Q-BridgeNet: A Quantization Network for Cross-Lingual Sign Language Translation

Liqian Feng[1], Lintao Wang[1], Xiaochen Liu[1], Anusha Withana[1], Ken-Tye Yong[1], Dehui Kong[2], Zhiyong Wang[1], and Kun Hu[3,✉]

[1] The University of Sydney, Darlington NSW, Australia
{lfen0902, lwan3720}@uni.sydney.edu.au
{xiaochen.liu, anusha.withana, ken.yong, zhiyong.wang}@sydney.edu.au
[2] Beijing University of Technology, Beijing, China
kdh@bjut.edu.cn
[3] Edith Cowan University, Joondalup WA, Australia
k.hu@ecu.edu.au

Abstract. Most sign language translation (SLT) methods focus on isolated native sign–spoken pairs (e.g., American Sign Language–English). Extending language-specific SLT models to multilingual translation would improve accessibility by enabling communication across diverse sign and spoken language communities. However, existing multilingual SLT approaches still struggle to learn a unified model that minimizes cross-lingual conflicts while capturing shared cross-lingual semantics and preserving language-specific variations across different sign languages. Therefore, we propose **Q-BridgeNet**, a unified framework for multilingual SLT that jointly mitigates cross-lingual conflicts across both the sign language and spoken language sides. On the sign language side, Q-BridgeNet learns discrete *Q-units* via adaptive segmentation and residual vector quantization: a shared base codebook provides language-agnostic semantic primitives, while language-specific residual codebooks refine heterogeneous signing semantics. On the spoken language side, a multilingual LLM is fine-tuned to operate in the Q-unit space, leveraging cross-lingual priors to enable a unified SLT model. Experiments on PHOENIX14T, How2Sign, and CSL-Daily show that Q-BridgeNet effectively mitigates cross-lingual conflicts, achieving state-of-the-art performance on native sign–spoken pairs while also demonstrating strong generalization to non-native pairs. Our source code is publicly available at: https://github.com/FengLiQ/Q-BridgeNet

Keywords: Sign Language Translation · Representation Learning · Multilingual Modeling

1 Introduction

Over 360 million people worldwide rely on sign languages for daily communication [24]. As visual languages with their own phonology, syntax, and discourse

✉ Corresponding author.

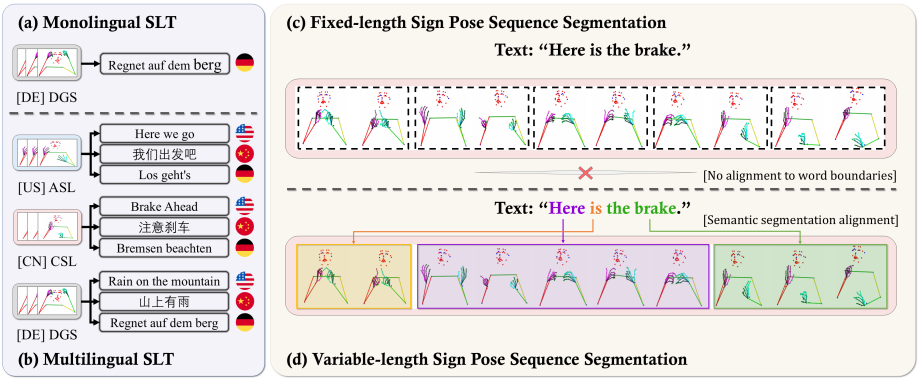

Fig. 1: Multilingual SLT and signing units. (a) Monolingual SLT translates a single sign–spoken language pair. (b) Multilingual SLT extends to many-to-many translation across multiple sign and spoken languages. (c) Fixed-length segmentation may cut across semantic/lexical boundaries, mixing heterogeneous signing patterns and yielding poor alignment to text. (d) Variable-length semantic Q-Unit forms coherent signing units that better align with textual units and helps reduce cross-lingual interference.

structure, sign languages differ fundamentally from spoken languages [23]. Bridging the two is therefore a core challenge in multimodal language understanding, with direct implications for accessibility and inclusion. Most prior work on *Sign Language Translation* (SLT) [1, 2, 8, 15, 20, 43] focuses on a *single* sign-spoken language pair, learning to map continuous sign observations (e.g., video/pose sequences) into spoken-language text. These approaches have achieved substantial progress in monolingual SLT by learning mappings between visual sign representations and textual outputs. While these methods have made strong progress in native (in-pair) settings, practical applications often require interaction across *multiple* sign languages (Figure 1 a,b) –e.g., American Sign Language (ASL), Chinese Sign Language (CSL), German Sign Language (DGS)–and *multiple* spoken languages.

This motivates multilingual SLT, which can broaden accessibility by enabling communication across diverse sign and spoken language communities. Recent efforts have begun to explore multilingual SLT [18, 33, 37]. However, these approaches still struggle to learn a *unified* sign representation that aligns cross-lingual semantics while retaining language-specific realizations. When multiple sign languages are forced into a single shared visual embedding space, differences in visual realization can cause cross-lingual interference (negative transfer): language-specific patterns compete for representational capacity and corrupt the shared manifold, undermining robust multilingual modeling [31].

Therefore, we propose *Q-BridgeNet*, a unified multilingual SLT framework that mitigates cross-lingual conflicts on *both* the sign and spoken language sides via a shared–private residual quantization scheme. Our key idea is to decompose sign representations into (i) *language-agnostic semantic primitives* shared

across sign languages and (ii) *language-specific refinements* that account for linguistic variation. Concretely, we first perform adaptive temporal segmentation (Figure 1 c,d) to partition continuous sign motion into semantically coherent, variable-length units. Each unit is then discretized with residual vector quantization (RQ), where a shared base codebook captures cross-lingual primitives and language-specific residual codebooks model sign-language-specific variations. Finally, we fine-tune a multilingual LLM to operate directly on the resulting discrete *Q-Unit* sequences, yielding a single many-to-many translation model across sign and spoken languages.

We evaluate our *Q-BridgeNet* on three standard datasets–PHOENIX14T (DGS–German) [7], How2Sign (ASL–English) [6], and CSL-Daily (CSL–Chinese) [12]–and construct many-to-many targets over German, English, and Chinese. *Q-BridgeNet* achieves state-of-the-art performance on native pairs and demonstrates strong transfer to non-native pairs, approaching native-pair accuracy in several cross-lingual settings.

Our main contributions are as follows:

- **Q-BridgeNet.** We introduce a unified many-to-many multilingual SLT framework based on a shared–private residual quantization scheme.
- **Signing segmentation for Q-Unit discovery.** We propose a temporal segmentation mechanism that decomposes signing streams into semantically coherent variable-length units for discrete tokenization into *Q-Units*.
- **LLM bridging in Q-unit space.** We fine-tune a multilingual LLM to operate in the discrete Q-Unit space for cross-lingual guidance.

