Unsupervised Sign Language Translation and Generation

Anonymous ACL submission

Abstract

Motivated by the success of unsupervised neural machine translation (UNMT), we introduce an unsupervised sign language translation and generation network (USLNet), which learns from abundant single-modality (text and video) data without parallel sign language data. USLNet comprises two main components: single-modality reconstruction modules (text and video) that rebuild the input from its noisy version in the same modality and cross-modality back-translation modules (text-video-text and video-text-video) that reconstruct the input from its noisy version in the different modality using back-translation procedure. Unlike the single-modality back-translation procedure in text-based UNMT, USLNet faces the cross-modality discrepancy in feature representation, in which the length and the feature dimension mismatch between text and video sequences. We propose a sliding window method to address the issues of aligning variable-length text with video sequences. To our knowledge, USLNet is the first unsupervised sign language translation and generation model capable of generating both natural language text and sign language video in a unified manner. Experimental results on the BBC-Oxford Sign Language dataset (BOBSL) and Open-Domain American Sign Language dataset (OpenASL) reveal that USLNet achieves competitive results compared to supervised baseline models, indicating its effectiveness in sign language translation and generation.

1 Introduction

Sign language translation and generation (SLTG) have emerged as essential tasks in facilitating communication between the deaf and hearing communities (Angelova et al., 2022a). Sign language translation involves the conversion of sign language videos into natural language, while sign language generation involves the generation of sign language videos from natural language.

Sign language translation and generation have achieved great progress in recent years. However, training an SLTG model requires a large parallel video-text corpus, which is known to be ineffective when the training data is insufficient (Müller et al., 2022a). Furthermore, manual and professional sign language annotations are expensive and time-consuming. Inspired by the successes of unsupervised machine translation (UNMT) (Artetxe et al., 2018; Lample et al.) and unsupervised image-to-image translation (Liu et al., 2017), we propose an unsupervised model for SLTG that does not rely on any parallel video-text corpus.

In this work, we propose an unsupervised SLTG network (USLNet), which learns from abundant single-modal (text and video) data without requiring any parallel sign language data. Similar to UNMT, USLNet consists of the following components: the text reconstruction module (§2.1) and the sign video reconstruction module (§2.2) that rebuild the input from its noisy version in the same modality, and cross-modality back-translation modules (§2.3) that reconstruct the input from its noisy version in the different modality using a back-translation procedure.

Unlike the single-modal back-translation in text-based UNMT, USLNet faces the challenge of cross-modal discrepancy. Sign and spoken languages exhibit distinct characteristics in terms of modality, structure, and expression. Sign language relies on visual gestures, facial expressions, and body movements to convey meaning, while spoken language depends on sequences of phonemes, words, and grammar rules (Chen et al., 2022). The cross-modal discrepancy in feature representation presents unique challenges for USLNet.

To address the cross-modal discrepancy in feature representation, a common practice is to use a linear projection to map the representations from the single-modal representation to a shared multi-modal embedding space (Radford et al., 2021).

This approach effectively bridges the gap between different feature representations, facilitating seamless integration of information and enhancing the overall performance of models in handling cross-modal translation tasks. In this work, we propose a sliding window method to address the issues of aligning the text with video sequences.

To the best of our knowledge, USLNet is the first unsupervised SLTG model capable of generating both text and sign language video in a unified manner. Experimental results on the BBC-Oxford Sign Language dataset (BOBSL) (Albanie et al., 2021) and Open-Domain American Sign Language dataset (OpenASL) (Shi et al., 2022) reveal that USLNet achieves competitive results compared to the supervised baseline model (Sincan et al., 2023; Shi et al., 2022) indicating its effectiveness in sign language translation and generation.

Our contributions are summarized below:

1. USLNet is the first unsupervised model for sign language translation and generation, addressing the challenges of scarce high-quality parallel sign language resources.

2. USLNet serves as a comprehensive and versatile model capable of performing both sign language translation and generation tasks efficiently in a unified manner.

3. USLNet demonstrates competitive performance compared to the previous supervised method on the BOBSL and OpenASL dataset.

2 Methodology

The proposed framework in this study consists of four primary components: a text encoder, a text decoder, a video encoder, and a video decoder. As illustrated in Figure 2, the USLNet framework encompasses four modules: **a text reconstruction module** (gray line in Figure 2), **a sign video reconstruction module** (blue line in Figure 2), **a text-video-text back-translation (T2V2T-BT) module** which initially translates input text into pseudo video (red line in Figure 2) and subsequently back-translates pseudo video into text (yellow line in Figure 2), and **a video-text-video back-translation (V2T2V-BT) module** which firstly translates input video into pseudo text (yellow line in Figure 2) and then back-translates pseudo text into video (red line in Figure 2). The latter two modules are considered cross-modality back-translation modules due to their utilization of the back-translation procedure. In this section, we will first describe each module and then introduce the training procedure.

Task Definition We formally define the setting of unsupervised sign language translation and generation. Specifically, we aim to develop a USLNet that can effectively perform both sign language translation and generation tasks, utilizing the available text corpus $\mathcal{T} = \{\mathbf{t}^i\}_{i=1}^M$, and sign language video corpus $\mathcal{V} = \{\mathbf{v}^j\}_{j=1}^N$, where M and N are the sizes of the text and video corpus, respectively.

2.1 Text Reconstruction Module

As shown in Figure 2, the text reconstruction module uses text encoder and text decoder to reconstruct the original text from its corrupted version. Following the implementation by Song et al. (2019), we employ masked sequence-to-sequence learning to implement the text reconstruction. Specifically, given an input text $\mathbf{t} = (\mathbf{t_1}, \ldots, \mathbf{t_n})$ with n words, we randomly mask out a sentence fragment $\mathbf{t^{u:v}}$ where $0 < u < v < n$ in the input text to construct the prediction sequence. The text encoder ENC-TEXT is utilized to encode the masked sequence $\mathbf{t}^{\backslash\mathbf{u:v}}$, and the text decoder DEC-TEXT is employed to predict the missing parts $\mathbf{t^{u:v}}$. The log-likelihood serves as the optimization objective function:

$$\mathcal{L}_{\text{text}} = \frac{1}{|\mathcal{T}|}\sum_{t\in\mathcal{T}} logP(\mathbf{t^{u:v}}|\mathbf{t}^{\backslash\mathbf{u:v}}) \tag{1}$$

This task facilitates the model's learning of the underlying text structure and semantics while enhancing its capacity to manage noisy or incomplete inputs.

