Gloss-Free Sign Language Translation: An Unbiased Evaluation of Progress in the Field

Ozge Mercanoglu Sincan, Jian He Low, Sobhan Asasi, Richard Bowden

[a]*Centre for Vision, Speech and Signal Processing (CVSSP), University of Surrey, Guildford, GU2 7XH, Surrey, UK*

Abstract

Sign Language Translation (SLT) aims to automatically convert visual sign language videos into spoken language text and vice versa. While recent years have seen rapid progress, the true sources of performance improvements often remain unclear. Do reported performance gains come from methodological novelty, or from the choice of a different backbone, training optimizations, hyperparameter tuning, or even differences in the calculation of evaluation metrics? This paper presents a comprehensive study of recent gloss-free SLT models by re-implementing key contributions in a unified codebase. We ensure fair comparison by standardizing preprocessing, video encoders, and training setups across all methods. Our analysis shows that many of the performance gains reported in the literature often diminish when models are evaluated under consistent conditions, suggesting that implementation details and evaluation setups play a significant role in determining results. We make the codebase publicly available here [1] to support transparency and reproducibility in SLT research.

Keywords: Sign language translation, gloss-free, assistive technology, codebase

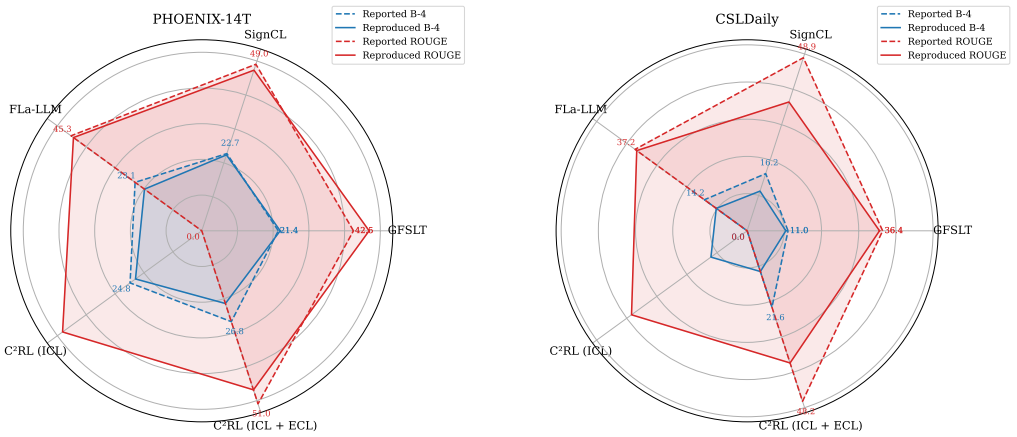


Figure 1: Reported vs reproduced results.

[1] https://github.com/ozgemercanoglu/sltbaselines

1. Introduction

Sign languages are visual languages that are mainly used by the deaf community. Like spoken languages, they vary between countries, and each sign language has its own lexicon and grammatical structures [1]. However, the structural and grammatical differences are even greater between sign and spoken languages. Sign languages involve multiple simultaneous channels of communication, such as hand gestures, facial expressions, mouthings and mouth gestures[2] and body posture. All of which occur in parallel during the signing process. In contrast, spoken and written languages are typically linear, where words and phrases are articulated one after the other. Moreover, the word order of spoken languages usually follows syntactic rules, such as subject-verb-object (SVO), whereas in sign languages, the order of signs often differs, and the use of space (both topological and syntactical) can result in profound variations in how something can be expressed in sign language.

Designing intelligent systems that can bridge the communication gap between deaf and hearing individuals is a critical step toward creating inclusive and accessible technologies. Sign Language Translation (SLT) plays a central role in this context, aiming to convert sign language into spoken language and vice versa. This task generally involves two main directions: (1) sign-to-text, where visual input (e.g., video of a signer) is translated into a spoken language sentence [2], and (2) text-to-sign, where an input spoken sentence is converted into sign language using avatars [3, 4], or photo-realistic video synthesis [5]. In this paper, we focus on the sign-to-text direction.

Sign-to-text translation remains challenging for several reasons: (i) the structural and grammatical differences between sign and spoken languages, (ii) the limited availability of large-scale, annotated datasets, and (iii) the visual complexity and variation in signing styles. Sign languages exhibit high intra- and inter- signer variability in terms of speed, articulation, and expression. Successful models must therefore learn both fine-grained visual representations (handshape trajectories, facial action units, body posture) and high-level semantic abstractions that generalize across signers and linguistic contexts.

Previous works have relied on intermediate gloss[3] annotations, where glosses serve as a linguistic bridge between sign videos and spoken-language sentences [2, 6, 7, 8, 9]. However, data annotation is expensive, and extreme care needs to be taken to ensure consistency in gloss annotation between annotators. It is labor-intensive and takes on average one hour per 90 seconds of video [10]. Therefore, recent research has shifted towards gloss-free approaches, eliminating the need for gloss annotation and learning a direct mapping from sign language videos to text. These gloss-free pipelines typically integrate advanced video encoders with sequence modeling and attention mechanisms to capture the full richness of sign language [11, 12, 13, 14, 15, 16, 17, 18]. One of the most influential models in this direction is GFSLT-VLP [11], which introduces a novel visual-language pretraining strategy to align visual and textual representations in a joint semantic space. Its public implementation has served as a foundation for many follow-up studies [15, 14, 13].

As in many machine learning domains, researchers are often competing to maximize benchmark performance. However, behind each reported number lies a multitude of design choices,

[2] In sign language linguistics, mouthing refers to the silent articulation of spoken language words alongside signs, often reflecting the local spoken language, while mouth gestures are non-verbal movements of the mouth that are part of the sign itself and convey grammatical or affective information.

[3] Gloss is a written representation of a sign using words from spoken language.

such as selection of backbone architectures, loss functions, and regularization techniques. All these decisions significantly influence the performance of SLT models. However, the papers often emphasize novel theoretical contributions over comprehensive reporting of implementation details and hyperparameter settings, making it difficult to disentangle the true impact of core design decisions.

Our work contributes to this direction by analyzing and fairly comparing recent gloss-free sign language translation methods–GFSLT-VLP [11], SignCL [15], Sign2GPT [19], FLa-LLM [14], and C2RL [13]– within a unified, modular codebase, most of which extend or refine the GFSLT-VLP framework. By standardizing data preprocessing, video encoders, and optimization schedules, we ensure all approaches are trained and evaluated under the same conditions. As a result, our study highlights the impact of key design choices on model performance. Figure 1 compares published performance results with those obtained in our re-implementation on the Phoenix-2014T [2] and CSL-daily [6] benchmarks, offering insights into the reproducibility and robustness of current SLT methods. Furthermore, our objective is to make this codebase and data processing framework available to the wider scientific community, so we can collectively progress the field.

The contributions of this paper can be summarized as:

- We provide a modular codebase containing five state-of-the-art gloss-free SLT models, standardizing core components such as data preprocessing pipelines, video encoders, and training schedules.
- We examine each model's key innovations –visual-language pretraining, contrastive learning on adjacent frames, pseudo-gloss pretraining, lightweight translation in pretraining, and hybrid CLIP-translation in pretraining– to measure their isolated impact on translation quality.
- We evaluate the methods on two widely used corpora, Phoenix-2014T and CSL-daily, ensuring identical evaluation protocols.

