Continual Learning in Multilingual Sign Language Translation

Shakib Yazdani and **Josef van Genabith** and **Cristina España-Bonet**
{Shakib.Yazdani,Josef.van_Genabith,cristinae}@dfki.de
German Research Center for Artificial Intelligence (DFKI GmbH)
Saarland Informatics Campus, Saarbrücken, Germany

Abstract

The field of sign language translation (SLT) is still in its infancy, as evidenced by the low translation quality, even when using deep learning approaches. Probably because of this, many common approaches in other machine learning fields have not been explored in sign language. Here, we focus on continual learning for multilingual SLT. We experiment with three continual learning methods and compare them to four more naive baseline and fine-tuning approaches. We work with four sign languages (ASL, BSL, CSL and DGS) and three spoken languages (Chinese, English and German). Our results show that incremental fine-tuning is the best performing approach both in terms of translation quality and transfer capabilities, and that continual learning approaches are not yet fully competitive given the current SOTA in SLT.

1 Introduction

Continual Learning, Incremental Learning and Life-long Learning are equivalent terms for machine learning approaches that learn sequentially or from dynamic data, where a sequence can be either a succession of tasks, datasets, languages, etc (Li and Hoiem, 2018; Ke and Liu, 2022; Gogoulou et al., 2023; Wang et al., 2024; Shi et al., 2024).

Data in real-world applications is rarely static; it typically arrives as a continuous stream, with only a fraction available at the beginning. This dynamic flow necessitates that deployed systems adapt incrementally to new information over time. Continual learning is critical for such scenarios, as it allows models to integrate new data—such as additional sign languages—without requiring complete retraining or redeployment. One of the main challenges in continual learning is catastrophic forgetting (Goodfellow et al., 2013; Kirkpatrick et al., 2016a; Lee et al., 2017), and tackling this issue is central to continual learning research. Researchers work on methods to enable models to adapt to a sequence of tasks while retaining previously acquired knowledge (Chen and Liu, 2017; van de Ven et al., 2022; Shi et al., 2024). While the phenomenon of catastrophic forgetting and its mitigation has been extensively studied in the field of continual learning, these investigations have primarily focused on traditional natural language processing (NLP) tasks and/or multilingual NLP contexts, and never before on sign language.

Wang et al. (2024) define a taxonomy for continual learning approaches based on their location within the machine learning pipeline, e.g., when feeding the data, at the architecture level, during optimization, etc. Their five classes correspond to approaches based on *replay* (saving, recovering or mimicking old data), *regularization* (either weight, feature or function regularization adding terms that take into account the old model), *optimization* (gradient modification or projection, meta-learning, etc.), *representation* (self-supervision, (continual) pre-training, adaptation of a fixed backbone) and *architecture* (including task-specific or adaptive parameters). In our work and for comparison purposes, we consider an example from each of the replay, regularization, and architecture classes.

Sign language translation (SLT) is a hard problem. It involves translating from an input video to text or speech with traditionally small amounts of parallel data such as the Phoenix2014T dataset (Forster et al., 2014) with 7096 video–text pairs for German sign language, and the CSL-Daily dataset (Zhou et al., 2021) with 18400 for Chinese sign language. These numbers are very far from the millions of parallel sentences used to train text-to-text transformer models for machine translation.

State-of-the-art SLT systems (Chen et al., 2022b) have traditionally relied on intermediate (manual) annotations called glosses and this has been limiting the size of the training data for a long time. Recently, the field has been significantly

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 10923–10938

advanced by the emergence of large, non-curated datasets (without glosses), such as BOBSL (Albanie et al., 2021) and YouTube-SL-25 (Tanzer and Zhang, 2024) both with more than 1 million video–text pairs, the first for British sign language and the second one for a combination of more than 25 sign languages. This new data, especially that coming from YouTube, is motivating sign language translation systems with a strong focus on pre-training.

To the best of our knowledge, we are the first to examine the effects of continual learning in multilingual sign language translation. This work is developed on the eve of forthcoming (and hopefully open) large multilingual sign language translation models, which, similar to machine translation, will enable fine-tuning and life-long learning. We present a gloss-free multilingual transformer model for sign language translation, pre-trained on Chinese Sign Language (CSL) to Chinese, German Sign Language (DGS) to German, and American Sign Language (ASL) to English using the Phoenix2014T, CSL-Daily, and How2Sign datasets. To evaluate its performance, we conduct experiments on three languages from the Spreadthesign-Ten (SP-10) dataset (Yin et al., 2022)—Chinese Sign Language, German Sign Language, and British Sign Language (BSL)—in a sequential learning setup, comparing continual and non-continual fine-tuning approaches.[1]

2 Related Work

Continual Learning Approaches. As explained in the introduction, multiple approaches exist in the literature to address the task. Relevant to our work, and based on the taxonomy of Wang et al. (2024), are regularization-based, replay-based, and architecture-based approaches. *Regularization-based* methods stabilize model parameters by adding regularization terms. These methods are simple to implement but require access to the previous model for reference, and can focus on weight or function regularization (Wang et al., 2024). Weight regularization adjusts parameters based on the importance of the old model, often using the Fisher information matrix (FIM), as in Elastic Weight Consolidation (EWC) (Kirkpatrick et al., 2016b) and its variants like Memory Aware Synapses (MAS) (Aljundi et al., 2018). Function regularization preserves the outputs of the previous model through knowledge distillation, with the old model as the teacher and the current model as the student. Techniques like LwF (Li and Hoiem, 2016), LwM (Dhar et al., 2018), iCaRL (Rebuffi et al., 2016), EEIL (Castro et al., 2018), and LUCIR (Hou et al., 2019) use either new or old training samples for distillation loss. *Replay-based* methods have proven to be the most effective among the three main continual learning (CL) approaches—regularization-based, replay-based, and architecture-based—as demonstrated by van de Ven and Tolias (2019). Replay-based methods retain a small set of old training samples in a memory buffer, which are later used for training alongside the current data. Some methods, like reservoir sampling (Chaudhry et al., 2019), ring buffer (Lopez-Paz and Ranzato, 2017), and mean-of-feature (Rebuffi et al., 2016), employ fixed strategies for sampling from memory. More sophisticated techniques focus on optimization, such as GSS (Aljundi et al., 2019), which maximizes sample diversity, and GEM (Lopez-Paz and Ranzato, 2017) and A-GEM (Chaudhry et al., 2018), which create individual constraints during training. *Architecture-based* methods, which focus on task-specific parameters, have gained renewed attention with the rise of parameter-efficient fine-tuning (PEFT). Early approaches, like progressive networks (Rusu et al., 2016), created a new neural network for each task, connected to previous ones via lateral links. Recent advancements include CoLoR (Wistuba et al., 2024), which trains task-specific LoRA modules, MoRAL (Yang et al., 2024), which combines mixture-of-experts with LoRA for life-long learning, and the approach by Ermis et al. (2024), which extends pre-trained Transformers with adapters.

