Overview of CCL24-Eval Task 10: Translation Quality Evaluation of Sign Language Avatar

Yuan Zhao[1*], Ruiquan Zhang[2,3*], Dengfeng Yao[1,4†], Yidong Chen[2,3†]
[1]Beijing Key Laboratory of Information Service Engineering, Beijing Union University
[2]School of Informatics, Xiamen University
[3]Key Laboratory of Digital Protection and Intelligent Processing of Intangible Cultural Heritage of Fujian and Taiwan, Ministry of Culture and Tourism, Xiamen University
[4]Lab of Computational Linguistics, School of Humanities, Tsinghua University
{annzy,tjtdengfeng}@buu.edu.cn, rqzhang@stu.xmu.edu.cn, ydchen@xmu.edu.cn

Abstract

Sign Language Avatar technology aims to create virtual agents capable of communicating with deaf individuals through sign language, similar to the text dialogue agent ChatGPT but focusing on sign language communication. Challenges in sign language production include limited dataset sizes, information loss due to reliance on intermediate representations, and insufficient realism in generated actions. In this event, we particularly focus on the ability of the Sign Language Avatar to translate spoken language text into sign language that is easily understood by deaf individuals. As the first sign language avatar event held by the China National Conference on Computational Linguistics(CCL), this event attracted wide attention from both industry and academia, with 14 teams registering and 10 of them submitting their system interfaces on time. We provided a dataset consisting of 1074 text-video parallel sentence pairs for training, and the evaluation team comprised proficient Chinese sign language users and professional sign language translators. The scoring method employed a comprehensive evaluation based on multiple metrics, focusing primarily on sign language grammar accuracy, naturalness, readability, and cultural adaptability. The final scores were determined by considering performance across these four aspects. The final scores, taking into account these four aspects, showed that four teams demonstrated good readability, with Vivo Mobile Communication Co., Ltd. ranking first with a score of 3.513 (out of a full score of 5), leading the baseline model by 1.394 points. According to the analysis of the results, most teams used the traditional method of converting text into Gloss sequences before generating sign language. Additionally, some teams experimented with emerging methods, including gloss-free end-to-end training and Large Language Model(LLMs) prompt learning, which also achieved promising results. We anticipate that this event will promote the development of sign language avatar technology and provide higher-quality communication tools for the deaf community. For more information on this task, please visit the website of the CCL24-Eval: Translation Quality Evaluation of Sign Language Avatar Task[1].

1 Introduction

Sign language, a rich and complex form of communication with its own unique vocabulary and grammar, is used by over 70 million deaf and hard of hearing people worldwide. Unlike spoken languages, sign language emphasizes body language, incorporating hand shapes, movements, positions, and palm orientations, as well as non-manual elements like body posture and facial expressions (Qiu et al., 2018; Yao et al., 2019). Despite its widespread use and significance, the distinct differences and unique modes of expression in sign language pose challenges to its dissemination and understanding. The purpose of this evaluation competition is to assess sign language avatars that can translate spoken text into sign language, enhancing comprehension for deaf or hard of hearing individuals who use sign language.

[*] Co-First Author
[†] Corresponding Author
[1]Our task website:https://github.com/ann-yuan/QESLAT-2024

©2023 China National Conference on Computational Linguistics
Published under Creative Commons Attribution 4.0 International License

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 405-415, Taiyuan, China, July 25 - 28, 2024.
Volume3: Evaluations

Recently, sign language research has become an active area within Computer Vision (CV) and Natural Language Processing (NLP)(Yin et al., 2021; Yu et al., 2023; Zhao et al., 2024), particularly in Sign Language Recognition (SLR)(Chen et al., 2022; Wei et al., 2023; Hu et al., 2023) and Translation (SLT)(Fu et al., 2023; Sun et al., 2024; Hu et al., 2024). However, research publications in Sign Language Production (SLP), which are closely related to this evaluation task, are scarce (Rastgoo et al., 2021; Yao, 2022).

The goal of SLP is to translate spoken or written content into sign language, making it accessible for the deaf. SLP faces challenges such as capturing detailed hand and body movements and dealing with unique visual semantics and grammar. Additionally, limited datasets, difficulties in simulating sign details, and technological constraints in producing contextually accurate signs complicate SLP tasks(Ren et al., 2024).