2. Related Work

Sign language research has progressed from basic sign recognition to continuous recognition and to translation, with each stage introducing more complexity and linguistic nuance. Sign Language Recognition (SLR) focuses on identifying individual signs from video frames, typically in an isolated setting [20, 21, 22, 23, 24, 25]. Building on isolated recognition, Continuous Sign Language Recognition (CSLR) seeks to segment and transcribe unsegmented signing streams into sequences of gloss tokens [26, 6, 27, 28]. Sign Language Translation (SLT), sign-to-text, aims to generate grammatically coherent sentences in a target spoken language given sign language videos.

In this section, we discuss relevant works on gloss-based and gloss-free sign language translation, as well as sign language datasets.

2.1. Gloss-based Sign Language Translation

Camgoz et al. [2] first formulated SLT as a sequence-to-sequence task using a CNN encoder and an RNN decoder, and released the first SLT dataset, Phoenix-2014T, which contained gloss sequences. Due to the sequential alignment between gloss annotations and the temporal

structure of sign language videos, early SLT research heavily relied on gloss-level supervision to simplify the translation task. For instance, SLRT [29] introduced a transformer encoder-decoder framework, integrating CTC loss [30] to soft-match sign representations and gloss sequences. Many subsequent works have leveraged gloss annotations to improve SLT performance, either by incorporating them into end-to-end transformer-based models or by decomposing the task into two sequential modules: a sign-to-gloss recognition stage followed by gloss-to-text translation [6, 9, 7, 31, 32, 8]. Some studies proposed modular pipelines that segment isolated signs using independently trained classifiers before feeding predictions into a translation module [33, 34]. On the other hand, several studies concentrate specifically on the gloss-to-text sub-task, aiming to enhance spoken sentence generation from gloss input by leveraging visual and contextual information, pre-trained language models, or non-autoregressive modeling [35, 36, 37].

In these approaches, glosses act as an intermediate representation, reducing the complexity of direct sign-to-text translation. However, they require large-scale gloss-annotated datasets, which are expensive and time-consuming to create. In addition, the two-stage approach is also prone to error propagation, where inaccuracies in gloss prediction can adversely affect the downstream translation. The reliance on fixed-vocabulary gloss classifiers may also constrain the model's capacity to generalize to unseen or out-of-vocabulary signs.

2.2. Gloss-free Sign Language Translation

More recently, research has shifted towards gloss-free SLT that aims to directly translate sign language videos into spoken language without relying on gloss annotations [38]. Initial efforts trained single-stage sequence-to-sequence models with enhanced visual encoders—e.g., 3D-CNNs or graph-based pose features—and attention/transformer architectures [12, 39, 17].

Recent advances have further improved gloss-free SLT through the use of pretraining strategies. GFSLT-VLP [11] was one of the pioneering studies, introducing a CLIP-style [40] contrastive learning strategy to learn alignments between sign videos and spoken language sentences. Due to its strong performance and publicly available implementation, it has become a widely adopted baseline, serving as the foundation for several models. Subsequent methods have explored various strategies, including contrastive learning between adjacent frames [15], enriching visual representations with additional modalities [16, 41], leveraging pseudo-glosses [19], adopting generative pretraining paradigms [14], combining visual-language alignment with generative objectives in a multi-task setup [13], and converting signs into discrete representations [38]. Including the aforementioned methods, recent approaches also integrate large language models (LLMs) or generative pretrained transformers (GPTs) to improve the quality of the generated translations [42, 43]. Recent efforts have also explored leveraging large-scale external sign language datasets to enhance representation learning [18, 44]. These approaches ultimately share a common objective, which is to improve translation performance.

In this work, we focus on a comparative analysis of five recent methods [11, 15, 19, 14, 13], which adopt different pretraining approaches, most of which extend or refine the GFSLT-VLP. This selection helps comparison to be fair and consistent. By evaluating these models under a consistent setting, we aim to better understand how various training strategies influence gloss-free SLT performance.

2.3. Datasets

Although numerous datasets have been developed for isolated and continuous sign language recognition [20, 21, 22, 45, 46], datasets specifically designed for sign language translation have

Dataset	Year	Source	Domain	Signers	Hours	Text Vocab.	Glosses	Max BLEU4	Popularity†
BSL Corpus [49, 47]	2008-present*	Lab	Conversational	249	125	-	~1800	-	187
Phoenix-2014T [2]	2018	TV	Weather	9	11	3K	1066	26.75 [13]	841
MeineDGS [48]	2008-20*	Lab	Conversational	330	50	18K	~10000	1.08 [33]	16
CSL-Daily [6]	2021	Lab	Daily	10	23	2K	2000	21.61 [13]	256
How2Sign [10]	2021	Lab	Instructional	11	79	16K	N/A	15.5 [44]	262
BOBSL [50]	2021	TV	Diverse	39	1467	77K	N/A	7.3 [43]	82
OpenASL [51]	2022	Web	Open	220	288	33K	N/A	13.21 [13]	62
Youtube-ASL [52]	2023	Web	Open	2519	984	60K	N/A	-	52
YouTube-SL-25 [53]	2024	Web	Open	3000	3207	-	N/A	-	6
CSL-News [18]	2025	TV	Diverse	-	1985	5K	N/A	-	1

Table 1: Summary statistics of sign language datasets. N/A demonstrates that gloss annotations are not available for the dataset, '-' indicates that the value was not reported. *The BSL Corpus, initiated in 2008, and its data annotation is still in progress. MeineDGS was initiated in 2008. †Popularity is the number of citations for the associated publication(s) as of May 13, 2025.

only recently seen notable growth. Although some linguistic projects were initiated in 2008, such as DGS Corpus and BSL Corpus, releases of data and annotations became available in recent years [47, 48]. An overview of SLT datasets is provided in Table 1, including their source, scale, annotation format, popularity, and highest performance in the literature. Additionally, a visual comparison of these datasets is presented in Fig. 2, providing a clear understanding of their characteristics.

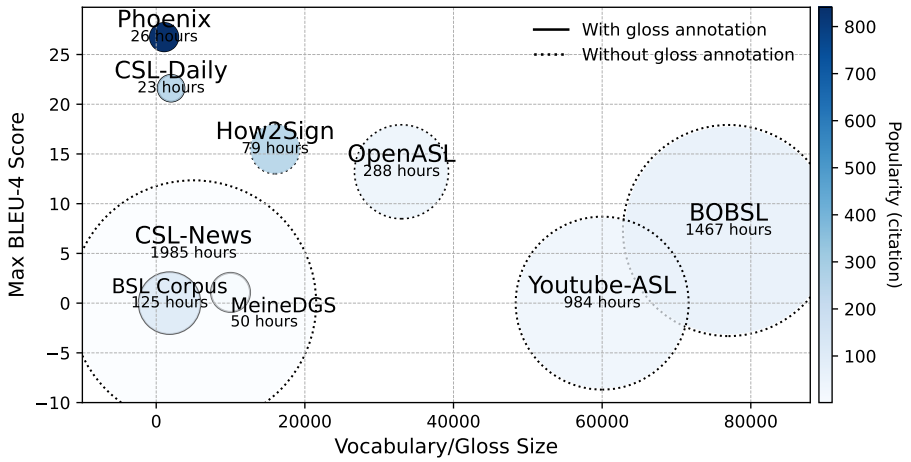


Figure 2: Comparison of sign language translation datasets in terms of gloss/vocabulary size, BLEU-4 score, dataset duration, and citation-based popularity. The x-axis represents the gloss size for datasets with gloss annotations (shown with solid borders), or vocabulary size from spoken language sentences for datasets without gloss annotations (shown with dashed borders). The y-axis shows the maximum BLEU-4 translation performance as reported in the literature. Circle sizes reflect the total number of hours. Color opacity encodes dataset popularity based on citation counts, with darker circles indicating more widely used datasets.