Continual Learning in Multilinguality and Machine Translation. Garcia et al. (2021) introduce a method for integrating new languages into multilingual translation models by a simple update of the vocabulary, allowing quick adaptation to languages with different scripts while maintaining minimal performance loss on existing language pairs. Coria et al. (2022) analyzes cross-lingual transfer in continual learning for sequence labeling using multilingual BERT, finding that despite some forgetting, forward transfer is retained, with most past language knowledge stored in the word representation encoder rather than the task-specific classifier. Related to this, Winata et al. (2023) examines catastrophic forgetting in a multilingual setting with up

[1] Our code is publicly available at https://github.com/shakibyzn/Multilingual-SLT-CL

to 51 languages. They introduce an effective learning rate scheduling method that reduces forgetting and performs well across various continual learning techniques. Similarly, Huang et al. (2023a) develops a two-stage approach to enhance pre-trained multilingual neural machine translation (MNMT) models, employing contrastive learning for adapting to new data and collaborative distillation for consolidating knowledge. They also introduce a knowledge transfer method that integrates external model insights into existing MNMT models and a dual importance-based model division technique that focuses on parameters crucial for incremental tasks, thus improving performance on new languages while preserving the quality of existing translations (Huang et al., 2023b; Liu et al., 2023). Two studies diverge from traditional continual learning between language pairs by focusing on multi-hop continual learning or evaluating continual learning under language shift, which simulates sequential learning of a single task across a stream of input from different languages. M'hamdi et al. (2023) analyzes various approaches for continually fine-tuning a pre-trained multilingual BERT model to adapt to emerging data from different languages. Recently, Gogoulou et al. (2023) examines the advantages and disadvantages of updating a language model when new data comes from new languages.

Sign Language Translation. Camgöz et al. (2018) is the first to approach sign language translation as a text generation task using deep learning, where input features coming from raw images are feed into a recurrent neural network. Their best model adds an intermediate gloss layer between the video features and the output text which adds gloss supervision to the final loss. With the creation of large datasets without gloss annotations, gloss-free models are getting close to the state of the art (Li et al., 2020b; Zhao et al., 2022; Yin et al., 2023; Lin et al., 2023). Zhou et al. (2023) leverages masked self-supervised learning with vision-language supervision by pre-training with CLIP (Radford et al., 2021). Chen et al. (2022a,b) also use pre-training for the action recognition and text generation subtasks and achieve the current state-of-the-art for the Phoenix2014T dataset. Hamidullah et al. (2024) pre-trains the textual part and Rust et al. (2024) the visual part to achieve the current state-of-the-art for the How2Sign dataset. Few works consider multilinguality (Yin et al., 2022; Hamidullah et al., 2024; Zhang et al., 2024). Recently, Zhang et al. (2024) extend prior SLT pretraining efforts by scaling data, model size, and translation directions, leveraging noisy multilingual SLT data, parallel text corpora, and augmented video captions to enhance cross-lingual and cross-modal transfer for open-domain SLT. Yin et al. (2022) present a transformer-based multilingual system with a dynamic routing mechanism that controls the ratio among languages. We train this system for our experiments as explained in Section 4.4.

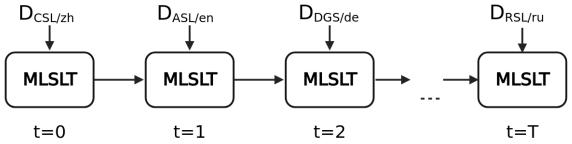


Figure 1: The multilingual sign language translation model (MLSLT) is incrementally fine-tuned with data from T different language pairs.

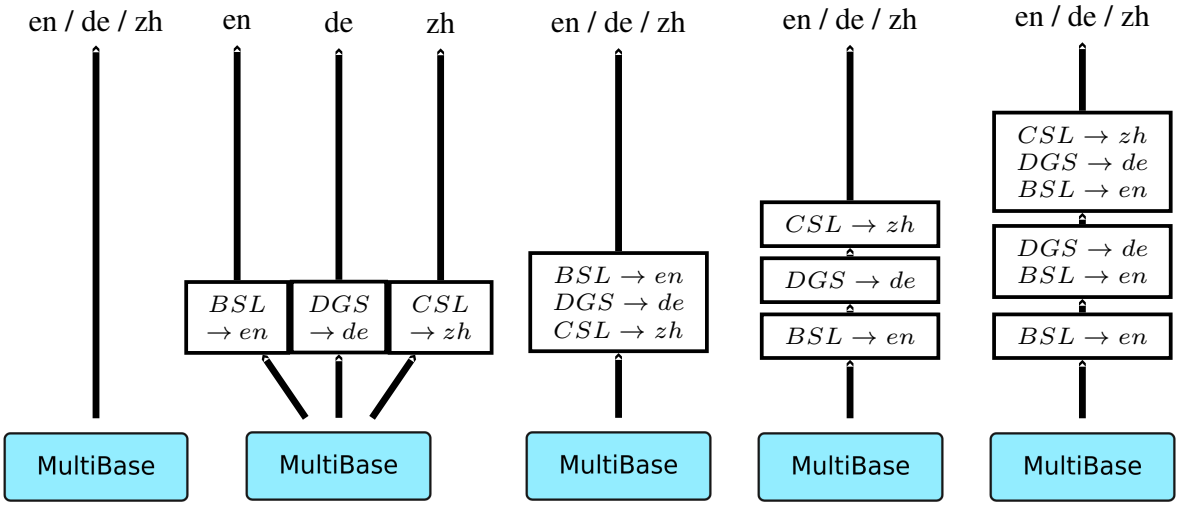
Video Feature Extraction. The most common method to represent an original sign language video is frame-level feature extraction using 2D CNNs or 3D CNNs. Camgöz et al. (Camgöz et al., 2018, 2020) use 2D CNNs to extract features from frames. Other commonly used features are those coming from inflated 3D Convnets (I3D), developed for action recognition (Carreira and Zisserman, 2017). Similarly, Chen et al. (2022a) uses S3D (Xie et al., 2018) features for transfer learning in the SLT domain. Recent studies highlight important findings in the transferability of sign language features across different languages using transfer learning techniques. For example, Kındıroglu et al. (2024) demonstrate that transfer learning with domain-specific attention and normalization techniques significantly improves performance when transferring knowledge between isolated sign language datasets. Töngi (2021) find substantial accuracy improvements when applying transfer learning from ASL to DGS using an inflated 3D deep convolutional neural network.

3 Continual Learning in Multilingual Sign Language Translation

In this section, we formally define continual learning for multilingual SLT and introduce the models that will be evaluated in subsequent sections.

3.1 Problem Definition

Continual Learning aims to tackle the ongoing challenges posed by sequentially arriving tasks. Each



(a) MultiBase (b) Bilingual FT (c) Multilingual FT (d) Naive FT (e) Incremental FT

Figure 2: Schematic representation of the non-continual fine-tuning architectures discussed in Section 3.2. Notice that there is only one language configuration for (a), (b) and (c), but 6 different language sequences for (d) and (e). All fine-tunings are performed using the SP-10 dataset. Notice further that (b) results in three distinct systems, each fine-tuned on only one of the three selected SP-10 languages.

task (language pair in our case) $T_t = \{\mathbf{x}_t, \mathbf{y}_t\}$, $t = 1, \ldots, T$ includes specific input–output pairs and a dataset size n_t. The goal is for the model to adapt to each new task T_t as it arrives, while also retaining its performance across all previously learned tasks. Notice that when a new task arrives, the data from the previous tasks might not be available to the system anymore. We utilize the setup from prior multilingual studies (Gogoulou et al., 2023; M'hamdi et al., 2023) and explore **continual fine-tuning (CFT)** of sign language translation models in which a pre-trained multilingual sign language translation model is incrementally fine-tuned using a sequential data stream $D = D_1, D_2, \ldots, D_n$. In this scenario depicted in Figure 1, each D_i contains data from a distinct language pair, such as $CSL \rightarrow zh$ for Chinese, $DGS \rightarrow de$ for German, etc. We conduct a series of experiments by varying the sequence in which the language pairs are introduced. Specifically, we experiment with datasets for three language pairs: Chinese $CSL \rightarrow zh$, German $DGS \rightarrow de$, and English from the United Kingdom $BSL \rightarrow en$, resulting in six different language pair sequences.