Early sign language processing (SLP) methods relied on one-to-one gloss-sign correspondences, producing only isolated sign words (Stoll et al., 2018). Subsequent attempts employed machine translation techniques to translate spoken text into continuous gloss sequences, which were then converted to sign language gestures (Saunders et al., 2021; Zhu et al., 2023). (Saunders et al., 2021; Zhu et al., 2023). However, this approach often failed to capture contextual information, leading to semantic losses. A more recent development is the end-to-end SLP method, which directly translates text into sign language videos without intermediate gloss, showing significant improvements with larger data volumes (Baltatzis et al., 2023).

Regarding avatar technologies, initial methods employed 2D or 3D skeletal models extracted from videos (Kapoor et al., 2021; Saunders et al., 2020). Recent advancements have focused on generating sign language from these models through rendering techniques (Saunders et al., 2020; Zelinka and Kanis, 2020; Xiao et al., 2020). The use of pretrained avatar models like SMPL-X has also been explored for enhanced sign language representation (Pavlakos et al., 2019; Stoll et al., 2022).

In China, sign language avatars are gaining traction. Initiatives like ZHIPU's "AI Sign Language Classmate" and vivo's "Sign Language Translator" highlight a deep understanding and proactive response to the deaf community's communication needs. These avatars are now visible at events such as sports games and news programs. With ongoing technological progress, sign language avatars are expected to play a vital role, enhancing communication for the deaf community.

This evaluation assesses the effectiveness of sign language avatars by using recorded videos and multiple-choice questions for feedback. We expect this round to enhance the avatar team's understanding of the deaf community's needs and pinpoint improvement areas.

2 Task Description

The purpose of this evaluation is to assess the naturalness and accuracy of sign language avatars translating Chinese into Chinese sign language(CSL), ensuring that the translations adhere to sign language grammatical rules and are understandable and acceptable to the deaf community. During the evaluation, we collect and construct a rich corpus covering various common scenarios for teams to train on. Additionally, we design a series of test sentences, four in total, each accompanied by the participating team's sign language avatar videos, along with four multiple-choice questions and optional evaluations. Each multiple-choice question provides four options, allowing evaluators to choose during the evaluation process to determine which option best matches the performance of the sign language avatar.

3 Evaluation Method

In the fields of SLR and SLT, common metrics such as Accuracy, Recall, Word Error Rate (WER), and translation-specific indicators like BLEU, CIDEr, ROUGE, METEOR often require comparison with ground truth videos, which are difficult to obtain(Rastgoo et al., 2021; Rastgoo et al., 2022). Therefore, automated machine evaluation is not used. Additionally, the diversity in styles of sign language avatars complicates the use of a standardized automated evaluation method. Instead, human evaluation is preferred due to its flexibility, accuracy, and thorough analytical capabilities, making it the method of choice for assessing the naturalness and fluency of sign language translations.

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 405-415, Taiyuan, China, July 25 - 28, 2024.
Volume3: Evaluations
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Multiple-choice: A Specific Example	
Digital Avatar Video Example	已经为您办理完成，这是您的机票。请问需要其他帮助吗? *I have completed the process for you. Here is your plane ticket. Do you need any other help?*
Accuracy Assessment Question	Please assess whether the digital avatar accurately performed the key signs for "办理(process)", "机票(plane ticket)", and "帮助(help)".
A. All included (5 points)	The signs are complete and precise, clearly including all the signs for "办理(process)", "机票(plane ticket)", and "帮助(help)".
B. Partially included (3 points)	The signs are partially correct, including at least 1 or 2 of the specified signs.
C. Signs semantically incorrect (1 point)	There are sign changes, but these signs do not accurately reflect the specific semantics of "办理(process)", "机票(plane ticket)", and "帮助(help)".
D. Signs unclear and unreadable (0 points)	The sign language does not correspond to the provided content, or the signs are too unclear to be recognizable.

Table 1: A specific example in multiple-choice questions

Participating teams are required to provide a sign language avatar interface for recording videos corresponding to the competition's test scenarios. These translation results will be scored by an evaluation team, consisting of 20 deaf individuals and professional translators certified by the Committee for Sign Language Research and Promotion of the Chinese Deaf Association, with each evaluation metric having a maximum score of 5 points. The final scores for each team will be calculated based on the feedback and scores from the evaluators, ensuring a comprehensive and fair assessment.

Next, we will introduce the evaluation metrics and scoring methods for this competition.

3.1 Metrics

Sign Language Grammar Accuracy Sign language grammar accuracy refers to the translation adhering to the semantics and grammatical rules of the target language. The translation must follow the word order and structural rules of CSL, as the subject-verb-object structure of Mandarin may need to be adjusted to a sequence more customary in sign language. The correctness of gesture forms is also crucial, ensuring the appropriate use of finger positions, palm directions, and hand movements. Additionally, grammatical markers in sign language, such as tense, negation, and questions, must also be accurately expressed in the translation, which is vital for conveying the complete meaning of sentences.