BSL Corpus is a collection of approximately 125 hours of British Sign Language data signed by 249 signers from 8 regions around the UK. It includes narratives, lexical elicitation, interviews, and conversational data. Lexical-level annotations and some of the English translations are available, and more annotations are being made available periodically. **MeineDGS** is also a linguistic dataset on German Sign Language, containing 50 hours of video from 330 signers. Similar to the BSL Corpus, it was designed with a focus on regional and linguistic diversity. The

data format is diverse, ranging from story-retelling to free-flowing conversations. The dataset provides both gloss and sentence-level annotations.

Phoenix-2014T [2] provides gloss annotations and spoken language sentences for German Sign Language (DGS) videos. Since its release, it has become a benchmark and played a central role in shaping SLT research. However, it is limited to a narrow domain of weather forecasts and a limited vocabulary. It has a vocabulary size of 1066 signs, 2887 German words, and 7K sentences in the training set. To address these limitations, subsequent datasets aimed to expand both linguistic and topical diversity. **CSL-Daily** [6] is designed for daily conversations in Chinese sign language, covering a diverse range of topics, such as family life, medical care, bank service, and shopping etc. The dataset includes 23 hours of video recorded in a laboratory environment with 10 deaf signers. It has a vocabulary size of 2000 signs, 2343 Chinese words, and 18K sentences in the training set.

How2Sign [10] expands the scope of SLT resources by providing multimodal data, including RGB, depth, speech, pose estimation, and multiview recordings for American Sign Language (ASL). Unlike earlier resources, How2Sign does not provide gloss annotations but instead contains English transcripts aligned with sign videos. The dataset consists of 80 hours of instructional videos in ASL in 10 categories, such as foods, hobbies, education, etc. The dataset is recorded in a laboratory environment with 11 professional hearing or deaf individuals.

In the meantime, efforts have been made to scale data collection and bring more real-world diversity into SLT [50, 18]. **BOBSL** [50], a large-scale British Sign Language dataset, is curated from interpreted BBC broadcast footage, totalling 1,467 hours with a wide range of topics. However, the dataset provides only audio-aligned English subtitles - manual alignment is performed for only the test set (31 hours). Similarly, the CSL-News [18], Chinese Sign Language (CSL) dataset, contains 1,985 hours of video paired with textual annotations from news content. **OpenASL** [51] and **Youtube-ASL** [52] are collected from online platforms to provide additional large-scale training data, though they include more noise. Youtube-ASL does not provide train, validation, and test splits; so, it serves as external data to increase sign language translation performance on other datasets like How2Sign.

Recently, **YouTube-SL-25** [53] has focused on building multilingual sign language datasets to support cross-lingual SLT models.

In this work, Phoenix-2014T and CSL-Daily datasets were chosen for their widespread use, and also due to their manageable size, which makes them suitable for reproducible and controlled experiments.

3. Overview of Methods Evaluated

Given an input video $V = (x_1, x_2, ..., x_T)$ with T frames, a sign language translation aims to translate the sign video into a spoken language sequence $S = (w_1, w_2, ..., w_U)$ with U words.

In recent years, one of the most significant advancements has been the introduction of visual-language pretraining techniques [11]. The availability of public implementation has further accelerated progress by enabling reproducibility and extensions, especially on small-scale datasets such as Phoenix-2014T [2].

Given the influence of GFSLT-VLP [11], we examine state-of-the-art methods that extend or refine this approach, including SignCL [15], FLa-LLM [14], C^2RL [13]. These methods build upon a shared architectural pipeline while introducing distinct training objectives or strategies.

Additionally, we include Sign2GPT [19], a recent approach that incorporates pseudo-gloss pretraining strategy. While it follows a distinct network design, incorporating pseudo-gloss training allows us to explore the impact of intermediate representations.

Figure 3 illustrates the common architectural pipeline across the models and highlights differences in their training objectives. Although the common structure is similar, methods differ in architectural choices and training setups. These differences are detailed in Table 2. Here, we summarize each method by highlighting their architectural designs, and training objectives.

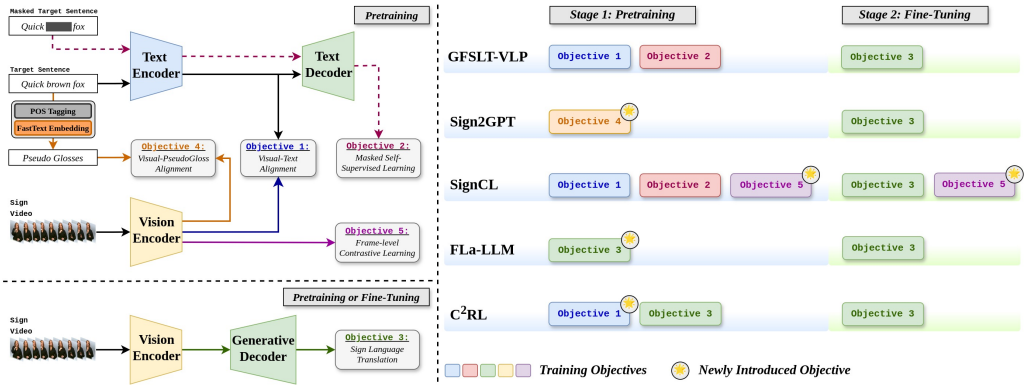


Figure 3: Architectural overview and training objectives of the compared sign language translation methods. The right side summarizes which objectives are used by each method.

3.1. GFSLT-VLP

GFSLT-VLP [11] introduces a two-stage SLT method with a novel visual-language pretraining strategy. In the first (pretraining) stage, sign language videos and their corresponding spoken language sentences are projected into a joint semantic space using contrastive learning. For the visual encoder, ResNet18 [55] (2D-CNN) is followed by a temporal block (Conv1D-BN-Relu-MaxPooling) and 3-layers of the mBART [56] encoder. Textual representations are obtained with 12-layers of the mBART [56] encoder. These visual and textual embeddings are projected via linear layers into a shared embedding space, where contrastive learning maximizes similarity between paired inputs and minimizes it for mismatched pairs. Simultaneously, a masked self-supervised learning strategy is utilized to pretrain a Text Decoder (3-layers of the mBART decoder), aiming to capture the syntactic and semantic knowledge of spoken language sentences. In the second stage, an encoder-decoder model is fine-tuned for the SLT task, initialized with the pretrained visual encoder and text decoder.

3.2. SignCL

SignCL [15] identifies a representation density problem as visually similar but semantically distinct signs tend to be mapped to nearby locations in the feature space. To address this, they incorporate a contrastive learning loss between adjacent frames, encouraging the separation of different signs and improving representation quality. They assume that if two frames are close enough, they are considered to belong to the same sign, otherwise different signs.