3.2 (Continual) Fine-tuning Methods and Architectures

As done in previous works (M'hamdi et al., 2023; Winata et al., 2023) and before implementing truly continual learning strategies, we define simple reference models based on fine-tuning a base multilingual model described later in Section 4.4. This base model is the first baseline where no adaptation to the new datasets is performed:

Multilingual baseline (MultiBase). MultiBase refers to the inference results of the pre-trained multilingual model on the downstream SP-10 dataset.

The next two methods below use all the new data available for the desired language pair with or without combining it with the other pairs but without any notion of sequentiality:

Bilingual fine-tuning (Bilingual FT). This method fine-tunes the base model with the new bilingual SLT datasets independently for each language pair, resulting in three distinct bilingual models corresponding to the three language pairs.

Multilingual fine-tuning (Multilingual FT). This approach involves fine-tuning the base model using the new data from the three language pairs combined.

The next two methods below fine-tune the base model sequentially without using a specific continual learning approach:

Naive fine-tuning (Naive FT). This strategy involves sequentially fine-tuning the base model on

each new task or language pair in the given sequence. At each step, the model is updated with the data from the current task while entirely disregarding previously learned tasks. This approach does not incorporate any mechanisms to preserve knowledge from earlier tasks, making it prone to catastrophic forgetting. Note again that for three language pairs, six sequences are possible and have been explored.

Incremental joint fine-tuning (Incremental FT). This approach involves fine-tuning the model by progressively incorporating the dataset for each previously learned language pair as a new language pair is introduced. Note that for three language pairs, six block sequences are possible and have been explored.

Figure 2 summarizes the multilingual baseline and four fine-tuning architectures in a schematic way. Finally, the last three methods in our study are representative of three different **continual learning** approaches:

Replay: Experience Replay (ER). ER (Chaudhry et al., 2019) allocates a small, equal memory budget for storing examples from previously encountered languages. These stored examples are revisited and incorporated into training while learning new tasks. This approach helps the model retain knowledge from earlier tasks by ensuring that previously learned information is reinforced during training on new data. In this approach, the loss is computed by jointly optimizing performance on the current task and retained examples from previous tasks. This ensures the model learns new information while maintaining knowledge from earlier tasks.

Regularization: Elastic Weight Consolidation (EWC). EWC (Kirkpatrick et al., 2016b) introduces a regularization term that penalizes significant changes in parameters identified as important for previous tasks. By leveraging information about the parameters' relevance, estimated using the Fisher Information Matrix, this method helps the model retain knowledge from earlier tasks.

Architecture: MAD-X adapters (Adapters) This model adds language-specific MAD-X adapters (Pfeiffer et al., 2020) in each Transformer encoder layer. During training, we fine-tune both the adapter modules and the pre-trained multilingual model. For inference, we use the pre-trained MultiBase model along with the adapter modules fine-tuned for each language pair, applied in sequence based on the order of the languages.

4 Experimental Setting

In this section, we describe the data and preprocessing we use for training and testing the above approaches, the evaluation metrics and methodology, and describe the base model on top of which all the experiments are performed.

4.1 Datasets

We train a gloss-free base multilingual SLT model (MultiBase) using three commonly used datasets from Chinese, German, and American sign languages.

How2Sign. A comprehensive American Sign Language (ASL) dataset designed for multimodal tasks (Duarte et al., 2021). It offers high-quality RGB and depth video recordings of signers in various scenarios, with detailed transcriptions and ASL gloss annotations.

CSL-Daily. A Chinese Sign Language (CSL) dataset that focuses on daily communication contexts (Zhou et al., 2021). It provides high-definition videos of native signers along with glosses and Chinese text translations, ideal for continuous sign language recognition and studying natural, conversational sign language.

Phoenix2014T. A large-scale collection of German Sign Language (DGS) videos featuring interpreters translating weather forecasts (Forster et al., 2014). It includes gloss annotations and spoken German translations, making it valuable for sign language recognition and translation research.

For our downstream experiments —fine-tuning and continual learning approaches—, we choose SP-10, one of the few freely available multilingual datasets for sign language translation.

Spreadthesign-Ten (SP-10). A multilingual dataset containing videos and the corresponding spoken translations in ten language pairs without gloss information (Yin et al., 2022): $CSL \rightarrow zh$, $UKL \rightarrow uk$, $RSL \rightarrow ru$, $BQN \rightarrow bg$, $ICL \rightarrow is$, $DGS \rightarrow de$, $ISE \rightarrow it$, $SWL \rightarrow sv$, $LLS \rightarrow lt$, and $BSL \rightarrow en$, collected from Spreadthesign (Hilzensauer and Krammer, 2015). We only consider $BSL \rightarrow en$, $CSL \rightarrow zh$, and $DGS \rightarrow de$ for our experiments.

Dataset	Stage	Lang pairs	Train	Dev	Test
How2Sign	Pre-training	$ASL \rightarrow en$	31047	1739	2343
CSL-Daily	Pre-training	$CSL \rightarrow zh$	18400	1077	1176
Phoenix14T	Pre-training	$DGS \rightarrow de$	7096	519	642
SP-10	Fine-tuning	$BSL \rightarrow en$	830	142	214
	Fine-tuning	$CSL \rightarrow zh$	830	142	211
	Fine-tuning	$DGS \rightarrow de$	830	142	185

Table 1: Languages and the number of video-text segment pairs in the datasets used for the experiments.

Notice that while our multilingual base model is trained with ASL (American Sign Language), we use BSL (British Sign Language) for fine-tuning. While American and British English are dialects of the same language (English), American and British Sign Language are genetically unrelated languages (Jachova et al., 2008). The language pairs have been chosen to consider both few-shot (Chinese and German) and zero-shot (British sign language) fine-tunings. The statistics of the datasets introduced above are summarized in Table 1.

4.2 Data Processing

Video. We compare two feature extractors as frozen encoders to extract visual features. We use the I3D backbone pre-trained on the WLASL dataset for word-Level ASL recognition (Li et al., 2020a) and the S3D model pre-trained on both WLASL and the Kinetics-400 dataset for action recognition (Kay et al., 2017), following the approach outlined by (Chen et al., 2022a).