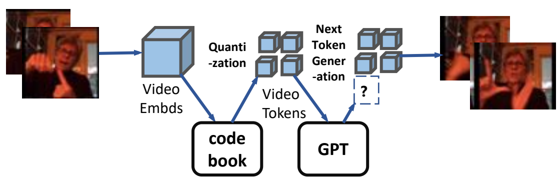

Figure 1: A figure describing sign video reconstruction module. This module is responsible for reconstructing the original video from the downsampled discrete latent representations of raw video data. In the quantization stage, the module transforms the video embeddings into discrete video tokens using a codebook. These video tokens are then input into GPT to generate the next visual token.

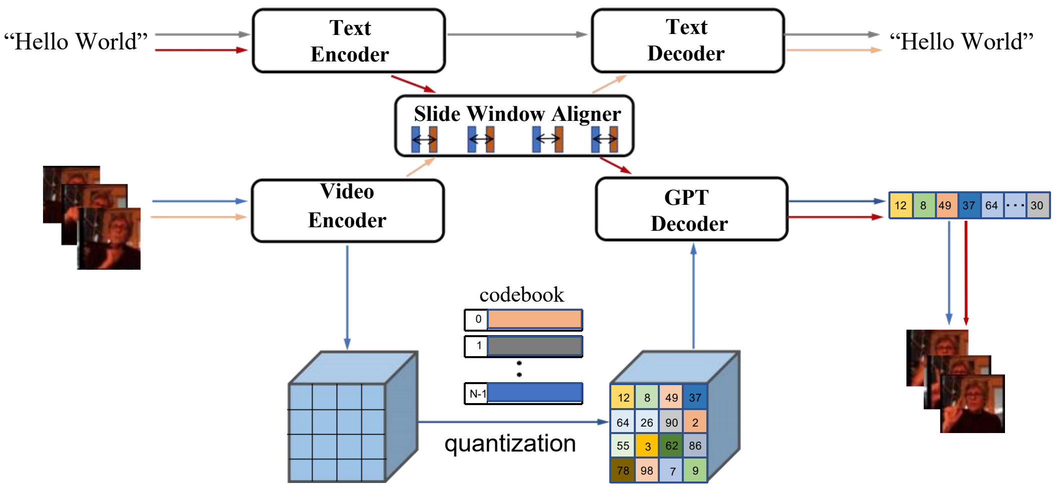

Figure 2: The overall framework of the proposed USLNet. The gray line denotes the text reconstruction procedure. The blue line denotes the video reconstruction procedure . The yellow line denotes the sign language translation procedure which translates video into the corresponding text. The red line denotes the sign language generation procedure which translates text into the corresponding video.

2.2 Sign Video Reconstruction Module

Shown in Figure 1, the sign video reconstruction module reconstructs the original video from the downsampled discrete latent representations of raw video data. In this work, we adopt the VideoGPT (Yan et al., 2021) architecture to build the sign video reconstruction module. VideoGPT consists of two sequential stages, i.e., quantization and video sequence generation.

Quantization VideoGPT employs 3D convolutions and transposed convolutions along with axial attention for the autoencoder in VQ-VAE, learning a downsampled set of discrete latent from raw pixels of the video frames.

Specifically in the quantization stage, given an input video $\mathbf{v} = (\mathbf{v_1}, \ldots, \mathbf{v_n})$ with n pixels, the video encoder encodes the input $\mathbf{v}$ into video embeddings $\mathbf{E_v} = (\mathbf{E_{v_1}}, \ldots, \mathbf{E_{v_n}})$, then $\mathbf{E_v}$ are discretized by performing a nearest neighbors lookup in a codebook of embeddings $\mathbf{C} = \{\mathbf{e_i}\}_{\mathbf{i=1}}^{\mathbf{N}}$, as shown in Eq.(2). Next, $\mathbf{E_v}$ can be represented as discrete encodings $\mathbf{E_v^q}$ which consists of the nearest embedding indexs in codebook, shown in Eq.(3). Finally, video decoder learns to reconstruct the input $\mathbf{v}$ from the quantized encodings.

$$\mathbf{E_{v_i}} = \mathbf{e_k}, \text{ where } \mathbf{k} = \mathrm{argmin}_\mathbf{j} \|\mathbf{E_{v_i}} - \mathbf{e_j}\|_\mathbf{2} \quad (2)$$

$$\begin{aligned} \mathbf{E_v} \rightarrow \mathbf{E_v^q} &= (\mathbf{k_1}, \ldots, \mathbf{k_n}), \\ \text{where} \quad k_i &= \mathrm{argmin}_j \|E_{v_i} - e_j\|_2 \end{aligned} \quad (3)$$

The similarity between $\mathbf{E_{v_i}}$ and $\mathbf{e_j}$ serves as the optimization objective function:

$$\mathcal{L}_{\text{codebook}} = \frac{1}{|\mathcal{C}|} \sum_{e_j \in \mathcal{C}} \|E_{v_i} - e_j\|_2 \quad (4)$$

Video Sequence Generation After quantization stage, the discrete video encodings $\mathbf{E_v^q} = (\mathbf{k_1}, \ldots, \mathbf{k_n})$ are feed into the GPT decoder, and generate the next video "word" $\mathbf{k_{n+1}}$. The similarity between autoregressively generated video $\mathbf{v_{recon}}$ and the original input video $\mathbf{v}$ serves as the optimization object function:

$$\mathcal{L}_{\text{video}} = \frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} \|v_{recon} - v\|_2 \quad (5)$$

2.3 Cross-modality Back-Translation Module

The cross-modality back-translation module consists of two tasks: text-video-text back-translation (T2V2T-BT) and video-text-video back-translation (V2T2V-BT). In contrast to conventional back-translation (Sennrich et al., 2016), which utilizes the same modality, cross-modal back-translation encounters the challenge of addressing discrepancies between different modalities (Ye et al., 2023b). Inspired by the recent work Visual-Language Mapper (Chen et al., 2022), we propose the implementation of a sliding window aligner to facilitate the mapping of cross-modal representations.

Sliding Window Aligner The sliding window aligner is proposed to address the discrepancies between text and video modal representations. Specif-

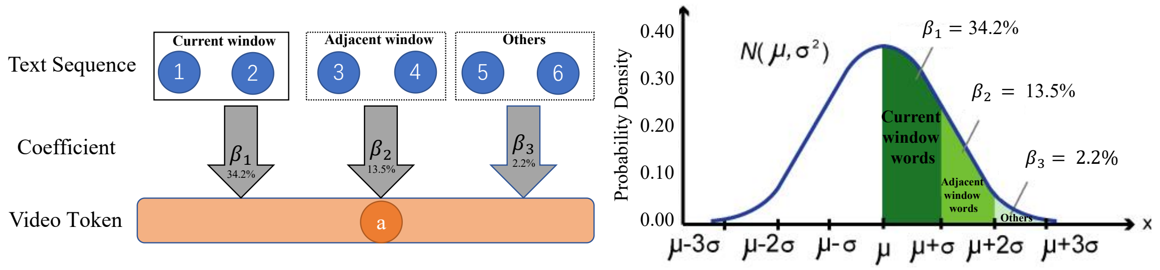

Figure 3: Left: A figure describing slide window aligner at step one. Right: Visualization of the probability distribution (Gaussian distribution) that satisfies the weight coefficients of words in different positions. At step one, we compute the first token "a" of pseudo video "sequence" by slide window aligner.

ically, two primary distinctions between text and video representation sequences are hidden dimensions and sequence length differences. Considering these differences, the aligner consists of two components: *length mapper* $\mathbf{M^L}$ and *dimension mapper* $\mathbf{M^D}$. Considering different back-translation directions (V2T2V and T2V2T), dimension mappers include text-to-video mapper $\mathbf{M^D_{T\to V}}$ and video-to-text mapper $\mathbf{M^D_{V\to T}}$.