2 Related Work

Sign Language Translation. Early SLT methods typically adopt a two-stage pipeline, where sign videos were first transcribed into intermediate symbolic glosses labels before being translated into spoken text [1,2,15,16]. However, gloss annotations can be costly and may act as an information bottleneck, which motivates recent gloss-free methods that directly translate sign into text. Gloss-free SLT approaches improve visual–linguistic alignment and representation learning through CTC-based modeling [32], dual-branch architectures [5], or multi-stream self-supervised pretraining [10]. More recent methods increasingly leverage LLMs to strengthen cross-modal semantic grounding. MMSLT [17] integrates multimodal LLMs to enhance contextual understanding of sign motion and linguistic structure, while SpaMo [13] combines motion–spatial modeling with LLM-based semantic mapping for translation. Although these advances improve SLT performance, they remain confined to single-language settings, limiting their ability to support communication across multiple sign and spoken language communities.
Quantized Sign Representations. Recent research [14, 30, 38, 39, 45] has explored representing sign and motion sequences using vector-quantized models [36] and motion tokenization approaches. And a data-driven representation [35] uses fixed-length VQ tokenization [29] for signs to map text to poses. [8] adopts LLMs for SLT by discretizing sign pose sequences via VQ-VAE. Yet,

their tokenization uses fixed-length downsampling, limiting its ability to adapt to the variable temporal structure of sign language semantics. Moreover, the learned discrete tokens are usually constructed within a single-language setting and do not explicitly model cross-lingual semantic sharing across different sign languages. In contrast, our approach introduces hierarchical quantization with shared and language-aware residual codebooks built upon adaptive variable-length segmentation, enabling structured and transferable sign representations for multilingual SLT.

Multilingual Sign Language Research. Recent advances in sign languages have explored how models can generalize across multiple sign–spoken language pairs. MGSLT [33] adapts LLMs for gloss-free multilingual SLT, translating sign features into several spoken languages through shared multimodal representations. IMSLT [41] improves cross-lingual transfer by clustering related sign languages and using these clusters as training priors. Signs as Tokens [46] extends sign language production to a multilingual setting by treating signs as discrete motion tokens and augmenting generation with retrieved cross-lingual examples. These works highlight an emerging shift toward unified multilingual sign language research that leverages linguistic similarity and shared visual patterns. Uni-Sign [18] introduces a unified multilingual translation framework by leveraging shared visual encoders and multilingual language models to enable many-to-many sign–spoken translation. However, such approaches mainly rely on continuous visual representations and do not explicitly model structured discrete sign units that separate cross-lingual semantic primitives from language-specific variations. Our approach Q-BridgeNet learns hierarchical discrete Q-units through shared–private quantization, enabling a structured sign representation that facilitates multilingual transfer.

3 Methodology

3.1 Overview

Architecture. We introduce Q-BridgeNet (Figure 2), a unified framework for many-to-many SLT that bridges visual signing and spoken language via a shared, discrete space. Any signing stream is represented as a sequence of quantized units (*Q-units*), which serve as the atomic token for translation. Q-units are learned by alternating between (i) *multi-codebook residual quantization* on variable-length sign pose segments, featuring a shared base codebook for language-agnostic signing primitives and factor-binned residual codebooks for language-specific variation refinement and (ii) *semantically temporal segmentation* of the pose stream based on previously trained residual codebooks. To connect Q-units to spoken languages, we fine-tune a pretrained multilingual LLM to operate in the Q-unit space, yielding a single model that supports both native (a sign language paired with its local spoken language) and cross-lingual (paired with a non-local spoken language) translation.

Notations. In SLT, signing pose sequences and spoken language sentences are involved. We denote a signing pose sequence as $\mathbf{P} = \{\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_T\}$, with T

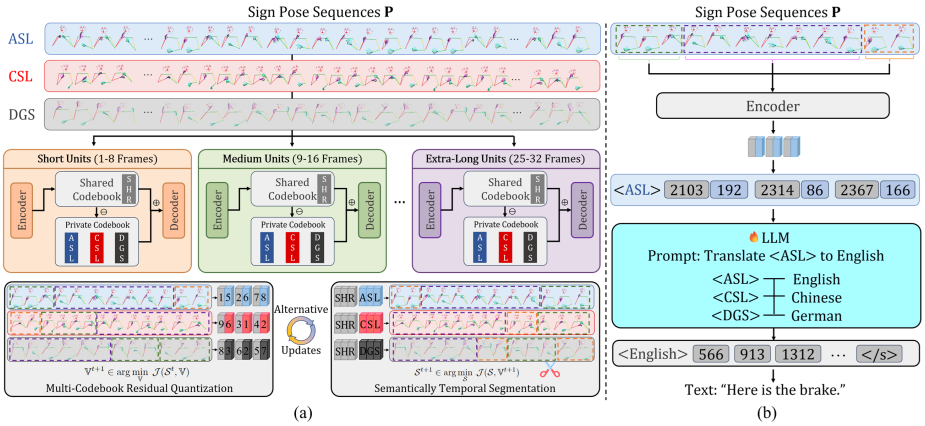

Fig. 2: Q-BridgeNet Overview. (a) *Iterative Q-Unit Discovery:* Segmentation alternates with residual VQ to produce variable-length discrete sign tokens from pose sequences. A *shared base* codebook captures language-agnostic primitives, while *language-specific* residual codebooks (ASL/CSL/DGS) add semantic detail. (b) *Multilingual LLM Re-Alignment:* A pretrained LLM is adapted to operate in the Q-unit space, enabling a unified *many-to-many* SLT model across different sign and spoken languages.

frames, representing a signer's motion trajectory. Each frame $\mathbf{p}_t$ consists of K joints: $\mathbf{p}_t = \{\mathbf{j}_{t,1}, \mathbf{j}_{t,2}, \ldots, \mathbf{j}_{t,K}\}$, where each joint $\mathbf{j}_{t,k} \in \mathbb{R}^d$ stores its coordinates. A signing clip is a contiguous subsequence of frames, denoted by the half-open interval $\mathbf{p}_{s:s'} = \{\mathbf{p}_s, \mathbf{p}_{s+1}, \ldots, \mathbf{p}_{s'-1}\}$ with $s < s'$. We denote the corresponding spoken-language sentence as $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_M)$ where $\mathbf{x}_m$ is the m-th word/token and $\mathbf{X}$ conveys the same meaning as $\mathbf{P}$.

Signing sequence partitioning. To accommodate the irregular temporal dynamics of signing, we partition $\mathbf{P}$ into non-overlapping variable-length segments. Formally, a segmentation is defined as

$$\mathcal{S}(\mathbf{P}) = \{\mathbf{p}_{s_0:s_1}, \mathbf{p}_{s_1:s_2}, \ldots, \mathbf{p}_{s_{N-1}:s_N}\},$$

where $1 = s_0 < s_1 < \cdots < s_N = T + 1$ are the segment boundaries. This segmentation is later coupled with quantization to produce discrete Q-units, via a segment-to-codebook alignment mechanism detailed in subsequent sections.