Naturalness Naturalness emphasizes the fluency and naturalness of the translation, making the sign language close to the natural communication methods of the deaf community. The gestures in the translation must be coherent and fluid, mimicking the fluency of natural sign language, avoiding stiff and disjointed movements. It is also crucial to assess whether the translation conforms to the everyday expressive habits of the deaf community, including the accurate application of common sign language phrases and idiomatic expressions. Non-verbal elements, such as facial expressions, body postures, and

CCL 2024

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 405-415, Taiyuan, China, July 25 - 28, 2024.
Volume3: Evaluations
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

spatial arrangement, should also be naturally integrated to achieve more natural and expressive communication.

Readability Readability ensures that the expression of sign language is clear and easy to understand, promoting effective communication. The clarity of gestures is crucial; they must be clear enough for observers to understand easily while avoiding any ambiguous movements that could cause confusion. Consistency is also key, ensuring that the same concept or vocabulary is expressed with the same gesture in different contexts, which helps observers better understand and remember the meaning of the sign language. Adaptability is also critical; translations need to adjust gestures and expressions according to different contexts to ensure effective communication.

Cultural Adaptability Cultural adaptability focuses on the suitability and accuracy of the translation across different cultural backgrounds, avoiding cultural misunderstandings. The translation must carefully consider cultural differences to avoid direct translations that might lead to cultural misunderstandings or inappropriate expressions. The translation of sign language avatars must not only be accurate on a literal level but also appropriate culturally, ensuring that the message is correctly understood across different cultural contexts. It also needs to adapt to specific social contexts, such as the appropriate use of polite expressions and industry-specific terminology. The accurate conveyance of emotional tones is also an important aspect of cultural adaptability; translations need to capture and convey the emotional aspects of the original text, such as sarcasm or humor, which is crucial for ensuring comprehensive message delivery and emotional resonance with the receiver.

3.2 Scoring Method

This assessment primarily utilizes manual evaluation to comprehensively assess the performance of digital sign language avatars across various indicators. Each indicator is scored from 1 to 5, where 5 is the best and 1 is the worst. We have designed multiple multiple-choice questions to facilitate the judges in scoring. A specific example in Table 1.

The total score is the arithmetic mean of all scores, excluding one highest and one lowest score. Assuming a set of scores $X = x_1, x_2, ..., x_n$, the specific calculation formula is

$$R = \frac{\sum_{x_i \subset X'} x_i}{n-2} \tag{1}$$

where n is the number of evaluators, and X' is the set of scores excluding the highest score $max(X)$ and the lowest score $min(X)$.

In the overall evaluation framework, the distribution of weights for individual indicators is as shown in Table 2. The composite score is calculated as the weighted sum of the scores for each indicator and their respective weight coefficients. The evaluation focuses on the sign language avatars' ability to accurately express sign language, emphasizing naturalness and readability while also considering cultural adaptability.

Metric Type	Metric Name	Weight Proportion(%)
Human Evaluation of Translation Quality	Sign Language Grammar Accuracy	30%
	Naturalness	25%
	Readability	25%
	Cultural Adaptability	20%

Table 2: Metrics corresponding to the human evaluation of translation quality

4 Dataset

Sign language, as a minority language, typically has a relatively small corpus, which is a common challenge in the field of SLR and SLT. Currently, among publicly available papers, the University of Science and Technology of China(USTC) holds two datasets: the USTC-CCSL dataset (Huang et al., 2018), which contains 25,000 sentences recorded by 50 sign language demonstrators; and the CSL-Daily dataset

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 405-415, Taiyuan, China, July 25 - 28, 2024.
Volume3: Evaluations
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Corpus	Count	Format	Owner	Scenario
XMU-CSL	500	Word-level	Xiamen University	Hospital, Services
BUU-CSL	500	Word-level, Gesture-level	Beijing Union University (College of Special Education)	Shopping, Dining, Accommodation, Tourism, Finance, Hospital, Security, Transportation, Legal, Employment, Public, Government Services, etc.
ZZSZY-CSL	74	Gesture-level	Zhuzhou Voice of Hand Information Technology Co., Ltd.	Legal, Services, Hospital, Services

Table 3: Datasets for the current evaluation task

(Zhou et al., 2021), containing 20,654 sentences recorded by 10 sign language demonstrators. The University of the Chinese Academy of Sciences(UCAS) also offers the RCSD dataset (Wang et al., 2019), though the exact number of videos has not been publicly disclosed, also recorded by 10 sign language demonstrators. Notably, the aforementioned three datasets are proprietary, and usage requires applying under the name of a university or research institute, which poses challenges for the evaluation task's progress.