Given the text encoder output $\mathbf{E_t}$, the text decoder input $\mathbf{D_t}$, the codebook reconstructed video embedding $\mathbf{E_v}$ and video GPT input $\mathbf{D_v}$, the feature dimension transformation procedure are as follows:

$$\mathbf{D_v} = \mathbf{M^L}(\mathbf{M^D_{T\to V}}(\mathbf{E_t})) \tag{6}$$

$$\mathbf{D_t} = \mathbf{M^L}(\mathbf{M^D_{V\to T}}(\mathbf{E_v})) \tag{7}$$

Aiming to solve the length discrepancy, we design **length mapper** $\mathbf{M^L}$ method, which uses the sliding window method. According to Sutton-Spence and Woll (1999), signing is particularly influenced by English word order when the signers sign while translating from a text. In the context of British Sign Language, presenters may adhere to a more English-like word order. Drawing upon this linguistic understanding, we propose a method wherein the source sequence is partitioned into distinct windows, allowing each word in the target sequence to align more closely with its corresponding source window.

Taking text-to-video for example, supposed that input text sequence $\mathbf{t} = (\mathbf{t_1}, \ldots, \mathbf{t_m})$ with m words, video sequence $\mathbf{v} = (\mathbf{v_1}, \ldots, \mathbf{v_n})$ with n frames and $\mathbf{m} > \mathbf{n}$, the sliding window method, Length Mapper $\mathbf{M^L}$ which can be described as follows:

$$\mathbf{v_i} = \sum_{\mathbf{i=1}}^{\mathbf{n}} \alpha_\mathbf{i} \mathbf{t_i} \tag{8}$$

$$\begin{bmatrix}\alpha_\mathbf{1} & \ldots & \alpha_\mathbf{n}\end{bmatrix} = \mathrm{softmax}\left(\begin{bmatrix}\beta_\mathbf{1} & \ldots & \beta_\mathbf{n}\end{bmatrix}\right) \tag{9}$$

$$\beta_\mathbf{i} \in \begin{Bmatrix} (p(\mu+\sigma), & p(\mu)], & i \in W_c \\ (p(\mu+2\sigma), & p(\mu+\sigma)], & i \in W_a \\ (p(\mu+3\sigma), & p(\mu+2\sigma)], & i \in W_o \end{Bmatrix} \tag{10}$$

Show Eq.(8), every video word accept all text words' information. However, each word in the target sequence aligns more closely with its corresponding window. For example, the beginning video frames conveys more information about the first some text words. Specifically, weight coefficient $[\alpha_\mathbf{1}, \alpha_\mathbf{2}, \ldots, \alpha_\mathbf{n}]$ comes from $X = [\beta_\mathbf{1}, \beta_\mathbf{2}, \ldots, \beta_\mathbf{n}]$. X follows a Gaussian distribution $N(\mu, \sigma^2)$. The value of $\beta_\mathbf{i}$ depends on where token i is and is divided into three probability intervals $(p(\cdot), p(\cdot)]$, shown in Eq.(10). W_c, W_a, W_o represent distinct positional intervals, namely the current window, adjacent window, and other positions. The value of token $\beta_\mathbf{i}$ exhibits an upward trend as its proximity to the current window increases. In the case where token i falls within the bounds of the current window W_c, the weight coefficient is assigned to the highest intervals.

For example, supposed text has 6 words $\mathbf{t} = (\mathbf{t_1}, \ldots, \mathbf{t_6})$ and video has 4 frames $\mathbf{v} = (\mathbf{v_a}, \mathbf{v_b}, \mathbf{v_c}, \mathbf{v_d})$. The window size can be computed as $\lceil 6/4 \rceil = 2$. As Figure 3 has shown, when generating the first video token v_a, it incorporates information from all text tokens while placing the highest weight coefficient β_1 on the first few text words W_c. Meanwhile, the value of token $\beta_\mathbf{i}$ exhibits a declining trend as its proximity to the current window diminishes ($\beta_1 > \beta_2 > \beta_3$).

We introduce **dimension mapper** $\mathbf{M^D}$ to address the differences in hidden dimensions of different modalities. For example, $\mathbf{M^D_{T\to V}}(\mathbf{E_t})$ transposes text embeddings' hidden dimensions into

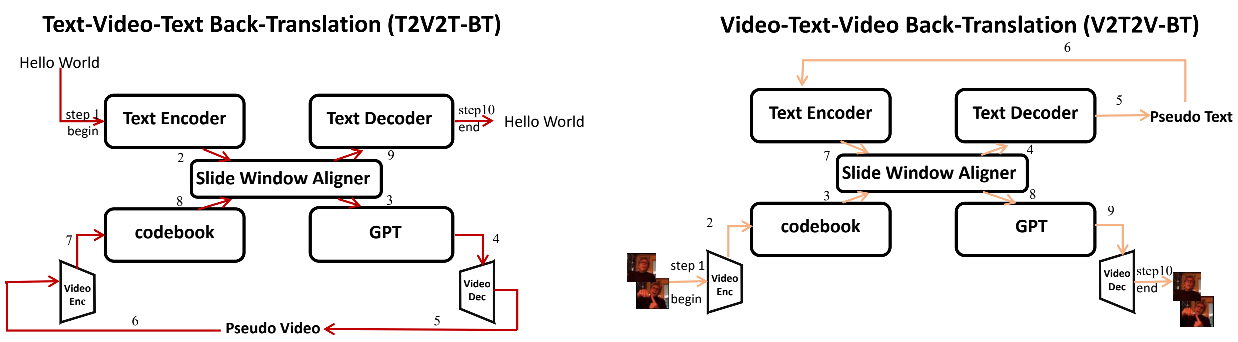

Figure 4: A figure describing the procedure of cross-modality back-translation. The left sub-figure depicts the Text-Video-Text Back-Translation (T2V2T-BT) procedure, while the right sub-figure showcases the Video-Text-Video Back-Translation (V2T2V-BT) procedure. Each sub-figure provides a step-by-step description of the respective back-translation process. The numbers assigned next to the arrows indicate the sequential order of the steps. For instance, "2" signifies that the step is the second step in the procedure.

video embeddings' hidden dimensions, facilitating the integration and alignment of textual and visual information for improved multimodal tasks.