	GFSLT-VLP	**Sign2GPT**	**SignCL**	**FLa-LLM**	**C^{2}RL**
Key idea	Visual-language pretraining	Pseudo-gloss pretraining	Contrastive learning on frames	Light-T in pretraining	Combine CiCO and Light-T
Code availability	Yes	Yes	Partially	No	No
Pretraining stage					
number of epochs	80	100	80 *	100	200
visual backbone	ResNet18†	DinoV2	Resnet18†	ResNet18†	ResNet18†
-TempBlock	Conv1D-BN-ReLU-MaxPool	–	Conv1D-BN-ReLU-MaxPool	Conv1D-BN-ReLU	Conv1D-BN-ReLU
-Temporal Encoder	1024 hid, 4096 FF	–	1024 hid, 4096 FF*	512 hid, 2048 FF	512 hid, 2048 FF
batch size	8	8	128	16	64
optimizer	SGD	adam	SGD*	SGD	SGD
scheduler	cosine	one-cycle cosine	cosine*	cosine	cosine
learning rate	10^{-2}	3×10^{-4}	10^{-2} *	10^{-2}	10^{-2}
label-smoothing	-	-	-	0.2	0.2
dropout	0.1		0.1*		0.1
Translation stage					
number of epochs	200	100	*	75	80
LLM backbone	3layers-Mbart	XGLM	3layers-Mbart	12layers-Mbart	12layers-Mbart
batch size	8	8	128	16	128
optimizer	SGD	adam	SGD*	adam	adam
scheduler	cosine	one-cycle cosine	cosine*	cosine	cosine
learning rate	10^{-2}	3×10^{-4}	10^{-2} *	10^{-5}, 10^{-3}	10^{-5}, 10^{-3}
label-smoothing	0.2	0.1	0.2*	0.2	0.2
dropout					0.3
Data processing					
frame sampling strategy	random	subsampling 25%	random*	subsampling 50%	subsampling 50%
max number of frames	300		300*		
Scores					
BLEU library	nlgeval	sacrableu	nlgeval*		
BLEU-1	43.71	49.54	49.76	46.29	52.81
BLEU-4	21.44	22.52	22.74	23.09	26.75
ROUGE	42.49	48.94	49.04	45.27	50.96

Table 2: Architectural design and hyperparameters of methods on the Phoenix-2014T [2]. Empty cells indicate that the information is not reported in the original papers. *VLP*: Visual-Language Pretraining, *Light-T*: Lightweight translation model, *CiCO*: Cross-Ingual COntrastive learning [54].
† A temporal module (Temporal Block + Temporal Encoder) follows ResNet18 in all cases. These temporal components differ across methods and are detailed in the sub-rows. * These details are not reported explicitly in SignCL, but the authors state that all training settings follow those used in GFSLT.

3.3. Sign2GPT

Sign2GPT [19] utilised pseudo-gloss pretraining, where glosses are generated from spoken sentences. During pretraining, visual features are aligned to pseudo-glosses instead of full sentences. For the visual encoder, Dino-V2 Vision Transformer is followed by a 4-layers transformer. In the second stage, XGLM, a multilingual Generative Pretrained Transformer (GPT) model is fine-tuned with LoRA (Low-Rank Adapters), a lightweight adaptation technique.

3.4. FLa-LLM

While the aforementioned approaches utilized contrastive learning to align video and text features in their pretraining stage, **FLa-LLM** [14] employs a lightweight translation (Light-T) model while pre-training the visual encoder. This Light-T corresponds to the second stage of GFSLT, where a 3-layer mBART decoder is used to generate translations from video representations. In the second stage, the pretrained visual encoder is frozen, and a larger MBart (12 layers) is employed. Compared to GFSLT, this model differs in several architectural aspects – such as applying frame subsampling instead of max pooling and reducing embedding dimensionality (see Table 2).

3.5. C^{2}RL

C^{2}RL [13] introduces multi-task pretraining framework by combining visual-text alignment and a lightweight translation objective. Similar to GFSLT, it aligns sign language videos and spoken language sentences in a shared semantic space, but further incorporates a Cross-Ingual COntrastive learning (CiCO) loss [54], referred to as Implicit Content Learning (ICL). Additionally, the Explicit Context Learning (ECL) focuses on translating video inputs to spoken language.

4. Evaluation Methodology

In this section, we describe how the main contributions of each model were incorporated into our unified codebase during the reproduction process. Rather than re-implementing every detail, we focused on the core ideas proposed by each method under a standardized training setup.

- **Visual-language pretraining:** Motivated by [11], we adopt a two-stage SLT approach, combining contrastive visual-language pretraining and encoder-decoder fine-tuning. In our setup, we retain its visual and textual encoder structure and pretraining objectives. The architecture of our baseline is illustrated in Figure 4.

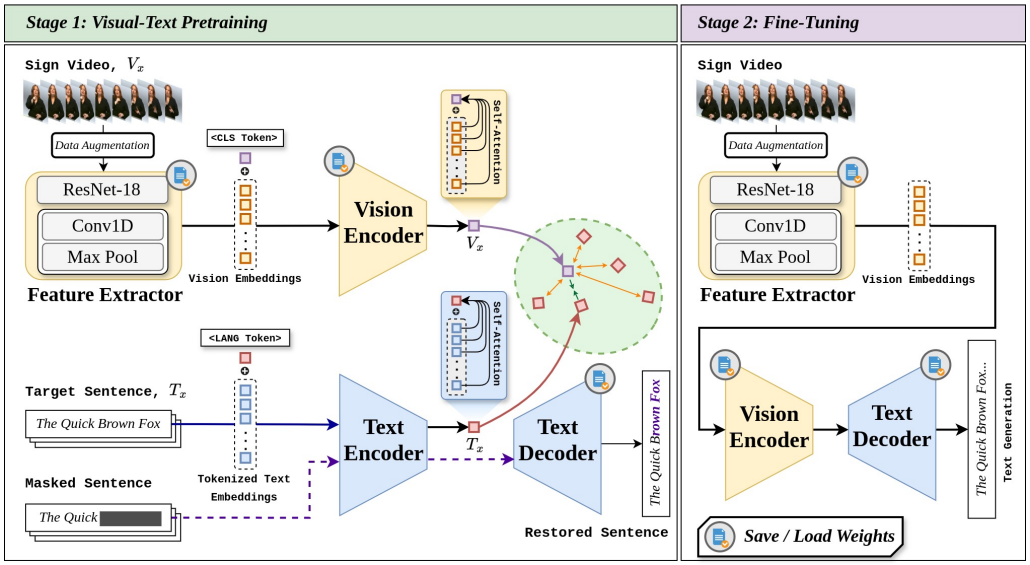


Figure 4: Overview of our two-stage architecture. ***Pretraining stage:*** Video frames are encoded using a ResNet-18 backbone, followed by temporal modeling with 1D convolutions and max pooling. A [CLS] token is appended and passed through a Transformer encoder to obtain a global video representation. Text is encoded similarly and summarized into a [LANG] token. Contrastive learning is then employed to match CLS-LANG token pairs. Meanwhile, a masked sentence modeling objective is used to train the text decoder. ***Fine-tuning stage:*** The pretrained visual encoder, text encoder, and decoder are loaded from pretraining and fine-tuned for sign translation using standard sequence generation.

- **Frame-level contrastive learning** is integrated into our baseline network to learn more discriminative feature representation in a self-supervised manner. We adopt the loss formulation from the official codebase provided by the authors [15]. The loss is added to both the pretraining and fine-tuning stages.