To address evaluation challenges, Table 3 outlines a corpus of 1,074 sentences with corresponding videos, provided by Xiamen University, Beijing Union University (College of Special Education), and Zhuzhou Voice of Hand Information Technology Co., Ltd. Initially, a text-based annotated corpus was supplied to meet the computational language processing needs of the teams. As training progressed, a composite dataset incorporating both text and video was introduced to support more comprehensive development in sign language avatar technology.

The training dataset used for this evaluation comprises 1,074 sentences from daily life scenarios relevant to the deaf community, such as transportation and medical services. It includes inputs from deaf individuals for whom sign language is the primary mode of communication, ensuring the data's authenticity and representativeness. The dataset incorporates written language, corresponding videos, and sign language glosses, with annotations in both word-level and gesture-level formats to provide expressive richness. These formats comply with the T/CADHOH0004-2023 standard specifications for Intelligent Sign Language Translation System Test. The XMU-CSL dataset features word-level annotations and is focused on hospital services. As shown in Table 3, the BUU-CSL dataset includes both annotation types, covering a broad range of practical scenarios, from shopping to government services. Meanwhile, the ZZSZY-CSL dataset, which uses gesture-level annotations, is tailored for legal and hospital services.

In the evaluation corpus for this event, specific scenarios were selected, and corresponding written language texts along with their sign language glosses are provided as reference examples. Table 4 includes part-of-speech tagging and gesture-level annotations for each reference example[2]. Additionally, sign language video demonstrations of the reference examples are provided[3] [4], with particular attention

[2]Note: Marks [1] denote words that are the same but have different meanings; marks ①② denote words that are the same in both term and meaning, but differ in sign language actions. Other reference materials for sign language include the 'National Common Sign Language Common Words List' and the 'National Common Sign Language Dictionary' APP, among others.

[3]Ex.1 sign language video: https://github.com/ann-yuan/QESLAT-2024/blob/main/video1.gif

[4]Ex.2 sign language video: https://github.com/ann-yuan/QESLAT-2024/blob/main/video2.gif

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 405-415, Taiyuan, China, July 25 - 28, 2024.
Volume3: Evaluations
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

	Written Language Texts	Sign Language Glosses	
Example 1 [3]	女儿可能生病了，快带她去医院。 My daughter may be sick. Take her to the hospital.	**Word-level**	女儿②/病/可能②，带[1]/医院/速度 daughter②/sick/might②, take[1]/hospital/quickly
		Gesture-level	女-矮/病/可能②【疑问】，带[1]/医生-家/速度 female-short/sick/might②【doubtful】,take[1]/doctor-home/quickly
Example 2 [4]	为了赢得比赛，儿子一直在专心训练。 In order to win the competition, my son has been concentrating on his training.	**Word-level**	为/比赛/赢②，儿子②/专心/训练/一直 in order to/competition/win②, son/dedicated/train/always
		Gesture-level	为/比较/胜利②，男-矮/认真【眼睛同时向下看】-心/练习/一直 in order to/compare/win②, son-short/careful[eyes looking down at the sametime]-heart/practice/all the time

Table 4: Reference examples of written language texts and sign language glosses

to the changes in facial expressions.

5 Registration and Evaluation Results

In this evaluation event, 14 teams, including 9 from academia and 5 from industry, registered and applied for the corpus, highlighting widespread interest in sign language avatar technology. By the submission deadline, 10 teams had submitted their interfaces. The event was supported by 18 professional evaluators from the Chinese Association of the Deaf's Committee for Sign Language Research and Promotion, split into an expert group and a collection group. The expert group, proficient in sign language, established benchmarks and facilitated comparisons across different groups. Meanwhile, the collection group, composed of members from 12 different regions, was responsible for identifying regional variations in sign language. The event also included 9 general evaluators—comprising university students, working adults, and retirees. A notable focus was placed on 24 experienced sign language users over the age of 35, 18 of whom are native users. Their extensive use and deep understanding of sign language provided crucial insights into its natural fluency, accuracy, and cultural nuances, thereby ensuring the evaluation's reliability and enhancing the assessment of the avatars' capability to capture nuanced and emotional expressions in sign language.