Cross-Modality Back-Translation The T2V2T-BT translates a given text sequence into a sign video, followed by translating the generated sign video back into text, shown in Figure 4. The objective of T2V2T-BT is to ensure consistency between the generated text and the original text while accurately translating the video back into the original text. This task assists the model in capturing the semantic and visual correspondence between text and video modalities and comprehending the input data's underlying structure and temporal dynamics. The similarity between back-translated text $\mathbf{t_{BT}}$ and the original input text $\mathbf{t}$ serves as the optimization object function:

$$\mathcal{L}_{\text{T2V2T}} = \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} \|t_{BT} - t\|_2 \tag{11}$$

Similarly, the V2T2V-BT task requires the model to translate a given video into its corresponding text description, and then translate the generated text back into a video, using the original video as a reference, shown in Figure 4. The similarity between back-translated video $\mathbf{v_{BT}}$ and the original input video $\mathbf{v}$ serves as the optimization object function:

$$\mathcal{L}_{\text{V2T2V}} = \frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} \|v_{BT} - v\|_2 \tag{12}$$

Overall, the cross-modality back-translation module of our proposed USLNet aims to improve the model's ability to translate between text and video modalities in an unsupervised manner, by learning a consistent and meaningful mapping between the two modalities.

2.4 Unsupervised Joint Training

The training objective of USLNet combines the aforementioned loss terms, enabling joint optimization of the text and video networks. The losses L_{text} and L_{video} encourage the generator network to reconstruct the input from its noisy version within the same modality, while the losses L_{T2V2T} and L_{V2T2V} encourage USLNet to reconstruct the input from its noisy version across different modalities. This joint training approach empowers USLNet to not only exhibit strong single-modality generation capabilities in text and video but also acquire cross-modality mapping abilities.

$$\begin{aligned}\mathcal{L}_{\text{overall}} = \alpha_1 \mathcal{L}_{\text{text}} + \alpha_2 \mathcal{L}_{\text{codebook}} + \alpha_3 \mathcal{L}_{\text{video}} + \\ \alpha_4 \mathcal{L}_{\text{T2V2T}} + \alpha_5 \mathcal{L}_{\text{V2T2V}}\end{aligned} \tag{13}$$

3 Experiment

Dataset We conduct a comprehensive evaluation of our approach using two distinct large-scale sign language translation datasets. BBC-Oxford British Sign Language Dataset (BOBSL) (Albanie et al., 2021) is the largest existing video collection of British sign language (BSL). It comprises 1,004K, 20K, and 168K samples in the train, dev, and test sets, respectively. The vocabulary size amounts to 78K, with an out-of-vocabulary (OOV) size of 4.8K in the test set. The second dataset we utilize is OpenASL (Shi et al., 2022), an expansive American Sign Language (ASL) - English dataset collected from various online video platforms. OpenASL boasts an impressive collection of 288 hours of ASL videos across multiple domains, featuring over 200 signers.

Metric The evaluation of USLNet comprises sign language translation (SLT) and sign language generation (SLG). For SLT task, we adopt the BLEU (Papineni et al., 2002) as the evaluation metric for the sign language translation. For SLG, we follow UNMT (Lample et al.) to utilize back-translation BLEU to assess the performance. Specifically, we back-translate the generated sign language video and use the input text as the reference to compute the BLEU score. Additionally,, we adopt Frechet Video Distance (FVD) (Unterthiner et al., 2019) scores to evaluate the quality of generated video.

Model USLNet integrates the MASS architecture (Song et al., 2019) as the foundational backbone for the text model, while the video model backbone is built upon VideoGPT (Yan et al., 2021). For the text model, we set the encoder and decoder layers to 6, and the hidden dimension to 1024. As for the video model, we build the VideoGPT with 8 layers and 6 heads,with a hidden dimension of 576. For the codebook, we set it with 2048 codes, wherein each code represents a feature tensor with a 256-dimensional. The training process comprises two stages: pre-training and unsupervised training. Firstly, we perform continued pre-training using the pre-trained MASS model (Song et al., 2019) on the text portion of the dataset. Then, we train the VideoGPT model (Yan et al., 2021) on the sign language video component of the dataset. Finally, we utilize the pre-trained MASS and VideoGPT models to initialize the USLNet and conduct unsupervised joint training, as described in Section 2.4. We train the whole network with a learning rate of 1e-3. Moreover, we use greedy decoding in evaluation procedure.

4 Results and Discussion

4.1 Main Result

Sign Language Translation In Table 1, we present a comparative analysis between our approach and state-of-the-art methods for SLT on the BOBSL and OpenASL dataset.

For unsupervised-based methods, given the fact that USLNet is the first unsupervised SLT method and BOBSL and openasl has no complete sentence-level gloss annotations datasets (Albanie et al., 2021; Shi et al., 2022; Lin et al., 2023), USLNet w/o, joint training is used to be unsupervised baseline. We observe an approximate improvement of 0.1 BLEU-4 on the BOBSL test set and 1.2 BLEU-4 on the OpenASL dataset. More results and analysis can be seen in Appendix A.1.

To ensure a fair evaluation of USLNet's effectiveness, we also present results for USLNet (S) , which represents USLNet in supervised settings, and USLNet (U+S) , where USLNet undergoes unsupervised training followed by supervised fine-tuning. We compare USLNet's performance in supervised settings against previous state-of-the-art methods. Remarkably, it is observed that USLNet attains new state-of-the-art (SOTA) performance on the BOBSL dataset, while also exhibiting competitive results on the OpenASL dataset. Importantly, USLNet (U+S) outperforms both USLNet and USLNet (S) in both the BOBSL and OpenASL datasets, underscoring the effectiveness of unsupervised training in enhancing the representation of the SLT system.

Sign Language Generation Since there are no existing results for sign language generation on the BOBSL dataset, we compare the use of unsupervised joint training in USLNet. As shown in Table 2, the unsupervised joint training in USLNet yields improvements in terms of back-translation BLEU and FVD scores, demonstrating the effectiveness of USLNet. More visual results can be seen in Appendix A.6.

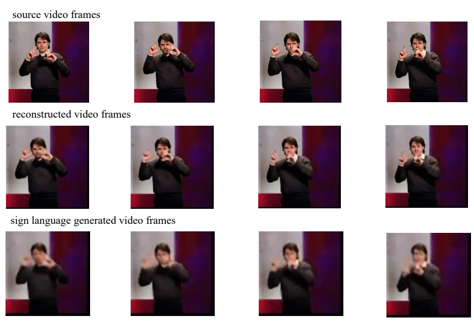

Figure 5: Case study of USLNet on BOBSL for sign language generation task. Examples are from test set.