3.2 Residual Quantization for Multilingual Signing Unit

Given a segmentation $\mathcal{S}(\mathbf{P})$, we discretize each sign into a quantized unit (Q-units) using a shared–private residual VQ-VAE (RQ-VAE).

Our design combines (i) a *shared base codebook* that captures language-agnostic signing primitives and (ii) *language-specific private residual codebooks* that refine language-dependent variations (ASL/CSL/DGS). This yields a compact discrete space that supports many-to-many learning across sign languages and spoken languages.

RQ Encoding for Sign Segments. Formally, we define a set of RQ-VAE modules $\mathbb{V} = \{\mathcal{V}_1, \ldots, \mathcal{V}_L\}$ for L sign languages, where each $\mathcal{V}_i = \{\mathcal{E}, \mathcal{D}, \mathcal{Q}_i^{0:R}, \mathcal{Z}_i^{0:R}\}$ consists of an encoder, decoder, a stack of $R+1$ quantization layers with R residuals, and their associated codebooks. Each module $\mathcal{V}_i$ handles segments with variable lengths for each sign language (ASL/CSL/DGS).

Given a segment $\mathbf{p}_{s_n:s_{n+1}} \in \mathcal{S}(\mathbf{P})$, the encoder produces a latent representation $\mathbf{z} = \mathcal{E}(\mathbf{p}_{s_n:s_{n+1}})$. The quantization process further represents $\mathbf{z}$ with R+1 ordered quantization vectors:

$$\mathrm{RQ}(\mathbf{z}) = [\mathbf{z}^r]_{r=0}^{R}, \tag{1}$$

where $\mathbf{z}^r$ denotes the quantized code at the r-th residual layer. Starting from the initial residual $\mathbf{r}^0 = \mathbf{z}$, the quantization proceeds recursively as

$$\begin{aligned} \mathbf{z}^r &= \mathcal{Q}_i^r(\mathbf{r}^r), \\ \mathbf{r}^{r+1} &= \mathbf{r}^r - \mathbf{z}^r, \end{aligned} \tag{2}$$

for $r = 0, \ldots, R$.

Q-Units. For cross-lingual representation, the first-layer quantizer $\mathcal{Q}_i^0$ and codebook $\mathcal{Z}_i^0$ are shared across all sign languages, reflecting the fact that the same meaning can span different lengths across sign languages. The residual quantizers $\{\mathcal{Q}_i^r\}_{r=1}^{R}$ employ *factor-binned* codebooks $\{\mathcal{Z}_i^r\}$ that are language-specific; language variation is injected at this stage. This provides language-aware refinement of segment representations and supports many-to-many learning within a unified, adaptable latent space. After the residual layer, the final quantized representation is the concatenation of all quantized residuals:

$$\hat{\mathbf{z}} = \mathbf{z}^0 \oplus \ldots \oplus \mathbf{z}^r \oplus \ldots \oplus \mathbf{z}^R, \tag{3}$$

which is then decoded as

$$\hat{\mathbf{p}}_{s_n:s_{n+1}} = \mathcal{D}_i(\hat{\mathbf{z}}). \tag{4}$$

Quantization Optimization. The RQ-VAEs are optimized with a reconstruction and commitment loss:

$$\begin{aligned} \mathcal{L}_{\mathrm{VQ}} = \|\mathbf{p}_{s_n:s_{n+1}} - \mathcal{D}_i(\hat{\mathbf{z}})\|_2^2 \\ + \beta \cdot \|\mathrm{sg}[\mathbf{z}] - \hat{\mathbf{z}}\|_2^2 + \gamma \cdot \|\mathbf{z} - \mathrm{sg}[\hat{\mathbf{z}}]\|_2^2, \end{aligned} \tag{5}$$

where sg[·] denotes the stop-gradient operator. The first term ensures accurate reconstruction, while the latter enforces consistency between the encoding and selected code.

After residual quantization, the full signing sequence is represented as an ordered sequence of discrete latent tokens:

$$\hat{\mathbf{Z}} = (\hat{\mathbf{z}}_1, \hat{\mathbf{z}}_2, \ldots, \hat{\mathbf{z}}_N). \tag{6}$$

The resulting Q-units are compositional symbolic units across temporal scales, serving as atomic tokens for downstream translation and sign-text alignment.

3.3 Joint Signing Segmentation-Quantization

Temporal Segmentation. Based on the learned RQ-VAE modules, we seek to construct the segmentation $\mathcal{S}(\mathbf{P})$. To this end, we define a global segmentation objective $\mathcal{L}_{\text{global}}$ that partitions the sequence into semantically coherent segments by minimizing the cumulative reconstruction error, where each segment is evaluated using an element-wise reconstruction loss $\mathcal{L}_{\text{elem}}$. Formally, for a continuous signing sequence $\mathbf{p}_{s_0:s_N}$, the global segmentation loss is defined as:

$$\mathcal{L}_{\text{global}}(\mathbf{p}_{s_0:s_N} \mid \mathbb{V}) = \min_{\mathcal{S}=\{s_0,\ldots,s_N\}} \sum_{n=0}^{N-1} \mathcal{L}_{\text{elem}}(\mathbf{p}_{s_n:s_{n+1}} \mid \mathbb{V}), \tag{7}$$

where $\mathcal{S}$ denotes a valid segmentation of the input sequence. Each $\mathcal{L}_{\text{elem}}$ measures the reconstruction error of an individual segment reconstructed by the corresponding RQ-VAE module in $\mathbb{V}$:

$$\mathcal{L}_{\text{elem}}(\mathbf{p}_{s_n:s_{n+1}} \mid \mathbb{V}) = \|\mathbf{p}_{s_n:s_{n+1}} - \hat{\mathbf{p}}_{s_n:s_{n+1}}\|_2^2. \tag{8}$$

Since the global segmentation involves a combinatorial search over all valid breakpoints, we adopt an approximation of the optimal segmentation. For a sequence ending at s_N, the optimal objective satisfies the following recurrence:

$$\begin{aligned}\mathcal{L}_{\text{global}}(\mathbf{p}_{s_0:s_N} \mid \mathbb{V}) = \min_{s_{N-1}} \Big[&\mathcal{L}_{\text{global}}(\mathbf{p}_{s_0:s_{N-1}} \mid \mathbb{V}) \\ &+\mathcal{L}_{\text{elem}}(\mathbf{p}_{s_{N-1}:s_N} \mid \mathbb{V})\Big].\end{aligned} \tag{9}$$

This recurrence decomposes the global optimization into a sequence of overlapping sub-problems. The base case is:

$$\mathcal{L}_{\text{global}}(\mathbf{p}_{0:1} \mid \mathbb{V}) = \mathcal{L}_{\text{elem}}(\mathbf{p}_{0:1} \mid \mathbb{V}). \tag{10}$$

By solving all sub-problems sequentially, the optimal segmentation minimizing the cumulative reconstruction loss can be recovered through backtracking.