We resize the resolution of all original video frames to 224×224 pixels. For the I3D method, we use only the RGB visual features, which are extracted using a sliding window of eight frames with a stride of two. The I3D model is initialized with the default WLASL2000 weights. For the S3D method, we utilize only the first four blocks of the S3D model. Each video is processed through the encoder to extract features, and the output from the final S3D block is spatially pooled to a dimension of $\frac{F}{4} \times 832$, where F is the total number of frames.

Text. We use the same tokenisation model as in MultiBPEmb (Heinzerling and Strube, 2018). MultiBEPemb uses SentencePiece (Kudo and Richardson, 2018) on Wikipedia texts to learn the BPE tokenization model. Our joint vocabulary for the three languages consists of 19,056 subword units.

4.3 Evaluation Metrics

For each language permutation, we sequentially train on each dataset, evaluating the resulting systems across all languages after training on each dataset. The model performance is assessed following (Müller et al., 2022; Müller et al., 2023) by using three machine translation evaluation metrics: chrF (Popović, 2015), BLEU (Papineni et al., 2002), and BLEURT (Sellam et al., 2020). We use sacreBLEU (Post, 2018) for BLEU[2] and chrF[3] for evaluation. Additionally, the Python library for BLEURT[4] is utilized.

In order to assess the effect of catastrophic forgetting in the continual learning framework, we use backward transfer (BWT) and forward transfer (FWT) metrics adapted from Lopez-Paz and Ranzato (2017). BWT measures how the learning of new tasks affects the performance on previously learned tasks. FWT measures how the knowledge from previous tasks influences the learning of new tasks. We formally define BWT and FWT as:

$$\text{BWT} = \frac{1}{N-1} \sum_{i=1}^{N-1} R_{N,i} - R_{i,i}$$

$$\text{FWT} = \frac{1}{N-1} \sum_{i=2}^{N} R_{i-1,i} - R_{0,i},$$

where N is the number of tasks, $R_{i,j}$ is the model's performance (using BLEURT) on task j after training on task i and $R_{0,i}$ is the performance of the model on task i before the start of training. In both cases, the higher the better.

4.4 Base Multilingual SLT Model, MultiBase

For our base model we train an end-to-end gloss-free multilingual sign language translation system presented in Yin et al. (2022). The system is based on a Transformer model (Vaswani et al., 2017) with three encoder and decoder layers. The MultiBase system differs from the standard transformer by introducing a dynamic routing mechanism that controls the ratio of data and the degree of parameter sharing between different languages. This system has been proven to be a strong baseline; in fact, it outperforms a multilingual system built by mixing all the data together and appending language

[2] `BLEU|nrefs:1|case:mixed|eff:yes|tok:13a|smooth:exp|version:1.4.22`

[3] `chrF|nrefs:1|case:mixed|eff:yes|nc:6|space:no|version:1.4.22`

[4] BLEURT using checkpoint BLEURT-20.

tags (Johnson et al., 2017), and in some cases, it surpasses the equivalent bilingual systems using the state-of-the-art approach from Camgöz et al. (2020).

We adapt the published code[5] and use the (continual) fine-tuning strategies for the SLT task presented above.

Experimental setup. We fix the learning rate to $5e{-}4$ for all experiments. We use a batch size of 64, $\beta_1 = 0.9$, $\beta_2 = 0.998$, and label smoothing of 0.4. For the multilingual pre-training stage, we report the results for three seeds of 42, 43, and 44. For the continual learning methods, unless stated otherwise, we set $\lambda = 100000$ for EWC and a memory size [6] of 200 training samples for the ER method. In all experiments, we run each algorithm for 100 epochs. However, if the BLEU score on the validation set fails to improve for 8 consecutive evaluations, the learning rate will be reduced by a factor of 0.5. This adjustment continues until the learning rate reaches a minimum of 1×10^{-7}. We also fix a seed of 44 for the random initialization of Numpy, random, and torch over all experiments in (continual) fine-tuning methods. All experiments are conducted on the same computing infrastructure, utilizing PyTorch version 1.4.0 and a single NVIDIA Quadro RTX 6000 GPU with CUDA version 10.1.

5 Results and Analysis

5.1 Base Multilingual SLT Model

Table 2 shows the automatic evaluation of the base multilingual model, MultiBase. To better analyze the efficacy of I3D and S3D for spatial embedding in SLT, we evaluate our pre-trained multilingual model on the test sets of Phoenix2014T, How2Sign, and CSL-Daily datasets. We observe that spatial embeddings from the S3D method yield better results. Based on this, we use the S3D method as our visual feature extractor for the rest of the experiments.

The performance of MultiBase is lower than that of bilingual transformer-based specialized systems, probably due to the fact that the training corpus is unbalanced and the intersection of the vocabulary between the three languages is small. A basic bilingual transformer achieves a BLEU score of 8 for How2Sign (Tarrés et al., 2023), 13 for CSL-Daily (Zhou et al., 2021), 10 for Phoenix2014T (Camgöz et al., 2018) and 1-5 depending on the language pair for SP-10 (Yin et al., 2022), with an average score of 4.35 across all languages for GASLT (Yin et al., 2023). In the following section, we study the effects of continual learning on SP-10.

Dataset	Feat.	chrF	BLEU	BLEURT
How2Sign	I3D	18.0±0.5	1.3±0.0	0.31±0.01
	S3D	18.3±0.8	1.5±0.1	0.31±0.01
CSL-Daily	I3D	4.0±0.3	2.6±0.4	0.22±0.00
	S3D	4.8±0.7	3.5±0.8	0.22±0.02
Phoenix14T	I3D	31.5±0.8	10.4±0.7	0.39±0.01
	S3D	34.3±2.3	12.2±1.3	0.42±0.03

Table 2: Performance of the base multilingual pre-trained model, MultiBase, according to evaluation metrics on the test set, including a comparison of the visual feature extraction methods I3D and S3D (Feat.). We report the mean and standard deviation over three seeds.

5.2 (Continual) Fine-tuning

We evaluate the learning techniques introduced in Section 3.2 across all languages at the end of the continual training pipeline. Table 3 shows a summary of translation quality and the transfer capability of the model (when possible) as the average —mean and standard deviation— across the six language pair combinations.

The non-continual learning techniques achieve the best average translation quality results according to all metrics. The Multilingual FT is the best option, indicating its capability to capitalize on combined dataset information. However, for highly multilingual settings, one should increase the capacity of the model to represent all of the languages properly and this might be too expensive computationally. Both Incremental FT and Bilingual FT perform similarly, with Incremental FT showing a slight advantage. Incremental FT provides a practical solution, particularly in situations where the dataset is continuously growing. This approach is resource-efficient, as it eliminates the need to manage and train multiple models separately. Instead, it allows for the gradual updating of a single model over time. The Bilingual FT approach involves having multiple bilingual sign language translation models which, can become inefficient as the number of language pairs increases.

Notice that the transfer capabilities for these basic techniques are slightly better (higher BWT and FWT) than for the continual learning techniques

[5] https://mlslt.github.io/

[6] For an ablation study on memory size, refer to Table 5.