4.2 Analysis

In this section, we aim to gain a deeper understanding of the improvements achieved by USLNet. To achieve this, we evaluate the effectiveness of the proposed novel sliding window aligner from two perspectives: order consistency and slider comparison.

Method	BOBSL			OpenASL		
	Dev	Test		Dev	Test	
	B@1↑	B@1↑	B@4↑	B@1↑	B@1↑	B@4↑
Supervised Approach						
Transformer (Albanie et al., 2021)	–	12.78	1.00	–	–	–
Context-Transformer (Sincan et al., 2023)	18.80	17.71	1.27	–	–	–
Conv-GRU (Shi et al., 2022)	–	–	–	16.72	16.11	4.58
Transformer (Shi et al., 2022)	–	–	–	20.10	20.92	6.72
USLNet (**S**)	19.60	15.50	1.00	15.40	16.90	4.30
USLNet (**U+S**)	24.60	27.00	1.40	19.30	20.90	6.30
Unsupervised Approach						
USLNet w/o. joint training	1.40	1.50	0.00	1.60	1.30	0.00
USLNet w. joint training	17.30	21.30	0.10	14.50	12.40	1.20

Table 1: Sign language translation performance in terms of BLEU on BOBSL and OpenASL test set. B@1 and B@4 denotes BLEU-1 and BLEU-4, respectively. **S** represents supervised settings; **U+S** represents firstly unsupervised training and then supervised fine-tuning.

Method	BOBSL					OpenASL				
	Dev		Test			Dev		Test		
	B@1↑	FVD↓	B@1↑	B@4↑	FVD↓	B@1↑	FVD↓	B@1↑	B@4↑	FVD↓
USLNet-P	0.50	892.8	0.70	0.00	872.7	1.50	886.4	1.30	0.00	890.2
USLNet	20.90	402.8	22.70	0.20	389.2	19.40	400.2	21.30	7.20	390.5

Table 2: Sign language generation performance in terms of back-translation BLEU and Frechet Video Distance (FVD) on BOBSL and OpenASL dataset. B@1 and B@4 denotes BLEU-1 and BLEU-4, respectively. USLNet-P is the comparison baseline, representing USLNet w/o. joint training. USLNet represenents USLNet w. joint training.

Order Validation Video and glosses are monotonically aligned. We hypothesis that video and text are roughly aligned. To verify this, we must first obtain the golden sign order. Because OpenASL don't have gloss annotation in train set (Shi et al., 2022), we only verify it in BOBSL. Moreover, BOBSL does not have human-evaluated sentence-level glosses annotations, we utilized the automatic gloss annotation released in (Momeni et al., 2022). This gloss annotation consists of word-level annotations, presented as [video name, global time, gloss, source, confidence]. We converted them into sentence-level annotations and assessed the consistency between the gloss (sign) and text orders. From Table 3, we can see the hypothesis that video and text are roughly aligned in BOBSL dataset.

Different Alignment Networks To further explore the advantages of the proposed sliding window aligner (soft connection) , we have designed two comparison aligner networks, altering only the length mapper component $\mathrm{M}^{\mathbf{L}}$. The first network is pooling, where the text sequence is padded to a fixed length and a linear network maps it to the video sequence length. The second network is the sliding window aligner with a hard connection, also utilizing a sliding window mechanism. However, α_i in Eq(8) is non-zero only if tokens are in the current window, indicating that it conveys information exclusively from tokens in the current window. As demonstrated in Table 4, our method achieves the best performance. Moreover, different alignment networks for SLG can be seen in Appendix A.2.

	Proportion
Strictly Consistency	0.83
Consistency with two gloss in disorder	0.87
Consistency with three gloss in disorder	0.91

Table 3: Validation between sign(gloss) and text order consistency for BOBSL.

Comparison between BOBSL and WMT USLNet's performance on the BOBSL dataset is inadequate, similar to the performance observed on

Method	Dev B@1↑	Test B@1↑	Test B@4↑
Pooling	10.70	12.00	0.00
Sliding Window Aligner (hard connection)	15.50	17.10	0.00
Sliding Window Aligner (soft connection)	17.30	21.30	0.10

Table 4: Sign language translation results of USLNet with different cross-modality mappers on BOBSL. B@1 and B@4 denotes BLEU-1 and BLEU-4, respectively.

ID	System	SLT B@1↑
1	Baseline	3.20
1.1	1+more text data	9.60
	Explore Multi-Task Learning	
2.1	1.1+ remove text reconstruction at training	5.40
2.2	1.1+ remove video reconstruction at training	8.30
2.3	1.1+remove cross-modality Back-Translation at training	0.70
	Adjust Data Distribution	
3	1.1+ 1M parallel video and text for unsupervised training	15.20
	Explore Different Freezing Strategy	
4.1	3+ freeze video decoder	10.80
4.2	3+ freeze text encoder	12.20
4.3	3+ freeze text decoder	12.60
4.1	3+ freeze video encoder	17.30

Table 5: Ablation study of USLNet on sign language translation (SLT) on the BOBSL dev set.

the WMT SLT task dataset where the state-of-the-art results showed low performance with a BLEU-4 score of 0.56 (Müller et al., 2022b). Our investigation revealed that the BOBSL dataset presents comparable difficulties to the WMT dataset. Notably, the BOBSL dataset possesses a substantially larger vocabulary of 72,000 words, compared to the WMT dataset's vocabulary of 22,000 words.

4.3 Ablation Study

We conduct our ablation studies on the BOBSL dataset, evaluating the SLT BLEU-1 score on the development set.

Adjust Data Distribution The transformation of un-parallel video and text data into parallel video and text data, employed in an unsupervised manner, has been demonstrated to significantly improve SLT (+5.60 BLEU-1 score).

Explore Different Freezing Strategy Inspired by Zhang et al., we we compare various freezing strategies by evaluating their impact on the performance of SLT.Our experimental results demonstrate that freezing video encoder can improve SLT effects (+2.10 BLEU-1 score).

5 Related Work

Sign Language Translation SLT involves translating sign language videos into text (Camgoz et al., 2018). Previous SLT methods can be categorized into two groups: those focusing on enhancing visual encoder representation (Yin et al., 2021; Zhou et al., 2021b; Yin and Read, 2020; Kan et al., 2022), and those aiming to improve text decoder quality (Camgoz et al., 2020; Chen et al., 2022; Ye et al., 2023b; Angelova et al., 2022b; He et al., 2022a, 2023; Ye et al., 2023a; Zhou et al., 2021a). For large-scale SLT datasets like BOBSL and openASL, Albanie et al. (2021) utilizes a standard transformer model, while Sincan et al. (2023) proposes a context-based approach to enhance quality. Additionally, Shi et al. (2022) incorporates pretraining and local feature modeling for capturing sign language features. To the best of our knowledge, our work represents the first exploration of unsupervised methods in the SLT domain.