Alternative Updates. Segmentation and discrete tokenization are inherently coupled. Accurate temporal boundaries reduce intra-segment variation, allowing each segment to correspond to a more coherent semantic unit and thereby simplifying quantization, while a better learned discrete codebook provides more stable semantic prototypes, which in turn facilitates more reliable boundary estimation. Therefore, segmentation and quantization should be optimized jointly rather than independently. However, jointly optimizing the segmentation boundaries and discrete codebooks is difficult because both variables are interdependent. We therefore formulate the following joint objective:

$$\begin{aligned}\mathcal{J}(\mathcal{S}, \mathbb{V}) = \sum_{n=0}^{N-1} \Big[&\|\mathbf{p}_{s_n:s_{n+1}} - \mathcal{D}_i(\hat{z}_n)\|_2^2 \\ &+ \beta\|\text{sg}[z_n] - \hat{z}_n\|_2^2 + \gamma\|z_n - \text{sg}[\hat{z}_n]\|_2^2\Big] + \lambda N,\end{aligned} \tag{11}$$

where λN regularizes the number of segments.

Since directly optimizing $\mathcal{J}(\mathcal{S}, \mathbb{V})$ is intractable, we adopt a block-coordinate descent strategy. The optimization is initialized from a randomly sampled valid segmentation, which serves as the starting point for the iterative refinement process. Each iteration alternates between updating the codebook and refining the segmentation, where the t-th update is as follows:
(A) Codebook Update:

$$\mathbb{V}^{t+1} \in \arg\min_{\mathbb{V}} \ \mathcal{J}(\mathcal{S}^t, \mathbb{V}). \tag{12}$$

(B) Segmentation Update:

$$\mathcal{S}^{t+1} \in \arg\min_{\mathcal{S}} \ \mathcal{J}(\mathcal{S}, \mathbb{V}^{t+1}). \tag{13}$$

Intuitively, each iteration first improves the representation quality under segmentation and then refines the segmentation according to the updated codebook, progressively aligning temporal boundaries with the learned semantic structure. This block-coordinate descent ensures monotonic decrease of $\mathcal{J}$, converging to a stable point and stopping when $(\mathcal{J}^t - \mathcal{J}^{t+1})/\mathcal{J}^t < \varepsilon$ or the change in breakpoints is below a threshold. The resulting segments are both temporally and semantically coherent, forming a discrete representation space well-suited for downstream SLT tasks.

3.4 LLM Re-Alignment with Unified Latent Space

To bridge signing and spoken language for downstream SLT, the trained RQ-VAE are frozen, an LLM is fine-tuned to align and adjust into the Q-unit space, leveraging its broad linguistic knowledge. For finetuning, we construct an extended multilingual, many-to-many sign-spoken dataset based on PHOENIX14T (DGS–German), How2Sign (ASL–English), and CSL-Daily (CSL–Chinese).

For each signing sequence $\mathbf{P}$, the original paired spoken sentence $\mathbf{X}^{(\mathrm{la})}$ corresponds to its native spoken language $\mathrm{la} \in \{\mathrm{de}, \mathrm{en}, \mathrm{zh}\}$. To enable many-to-many supervision, we employ GPT-5 [26] to translate each native spoken caption $\mathbf{X}^{(\mathrm{la})}$ to the remaining two languages and get a trilingual set

$$\mathbb{X} = \{\mathbf{X}^{(\mathrm{de})}, \mathbf{X}^{(\mathrm{en})}, \mathbf{X}^{(\mathrm{zh})}\}. \tag{14}$$

Each signing sequence $\mathbf{P}$ is paired with spoken-language sentences in all three languages, forming $(\mathbf{P}, \mathbb{X})$. This multilingual augmentation facilitates cross-lingual alignment and supports many-to-many learning across diverse sign and spoken modalities. Given a signing pose sequence, we segment and quantize it into a sequence of discrete tokens $\hat{\mathbf{Z}}$ via the frozen RQ-VAEs. The LLM is then finetuned to map between $\hat{\mathbf{Z}}$ and a corresponding spoken language sentence $\mathbf{X} \in \mathbb{X}$.

In SLT, the LLM is fine-tuned to generate $\mathbf{X}$ from $\hat{\mathbf{Z}}$ autoregressively:

$$\mathcal{L}_{\mathrm{SLT}} = -\sum_{k=1}^{M} \log P(\mathbf{x}_k \mid \mathbf{x}_{<k}, \hat{\mathbf{Z}}). \tag{15}$$

Method	How2Sign B-4	How2Sign B-1	How2Sign RG	Method	CSL-Daily B-4	CSL-Daily B-1	CSL-Daily RG	Method	PHOENIX14T B-4	PHOENIX14T B-1	PHOENIX14T RG
SL-Trans.	9.93	28.37	28.94	SL-Trans.	14.15	39.72	38.44	SL-Trans.	19.45	43.72	45.44
Uni-Sign	14.50	40.40	34.30	Uni-Sign	25.61	53.86	54.92	TwoStr.	28.42	54.22	53.19
GloFE	2.24	14.94	12.61	MSLU	11.42	33.97	33.80	CV-SLT	29.27	54.88	54.33
YT-ASL	12.39	37.82	–	MSKA	25.52	56.37	54.04	MSKA	29.03	54.79	53.54
ShuBert	16.20	–	–	MMSLT	21.11	49.87	48.92	MMSLT	25.73	48.92	47.97
				ICTCA	22.47	52.37	51.87	ICTCA	28.42	53.95	53.34
								SCOPE	32.84	61.74	60.06
Ours (Multilingual)											
Avg.	16.44	41.73	35.20	Avg.	26.33	57.09	55.04	Avg.	33.40	62.56	61.39
ASL-EN♣	16.55	41.81	34.97	CSL-EN	26.40	56.98	54.30	DGS-EN	33.20	62.21	60.73
ASL-ZH	**16.83**	**42.04**	**35.52**	CSL-ZH♣	**26.91**	**57.39**	54.85	DGS-ZH	33.36	62.67	**61.67**
ASL-DE	15.96	41.34	35.11	CSL-DE	25.68	56.89	**55.97**	DGS-DE♣	**33.62**	**62.80**	61.76
Ours (Monolingual)♣											
ASL-EN♣	14.35	37.81	32.77	CSL-ZH♣	24.08	52.37	50.27	DGS-DE♣	30.16	56.02	53.75

Table 1: Performance comparison on the SLT task. ♣ indicates the native sign-spoken language pair in the dataset.

This unified next-token prediction design enables translation between multiple signing and natural language, grounded in a shared latent space and facilitated by the LLM's semantic prior.