Method	chrF (↑)	BLEU (↑)	BLEURT (↑)	BWT (↑)	FWT (↑)
MultiBase	9.5 ± 7.0	0.0 ± 0.0	0.13 ± 0.11	–	–
Bilingual FT	10.6 ± 3.8	4.4 ± 2.5	0.24 ± 0.07	–	–
Multilingual FT	12.5 ± 4.8	5.2 ± 2.9	0.25 ± 0.07	–	–
Naive FT	5.2 ± 2.2	1.7 ± 0.9	0.15 ± 0.01	-0.18 ± 0.01	-0.02 ± 0.03
Incremental FT	11.4 ± 0.4	4.5 ± 0.6	0.23 ± 0.01	-0.01 ± 0.01	-0.03 ± 0.04
ER	9.4 ± 0.8	2.7 ± 0.4	0.20 ± 0.01	-0.04 ± 0.01	-0.03 ± 0.04
EWC	5.0 ± 1.6	1.2 ± 0.5	0.14 ± 0.02	-0.12 ± 0.04	-0.06 ± 0.05
Adapters	9.5 ± 0.1	0.0 ± 0.0	0.14 ± 0.00	-0.01 ± 0.01	0.00 ± 0.00

Table 3: Translation quality (chrF, BLEU, BLEURT) and transfer capability of the methods (BWT, FWT) on SP-10 as the average —mean and standard deviation— across the six language pair combinations. The upward arrow (↑) indicates that a higher value is better for the corresponding metric.

(ER, EWC and Adapters) and the average final translation quality is better. However, to a greater or lesser extent, all of them suffer from catastrophic forgetting (BWT<0). This is especially true for Naive FT as expected, but also for EWC. These two methods are even worse than MultiBase with respect to translation quality. The continual learning approaches show matching or marginally improved translation quality compared to MultiBase. Confirming the trends observed in the previous machine learning literature (van de Ven and Tolias, 2019), ER is the best option for continual learning. We show translation quality per language in Appendix B.

Order	chrF	BLEU	BLEURT
BSL-DGS-CSL	9.7 ± 3.4	3.1 ± 2.9	0.20 ± 0.06
BSL-CSL-DGS	9.5 ± 4.3	2.8 ± 0.9	0.20 ± 0.05
DGS-BSL-CSL	10.3 ± 2.5	2.5 ± 4.4	0.20 ± 0.05
DGS-CSL-BSL	8.0 ± 4.4	2.8 ± 2.6	0.21 ± 0.08
CSL-DGS-BSL	9.6 ± 4.8	2.1 ± 0.7	0.20 ± 0.03
CSL-BSL-DGS	9.5 ± 5.0	3.1 ± 2.8	0.22 ± 0.06

Table 4: Impact of language order on the final performance for the ER method. Results are reported as the mean and standard deviation over three languages: English, Chinese, and German.

5.3 The Role of Language Pair Order

First, we evaluate the impact of ER —our best continual learning approach— when applied to different sequences of language pairs. This helps us determine whether the order in which languages are presented affects the quality and efficiency of the replay mechanism. Table 4 displays the final performance averaged across the three language pairs. While the BLEURT scores are quite similar, a more nuanced view emerges when considering both BLEU and BLEURT scores together. In this context, the CSL-BSL-DGS order demonstrates a superior overall final score.

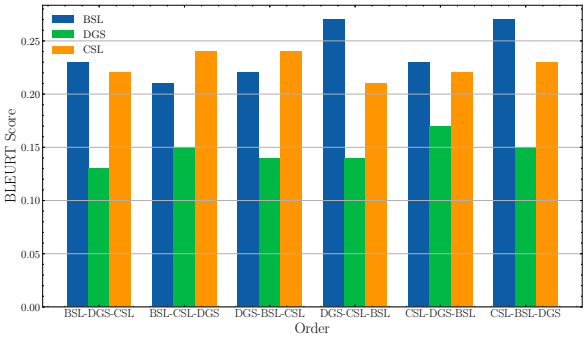


Figure 3: Language-specific performance comparison based on each order for the ER method.

Additionally, we analyze language-specific scores to understand how experience replay affects performance on individual languages with different language orders. This level of granularity allows us to better analyze balanced performance and order sensitivity. Figure 3 shows that our model does not perform as well on German as compared to Chinese or English. This is due to the fact that the MultiBase model has been trained with few German data from a very specific domain (weather forecasts) and therefore exhibits a rather limited capability to generalize to other domains. From order sensitivity, we can expect that a language pair that is presented early in the order might benefit differently from experience replay compared to a language pair presented later. This is almost true, as Figure 3 shows that a language pair tends to perform better when it does not appear in the first place.

ER is not the only method that is sensitive to the input order of the datasets. Figure 4 shows the average BLEU scores across the three language pairs for different methods—ER, Naive FT, and

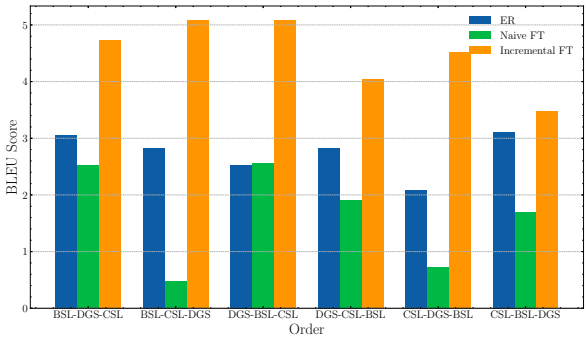


Figure 4: Performance comparison as measured by BLEU between different methods on each language order.

Incremental FT—across the six different orderings. As highlighted earlier in Table 3, Incremental FT yields the highest BLEU scores, both overall and for each specific language pair order. Experience Replay (ER) ranked second, while Naive FT produced nearly identical scores for each language pair and displayed similar trends on the graph. We also present the translation quality results for Multilingual FT, Incremental FT, and ER along with the reference translations in Appendix C.

6 Summary and Conclusion

We presented the first study on continual learning for multilingual sign language translation. Our multilingual setting involves three language pairs: $BSL \rightarrow en$, $CSL \rightarrow zh$, and $DGS \rightarrow de$. We first pre-trained a multilingual SLT model using two different video feature extraction methods, I3D and S3D, with S3D achieving better translation quality under the same conditions. We then compare a range of standard fine-tuning and continual learning methods. Multilingual and Incremental fine-tuning are the methods that offer the best average translation quality and Incremental fine-tuning is also able to minimize the effect of catastrophic forgetting. Experience replay is the best within the continual learning methods irrespective of the size of the memory buffer.

The language pair order plays a role in our experiments. We observe that languages seen later during training are improved compared to an early appearance. We also observe that the adaptation to $DGS \rightarrow de$ is the most difficult one as the base multilingual model does not cover the language pair properly (few and narrow-domain data). Curiously, translation quality into English is good even if the multilingual base model is trained with ASL data and the continual fine-tuning is done with a BSL corpus. This indicates that, given the low overall translation quality, the decoder performance (text generation into English in this case) is the most important factor. Stronger base multilingual models are needed for sign language to achieve comparable results to their textual machine translation counterparts.