Sign Language Generation Sign language generation focuses on generating highly reliable sign language videos (Bragg et al., 2019; Cox et al., 2002). Previous research predominantly relied on high-quality parallel sign language video and text corpora (Glauert et al., 2006a; Cox et al., 2002; Inan et al., 2022). In our work, we aim to explore an unsupervised approach (Lample et al.; Artetxe et al., 2018; He et al., 2022b) that leverages unlabeled data for training the first SLG model.

6 Conclusion

In this paper, we present an unsupervised sign language translation and generation network, USLNet. USLNet is the first bi-directional (translation/generation) sign language approach trained in unsupervised manner. Experimental results on the large-scale sign dataset such as BOBSL and OpenASL reveal that USLNet achieves competitive performance compared to the supervised approach.

7 Limitations

Our USLNet for unsupervised sign language translation and generation has the following limitations:

- **Performance on sign language translation and generation**: As the pioneering unsupervised Sign Language Translation and Generation (SLTG) model, we acknowledge that USLNet's performance is not flawless and further advancements are needed, particularly in the realm of large-scale sign language. We recognize the significance of ongoing breakthroughs required to enhance USLNet's capabilities in this domain.

- **Model Structure**: USLNet has been designed with the objective of exploring a unified model that is capable of both sign language translation and generation. To achieve this, USLNet adopts a twin tower model, comprising separate components for text and video processing. Additionally, to treat videos as sequences, we have incorporated a video quantization model. These factors contribute to the complexity of the USLNet model, which necessitates substantial resources for training.

References

Samuel Albanie, Gül Varol, Liliane Momeni, Hannah Bull, Triantafyllos Afouras, Himel Chowdhury, Neil Fox, Bencie Woll, Rob Cooper, Andrew McParland, et al. 2021. Bbc-oxford british sign language dataset. *arXiv preprint arXiv:2111.03635*.

Galina Angelova, Eleftherios Avramidis, and Sebastian Möller. 2022a. Using neural machine translation methods for sign language translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 273–284.

Galina Angelova, Eleftherios Avramidis, and Sebastian Möller. 2022b. Using neural machine translation methods for sign language translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 273–284.

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018. Unsupervised neural machine translation. In *6th International Conference on Learning Representations, ICLR 2018*.

Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi, et al. 2023. Audiolm: a language modeling approach to audio generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Danielle Bragg, Oscar Koller, Mary Bellard, Larwan Berke, Patrick Boudreault, Annelies Braffort, Naomi Caselli, Matt Huenerfauth, Hernisa Kacorri, Tessa Verhoef, et al. 2019. Sign language recognition, generation, and translation: An interdisciplinary perspective. In *Proceedings of the 21st International ACM SIGACCESS Conference on Computers and Accessibility*, pages 16–31.

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

Necati Cihan Camgoz, Oscar Koller, Simon Hadfield, and Richard Bowden. 2020. Sign language transformers: Joint end-to-end sign language recognition and translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10023–10033.

Yutong Chen, Fangyun Wei, Xiao Sun, Zhirong Wu, and Stephen Lin. 2022. A simple multi-modality transfer learning baseline for sign language translation. In *CVPR*.

Stephen Cox, Michael Lincoln, Judy Tryggvason, Melanie Nakisa, Mark Wells, Marcus Tutt, and Sanja Abbott. 2002. Tessa, a system to aid communication with deaf people. In *Proceedings of the fifth international ACM conference on Assistive technologies*, pages 205–212.

John RW Glauert, Ralph Elliott, Stephen J Cox, Judy Tryggvason, and Mary Sheard. 2006a. Vanessa–a system for communication between deaf and hearing people. *Technology and disability*, 18(4):207–216.

John RW Glauert, Ralph Elliott, Stephen J Cox, Judy Tryggvason, and Mary Sheard. 2006b. Vanessa–a system for communication between deaf and hearing people. *Technology and disability*, 18(4):207–216.

Di He, Yingce Xia, Tao Qin, Liwei Wang, Nenghai Yu, Tie-Yan Liu, and Wei-Ying Ma. 2016. Dual learning for machine translation. *Advances in neural information processing systems*, 29.

Zhiwei He, Tian Liang, Wenxiang Jiao, Zhuosheng Zhang, Yujiu Yang, Rui Wang, Zhaopeng Tu, Shuming Shi, and Xing Wang. 2023. Exploring human-like translation strategy with large language models.

Zhiwei He, Xing Wang, Zhaopeng Tu, Shuming Shi, and Rui Wang. 2022a. Tencent AI lab - shanghai jiao tong university low-resource translation system for the WMT22 translation task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 260–267, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Zhiwei He, Xing Wang, Rui Wang, Shuming Shi, and Zhaopeng Tu. 2022b. Bridging the data gap between training and inference for unsupervised neural machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6611–6623, Dublin, Ireland. Association for Computational Linguistics.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460.

Mert Inan, Yang Zhong, Sabit Hassan, Lorna Quandt, and Malihe Alikhani. 2022. Modeling intensification for sign language generation: A computational approach. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2897–2911.

Jichao Kan, Kun Hu, Markus Hagenbuchner, Ah Chung Tsoi, Mohammed Bennamoun, and Zhiyong Wang. 2022. Sign language translation with hierarchical spatio-temporal graph neural network. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3367–3376.

Kostas Karpouzis, George Caridakis, S-E Fotinea, and Eleni Efthimiou. 2007. Educational resources and implementation of a greek sign language synthesis architecture. *Computers & Education*, 49(1):54–74.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. Unsupervised machine translation using monolingual corpora only. In *International Conference on Learning Representations*.

Kezhou Lin, Xiaohan Wang, Linchao Zhu, Ke Sun, Bang Zhang, and Yi Yang. 2023. Gloss-free end-to-end sign language translation. *arXiv preprint arXiv:2305.12876*.

Ming-Yu Liu, Thomas Breuel, and Jan Kautz. 2017. Unsupervised image-to-image translation networks. *Advances in neural information processing systems*, 30.

Ping Luo, Guangrun Wang, Liang Lin, and Xiaogang Wang. 2017. Deep dual learning for semantic image segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 2718–2726.

John McDonald, Rosalee Wolfe, Jerry Schnepp, Julie Hochgesang, Diana Gorman Jamrozik, Marie Stumbo, Larwan Berke, Melissa Bialek, and Farah Thomas. 2016. An automated technique for real-time production of lifelike animations of american sign language. *Universal Access in the Information Society*, 15:551–566.