4 Experiments

4.1 Experimental Settings

Dataset. We evaluate our method on a multilingual corpus aggregated from three widely used sign language datasets: PHOENIX14T [7], How2Sign [6], and CSL-Daily [12]. PHOENIX14T contains approximately 8K German Sign Language (DGS) clips from weather forecast broadcasts. How2Sign consists of around 35K American Sign Language (ASL) samples extracted from instructional videos featuring diverse signers and content. CSL-Daily offers roughly 20K Chinese Sign Language (CSL) samples of daily conversations, supporting evaluation under a low-resource, non-western setting. Following prior work, we adopt the standard evaluation protocol defined for each dataset to ensure fair comparison with existing methods.

Implementation Details. Following existing studies, pose/skeleton pipelines (landmarks of hands–body–face) are a prevalent choice, as they preserve articulatory structure while reducing appearance bias and privacy risk. We represent each sign pose frame $\mathbf{p}_i$ using 79 keypoints (21 per hand, 11 for body, 26 for face), with a maximum sequence length $T = 256$ for efficiency. For preliminary segmentation, we limit each segment to maximum 32 frames. For multi-codebook design, we employ three language-specific RQ-VAEs, with a 96-dimensional latent space, corresponding to the ASL, CSL, DGS datasets. The shared codebook

across these three RQ-VAEs consists of 1,024 embeddings, each RQ-VAE maintains a private codebook of 512 embeddings. We train the RQ-VAEs for 1,000 epochs with hyperparameters $\beta = 1$, $\gamma = 0.25$ empirically. Segmentation and quantization are co-optimized in an alternating manner, which is repeated for 5 times and the initial segmentation is created randomly. For LLM realignment, we fine-tuned Qwen3-1.7B [40] for both SLT using a shared prompting schema with LoRA [11]. Sign tokens are tagged with language labels and paired with their corresponding spoken texts. Prompts like Translate <DGS> to German guide the sign language translation task. We finetune for 100 epochs across the extended multilingual dataset on four NVIDIA RTX 4090 GPUs. To enable many-to-many supervision, we translate each native spoken caption into the remaining two languages using GPT-5 [26], forming a trilingual caption set for every sign instance. Furthermore, As the keypoints in datasets are extracted from raw video frames using human pose estimation methods (e.g., OpenPose [3]) and these pose estimation methods are inherently imperfect, both training and test data contain noisy keypoint measurements. This aligns with practical scenarios in real-world applications, where visual modality data inevitably carries noise due to occlusions, varying lighting conditions, and estimation errors. To improve robustness and simulate realistic variations, we apply standard data augmentation techniques and additionally inject i.i.d. Gaussian noise with standard deviation 0.002 to the normalized keypoint coordinates during training. Supplementary Material provides more details.

Evaluation Metrics. For evaluation, we follow standard protocols established in prior SLT work. We assess the quality of generated spoken language using BLEU [27, 28] and ROUGE [19], two widely adopted metrics for machine translation [33]. For native SLT evaluation, we report results against the original ground-truth of each benchmark. For cross-lingual sign–spoken pairs where no original ground-truth references exist, we evaluate against the corresponding GPT-5 [26] translated captions.

Baselines. For baselines, we compare our method against existing state-of-the-art SLT models, including SL-Transformers [2]$^{\text{CVPR'20}}$, TwoStreamSLT [4]$^{\text{NIPS'22}}$, GloFE [20]$^{\text{ACL'23}}$, YT-ASL [34]$^{\text{NIPS'23}}$, CV-SLT [42]$^{\text{AAAI'24}}$,MSLU [44]$^{\text{TPAMI'24}}$, MSKA [9]$^{\text{PR'25}}$, SCOPE [21]$^{\text{AAAI'25}}$, MMSLT [17]$^{\text{ICCV'25}}$, ICTCA [32]$^{\text{COLING'25}}$, ShuBert [10]$^{\text{ACL'25}}$, and Uni-Sign [18]$^{\text{ICLR'25}}$.

4.2 Comparison with State-of-the-Art

Our method achieves state-of-the-art results across all evaluation metrics for SLT on *native* PHOENIX14T, How2Sign, and CSL-Daily, demonstrating the robustness and versatility of our unified framework.

Q-BridgeNet establishes a consistent lead on *BLEU-4* across all three benchmarks, with clear margins over the strongest prior on each dataset. Crucially, these gains do not come at the expense of surface precision: wherever baselines report *BLEU-1* and *ROUGE*, Q-BridgeNet is also best, indicating that the improvements extend beyond long n-grams to token-level accuracy and summary overlap. The advantage holds against diverse families (seq2seq, multi-stream,

Method	How2Sign			CSL-Daily			PHOENIX14T		
	B-4	B-1	RG	B-4	B-1	RG	B-4	B-1	RG
Ours	**16.44**	**41.73**	**35.20**	**26.33**	**57.09**	**55.04**	**33.40**	**62.56**	**61.39**
Uni-codebook VQ	7.23	21.61	18.09	16.18	33.97	35.12	21.38	42.34	39.22
Non-shared VQ	9.29	25.60	22.52	18.90	37.82	39.01	22.62	42.03	43.14
Non-residual VQ	12.04	31.48	27.75	21.49	41.44	44.32	27.32	50.98	52.45
Single-pass Seg. & RQ	13.39	33.24	31.56	24.11	46.47	47.36	28.61	52.03	49.46
GPT-5 Translation	14.20	36.61	31.82	23.18	51.47	50.02	28.36	54.13	52.26
GPT-4.1 Created Caption	15.58	41.36	34.07	26.30	55.92	54.48	33.07	62.34	60.79
Qwen3-0.6B Fine-Tuning	15.64	41.01	34.20	25.84	55.41	53.54	32.25	60.76	59.01

Table 2: Ablation study results for SLT.

alignment-oriented, and recent large-model pipelines), suggesting robustness to architectural choices. Q-units, formed by adaptive segmentation and residual VQ, provide a more stable discrete interface for the LLM to compose multi-token semantics, especially when signing tempo and segment length distributions differ across datasets.

4.3 Cross-Lingual Non-Native Performance

Most SLT benchmarks evaluate only native sign–spoken pairs, leaving cross-lingual generalization largely unexplored. Q-BridgeNet uniquely enables and excels in such non-native settings, demonstrating true many-to-many communication capability.

Across all three corpora, the multilingual model outperforms its monolingual counterpart trained only on the native pair, indicating that cross-lingual supervision regularizes the discrete signing space and improves potential lexical/phrase coverage. Strikingly, the native target is *not* always optimal: on DGS and ASL, non-native targets (e.g., ASL→zh) edge out the native ones, suggesting that a shared Q-unit inventory with multilingual decoding can recover semantics underrepresented in native captions—or easier to learn in certain target languages. On CSL, the native target remains the best single direction, yet the multilingual average still exceeds monolingual training, pointing to a net benefit from many-to-many constraints even when the native data are strong. Overall, these trends support our design goal: language-agnostic Q-units enable transfer across spoken targets without re-learning sign units per pair.

4.4 Qualitative Evaluation

We present representative examples in Figure 3 for SLT, highlighting both native and non-native translation scenarios. Q-BridgeNet consistently preserves the core semantics of the input sign sequences across English, Chinese, and German, demonstrating the effectiveness of our shared–private Q-unit representation in capturing cross-lingual semantic primitives while maintaining language-specific nuances.