Limitations

Our work has several limitations. Firstly, the number of language pairs we investigated is limited, and a more diverse set of language pairs should be explored. To achieve this, we need a multilingual sign language translation model capable of understanding a broader range of sign languages. Recently, Tanzer and Zhang (2024) introduced YouTube-SL-25, a large-scale, open-domain multilingual sign language parallel corpus that includes at least 55 sign languages. This resource could facilitate the study of a wider variety of language pair orders, including those from high-resource to low-resource languages. Secondly, while our multilingual base model provides a strong foundation, achieving a BLEU score of 6.3 on English and 7.4 on Chinese, which surpasses the results reported in the base paper by Yin et al. (2022), there is still significant room for improvement. Recent works, such as Hamidullah et al. (2024), have introduced advancements in multilingual sign language translation models. Such improvements could lead to more robust comparisons between continual learning methods and yield clearer, more interpretable results, advancing the field of multilingual sign language translation.

References

Samuel Albanie, Gül Varol, Liliane Momeni, Hannah Bull, Triantafyllos Afouras, Himel Chowdhury, Neil Fox, Bencie Woll, Rob Cooper, Andrew McParland, and Andrew Zisserman. 2021. BOBSL: BBC-Oxford British Sign Language Dataset. *CoRR*, abs/2111.03635.

Rahaf Aljundi, Francesca Babiloni, Mohamed Elhoseiny, Marcus Rohrbach, and Tinne Tuytelaars. 2018. Memory aware synapses: Learning what (not) to forget. In *Computer Vision – ECCV 2018: 15th European Conference, Munich, Germany, September 8–14, 2018, Proceedings, Part III*, page 144–161, Berlin, Heidelberg. Springer-Verlag.

Rahaf Aljundi, Min Lin, Baptiste Goujaud, and Yoshua Bengio. 2019. Gradient based sample selection for

online continual learning. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Necati Cihan Camgöz, Simon Hadfield, Oscar Koller, Hermann Ney, and R. Bowden. 2018. Neural sign language translation. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7784–7793.

Necati Cihan Camgöz, Oscar Koller, Simon Hadfield, and R. Bowden. 2020. Sign language transformers: Joint end-to-end sign language recognition and translation. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10020–10030.

João Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4724–4733.

Francisco Manuel Castro, Manuel J. Marín-Jiménez, Nicolás Guil Mata, Cordelia Schmid, and Alahari Karteek. 2018. End-to-end incremental learning. In *European Conference on Computer Vision*.

Arslan Chaudhry, Marc'Aurelio Ranzato, Marcus Rohrbach, and Mohamed Elhoseiny. 2018. Efficient Lifelong Learning with A-GEM. *ArXiv*, abs/1812.00420.

Arslan Chaudhry, Marcus Rohrbach, Mohamed Elhoseiny, Thalaiyasingam Ajanthan, Puneet Kumar Dokania, Philip H. S. Torr, and Marc'Aurelio Ranzato. 2019. Continual learning with tiny episodic memories. *CoRR*, abs/1902.10486.

Yutong Chen, Fangyun Wei, Xiao Sun, Zhirong Wu, and Stephen Lin. 2022a. A simple multi-modality transfer learning baseline for sign language translation. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5110–5120.

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

Zhiyuan Chen and Bing Liu. 2017. *Lifelong Machine Learning*. Online access: Morgan & Claypool Synthesis Collection Seven. Morgan & Claypool Publishers.

Juan Manuel Coria, Mathilde Veron, Sahar Ghannay, Guillaume Bernard, Hervé Bredin, Olivier Galibert, and Sophie Rosset. 2022. Analyzing BERT cross-lingual transfer capabilities in continual sequence labeling. In *Proceedings of the First Workshop on Performance and Interpretability Evaluations of Multimodal, Multipurpose, Massive-Scale Models*, pages 15–25. International Conference on Computational Linguistics.

Prithviraj Dhar, Rajat Vikram Singh, Kuan-Chuan Peng, Ziyan Wu, and Rama Chellappa. 2018. Learning without memorizing. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5133–5141.

Amanda Duarte, Shruti Palaskar, Lucas Ventura, Deepti Ghadiyaram, Kenneth DeHaan, Florian Metze, Jordi Torres, and Xavier Giró-i Nieto. 2021. How2Sign: A Large-Scale Multimodal Dataset for Continuous American Sign Language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2735–2744.

Beyza Ermis, Giovanni Zappella, Martin Wistuba, Aditya Rawal, and Cédric Archambeau. 2024. Memory efficient continual learning with transformers. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Jens Forster, Christoph Schmidt, Oscar Koller, Martin Bellgardt, and Hermann Ney. 2014. Extensions of the sign language recognition and translation corpus RWTH-PHOENIX-weather. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1911–1916.

Xavier Garcia, Noah Constant, Ankur Parikh, and Orhan Firat. 2021. Towards continual learning for multilingual machine translation via vocabulary substitution. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1184–1192. Association for Computational Linguistics.

Evangelia Gogoulou, Timothée Lesort, Magnus Boman, and Joakim Nivre. 2023. Continual learning under language shift. *ArXiv*, abs/2311.01200.

Ian J. Goodfellow, Mehdi Mirza, Xia Da, Aaron C. Courville, and Yoshua Bengio. 2013. An empirical investigation of catastrophic forgeting in gradient-based neural networks. *CoRR*, abs/1312.6211.

Yasser Hamidullah, Josef van Genabith, and Cristina España-Bonet. 2024. Sign language translation with sentence embedding supervision. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 425–434, Bangkok, Thailand. Association for Computational Linguistics.

Benjamin Heinzerling and Michael Strube. 2018. BPEmb: Tokenization-free pre-trained subword embeddings in 275 languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Marlene Hilzensauer and Klaudia Krammer. 2015. A multilingual dictionary for sign languages: "spreadthesign". In *ICERI2015 Proceedings*, pages 7826–7834.

Saihui Hou, Xinyu Pan, Chen Change Loy, Zilei Wang, and Dahua Lin. 2019. Learning a unified classifier incrementally via rebalancing. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 831–839.

Kaiyu Huang, Peng Li, Junpeng Liu, Maosong Sun, and Yang Liu. 2023a. Learn and consolidate: Continual adaptation for zero-shot and multilingual neural machine translation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13938–13951. Association for Computational Linguistics.

Kaiyu Huang, Peng Li, Jin Ma, Ting Yao, and Yang Liu. 2023b. Knowledge transfer in incremental learning for multilingual neural machine translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15286–15304. Association for Computational Linguistics.

Zora Jachova, Olivera Kovacheva, and Aleksandra Karovska. 2008. Differences between american sign language (asl) and british sign language (bsl). *Journal of Special Education and Rehabilitation*, 9(1-2):41–54.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Will Kay, João Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Apostol Natsev, Mustafa Suleyman, and Andrew Zisserman. 2017. The kinetics human action video dataset. *ArXiv*, abs/1705.06950.

Zixuan Ke and Bin Liu. 2022. Continual learning of natural language processing tasks: A survey. *ArXiv*, abs/2211.12701.

James Kirkpatrick, Razvan Pascanu, Neil C. Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. 2016a. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114:3521 – 3526.

James Kirkpatrick, Razvan Pascanu, Neil C. Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. 2016b. Overcoming catastrophic forgetting in neural networks. *CoRR*, abs/1612.00796.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Ahmet Alp Kındıroglu, Ozgur Kara, Ogulcan Özdemir, and Lale Akarun. 2024. Transfer learning for cross-dataset isolated sign language recognition in under-resourced datasets. *2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG)*, pages 1–8.

