten C, Khoshgoftaar TM. A survey on image data augmentation for deep learning. J Big Data. 2019;6(1):60. doi:10.1186/s40537-019-0197-0.
14. Walsh H, Ivashechkin M, Bowden R. Using sign language production as data augmentation to enhance sign language translation. arXiv:2506.09643. 2025.
15. Perea-Trigo M, López-Ortiz EJ, Soria-Morillo LM, Álvarez-García JA, Vegas-Olmos JJ. Impact of face swapping and data augmentation on sign language recognition. Univers Access Inf Soc. 2025;24(2):1283–94. doi:10.1007/s10209-024-01133-y.
16. Moryossef A, Yin K, Neubig G, Goldberg Y. Data augmentation for sign language gloss translation. In: Proceedings of the 1st International Workshop on Automatic Translation for Signed and Spoken Languages (AT4SSL); 2021 Aug 20; Virtual. p. 1–11.

17. Ministry of Culture, Sports and Tourism. Culture open data portal [Internet]. Seoul, Republic of Korea: MCST; [cited 2025 Sep 26]. Available from: https://www.culture.go.kr.
18. Jeong SY, Jeong HY, Ihm SY. Korean sign language recognition using LSTM and video datasets. J Inf Process Syst. 2025;21(4):427–38. doi:10.3745/JIPS.04.0356.