- **Pseudo-gloss pretraining:** We first create gloss-like intermediate representations, pseudo-glosses from spoken language sentences utilizing the official codebase [19]. Then, we follow the pseudo-gloss pretraining strategy instead of the base CLIP-style loss. However, we keep our visual backbone consistent, such as Resnet18, Conv1D-BN-Relu-MaxPooling, and a transformer encoder.

- **Lightweight translation, Light-T:** To implement the key idea of factorized learning [14] within our unified framework, we first follow the translation task baseline setup, which

aligns with the lightweight translation (Light-T) objective, employing a 3-layer mBART decoder to generate translations from video features. In the second stage, we freeze the pretrained visual encoder and replace the decoder with a larger, 12-layer mBART to enhance language modeling capacity. As opposed to the original paper, which uses a lower embedding size in the visual encoder, applies a 25% frame sampling strategy, and trains with a larger batch size, we retain our standardized settings.

- **Cross-lingual Contrastive learning (CiCO) loss in pretraining:** As the first contribution of C^2RL [13], the Implicit Content Learning (ICL), we replace the original CLIP-style contrastive loss with CiCO loss [54] in pretraining. In the second stage, we perform standard SLT fine-tuning using the pretrained visual encoder and a 3-layer mBART decoder.

- **Multi-task pretraining:** We refer to the core contribution of C^2RL, namely the combination of Implicit Content Learning (ICL) and Explicit Content Learning (ECL), as a multi-task pretraining strategy. Specifically, we jointly apply the CiCO loss for visual-text alignment and a lightweight translation objective using a 3-layer mBART decoder. For the second stage, we load only Resnet18, Conv1D-BN-Relu-MaxPooling as the visual encoder. We fine-tune a 12-layer mBART decoder while keeping the pretrained visual encoder frozen.

5. Results

This section presents the results of evaluating several state-of-the-art gloss-free sign language translation (SLT) approaches. The goal of this evaluation is to provide a critical comparison of existing methods. We trained and tested all models under identical conditions, ensuring a fair and consistent comparison to measure the effectiveness of their design decisions.

5.1. Evaluation Metrics

We use BLEU [57] and ROUGE [58] metrics to evaluate the performance of our SLT approach. BLEU (Bilingual Evaluation Understudy) measures the overlap between generated translations and reference sentences by computing precision over n-grams (consecutive word sequences). Particularly, BLEU-4 becomes the most commonly used variant in SLT. ROUGE-L (Recall-Oriented Understudy for Gisting Evaluation-Longest Common Subsequence), on the other hand, evaluates the quality of a generated sentence by measuring the length of the longest common subsequence between the candidate and reference texts. Unlike n-gram based metrics, ROUGE-L captures sentence-level structure, making it suitable for evaluating translations with flexible word order.

5.2. Implementation Details

We train all models on a single NVIDIA A100 GPU with a batch size of 8. We begin our experiments on the Phoenix-2014T dataset, initially adopting hyperparameters from the GFSLT-VLP implementation [11]. In pretraining, the original GFSLT used a batch size of 16 with a learning rate of 10^{-2}. Since we reduce the batch size to 8 due to hardware constraints, we proportionally lower the learning rate to 5×10^{-3} during pretraining. For the translation stage, where both the original and our setup use a batch size of 8, we maintain the original learning rate of 10^{-2}. This configuration serves as our baseline, and we strive to maintain consistency across all methods and datasets.

However, *exact consistency was not achievable across different datasets* as seen in Table 3. While reproducing the baseline results on CSL-Daily, we observed that validation loss began to rise early in training (Figure 5), suggesting the **learning rate** was too high. We empirically reduced the learning rate to 10^{-3} for both pretraining and translation stages on CSL-Daily.

	Phoenix-2014T			**CSL-Daily**		
	GFSLT, SignCL	**FLa-LLM**	**C^{2}RL**	**GFSLT, SignCL**	**FLa-LLM**	C^{2}RL
Pretraining stage						
number of epochs	80	100	200	80	100	100
learning rate	5×10^{-3}	10^{-2}	10^{-2}	5×10^{-3}, $\underline{10^{-3}}$	$\underline{5\times10^{-3}}$, 10^{-3}	$\underline{5\times10^{-3}}$, 10^{-3}
Translation stage						
number of epochs	200	75	80	80	75	80
learning rate	10^{-2}	10^{-2}, $\underline{5\times10^{-3}}$	$\underline{10^{-2}}$, 5×10^{-3}	10^{-2}, $\underline{10^{-3}}$	5×10^{-3}	$\underline{5\times10^{-3}}$, 10^{-3}

Table 3: Hyperparameters of different datasets across methods. Underlined values indicate the best-performing learning rate.

(a) Pretraining stage

(b) Translation stage

Figure 5: Validation loss and BLEU-4 score graphs for the GFSLT-VLP baseline [11]. For CSL-Daily, validation loss remained stable or began to rise early in training when using the same learning rate as for Phoenix-2014T, suggesting that the learning rate was too high.

In addition, the **training time** varied significantly across datasets. We observed that models converged much earlier on CSL-Daily, with minimal improvement beyond 50 epochs. Therefore, we trained models on CSL-Daily up to 80 epochs in the second phase. We hypothesize that early convergence is because CSL-Daily is larger, has cleaner videos recorded in a lab setting, and includes some sentences that are repeated by different signers.

Another consideration is the variation in **sequence lengths**. Both Phoenix-2014T and CSL-Daily contain videos typically fewer than 300 frames. While GFSLT-VLP and SignCL restricted the maximum number of frames to 300 by random sampling, other methods did not report such details. We applied the same restriction to all methods to ensure a fair comparison.

All models are trained and evaluated three times with different random seeds (0, 42, and 100) to ensure robustness and reproducibility of results.

5.3. Experimental Comparison of Models

We successfully reproduce GFSLT-VLP [11] results on Phoenix-2014T, with our BLEU-2, BLEU-3, and BLEU-4 scores closely aligning with the originally reported values (Table 4), and our reproduction achieves higher BLEU-1 (43.71 vs 46.44 ± 0.48) and ROUGE (42.49 vs 46.87 ± 0.58). For CSL-Daily, we use the adjusted learning rate and shortened training time as described in Section 5.2. However, our results on CSL-Daily are slightly below those reported in the original paper, as seen in Table 5 (BLEU-4: 11.00 vs 10.41).

Next, we added the contrastive learning loss for adjacent frames from SignCL [15]. Our reproduced results showed improvements over the GFSLT-VLP baseline, but did not reach the

Method	BLEU1	BLEU2	BLEU3	BLEU4	ROUGE-L
Reported					
GFSLT-VLP [11]	43.71	33.18	26.11	21.44	42.49
Sign2GPT [19]	49.54	35.96	28.83	22.52	48.94
SignCL [15]	49.76	36.85	29.97	22.74	49.04
FLa-LLM [14]	46.29	35.33	28.03	23.09	45.27
C^2RL (ICL only) [13]	-	-	-	24.83	
C^2RL (ICL + ECL) [13]	52.81	40.20	32.20	26.75	50.96
Reproduced					
Visual-language pretraining [11]	46.44 ± 0.48	33.96 ± 0.40	26.69 ± 0.32	21.97 ± 0.27	46.87 ± 0.58
Contrastive learning on frames [15]	46.62 ± 0.28	34.41 ± 0.15	27.06 ± 0.22	22.35 ± 0.28	47.30 ± 0.26
Lightweight translation (Light-T) in pretraining [14]	45.29 ± 0.26	32.08 ± 0.29	24.53 ± 0.33	19.84 ± 0.28	44.43 ± 0.15
Utilize CICO loss in pretraining [13]	47.30 ± 0.47	**35.05 ± 0.51**	**27.72 ± 0.59**	**22.95 ± 0.63**	**48.20 ± 0.37**
CICO loss and Light-T task in pretraining [13]	**47.61 ± 0.35**	34.41 ± 0.42	26.48 ± 0.43	21.41 ± 0.43	46.84 ± 0.72

Table 4: Reported and reproduced results on the Phoenix-2014T.