Sang-Woo Lee, Jin-Hwa Kim, Jaehyun Jun, Jung-Woo Ha, and Byoung-Tak Zhang. 2017. Overcoming catastrophic forgetting by incremental moment matching. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 4655–4665, Red Hook, NY, USA. Curran Associates Inc.

Dongxu Li, Cristian Rodriguez, Xin Yu, and Hongdong Li. 2020a. Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. In *The IEEE Winter Conference on Applications of Computer Vision*, pages 1459–1469.

Dongxu Li, Chenchen Xu, Xin Yu, Kaihao Zhang, Ben Swift, Hanna Suominen, and Hongdong Li. 2020b. TSPNet: Hierarchical Feature Learning via Temporal Semantic Pyramid for Sign Language Translation. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA. Curran Associates Inc.

Zhizhong Li and Derek Hoiem. 2016. Learning without forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40:2935–2947.

Zhizhong Li and Derek Hoiem. 2018. Learning without forgetting. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(12):2935–2947.

Kezhou Lin, Xiaohan Wang, Linchao Zhu, Ke Sun, Bang Zhang, and Yi Yang. 2023. 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)*, pages 12904–12916, Toronto, Canada. Association for Computational Linguistics.

Junpeng Liu, Kaiyu Huang, Hao Yu, Jiuyi Li, Jinsong Su, and Degen Huang. 2023. Continual learning for multilingual neural machine translation via dual importance-based model division. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12011–12027. Association for Computational Linguistics.

David Lopez-Paz and Marc' Aurelio Ranzato. 2017. Gradient episodic memory for continual learning. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Meryem M'hamdi, Xiang Ren, and Jonathan May. 2023. Cross-lingual continual learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3908–3943. Association for Computational Linguistics.

Mathias Müller, Malihe Alikhani, Eleftherios Avramidis, Richard Bowden, Annelies Braffort, Necati Cihan Camgöz, Sarah Ebling, Cristina España-Bonet, Anne Göhring, Roman Grundkiewicz, Mert Inan, Zifan Jiang, Oscar Koller, Amit Moryossef, Annette Rios, Dimitar Shterionov, Sandra Sidler-Miserez, Katja Tissi, and Davy Van Landuyt. 2023. Findings of the second WMT shared task on sign language translation (WMT-SLT23). In *Proceedings of the Eighth Conference on Machine Translation*, pages 68–94, Singapore. Association for Computational Linguistics.

Mathias Müller, Sarah Ebling, Eleftherios Avramidis, Alessia Battisti, Michèle Berger, Richard Bowden, Annelies Braffort, Necati Cihan Camgöz, Cristina España-Bonet, Roman Grundkiewicz, Zifan Jiang, Oscar Koller, Amit Moryossef, Regula Perrollaz, Sabine Reinhard, Annette Rios, Dimitar Shterionov, Sandra Sidler-Miserez, Katja Tissi, and Davy Van Landuyt. 2022. Findings of the First WMT Shared Task on Sign Language Translation (WMT-SLT22). In *Proceedings of the Seventh Conference on Machine Translation*, pages 744–772, Abu Dhabi. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020. MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7654–7673.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

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.

Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, G. Sperl, and Christoph H. Lampert. 2016. iCaRL: Incremental Classifier and Representation Learning. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5533–5542.

Phillip Rust, Bowen Shi, Skyler Wang, Necati Cihan Camgoz, and Jean Maillard. 2024. 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)*, pages 8624–8641, Bangkok, Thailand. Association for Computational Linguistics.

Andrei A. Rusu, Neil C. Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, and Raia Hadsell. 2016. Progressive neural networks. *ArXiv*, abs/1606.04671.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892. Association for Computational Linguistics.

Haizhou Shi, Zihao Xu, Hengyi Wang, Weiyi Qin, Wenyuan Wang, Yibin Wang, and Hao Wang. 2024. Continual learning of large language models: A comprehensive survey. *ArXiv*, abs/2404.16789.

Garrett Tanzer and Biao Zhang. 2024. YouTube-SL-25: A Large-Scale, Open-Domain Multilingual Sign Language Parallel Corpus. *ArXiv*, abs/2407.11144.

Laia Tarrés, Gerard I. Gallego, Amanda Duarte, Jordi Torres, and Xavier Giró i Nieto. 2023. Sign language translation from instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR): Workshops*, pages 5625–5635.

Roman Töngi. 2021. Application of transfer learning to sign language recognition using an inflated 3d deep convolutional neural network. *ArXiv*, abs/2103.05111.

Gido van de Ven, Tinne Tuytelaars, and Andreas Tolias. 2022. Three types of incremental learning. *Nature Machine Intelligence*, 4:1–13.

Gido M. van de Ven and Andreas Savas Tolias. 2019. Three scenarios for continual learning. *ArXiv*, abs/1904.07734.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Liyuan Wang, Xingxing Zhang, Hang Su, and Jun Zhu. 2024. A Comprehensive Survey of Continual Learning: Theory, Method and Application. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(8):5362–5383.

Genta Winata, Lingjue Xie, Karthik Radhakrishnan, Shijie Wu, Xisen Jin, Pengxiang Cheng, Mayank Kulkarni, and Daniel Preotiuc-Pietro. 2023. Overcoming catastrophic forgetting in massively multilingual continual learning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 768–777. Association for Computational Linguistics.

Martin Wistuba, Prabhu Teja S, Lukas Balles, and Giovanni Zappella. 2024. Continual learning with low rank adaptation. In *NeurIPS 2023 Workshop on Distribution Shifts: New Frontiers with Foundation Models*.

Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. 2018. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Shu Yang, Muhammad Asif Ali, Cheng-Long Wang, Lijie Hu, and Di Wang. 2024. MoRAL: MoE Augmented LoRA for LLMs' Lifelong Learning. *ArXiv*, abs/2402.11260.

Aoxiong Yin, Zhou Zhao, Weike Jin, Meng Zhang, Xingshan Zeng, and Xiaofei He. 2022. MLSLT: Towards Multilingual Sign Language Translation. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5099–5109.

Aoxiong Yin, Tianyun Zhong, Lilian H. Y. Tang, Weike Jin, Tao Jin, and Zhou Zhao. 2023. Gloss attention for gloss-free sign language translation. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2551–2562.

Biao Zhang, Garrett Tanzer, and Orhan Firat. 2024. Scaling sign language translation. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Jian Zhao, Weizhen Qi, Wengang Zhou, Nan Duan, Ming Zhou, and Houqiang Li. 2022. Conditional sentence generation and cross-modal reranking for sign language translation. *IEEE Transactions on Multimedia*, 24:2662–2672.

Benjia Zhou, Zhigang Chen, Albert Clapés, Jun Wan, Yanyan Liang, Sergio Escalera, Zhen Lei, and Du Zhang. 2023. Gloss-free sign language translation: Improving from visual-language pretraining. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 20871–20881.

Hao Zhou, Wen gang Zhou, Weizhen Qi, Junfu Pu, and Houqiang Li. 2021. Improving sign language translation with monolingual data by sign back-translation. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1316–1325.

A Experience Replay Effectiveness

We conduct additional experiments to better understand the effectiveness of experience replay in multilingual sign language translation focusing on the memory size.