After preliminary screening by the evaluation team, we found that the system interfaces submitted by 4 of the teams listed in Table 5 not only function effectively but also well reflect the core characteristics of sign language avatars. Table 6 provides a more detailed breakdown of the scores. This outcome indicates that, despite many challenges, the participating teams have made significant progress in the development and application of sign language avatars. However, the performances of the other 6 teams were not satisfactory. Two of these teams' systems could produce sign language sequences, but upon preliminary review, these sequences did not comply with sign language grammar and were unsuitable for scoring by evaluators. Given that these two teams could produce sign language sequences, we awarded them 1 point

CCL 2024

Rank	Competing Team	Weighted Total Score
1	VIVO: Vivo Mobile Communication Co., Ltd.	3.513
2	GBAT: GoBetter AccessTech (SuZhou) Co., Ltd.	2.447
3	BJTT: Beijing Tian Tang Technology Co., Ltd. **(Baseline)**	2.119
4	QDHDT: Qingdao Heshi Digital Technology Co., Ltd.	1.806

Table 5: Effective Team Rankings

Expert Group Scores (9 people)				
Sign Language Avatar	Accuracy	Naturalness	Readability	Cultural Adaptability
VIVO	3.50	3.75	3.43	2.36
GBAT	2.21	2.36	2.00	1.79
BJTT	2.61	2.25	2.21	2.07
QDHDT	1.04	2.14	1.68	1.75
Collection Group Scores (9 people)				
Sign Language Avatar	Accuracy	Naturalness	Readability	Cultural Adaptability
VIVO	3.39	3.43	2.86	2.43
GBAT	1.11	2.07	1.86	1.29
BJTT	2.14	2.07	1.96	1.11
QDHDT	0.25	1.79	1.54	0.82
General Group Scores (9 people)				
Sign Language Avatar	Accuracy	Naturalness	Readability	Cultural Adaptability
VIVO	4.14	4.14	3.75	3.11
GBAT	1.75	2.89	2.43	2.50
BJTT	2.93	2.89	2.89	2.57
QDHDT	1.79	2.86	2.54	2.04

Table 6: Individual scores to each team given by three groups of evaluators

for encouragement in naturalness and ranked them jointly in fifth place. The other four teams, due to repetitive content that did not meet the requirements of our evaluation task, were given zero points and ranked at the bottom.

In a sign language avatar evaluation, Beijing Tian Tang Technology Co., Ltd. (BJTT) was used as the baseline team, using a classical algorithm with two Transformers for text-to-gloss and gloss-to-3D motion translation. While BJTT showed moderate success in the collection group, it rated higher in naturalness in the general group.

Vivo Mobile Communication Co., Ltd. (VIVO) topped the competition with a score of 3.513, outperforming BJTT by 1.394 points. VIVO's success was credited to its pre-trained multilingual model fine-tuning, back-translation for data augmentation, and strategies to smooth animation, leading to high marks in accuracy, naturalness, and readability from general users.GoBetter AccessTech (SuZhou) Co., Ltd. (GBAT) placed second with a score of 2.447, performing slightly above the baseline and noted for its accuracy and naturalness by general users. Qingdao Heshi Digital Technology Co., Ltd. (QDHDT) ranked fourth, facing difficulties in accuracy and readability in the collection group.

In the evaluation, VIVO excelled, while GBAT, BJTT, and QDHDT showed potential for improvement in accuracy and cultural adaptability. Future efforts should aim at enhancing sign language avatar performance and user satisfaction.

To gather specific feedback, optional evaluations were added to the questionnaire for each avatar. VIVO's avatar was praised for its clear, natural movements and coordinated expressions, though it could improve in expression richness and contextual adaptability. GBAT's avatar stood out for clear hand-

CCL 2024

shapes but needs better movement naturalness, expression richness, and sign language fluency. BJTT's avatar was noted for clear hand movements and expressions but required improvements in accuracy and naturalness. QDHDT's avatar was recognized for good visual design and movement coordination but needs better vocabulary accuracy, expression fluency, and non-manual element depiction.

In summary, each team has made certain progress in the development of sign language avatars, but they also face challenges in accuracy, naturalness, and cultural adaptability. Future research and development should focus on these challenges to continuously optimize the technology, allowing sign language avatar to naturally display sign language translations.

6 Overview of Methods Used by Participating Teams

Currently, the participating teams generally adopt the text-gloss-video technology route. This method first uses a translation model to convert spoken text into gloss. Then, it retrieves corresponding videos from a gloss-video database and uses smoothing techniques to produce fluid, continuous sign language videos. Figure 1 shows the basic process of generating Avatar for sign language translation Below, we will discuss the methods from two aspects: Text2Gloss translation and avatar synthesis.