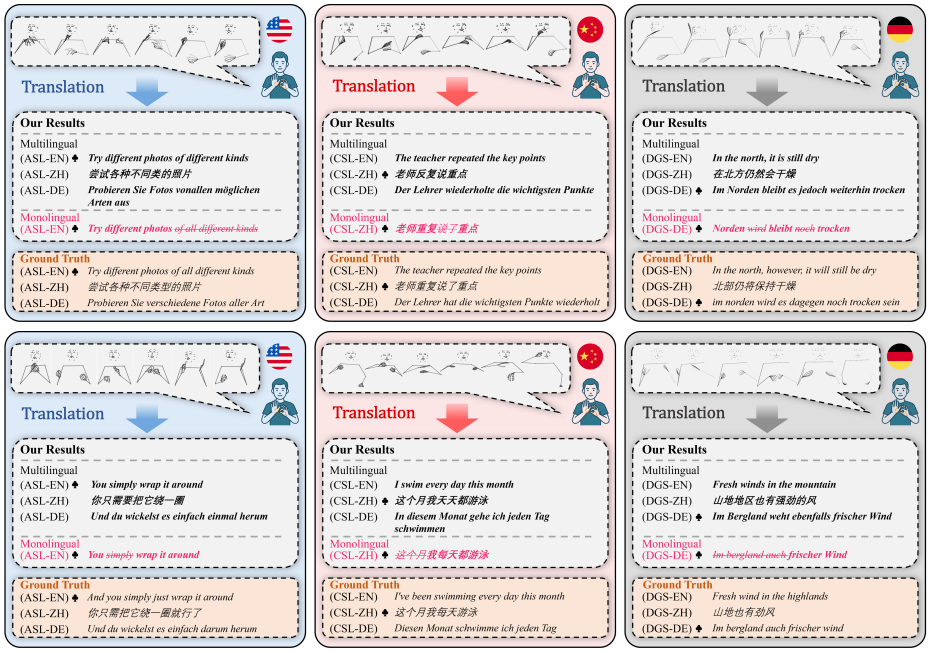

Fig. 3: Multilingual qualitative examples for SLT. **Bold**: semantically aligned with GT; and ~~Strike~~: missing semantics.

In native settings, the model produces fluent and accurate translations that align closely with the ground-truth text, effectively leveraging both the shared base codebook and language-specific residual codebooks to handle linguistic variations. In non-native scenarios—where sign languages are paired with a spoken language not seen during training—Q-BridgeNet still maintains semantic coherence, often producing translations that retain the meaning of the source signs, whereas a monolingual ablation baseline frequently misses key information or produces incomplete translations.

These observations suggest that our variable-length semantic segmentation, combined with hierarchical quantization, not only generates semantically meaningful discrete units but also provides a structured representation that can be effectively interpreted by the multilingual LLM. Overall, this evaluation illustrates the benefits of unified many-to-many modeling and highlights the potential of Q-unit representations to facilitate robust cross-lingual transfer in SLT.

4.5 Ablation Study

We analyze variants to demonstrate the effectiveness of the proposed mechanisms. Results are reported in Table 2.

Uni-Codebook VQ (Vanilla VQ-VAE). One global codebook is used for all segments regardless of the language. This variant is consistently the weakest

across datasets and tasks. BLEU-4 collapses (e.g., PHOENIX14T from 33.40 to 21.38). A single codebook is too coarse to capture multi-lingual complexity; language-specific factors should be modeled explicitly in a many-to-many setting.

Non-Shared VQ (No Shared Base). Remove the shared base codebook; each language bin learns its own first-layer codebook. Performance improves over single–codebook on SLT, while remains behind the full model. The latent space becomes fragmented across bins, limiting cross-bin/unit multilingual reuse. Without a language-agnostic pivot, units learned in different bins fail to align semantically, hindering transfer and compositionality.

Non–Residual VQ (Flat Multi-Codebook, No Residual Stack). Replace residual stacking with parallel (concatenated) codebooks per language bin, akin to multi-codebook VQ-VAE. This is a mid-point: it recovers a large portion of SLT text performance, but still lags the full design on all three datasets. The gap is stable across metrics. Language-specific factorization helps, but without residual "error peeling," the representation lacks the fine-grained refinement provided by stacked residuals.

Single-pass Segmentation & Quantization. Using the full residual quantizer but reduce to a single segmentation–quantization pass lags behind the full model, as it lacks the boundary–code feedback: iterative refinement reconciles mismatches between initial segments and token assignments. In short, single-pass+RQ is a strong first cut; alternating refinement closes the remaining gap.

GPT-5 Translation. To verify whether the multilingual gains come from cross-lingual sign modeling rather than GPT-5 [26] translation, we construct a control baseline that first performs monolingual SLT and then translates the predicted output into non-native target languages using GPT. Although this baseline is evaluated against the same GPT-generated references, it does not perform cross-lingual sign modeling. As shown in Table 2, its consistently lower performance suggests that the multilingual gains stem from learning cross-lingual sign representations rather than GPT-based post-hoc translation.

GPT-4.1 Created Caption. To assess the effect of translation quality in multilingual supervision, we replace GPT-5 [26] with GPT-4.1 [25] while keeping all other settings unchanged. GPT-4.1 yields slightly lower SLT performance than GPT-5, indicating that higher-quality translations provide more reliable multilingual supervision. Nevertheless, the results remain comparable, suggesting that the proposed framework is not overly sensitive to moderate variations in translation quality.

Qwen3-0.6B Fine-Tuning. To evaluate the effect of LLM scale, we replace Qwen3-1.7B [40] with the lightweight Qwen3-0.6B while keeping the remaining configurations fixed. As shown in Table 2, the smaller backbone leads to only a modest degradation across all evaluation metrics. The consistently competitive performance indicates that the proposed Q-unit representation generalizes well across different LLM scales, making the framework suitable for resource-constrained deployment.

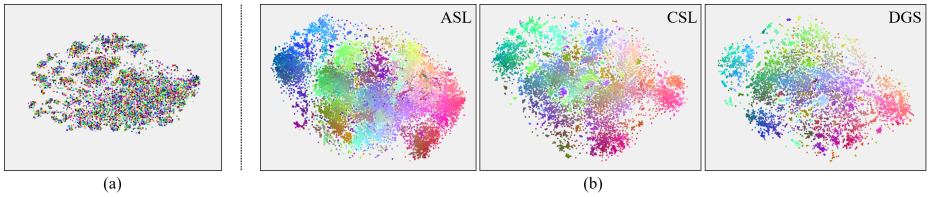

Fig. 4: t-SNE visualization of discrete sign tokens. (a) Fixed-length segmentation with a single VQ-VAE results in scattered and semantically entangled clusters. (b) Our variable-length segmentation with three language-specific codebooks ($\mathcal{V}_{ASL}$, $\mathcal{V}_{CSL}$, and $\mathcal{V}_{DGS}$) yields well-separated clusters, where similar motion units are grouped and color-coded by semantic similarity.