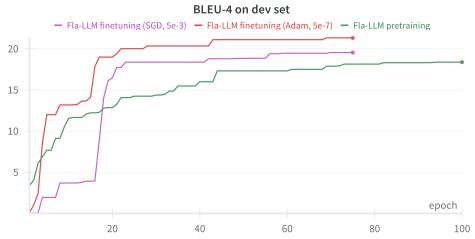
Method	BLEU1	BLEU2	BLEU3	BLEU4	ROUGE-L
Reported					
GFSLT-VLP [11]	39.37	24.93	16.26	11.00	36.44
Sign2GPT [19]	41.75	28.73	20.60	15.40	42.36
SignCL [15]	47.47	32.53	22.62	16.16	48.92
FLa-LLM [14]	37.13	25.12	18.38	14.20	37.25
C^2RL (ICL + ECL) [13]	49.32	36.28	27.54	21.61	48.21
Reproduced					
Visual-language pretraining [11]	36.78 ± 0.67	23.23 ± 0.50	15.25 ± 0.34	10.41 ± 0.17	35.60 ± 0.65
Contrastive learning on frames [15]	37.31 ± 0.26	23.97 ± 0.26	16.11 ± 0.21	11.22 ± 0.18	36.42 ± 0.12
Lightweight translation (Light-T) in pretraining [14]	37.41 ± 1.33	23.30 ± 1.07	15.22 ± 0.70	10.26 ± 0.40	36.69 ± 1.11
Utilize CICO loss in pretraining [13]	**39.98 ± 0.18**	**26.06 ± 0.13**	**17.45 ± 0.06**	**12.03 ± 0.06**	**38.42 ± 0.17**
CICO loss and Light-T task in pretraining [13]	39.59 ± 0.10	25.39 ± 0.19	16.77 ± 0.25	11.51 ± 0.27	37.36 ± 0.34

Table 5: Reported and reroduced results on the CSL-Daily.

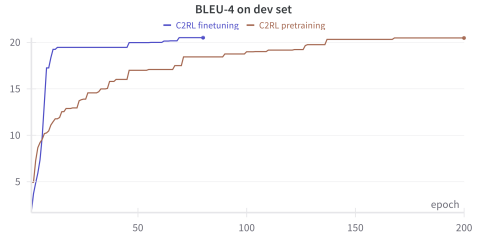
reported results in the original paper. This issue has also been noted by other researchers. We believe the possible reason for this discrepancy is the use of a smaller batch size (8 rather than 128 as reported), which may have limited the effectiveness of the contrastive loss. Nevertheless, additional contrastive learning yielded a lower standard deviation across seeds, indicating better training stability.

We then attempted to implement pseudo-gloss pretraining into our architecture. However, this configuration yielded significantly lower performance compared to the original paper and was therefore excluded from our results table. To validate the reproducibility of Sign2GPT [19], we also ran their original codebase, which includes components such as DINOv2 [59], LoRA (Low-Rank Adapters), and the XGLM language model [60]. Using their official setup, we were able to reproduce results that align closely with those reported in the original paper. This suggests that the performance of Sign2GPT [19] depends on its specific architectural and training choices.

Our next experiment explores the use of lightweight-translation during pretraining stage, which is the key contribution proposed in FLa-LLM [14]. To this end, we used the translation task setup from the GFSLT-VLP baseline and obtained a BLEU-4 score of 19.50 (Table 6). This score is expected, as it reflects the performance of the SLT model without any visual-language pretraining. In the subsequent fine-tuning stage, although learning progressed quickly at the beginning, BLEU scores plateaued shortly after (see Fig. 6). The final BLEU-4 score of 19.84 ± 0.28 remained below the baseline, likely due to early overfitting caused by the increased model capacity. These results were obtained using the same optimizer, SGD, as in our earlier experiments. To better understand the performance gap, we conducted an ablation study using the



(a) FLa-LLM: Light translation during pretraining



(b) C^{2}RL: CiCO loss and light translation during pretraining

Figure 6: BLEU-4 progression on the Phoenix-2014T dev set for FLa-LLM [14] and C^{2}RL [13] during training. Each model is first pretrained and then fine-tuned starting from the best pretrained checkpoint. Fine-tuning exhibits rapid improvement, but then saturates. This behavior causes the gap between our reproduced results and those reported in the original papers.

Stage	Optimizer	Learning rate	BLEU1	BLEU2	BLEU3	BLEU4	ROUGE-L
Pretraining	SGD	10^{-2}	44.63	31.72	24.15	19.50	43.69
Fine-tuning	SGD	5×10^{-3}	45.29 ± 0.26	32.08 ± 0.29	24.53 ± 0.33	19.84 ± 0.28	44.43 ± 0.15
	Adam	5×10^{-6}	44.44	32.13	24.89	20.24	44.47
	Adam	LLM: 5×10^{-6}, LLM Adapter: 5×10^{-4}	x	x	x	x	x
	Adam	5×10^{-7}	46.55	33.76	26.16	21.09	44.86
	Adam	LLM: 5×10^{-7}, LLM Adapter: 5×10^{-5}	45.37	32.75	25.14	20.23	43.51

Table 6: Performance of FLa-LLM [14] on Phoenix-2014T using different learning rate variations with Adam optimizer. The experiment marked with 'x' was aborted early due to unstable training caused by a high learning rate.

Adam optimizer, as employed in the original paper, on the Phoenix-2014T dataset. A comparison of BLEU-4 scores across different learning rates using Adam is provided in Table 6. While Adam leads to better performance than SGD, it was still insufficient to close the gap with the reported results. Several other factors may have contributed to this discrepancy. For instance, FLa-LLM employs a lower embedding size in the visual encoder, applies a 25% frame sampling strategy, and uses a larger batch size. The authors' own ablation studies show that both of these choices positively affect performance.

In our first experiment with the C^{2}RL setup, we replaced the original CLIP-based visual-text alignment loss with the CiCO loss [54], which is designed to better capture semantic relationships across sign and spoken modalities. This substitution resulted in improved translation performance, achieving BLEU-4 scores of 22.95 (vs. 21.97) on Phoenix-2014T and 12.03 (vs. 10.41) on CSL-Daily. It suggests that CiCo provides a more effective supervision signal for aligning video and text representations.

Next, we combined CiCo-based alignment with lightweight translation during pretraining. Compared to the single-task setup used in FLa-LLM—where only lightweight translation was performed—this multi-task configuration yielded better initial performance. However, as with FLa-LLM, further improvements could not be achieved in the second stage. Although learning initially progressed rapidly, BLEU scores quickly plateaued. This behavior may point to overfitting or limited transferability beyond the pretraining stage.

We therefore believe that the design choices in the original paper, such as applying dropout in the second stage, and adopting a larger batch size, may have contributed to mitigating overfitting and achieving higher BLEU scores. In contrast, we deliberately refrained from making such architectural modifications introduced in their work to ensure consistency across our own experiments and to isolate the effect of the key contribution, the multi-task pretraining strategy.