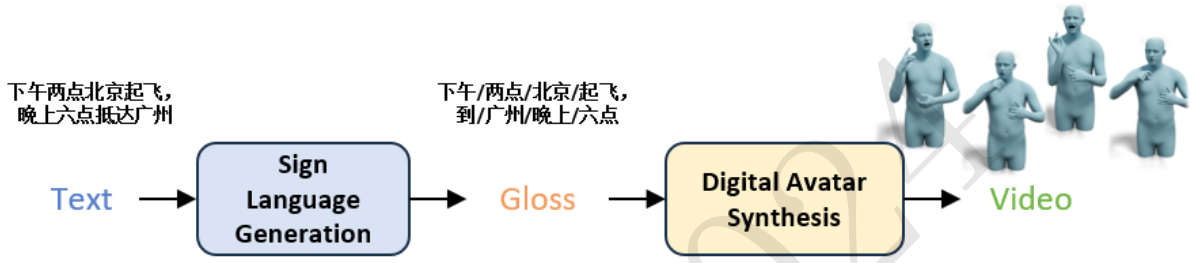

Figure 1: Flowchart of Sign Language Avatar Translation System

6.1 Text2Gloss Translation

Text2Gloss involves translating text into gloss sequences for sign language videos, traditionally using rule-based systems or models trained on text-gloss datasets.However, current methods face several limitations, including dynamic context handling, information loss, data scarcity, and generalization of large models. To address these issues, the participating teams have developed various strategies:

Dynamic Context Handling: Traditional text-to-gloss methods often falter with dynamic contexts, where the meaning of the text depends heavily on its surrounding context. VIVO has fine-tuned a pre-trained multilingual model (mRASP) with one million text-gloss parallel sentences, exploiting its multilingual capabilities for superior context comprehension. This method enhances translation accuracy by considering the context across multiple languages. In a similar vein, BJTT utilized a Transformer model trained on 300,000 parallel sentences, incorporating syntactic tree-based methods to improve contextual understanding, thus preserving the contextual coherence in the gloss translation.

Information Loss: The process of converting text to gloss often leads to information being lost, especially when direct equivalents in sign language are absent. VIVO countered this by implementing a back-translation technique for data enrichment, creating numerous gloss pseudo labels to ensure data variety. This strategy boosts the model's resilience by introducing a broader range of data, minimizing information loss. Conversely, QDHDT adopted an end-to-end approach by integrating text encoding with a pre-trained BERT model and a decoding process via an LSTM sequence model to produce 3D animations directly. This method effectively preserves more information by omitting the intermediate gloss translation.

Data Scarcity: The scarcity of sign language datasets poses a significant challenge for model training and generalization. VIVO addressed this by using a back-translation strategy for data augmentation, significantly expanding the training dataset. BJTT employed data augmentation techniques, including using ChatGPT to generate additional training data. These approaches are effective as they increase the

volume and diversity of training data, which is crucial for training robust models. GBAT prepared high-quality datasets and utilized few-shot learning techniques, leveraging the strong performance of LLMs on small datasets. This method allows the model to learn effectively even with limited data, enhancing its generalization capability.

Generalization of Large Models: While large models perform well on many tasks, their generalization ability in specific domains is still limited. GBAT enhanced generalization by meticulously selecting and preparing parallel sentence pairs, using large language models for few-shot learning to build an agent specifically for text-gloss conversion. This approach leverages the adaptability of large language models to learn specific domain tasks effectively. QDHDT explored end-to-end methods to directly generate 3D sign language animations from text, aiming to overcome generalization issues by avoiding intermediate gloss steps, thus creating a more direct and potentially more accurate translation process.

By addressing these challenges, the participating teams have made significant advancements in Text2Gloss translation and avatar synthesis, contributing to more accurate and fluid sign language translation systems.

6.2 Avatar Synthesis

Avatar synthesis includes action synthesis and smoothing, as well as action rendering. Competing teams mapped gloss to 3D skeleton videos, then used rendering technologies to bind the skeletons to specific avatars to produce the final avatar videos.

VIVO used motion capture technology to acquire common gloss gestures, joint rotation angles, and other skeletal information. They applied a general method proposed by Slot (Lin et al., 2020) for action fusion, achieving seamless transitions between two actions. This method extracts several frames from both the preceding and following skeleton videos, calculates the spatial distances between skeletal joints frame by frame, constructs a cost matrix, and finds the skeleton synthesis plan by calculating the path of minimum total cost.