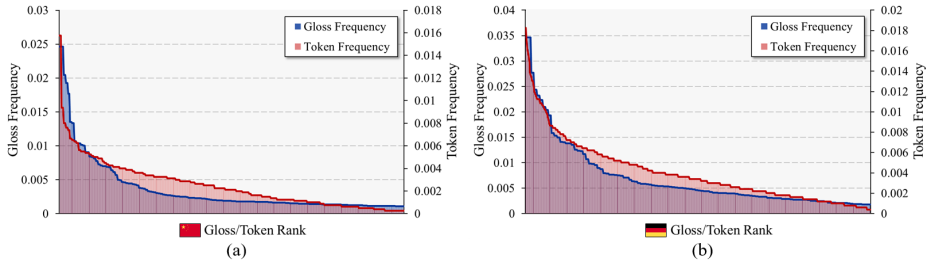

Fig. 5: Comparison between gloss frequency distributions (blue) and private Q-unit token distributions (red) on CSL and DGS. The similar long-tailed trends indicate structural resemblance between learned Q-units and linguistic-level annotation units.

4.6 Analysis of Q-unit Space Distribution

To further examine the structural properties of the learned Q-unit space, we visualize segment-level latent representations using t-SNE [22]. Each motion segment is projected into a 2D space and color-coded based on its assigned discrete token.

We compare our multi-codebook design—based on variable-length segmentation and language-specific codebooks ($\mathcal{V}_{ASL}$, $\mathcal{V}_{CSL}$, and $\mathcal{V}_{DGS}$)—with a baseline that employs fixed-length segmentation and a single codebook, consistent with our ablation setup. As shown in Figure 4, the baseline exhibits a relatively scattered distribution with limited cluster structure, suggesting weaker organization in the learned latent space.

In contrast, our approach produces more compact and visually separable clusters, where segments assigned to the same discrete token tend to group together. This improved spatial coherence indicates that the hierarchical segmentation and shared–private quantization strategy encourages structurally organized and semantically consistent representations.

To investigate whether the learned Q-units exhibit meaningful structural properties, we analyze their frequency distributions and compare them with gloss annotations available in the CSL and DGS datasets. The ASL dataset does not provide gloss annotations and is therefore excluded from this comparison.

Gloss annotations are commonly regarded as linguistic-level symbolic units that reflect semantic decomposition in sign language corpora. As shown in Figure 5, we compare the frequency distributions of gloss tokens with those of private Q-unit tokens learned by our model. Specifically, the x-axis represents gloss words and language-specific Q-unit tokens sorted by descending frequency, while the left y-axis shows the occurrence frequency of gloss words and the right y-axis shows the occurrence frequency of Q-unit tokens. Both glosses and Q-unit tokens exhibit clear long-tailed distributions, and their decay trends are structurally similar across both CSL and DGS datasets.

Although Q-units are learned without gloss supervision, their usage patterns resemble those of manually annotated gloss units in terms of distributional structure. This suggests that the hierarchical quantization mechanism captures reusable structural patterns rather than arbitrary discretization artifacts. Importantly, this similarity emerges purely from data-driven segmentation and residual quantization, indicating that the learned token space exhibits semantic compactness and organized structure.

These observations support our design choice of modeling sign language through hierarchical discrete units and validate the effectiveness of the shared-private decomposition in learning transferable yet expressive representations.

5 Conclusion

We presented **Q-BridgeNet**, a unified multilingual SLT framework that mitigates cross-lingual conflicts across sign and spoken languages by learning discrete *Q-units* via adaptive variable-length segmentation and hierarchical shared–private residual vector quantization, capturing both cross-lingual semantic primitives and language-specific variations. A multilingual LLM is fine-tuned to operate in the Q-unit space, enabling a single many-to-many SLT model. Experiments on PHOENIX14T, How2Sign, and CSL-Daily demonstrate state-of-the-art performance on native pairs and strong generalization to non-native pairs.

While this work focuses on multilingual many-to-many SLT, a promising direction is to extend the framework to cross-lingual sign language production (SLP), where spoken language inputs are mapped back to sign sequences. Since Q-units provide a structured and transferable discrete sign space, a unified Q-unit representation could serve as a bidirectional bridge between multilingual spoken and sign languages, enabling scalable and symmetric sign–text modeling.

Acknowledgements

This research was supported by the Australian Research Council (ARC) under Project LP230100294 and the Edith Cowan University (ECU) Early-Mid Career Researcher (EMCR) Grant Scheme. Dr. Anusha Withana is the recipient of a Future Fellowship (FT250100813) funded by ARC.

References

1. Camgoz, N.C., Hadfield, S., Koller, O., Ney, H., Bowden, R.: Neural sign language translation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
2. Camgoz, N.C., Koller, O., Hadfield, S., Bowden, R.: Sign language transformers: Joint end-to-end sign language recognition and translation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
3. Cao, Z., Hidalgo, G., Simon, T., Wei, S.E., Sheikh, Y.: Openpose: Realtime multi-person 2d pose estimation using part affinity fields. IEEE Transactions on Pattern Analysis and Machine Intelligence (2021)
4. Chen, Y., Zuo, R., Wei, F., Wu, Y., Liu, S., Mak, B.: Two-stream network for sign language recognition and translation. In: International Conference on Neural Information Processing Systems (NeurIPS) (2022)
5. Chen, Z., Zhou, B., Huang, Y., Wan, J., Hu, Y., Shi, H., Liang, Y., Lei, Z., Zhang, D.: C2rl: Content and context representation learning for gloss-free sign language translation and retrieval. IEEE Transactions on Circuits and Systems for Video Technology (2025)
6. Duarte, A., Palaskar, S., Ventura, L., Ghadiyaram, D., DeHaan, K., Metze, F., Torres, J., Giro-i Nieto, X.: How2sign: a large-scale multimodal dataset for continuous american sign language. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2021)
7. Forster, J., Schmidt, C., Koller, O., Bellgardt, M., Ney, H.: Extensions of the sign language recognition and translation corpus rwth-phoenix-weather. In: International Conference on Language Resources and Evaluation (LREC) (2014)
8. Gong, J., Foo, L.G., He, Y., Rahmani, H., Liu, J.: Llms are good sign language translators. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024)
9. Guan, M., Wang, Y., Ma, G., Liu, J., Sun, M.: Mska: Multi-stream keypoint attention network for sign language recognition and translation. Pattern Recognition (2025)
10. Gueuwou, S., Du, X., Shakhnarovich, G., Livescu, K., Liu, A.H.: Shubert: Self-supervised sign language representation learning via multi-stream cluster prediction. In: Annual Meeting of the Association for Computational Linguistics (ACL) (2025)
11. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. In: International Conference on Learning Representations (ICLR) (2022)
12. Huang, J., Zhou, W., Zhang, Q., Li, H., Li, W.: Video-based sign language recognition without temporal segmentation. In: AAAI Conference on Artificial Intelligence (AAAI) (2018)
13. Hwang, E.J., Cho, S., Lee, J., Park, J.C.: An efficient gloss-free sign language translation using spatial configurations and motion dynamics with llms. In: Annual Meeting of the Association for Computational Linguistics (ACL) (2025)
14. Jiang, B., Chen, X., Liu, W., Yu, J., Yu, G., Chen, T.: Motiongpt: Human motion as a foreign language. In: Advances in Neural Information Processing Systems (NeurIPS) (2023)
15. Jiao, P., Min, Y., Chen, X.: Visual alignment pre-training for sign language translation. In: European Conference on Computer Vision (ECCV) (2024)