5.4. Metric Inconsistency in Evaluation

Dataset-Specific Conventions: We observed that metric calculations differ between datasets. For instance, although Phoenix-2014T does not contain punctuation in its original sentences, many works append a dot with a space (‘ .’) to each sentence, artificially inflating BLEU scores. On the other hand, in CSL-Daily, character-based BLEU is often reported, which produces again higher scores than word-based evaluation. These implementation-specific conventions are rarely stated in papers and need to be inferred from qualitative results or the available implementations. To ensure a fair comparison with prior works, we follow the same conventions when reporting results on each dataset.

Library-Specific Variability: We also found that some metric scores may not be consistent across libraries, particularly between popular toolkits such as `nlg-eval` and `sacreBLEU`. This discrepancy arises in CSL-Daily, where character-based segmentation is used. For example, the same baseline model [11] can yield a BLEU-4 score of 10.41 using `nlg-eval` (v2.4.1) and 11.66 using `sacreBLEU` [4]. In contrast, Phoenix-2014T does not exhibit such variation in our experiments. For all aforementioned experiments, we used `nlg-eval` to ensure a fair comparison.

6. Conclusion

This paper presented a systematic comparison of recent gloss-free sign language translation models, which differ in their pretraining strategies. These strategies include vision-language pretraining, contrastive learning on adjacent frames, pseudo-gloss supervision, lightweight translation modules during pretraining, and the combination of visual-language alignment with lightweight translation.

By re-implementing all models within a unified, modular codebase and evaluating them under identical training conditions, we isolate the contribution of each pretraining approach from other factors such as data preprocessing and optimization schedules. Our experiments show that performance gaps reported in the literature often narrow under controlled conditions. Still, we observe that some pretraining designs, such as employing cross-lingual contrastive learning to align sign videos with textual representations, yield better results.

We release our codebase to promote transparency, reproducibility, and future experimentation in sign language translation research. This work provides a standardized and modular framework that can serve as a helpful starting point for further studies. Future directions include scaling the models and evaluation to larger sign language datasets to improve generalization across different signers and domains.

Acknowledgments

This work was supported by the SNSF project ‘SMILE II’ (CRSII5 193686), the Innosuisse IICT Flagship (PFFS-21-47), EPSRC grant APP24554 (SignGPT-EP/Z535370/1), and through funding from Google.org via the AI for Global Goals scheme. This work reflects only the author’s views and the funders are not responsible for any use that may be made of the information it contains.

[4] The exact signature for the sacreBLEU is: `nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.2.0`. `tok:zh` also produced the same results.

References

[1] R. Sutton-Spence, B. Woll, The linguistics of British Sign Language: an introduction, Cambridge University Press, 1999.

[2] N. C. Camgoz, S. Hadfield, O. Koller, H. Ney, R. Bowden, Neural sign language translation, in: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 2018, pp. 7784–7793.

[3] S. Cox, M. Lincoln, J. Tryggvason, M. Nakisa, M. Wells, M. Tutt, S. Abbott, Tessa, a system to aid communication with deaf people, in: Proceedings of the fifth international ACM conference on Assistive technologies, 2002, pp. 205–212.

[4] J. McDonald, R. Wolfe, J. Schnepp, J. Hochgesang, D. G. Jamrozik, M. Stumbo, L. Berke, M. Bialek, F. Thomas, An automated technique for real-time production of lifelike animations of american sign language, Universal Access in the Information Society 15 (2016) 551–566.

[5] B. Saunders, N. C. Camgoz, R. Bowden, Signing at scale: Learning to co-articulate signs for large-scale photo-realistic sign language production, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 5141–5151.

[6] H. Zhou, W. Zhou, W. Qi, J. Pu, H. Li, Improving sign language translation with monolingual data by sign back-translation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 1316–1325.

[7] Y. Chen, R. Zuo, F. Wei, Y. Wu, S. Liu, B. Mak, Two-stream network for sign language recognition and translation, Advances in Neural Information Processing Systems 35 (2022) 17043–17056.

[8] A. Yin, Z. Zhao, J. Liu, W. Jin, M. Zhang, X. Zeng, X. He, Simulslt: End-to-end simultaneous sign language translation, in: Proceedings of the 29th ACM International Conference on Multimedia, 2021, pp. 4118–4127.

[9] H. Zhou, W. Zhou, Y. Zhou, H. Li, Spatial-temporal multi-cue network for sign language recognition and translation, IEEE Transactions on Multimedia 24 (2021) 768–779.

[10] A. Duarte, S. Palaskar, L. Ventura, D. Ghadiyaram, K. DeHaan, F. Metze, J. Torres, X. Giro-i Nieto, How2sign: a large-scale multimodal dataset for continuous american sign language, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 2735–2744.

[11] B. Zhou, Z. Chen, A. Clapés, J. Wan, Y. Liang, S. Escalera, Z. Lei, D. Zhang, Gloss-free sign language translation: Improving from visual-language pretraining, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (CVPR), 2023, pp. 20871–20881.

[12] A. Yin, T. Zhong, L. Tang, W. Jin, T. Jin, Z. Zhao, Gloss attention for gloss-free sign language translation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 2551–2562.

[13] Z. Chen, B. Zhou, Y. Huang, J. Wan, Y. Hu, H. Shi, Y. Liang, Z. Lei, D. Zhang, C2rl: Content and context representation learning for gloss-free sign language translation and retrieval, IEEE Transactions on Circuits and Systems for Video Technology (2025).

[14] Z. Chen, B. Zhou, J. Li, J. Wan, Z. Lei, N. Jiang, Q. Lu, G. Zhao, Factorized learning assisted with large language model for gloss-free sign language translation, the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING) (2024).

[15] J. Ye, X. Wang, W. Jiao, J. Liang, H. Xiong, Improving gloss-free sign language translation by reducing representation density, in: The Thirty-eighth Annual Conference on Neural Information Processing Systems, 2024.

[16] J. Kim, H. Jeon, J. Bae, H. Y. Kim, Leveraging the power of mllms for gloss-free sign language translation, arXiv preprint arXiv:2411.16789 (2024).

[17] O. M. Sincan, N. C. Camgoz, R. Bowden, 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, 2023, pp. 1955–1965.

[18] Z. Li, W. Zhou, W. Zhao, K. Wu, H. Hu, H. Li, Uni-sign: Toward unified sign language understanding at scale, in: The Thirteenth International Conference on Learning Representations, 2025.

[19] R. Wong, N. C. Camgoz, R. Bowden, Sign2GPT: Leveraging large language models for gloss-free sign language translation, in: The Twelfth International Conference on Learning Representations, 2024.

[20] H. R. V. Joze, O. Koller, Ms-asl: A large-scale data set and benchmark for understanding american sign language, arXiv preprint arXiv:1812.01053 (2018).

[21] D. Li, C. Rodriguez, X. Yu, H. Li, Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2020, pp. 1459–1469.

[22] O. M. Sincan, H. Y. Keles, Autsl: A large scale multi-modal turkish sign language dataset and baseline methods, IEEE access 8 (2020) 181340–181355.

[23] S. Jiang, B. Sun, L. Wang, Y. Bai, K. Li, Y. Fu, Skeleton aware multi-modal sign language recognition, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 3413–3423.