GBAT employed a more refined approach to constructing individual gloss skeleton videos with 3D keyframe animation. They used keyframe technology to capture critical states and transitions in motion, achieving continuous and natural transitions while solving the challenge of smooth transitions between different sign language videos. In rendering, the team used ThreeJS to implement the sign language avatar, supporting WebGL1 API, suitable for both PC and mobile platforms.

Unlike the first two teams, BJTT used an end-to-end Transformer model to directly generate 3D skeleton videos from text, creating semantically consistent and smoothly acted videos without needing retrieval or smoothing. This approach robustly handles complex sign expressions but is constrained by data and computational resources. Meanwhile, QDHDT optimized animation with state machines and montage techniques, boosting efficiency through graphical state management and flexible code control.

7 Conclusion and Future Prospects

This evaluation involved 14 teams—9 from academia and 5 from industry—highlighting significant interest in sign language avatar technology. By the submission deadline, 10 teams had successfully submitted their interfaces, with 4 advancing past preliminary screening due to their effectiveness and readability.

VIVO stood out by ranking first, excelling across expert, collection, and general groups. Despite all teams following a text-gloss-video route, varied approaches were seen in Text2Gloss translation and avatar synthesis, including fine-tuning pre-trained models, utilizing LLM prompts, ensuring smooth transitions, and implementing end-to-end synthesis.

Evaluators consistently found that while some avatars showcased fluent, clear, and accurate sign language expressions, many still had issues like missing information, inaccurate gestures, and stiff movements. These shortcomings impact the accuracy, readability, and overall viewing experience. Future work should concentrate on refining these technologies to better serve sign language users, aiming for continuous improvement and enhanced user satisfaction.

- **Improving accuracy and readability:** Further optimize Text2Gloss translation and avatar synthesis methods to ensure accurate conveyance of sign language information.

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 405-415, Taiyuan, China, July 25 - 28, 2024.
Volume3: Evaluations
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

- **Enhancing naturalness and cultural adaptability:** Enhance the simulation of sign language expressions, rhythms, and intonations, taking into account the unique sign language habits and cultural characteristics of various regions and groups to better cater to users' communication needs.
- **Exploring more methods:** Explore advanced avatar production methods, including direct end-to-end models, generation strategies like diffusion, and the integration of reinforcement learning or transfer learning to enhance the efficiency and quality of sign language avatar production.
- **Continual improvement and optimization:** Continuously adjust and improve system interfaces to adapt to changing user needs and evolving technologies.

We look forward to sign language avatars serving the deaf community better in the future and providing them with a more convenient and friendly communication experience.

8 Acknowledgements

This research was supported by the National Natural Science Foundation of China [62036001; 62076211]; National Social Science Foundation of China [21BYY106]; General Project of the National Language Committee [YB145-25]; and the Support Plan for Beijing Municipal University Faculty Construction - High-Level Scientific Research and Innovation Team Project [BPHR20220121].

References

Yunfeng Qiu, Dengfeng Yao, Rong Li, and Chunda Liu. 2018. *Introduction to Chinese Sign Language Linguistics*. China International Broadcasting Press.

Dengfeng Yao, Minghu Jiang, Hong Bao, Hanjing Li, and others. 2019. Thirty Years Beyond Sign Language Computing: Retrospect and Prospect. *Chinese Journal of Computers*, volume 42, number 1, pages 111–135.

Tianyu Ren, Dengfeng Yao, Chaoran Yang, and Xinchen Kang. 2024. The Influence of Chinese Characters on Chinese Sign Language. *ACM Transactions on Asian and Low-Resource Language Information Processing*, volume 23, number 1, pages 1–31.

Kayo Yin, Amit Moryossef, Julie Hochgesang, Yoav Goldberg, and Malihe Alikhani. 2021. Including signed languages in natural language processing. *arXiv preprint arXiv:2105.05222*.

Pei Yu, Liang Zhang, Biao Fu, and Yidong Chen. 2023. Efficient sign language translation with a curriculum-based non-autoregressive decoder. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pages 5260–5268.

Rui Zhao, Liang Zhang, Biao Fu, Cong Hu, Jinsong Su, and Yidong Chen. 2024. *Conditional variational autoencoder for sign language translation with cross-modal alignment*. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, number 17, pages 19643–19651.