16. Kan, J., Hu, K., Hagenbuchner, M., Tsoi, A.C., Bennamoun, M., Wang, Z.: Sign language translation with hierarchical spatio-temporal graph neural network. In: IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) (2022)
17. Kim, J., Jeon, H., Bae, J., Kim, H.Y.: Leveraging the power of mllms for gloss-free sign language translation. In: IEEE/CVF International Conference on Computer Vision (ICCV) (2025)
18. Li, Z., Zhou, W., Zhao, W., Wu, K., Hu, H., Li, H.: Uni-sign: Toward unified sign language understanding at scale. In: International Conference on Learning Representations (ICLR) (2025)
19. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: Text Summarization Branches Out (2004)
20. Lin, K., Wang, X., Zhu, L., Sun, K., Zhang, B., Yang, Y.: Gloss-free end-to-end sign language translation. In: Annual Meeting of the Association for Computational Linguistics (ACL) (2023)
21. Liu, Y., Zhang, W., Ren, S., Huang, C., Yu, J., Xu, L.: Scope: Sign language contextual processing with embedding from llms. In: AAAI Conference on Artificial Intelligence (AAAI) (2025)
22. Maaten, L.v.d., Hinton, G.: Visualizing data using t-sne. Journal of Machine Learning Research (2008)
23. Mann, W., Marshall, C.R., Mason, K., Morgan, G.: The acquisition of sign language: The impact of phonetic complexity on phonology. Language Learning and Development (2010)
24. Núñez-Marcos, A., de Viñaspre, O.P., Labaka, G.: A survey on sign language machine translation. Expert Systems with Applications (2023)
25. OpenAI: Introducing GPT-4.1. `https://openai.com/index/gpt-4-1/` (April 2025), accessed: 2025-4-16
26. OpenAI: Introducing GPT-5. `https://openai.com/index/introducing-gpt-5/` (August 2025), accessed: 2025-8-10
27. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Annual Meeting of the Association for Computational Linguistics (ACL) (2002)
28. Post, M.: A call for clarity in reporting BLEU scores. In: Conference on Machine Translation: Research Papers (2018)
29. Shen, F., Jiang, X., He, X., Ye, H., Wang, C., Du, X., Li, Z., Tang, J.: Imagdressing-v1: Customizable virtual dressing. In: AAAI Conference on Artificial Intelligence (AAAI) (2025)
30. Shen, F., Tang, J.: Imagpose: A unified conditional framework for pose-guided person generation. Advances in neural information processing systems (2024)
31. Shi, C., Li, S., Guo, S., Xie, S., Wu, W., Dou, J., Wu, C., Xiao, C., Wang, C., Cheng, Z., Shen, F., Chua, T.S.: Where culture fades: Revealing the cultural gap in text-to-image generation. In: IEEE/CVF International Conference on Computer Vision (CVPR) (2026)
32. Tan, S., Miyazaki, T., Khan, N., Nakadai, K.: Improvement in sign language translation using text ctc alignment. In: International Conference on Computational Linguistics (COLING) (2025)
33. Tan, S., Miyazaki, T., Nakadai, K.: Multilingual gloss-free sign language translation: Towards building a sign language foundation model. In: Annual Meeting of the Association for Computational Linguistics (ACL) (2025)

34. Uthus, D., Tanzer, G., Georg, M.: Youtube-asl: A large-scale, open-domain american sign language-english parallel corpus. In: Advances in Neural Information Processing Systems (NeurIPS) (2023)
35. Walsh, H., Ravanshad, A., Rahmani, M., Bowden, R.: A data-driven representation for sign language production. In: IEEE International Conference on Automatic Face and Gesture Recognition (2024)
36. Wu, Q., Huang, H., Su, K., Wang, Z., Hu, K.: Dc-pcn: Point cloud completion network with dual-codebook guided quantization. In: AAAI Conference on Artificial Intelligence (AAAI) (2025)
37. Yin, A., Zhao, Z., Jin, W., Zhang, M., Zeng, X., He, X.: Mlslt: Towards multilingual sign language translation. In: IEEE/CVF conference on computer vision and pattern recognition (CVPR) (2022)
38. Yu, Z., Huang, S., Cheng, Y., Birdal, T.: Signavatars: A large-scale 3d sign language holistic motion dataset and benchmark. In: European Conference on Computer Vision (ECCV) (2024)
39. Zhang, J., Zhang, Y., Cun, X., Zhang, Y., Zhao, H., Lu, H., Shen, X., Shan, Y.: Generating human motion from textual descriptions with discrete representations. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2023)
40. Zhang, L., Liu, Y., Luo, Y., Gao, F., Gu, J.: Qwen-ig: A qwen-based instruction generation model for llm fine-tuning. In: International Conference on Computing and Pattern Recognition (ICCPR) (2025)
41. Zhang, R., Hu, C., Yu, P., Chen, Y.: Improving multilingual sign language translation with automatically clustered language family information. In: International Conference on Computational Linguistics (COLING) (2025)
42. Zhao, R., Zhang, L., Fu, B., Hu, C., Su, J., Chen, Y.: Conditional variational autoencoder for sign language translation with cross-modal alignment. In: AAAI Conference on Artificial Intelligence (AAAI) (2024)
43. Zhou, B., Chen, Z., Clapés, A., Wan, J., Liang, Y., Escalera, S., Lei, Z., Zhang, D.: Gloss-free sign language translation: Improving from visual-language pretraining. In: IEEE/CVF International Conference on Computer Vision (ICCV) (2023)
44. Zhou, W., Zhao, W., Hu, H., Li, Z., Li, H.: Scaling up multimodal pre-training for sign language understanding. IEEE Transactions on Pattern Analysis and Machine Intelligence (2025)
45. Zhou, Z., Wan, Y., Wang, B.: Avatargpt: All-in-one framework for motion understanding planning generation and beyond. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024)
46. Zuo, R., Potamias, R.A., Ververas, E., Deng, J., Zafeiriou, S.: Signs as tokens: A retrieval-enhanced multilingual sign language generator. In: IEEE/CVF International Conference on Computer Vision (ICCV) (2025)