Liliane Momeni, Hannah Bull, KR Prajwal, Samuel Albanie, Gül Varol, and Andrew Zisserman. 2022. Automatic dense annotation of large-vocabulary sign language videos. In *European Conference on Computer Vision*, pages 671–690. Springer.

Mathias Müller, Sarah Ebling, Eleftherios Avramidis, Alessia Battisti, Michèle Berger, Richard Bowden, Annelies Braffort, Necati Cihan Camgöz, Cristina España-Bonet, Roman Grundkiewicz, et al. 2022a. Findings of the first wmt shared task on sign language translation (wmt-slt22). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 744–772.

Mathias Müller, Sarah Ebling, Eleftherios Avramidis, Alessia Battisti, Michèle Berger, Richard Bowden, Annelies Braffort, Necati Cihan Camgöz, Cristina España-Bonet, Roman Grundkiewicz, et al. 2022b. Findings of the first wmt shared task on sign language translation (wmt-slt22). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 744–772.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. 2020a. Adversarial training for multi-channel sign language production. *arXiv preprint arXiv:2008.12405*.

Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. 2020b. 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.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *ACL*.

Bowen Shi, Diane Brentari, Gregory Shakhnarovich, and Karen Livescu. 2022. Open-domain sign language translation learned from online video. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6365–6379.

Ozge Mercanoglu Sincan, Necati Cihan Camgoz, and Richard Bowden. 2023. Is context all you need? scaling neural sign language translation to large domains of discourse. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1955–1965.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2019. Mass: Masked sequence to sequence pre-training for language generation. In *International Conference on Machine Learning*, pages 5926–5936. PMLR.

Rachel Sutton-Spence and Bencie Woll. 1999. *The linguistics of British Sign Language: an introduction*. Cambridge University Press.

Sarah L Taylor, Moshe Mahler, Barry-John Theobald, and Iain Matthews. 2012. Dynamic units of visual speech. In *Proceedings of the 11th ACM SIGGRAPH/Eurographics conference on Computer Animation*, pages 275–284.

Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. 2019. Fvd: A new metric for video generation.

Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, et al. 2023. Neural codec language models are zero-shot text to speech synthesizers. *arXiv preprint arXiv:2301.02111*.

Yingce Xia, Jiang Bian, Tao Qin, Nenghai Yu, and Tie-Yan Liu. 2017a. Dual inference for machine learning. In *IJCAI*, pages 3112–3118.

Yingce Xia, Tao Qin, Wei Chen, Jiang Bian, Nenghai Yu, and Tie-Yan Liu. 2017b. Dual supervised learning. In *International conference on machine learning*, pages 3789–3798. PMLR.

Yingce Xia, Xu Tan, Fei Tian, Tao Qin, Nenghai Yu, and Tie-Yan Liu. 2018. Model-level dual learning. In *International Conference on Machine Learning*, pages 5383–5392. PMLR.

Wilson Yan, Yunzhi Zhang, Pieter Abbeel, and Aravind Srinivas. 2021. Videogpt: Video generation using vq-vae and transformers. *arXiv preprint arXiv:2104.10157*.

Jinhui Ye, Wenxiang Jiao, Xing Wang, and Zhaopeng Tu. 2023a. Scaling back-translation with domain text generation for sign language gloss translation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 463–476, Dubrovnik, Croatia. Association for Computational Linguistics.

Jinhui Ye, Wenxiang Jiao, Xing Wang, Zhaopeng Tu, and Hui Xiong. 2023b. Cross-modality data augmentation for end-to-end sign language translation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13558–13571, Singapore. Association for Computational Linguistics.

Zili Yi, Hao Zhang, Ping Tan, and Minglun Gong. 2017. Dualgan: Unsupervised dual learning for image-to-image translation. In *Proceedings of the IEEE international conference on computer vision*, pages 2849–2857.

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

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

Biao Zhang, Mathias Müller, and Rico Sennrich. Sltunet: A simple unified model for sign language translation. In *International Conference on Learning Representations*.

Hang Zhang, Xin Li, and Lidong Bing. 2023. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*.

Sibo Zhang, Jiahong Yuan, Miao Liao, and Liangjun Zhang. 2022. Text2video: Text-driven talking-head video synthesis with personalized phoneme-pose dictionary. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2659–2663. IEEE.

Hao Zhou, Wengang Zhou, Weizhen Qi, Junfu Pu, and Houqiang Li. 2021a. Improving sign language translation with monolingual data by sign back-translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1316–1325.

Hao Zhou, Wengang Zhou, Yun Zhou, and Houqiang Li. 2021b. Spatial-temporal multi-cue network for sign language recognition and translation. *IEEE Transactions on Multimedia*, 24:768–779.

A APPENDIX

A.1 QULITATIVE RESULTS AND FAILURE ANALYSIS

Overall the results in Table 1 are seemingly poor in BOBSL dataset. We dig deep into 'why' the results are poor and to work towards building an understanding for "how" they can be improved significantly.

Regarding the "Why" Aspect We conduct a thorough analysis of the results, identifying the areas in which our approach performs well and those that require further improvement.

Initially, we conduct thorough case study including good cases, bad cases and comparision case between USLNet (unsupervised setting) and Albanie et al. (2021) which is one supervised model. From digging into our results in Table 6 , we find that we can do relatively better in Main ingredients (eg: bus, I, anything), but always fail in other detail, such as proper noun (eg: Ma Effanga), and complex sentence (which is that).

Furthermore, we present a comparative analysis between USLNet in the unsupervised setting and the approach proposed by Albanie et al. (2021). From the Table 6 , we observe that our outcomes are competitive with those of supervised methods. Furthermore, in certain instances, we can achieve more accurate output (for example, particularly in specific cases).

Regarding the "How" Aspect We propose a two-fold approach. Firstly, we suggest allowing unsupervised learning to serve as a representation learning stage. From the Table 1, we can use unspervised training way can provide one good representation and is significant for improve supervised translation method, resulting in a substantial increase in the BLEU-4 score from 1.0 to 1.4. Secondly, we recommend enhancing USLNet by focusing on improvements in both the pre-training and aligner components.

USLNet can be divided into two primary components: the pre-training module (comprising the text pre-training module and the video pre-training module) and the mapper part (slide window aligner). Consequently, the paths to success can be categorized into two aspects. The first aspect involves pre-training, where we can adapt our method using multi-modal models, such as videoLLama (Zhang et al., 2023). The second aspect focuses on designing an effective mapper (Saunders et al., 2020b,a).

A.2 DIFFERENT ALIGNMENT NETWORKS

The effects of different alignment networks for sign language generation are in Table 7. Our method outperforms all other approaches, demonstrating the remarkable effectiveness of USLNet in achieving superior performance.