[24] H. Hu, W. Zhou, H. Li, Hand-model-aware sign language recognition, in: Proceedings of the AAAI conference on artificial intelligence, Vol. 35, 2021, pp. 1558–1566.
[25] R. Zuo, F. Wei, B. Mak, Natural language-assisted sign language recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 14890–14900.
[26] Y. Min, A. Hao, X. Chai, X. Chen, Visual alignment constraint for continuous sign language recognition, in: proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 11542–11551.
[27] F. Wei, Y. Chen, Improving continuous sign language recognition with cross-lingual signs, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 23612–23621.
[28] L. Hu, L. Gao, Z. Liu, W. Feng, Continuous sign language recognition with correlation network, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 2529–2539.
[29] N. C. Camgoz, O. Koller, S. Hadfield, R. Bowden, 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, 2020, pp. 10023–10033.
[30] A. Graves, S. Fernández, F. Gomez, J. Schmidhuber, Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks, in: Proceedings of the 23rd international conference on Machine learning, 2006, pp. 369–376.
[31] Y. Chen, F. Wei, X. Sun, Z. Wu, S. Lin, A simple multi-modality transfer learning baseline for sign language translation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 5120–5130.
[32] B. Zhang, M. Müller, R. Sennrich, Sltunet: A simple unified model for sign language translation, arXiv preprint arXiv:2305.01778 (2023).
[33] O. M. Sincan, N. C. Camgoz, R. Bowden, Using an llm to turn sign spottings into spoken language sentences, arXiv preprint arXiv:2403.10434 (2024).
[34] R. Zuo, F. Wei, B. Mak, Towards online continuous sign language recognition and translation, in: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, 2024, pp. 11050–11067.
[35] L. Jing, X. Song, X. Zu, N. Zheng, Z. Zhao, L. Nie, Vk-g2t: Vision and context knowledge enhanced gloss2text, in: ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2024, pp. 7860–7864.
[36] P. Fayyazsanavi, A. Anastasopoulos, J. Kosecka, Gloss2text: Sign language gloss translation using llms and semantically aware label smoothing, in: Findings of the Association for Computational Linguistics: EMNLP 2024, 2024, pp. 16162–16171.
[37] F. Zhou, T. Van de Cruys, Non-autoregressive modeling for sign-gloss to texts translation, Proceedings of Machine Translation Summit XX Volume 1 (2025) 220–230.
[38] J. Gong, L. G. Foo, Y. He, H. Rahmani, J. Liu, Llms are good sign language translators, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (CVPR), 2024, pp. 18362–18372.
[39] K. Lin, X. Wang, L. Zhu, K. Sun, B. Zhang, Y. Yang, Gloss-free end-to-end sign language translation, in: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2023, pp. 12904–12916.
[40] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: International Conference on Machine Learning, PmLR, 2021, pp. 8748–8763.
[41] E. J. Hwang, S. Cho, J. Lee, J. C. Park, An efficient sign language translation using spatial configuration and motion dynamics with llms, arXiv preprint arXiv:2408.10593 (2024).
[42] H. Liang, C. Huang, Y. Xu, C. Tang, W. Ye, J. Zhang, X. Chen, J. Yu, L. Xu, Llava-slt: Visual language tuning for sign language translation, arXiv preprint arXiv:2412.16524 (2024).
[43] Y. Jang, H. Raajesh, L. Momeni, G. Varol, A. Zisserman, Lost in translation, found in context: Sign language translation with contextual cues, in: Proceedings of the Computer Vision and Pattern Recognition Conference, 2025, pp. 8742–8752.
[44] P. Rust, B. Shi, S. Wang, N. C. Camgöz, J. Maillard, Towards privacy-aware sign language translation at scale, in: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2024, pp. 8624–8641.
[45] J. Forster, C. Schmidt, T. Hoyoux, O. Koller, U. Zelle, J. Piater, H. Ney, Rwth-phoenix-weather: A large vocabulary sign language recognition and translation corpus, in: Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12), 2012, pp. 3785–3789.
[46] A. Desai, L. Berger, F. Minakov, N. Milano, C. Singh, K. Pumphrey, R. Ladner, H. Daumé III, A. X. Lu, N. Caselli, et al., Asl citizen: a community-sourced dataset for advancing isolated sign language recognition, Advances in Neural Information Processing Systems 36 (2023) 76893–76907.
[47] A. Schembri, F. Jordan, R. Rentelis, K. Cormier, British sign language corpus project: A corpus of digital video data and annotations of british sign language 2008-2017 (third edition). (2017).

URL https://www.bslcorpusproject.org
[48] R. Konrad, T. Hanke, G. Langer, D. Blanck, J. Bleicken, I. Hofmann, O. Jeziorski, L. König, S. König, R. Nishio, A. Regen, U. Salden, S. Wagner, S. Worseck, O. Böse, E. Jahn, M. Schulder, Meine dgs – annotiert. öffentliches korpus der deutschen gebärdensprache, 3. release / my dgs – annotated. public corpus of german sign language, 3rd release (2020).
URL https://doi.org/10.25592/dgs.corpus-3.0
[49] A. Schembri, J. Fenlon, R. Rentelis, S. Reynolds, K. Cormier, Building the british sign language corpus, University of Hawaii Press (2013).
[50] S. Albanie, G. Varol, L. Momeni, H. Bull, T. Afouras, H. Chowdhury, N. Fox, B. Woll, R. Cooper, A. McParland, et al., Bbc-oxford british sign language dataset, arXiv preprint arXiv:2111.03635 (2021).
[51] B. Shi, D. Brentari, G. Shakhnarovich, K. Livescu, Open-domain sign language translation learned from online video, in: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, 2022, pp. 6365–6379.
[52] D. Uthus, G. Tanzer, M. Georg, Youtube-asl: A large-scale, open-domain american sign language-english parallel corpus, Advances in Neural Information Processing Systems 36 (2023) 29029–29047.
[53] G. Tanzer, B. Zhang, Youtube-sl-25: A large-scale, open-domain multilingual sign language parallel corpus, in: The Thirteenth International Conference on Learning Representations, 2025.
[54] Y. Cheng, F. Wei, J. Bao, D. Chen, W. Zhang, Cico: Domain-aware sign language retrieval via cross-lingual contrastive learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 19016–19026.
[55] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.
[56] Y. Liu, J. Gu, N. Goyal, X. Li, S. Edunov, M. Ghazvininejad, M. Lewis, L. Zettlemoyer, Multilingual denoising pre-training for neural machine translation, Transactions of the Association for Computational Linguistics 8 (2020) 726–742.
[57] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, Bleu: a method for automatic evaluation of machine translation, in: Proceedings of the 40th annual meeting of the Association for Computational Linguistics, 2002, pp. 311–318.
[58] C.-Y. Lin, Rouge: A package for automatic evaluation of summaries, in: Text summarization branches out, 2004, pp. 74–81.
[59] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, et al., Dinov2: Learning robust visual features without supervision, arXiv preprint arXiv:2304.07193 (2023).
[60] X. V. Lin, T. Mihaylov, M. Artetxe, T. Wang, S. Chen, D. Simig, M. Ott, N. Goyal, S. Bhosale, J. Du, R. Pasunuru, S. Shleifer, P. S. Koura, V. Chaudhary, B. O'Horo, J. Wang, L. Zettlemoyer, Z. Kozareva, M. Diab, V. Stoyanov, X. Li, Few-shot learning with multilingual generative language models, in: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022.