In ER, one can assess how memory capacity influences the model's performance in retaining and applying learned knowledge across languages by adjusting the amount of stored experience. Table 5 presents a comparison of the performance across various memory sizes, ranging from 100 to 800 training samples. The BLEU scores indicate that performance peaks with a memory size of 800; however, the improvement is marginal and not significantly different from a memory size of 200 samples. Additionally, when considering BLEURT scores, there is no difference observed across the different memory sizes.

Method	chrF	BLEU	BLEURT
ER-100	9.1 ± 0.3	2.4 ± 0.5	0.20 ± 0.00
ER-200	9.4 ± 0.8	2.7 ± 0.4	0.20 ± 0.01
ER-400	9.5 ± 0.7	2.7 ± 0.7	0.20 ± 0.01
ER-800	9.4 ± 0.4	2.8 ± 0.4	0.20 ± 0.01

Table 5: Experiments with different memory sizes (number of past examples) per language in experience replay.

B Performance per Language Pair

Table 6 presents the average final performance of the model, as measured by chrF, BLEU, and BLEURT, across the six language pair combinations for each language. When considering all metrics, it is evident that the fine-tuning methods perform better in English and Chinese than in German, as shown in Figure 3. Among the continual fine-tuning methods, ER demonstrates the best overall performance for each language. When considering all methods, Multilingual FT clearly delivers the best results, achieving an average chrF of 12.5, BLEU of 5.2, and BLEURT of 0.25, with the final scores penalized by the performance on the German data.

C Translation Quality Comparison

In this section, we present a comparative analysis of the translation quality achieved using three methods, which have demonstrated the best performance: Multilingual FT, Incremental FT, and ER. For both Incremental FT and ER, the translation quality is shown across six different language permutations. As discussed previously in Appendix B, the translation quality in English and Chinese is superior to that in German, which is also reflected in Table 7. We believe this is because the pre-trained model has encountered more samples in English and Chinese than in German. Another noteworthy observation is the difference in translation quality across permutations, which further emphasizes the importance of stronger continual learning approaches to achieve balanced translation quality across all sign language permutations.

Method	de			en			zh			avg		
	chrF	**BLEU**	**BLEURT**	**chrF**	**BLEU**	**BLEURT**	**chrF**	**BLEU**	**BLEURT**	**chrF**	**BLEU**	**BLEURT**
MultiBase	14.6	0.0	0.06	12.5	0.0	0.26	1.5	0.0	0.09	9.5 ± 7.0	0.0 ± 0.0	0.13 ± 0.11
Bilingual FT	12.0	1.5	0.16	13.6	5.5	0.27	6.3	6.2	0.28	10.6 ± 3.8	4.4 ± 2.5	0.24 ± 0.07
Multilingual FT	14.1	1.9	0.17	16.3	6.3	0.30	7.1	7.4	0.28	12.5 ± 4.8	5.2 ± 2.9	0.25 ± 0.07
Naive FT	6.2 ± 5.3	0.6 ± 1.0	0.11 ± 0.04	6.9 ± 6.5	1.8 ± 2.8	0.19 ± 0.07	2.6 ± 3.7	2.5 ± 3.9	0.15 ± 0.08	5.2 ± 2.2	1.7 ± 0.9	0.15 ± 0.01
Incremental FT	13.3 ± 0.5	1.8 ± 0.4	0.16 ± 0.01	14.4 ± 0.8	5.1 ± 0.8	0.28 ± 0.01	6.5 ± 0.7	6.6 ± 1.1	0.24 ± 0.01	11.4 ± 0.4	4.5 ± 0.6	0.23 ± 0.01
ER	10.3 ± 2.6	0.6 ± 1.0	0.15 ± 0.01	12.6 ± 1.5	3.0 ± 2.1	0.24 ± 0.03	5.4 ± 1.3	4.6 ± 1.8	0.23 ± 0.01	9.4 ± 0.8	2.7 ± 0.4	0.20 ± 0.01
EWC	6.0 ± 5.0	0.6 ± 1.0	0.10 ± 0.05	6.7 ± 5.6	1.2 ± 1.9	0.17 ± 0.07	2.2 ± 3.2	1.8 ± 2.8	0.14 ± 0.07	5.0 ± 1.6	1.2 ± 0.5	0.14 ± 0.02
Adapters	14.4 ± 0.3	0.0 ± 0.0	0.06 ± 0.01	12.5 ± 0.0	0.0 ± 0.0	0.26 ± 0.00	1.5 ± 0.0	0.0 ± 0.0	0.09 ± 0.00	9.5 ± 0.1	0.0 ± 0.0	0.14 ± 0.00

Table 6: The average final performance across the six language pair combinations for each language and method. We report the mean and standard deviation for the last five methods over six language pair combinations.

Method	Translation
Reference(en)	can i offer you anything to eat?
Multilingual FT	can i offer you anything to drink?
Incremental FT	
BSL-CSL-DGS	can i offer you anything to drink?
BSL-DGS-CSL	can i offer you anything to drink?
CSL-BSL-DGS	can i offer you anything to drink?
CSL-DGS-BSL	can i offer you anything to drink?
DGS-BSL-CSL	can i offer you anything to drink?
DGS-CSL-BSL	can i offer you anything to drink?
ER	
BSL-CSL-DGS	do you have been there.
BSL-DGS-CSL	c<unk>, please.
CSL-BSL-DGS	come on!
CSL-DGS-BSL	can i offer you anything to drink?
DGS-BSL-CSL	do you have any<unk>?
DGS-CSL-BSL	can i offer you anything to drink?
Reference(de)	wo ist das nächste postamt?
Multilingual FT	wo ist die post?
Incremental FT	
BSL-CSL-DGS	wo ist die g<unk><unk>?
BSL-DGS-CSL	wo ist die post?
CSL-BSL-DGS	wo ist die g<unk><unk>?
CSL-DGS-BSL	wie ist die post?
DGS-BSL-CSL	wie<unk> g<unk><unk>?
DGS-CSL-BSL	wo ist die pr<unk>ungen?
ER	
BSL-CSL-DGS	wennst du die s<unk>?
BSL-DGS-CSL	wo ist ein wo ist ein b<unk><unk><unk>?
CSL-BSL-DGS	was möchten sie<unk>?
CSL-DGS-BSL	wo ist der s<unk>?
DGS-BSL-CSL	ich habe einen<unk>men <unk>.
DGS-CSL-BSL	ich habe diesen f<unk><unk>.
Reference(zh)	你想点什么吗
Multilingual FT	你要点菜了吗?
Incremental FT	
BSL-CSL-DGS	你要点菜?
BSL-DGS-CSL	你要哪里想的?
CSL-BSL-DGS	你要点了吗?
CSL-DGS-BSL	你要点菜了吗?
DGS-BSL-CSL	你要点菜了吗?
DGS-CSL-BSL	你想要什么?
ER	
BSL-CSL-DGS	你准备点餐吗?
BSL-DGS-CSL	你要在这哪里?
CSL-BSL-DGS	你最近过得怎么样?
CSL-DGS-BSL	你想理发
DGS-BSL-CSL	你要点菜了吗?
DGS-CSL-BSL	你怎么了?

Table 7: Translation Quality Comparison Among Multilingual FT, Incremental FT, and ER