A.3 ADDITIONAL RELATED WORK

Text-to-Video Aligner Text-to-video aligners in sign language domain can be broadly classified into two main categories. The first category involves the use of animated avatars to generate sign language, relying on a predefined text-sign dictionary that converts text phrases into sign pose sequences (Glauert et al., 2006b; Karpouzis et al., 2007; McDonald et al., 2016). The second category encompasses deep learning approaches applied to text-video mapping. Saunders et al. (2020b,a) adapt the transformer architecture to the text-video domain and employ a linear embedding layer to map the visual embedding into the corresponding space. Unlike these methods, which can only decode pose images, our Unsupervised Sequence Learning Network (USLNet) is capable of generating videos. We address the length and dimension mismatch issues by utilizing a simple sliding window aligner.

In various domains, there have been other proposed text-to-video aligners. For instance, Taylor et al. (2012) presented a method that focuses on automatic redubbing of videos. Their approach leverages the many-to-many mapping between phoneme sequences and lip movements, which is modeled as dynamic visemes. The Text2Video approach Zhang et al. (2022) employs a phoneme-to-pose dictionary to generate key poses and high-quality videos from phoneme-poses. This phoneme-pose dictionary can be considered as a token-token mapper. Similarly, USLNet adopts the practice of quantization and extracting discrete video tokens, a widely recognized technique commonly employed in the audio domain, as demonstrated in studies such as (Hsu et al., 2021; Wang et al., 2023; Borsos et al., 2023). Consequently, the sliding window aligner also serves as a token-token aligner. However, unlike the Text2Video method, which performs a lookup action to obtain target tokens, our approach decodes the target token using all source tokens.

Golden Text:	It's quite a journey **especially** if **I get the bus**.
USLNet:	It's **especially** long if I **get the bus**.
Golden Text:	It's hell of a **difference** yeah.
USLNet:	It's **different completely**.
Golden Text:	Oh, **Ma Effanga** is going to be green.
USLNet:	It's not going to be green.
Golden Text:	They started challenging the sultan in a very important aspect, **which is that he is not Muslim enough**.
USLNet:	This is a very important aspect.
Golden Text:	It's quite a journey **especially** if **I get the bus**.
USLNet:	It's **especially** long if **I get the bus**.
Albanie et al. (2021):	How long have you been in the **bus** now.
Golden Text:	It's hell of a **difference** yeah.
USLNet:	It's **different completely**.
Albanie et al. (2021):	It was like trying to be **different** to the world.

Table 6: A case study of USLNet on the BOBSL dataset is presented, featuring six examples taken from the test set. The first and second examples highlight the successful decoding achieved by USLNet, demonstrating its efficacy in these instances. On the other hand, the third and fourth cases reveal the limitations of USLNet, showcasing areas where improvements are needed. Finally, the last two cases demonstrate the competitive performance of our unsupervised model when compared to the supervised model, further validating the effectiveness of USLNet in sign language translation.

Method	**Dev**	**Test**	
	B@1 ↑	B@1 ↑	B@4↑
Pooling	7.00	6.60	0.00
Sliding Window Aligner (hard connection)	11.70	11.70	0.00
Sliding Window Aligner (soft connection)	20.90	22.70	0.20

Table 7: Sign language generation results in terms of back-translation BLEU of USLNet with different cross-modality mappers on BOBSL. B@1 and B@4 denotes BLEU-1 and BLEU-4, respectively.

Dual Learning He et al. (2016) propose dual learning to reduce the requirement on labeled data aiming to train English-to-French and French-to-English translators. It regards that French-to-English translation is the dual task to English-to-French translation. Thus, it designs to set up a dual-learning game which two agents , each of whom only understands one language and can evaluate how likely the translated are natural sentences in targeted language and to what to extent the reconstructed are consistent with the original. Moreover, researchers exploit the duality between two tasks in training (Xia et al., 2017b) and inference (Xia et al., 2017a) stage , so as to achieve better performance. Dual learning algorithms have been proposed for different tasks, such as translation (He et al., 2016), sentence analysis (Xia et al., 2018), image-image translation (Yi et al., 2017), image segmentation (Luo et al., 2017). USLNet extend dual learning to sign language realm and design dual cross-modality back-translation to learn sign language translation and generation tasks in one unified way.

A.4 ADDITIONAL ANALYSIS

MASS Text Pre-Training Method Outperform than MLM Method In this study, we conduct a comparative analysis of various text pre-training methods to assess their impact on sign language translation task shown in Table 8. Specifically, we focus on comparing the performance of the masked language modeling (MLM) (Kenton and Toutanova, 2019) method and the recently proposed masked sequence-to-sequence (MASS) (Song et al., 2019). Our findings reveal that the MASS method outperforms the MLM method (+1.00 BLEU-1 score) in terms of enhancing the model's ability to capture semantic relationships and improve the overall quality of the learned representations.

ID	System	SLT B@1↑
1	Baseline	3.20
1.1	1+more text data	9.60
	Adjust Data Distribution	
2	1.1+ 1M parallel video and text for unsupervised training	15.20
	Explore Different Text Pretraining Method	
3.1	2+ MLM text pretrain method	15.20
3.2	2+ MASS text pretrain	16.20

Table 8: Additional Ablation study of USLNet on sign language translation (SLT) on the BOBSL dev set. B@1 denotes BLEU-1.

A.5 DISCUSSION ABOUT Albanie et al. (2021).

In terms of model architecture, both Albanie 2021 and USLNet employ a standard transformer encoder-decoder structure. In the Albanie method, the encoder and decoder comprise two attention layers, each with two heads. Conversely, USLNet adopts a large model architecture, setting the encoder and decoder layers to six. Regarding methodology, Albanie 2021 utilizes a supervised approach for learning sign language translation. In contrast, USLNet employs an unsupervised method, leveraging an abundant text corpus to learn text generation capabilities and employing video-text-video back-translation to acquire cross-modality skills. Concerning model output, Albanie 2021 has released several qualitative examples. We have compared these with the results from USLNet, which demonstrate that USLNet achieves competitive outcomes in comparison to the supervised method.

A.6 QUALITATIVE VISUAL RESULTS

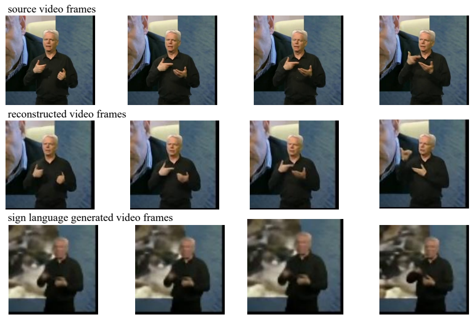

Figure 6: Case study of USLNet on BOBSL for sign language generation task. Examples are from test set.

Figure 7: Case study of USLNet on BOBSL for sign language generation task. Examples are from test set.