Yutong Chen, Ronglai Zuo, Fangyun Wei, Yu Wu, Shujie Liu, and Brian Mak. 2022. Two-stream network for sign language recognition and translation. *Advances in Neural Information Processing Systems*, volume 35, pages 17043–17056.

Fangyun Wei and Yutong Chen. 2023. Improving continuous sign language recognition with cross-lingual signs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23612–23621.

Lianyu Hu, Liqing Gao, Zekang Liu, and Wei Feng. 2023. *Continuous sign language recognition with correlation network*. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2529–2539.

Biao Fu, Peigen Ye, Liang Zhang, Pei Yu, Cong Hu, Xiaodong Shi, and Yidong Chen. 2023. A token-level contrastive framework for sign language translation. In *ICASSP 2023 - IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Tong Sun, Biao Fu, Cong Hu, Liang Zhang, Ruiquan Zhang, Xiaodong Shi, Jinsong Su, and Yidong Chen. 2024. Adaptive Simultaneous Sign Language Translation with Confident Translation Length Estimation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 372–384.

Cong Hu, Biao Fu, Pei Yu, Liang Zhang, Xiaodong Shi, and Yidong Chen. 2024. An Explicit Multi-Modal Fusion Method for Sign Language Translation. In *ICASSP 2024 - IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3860–3864.

Razieh Rastgoo, Kourosh Kiani, Sergio Escalera, and Mohammad Sabokrou. 2021. Sign language production: A review. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3451–3461.

Dengfeng Yao. 2022. *A Guide to Sign Language Computing*. Science Press.

Stephanie Stoll, Necati Cihan Camgöz, Simon Hadfield, and Richard Bowden. 2018. *Sign language production using neural machine translation and generative adversarial networks*. In *Proceedings of the 29th British Machine Vision Conference (BMVC 2018)*. British Machine Vision Association.

Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. 2021. *Mixed signals: Sign language production via a mixture of motion primitives*. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1919–1929.

Dele Zhu, Vera Czehmann, and Eleftherios Avramidis. 2023. *Neural Machine Translation Methods for Translating Text to Sign Language Glosses*. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12523–12541.

Vasileios Baltatzis, Rolandos Alexandros Potamias, Evangelos Ververas, Guanxiong Sun, Jiankang Deng, and Stefanos Zafeiriou. 2023. *Neural Sign Actors: A diffusion model for 3D sign language production from text*. *arXiv preprint arXiv:2312.02702*.

Parul Kapoor, Rudrabha Mukhopadhyay, Sindhu B. Hegde, Vinay Namboodiri, and C. V. Jawahar. 2021. *Towards automatic speech to sign language generation*. *arXiv preprint arXiv:2106.12790*.

Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. 2020. *Progressive transformers for end-to-end sign language production*. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pages 687–705. Springer.

Jan Zelinka and Jakub Kanis. 2020. Neural sign language synthesis: Words are our glosses. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3395–3403.

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

Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. 2019. *Expressive body capture: 3d hands, face, and body from a single image*. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10975–10985.

Stephanie Stoll, Armin Mustafa, and Jean-Yves Guillemaut. 2022. *There and back again: 3d sign language generation from text using back-translation*. In *2022 International Conference on 3D Vision (3DV)*, pages 187–196. IEEE.

Razieh Rastgoo, Kourosh Kiani, and Sergio Escalera. 2021. Sign language recognition: A deep survey. *Expert Systems with Applications*, volume 164, pages 113794. Elsevier.

Razieh Rastgoo, Kourosh Kiani, Sergio Escalera, Vassilis Athitsos, and Mohammad Sabokrou. 2022. All You Need In Sign Language Production. *arXiv preprint arXiv:2201.01609*.

Jie Huang, Wengang Zhou, Qilin Zhang, Houqiang Li, and Weiping Li. 2018. Video-based sign language recognition without temporal segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, number 1.

Hao Zhou, Wengang Zhou, Weizhen Qi, Junfu Pu, and Houqiang Li. 2021. Improving Sign Language Translation With Monolingual Data by Sign Back-Translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1316–1325, June.

Hanjie Wang, Xiujuan Chai, and Xilin Chen. 2019. A novel sign language recognition framework using hierarchical grassmann covariance matrix. *IEEE Transactions on Multimedia*, volume 21, number 11, pages 2806–2814.

Zehui Lin, Xiao Pan, Mingxuan Wang, Xipeng Qiu, Jiangtao Feng, Hao Zhou, and Lei Li. 2020. *Pretraining multilingual neural machine translation by leveraging alignment information*. arXiv preprint arXiv:2010.03142.

CCL 2024