Paper Review

MLCF

Published as a conference paper at ICLR 2024

SIGN2GPT: LEVERAGING LARGE LANGUAGE MODELS FOR GLOSS-FREE SIGN LANGUAGE TRANSLATION

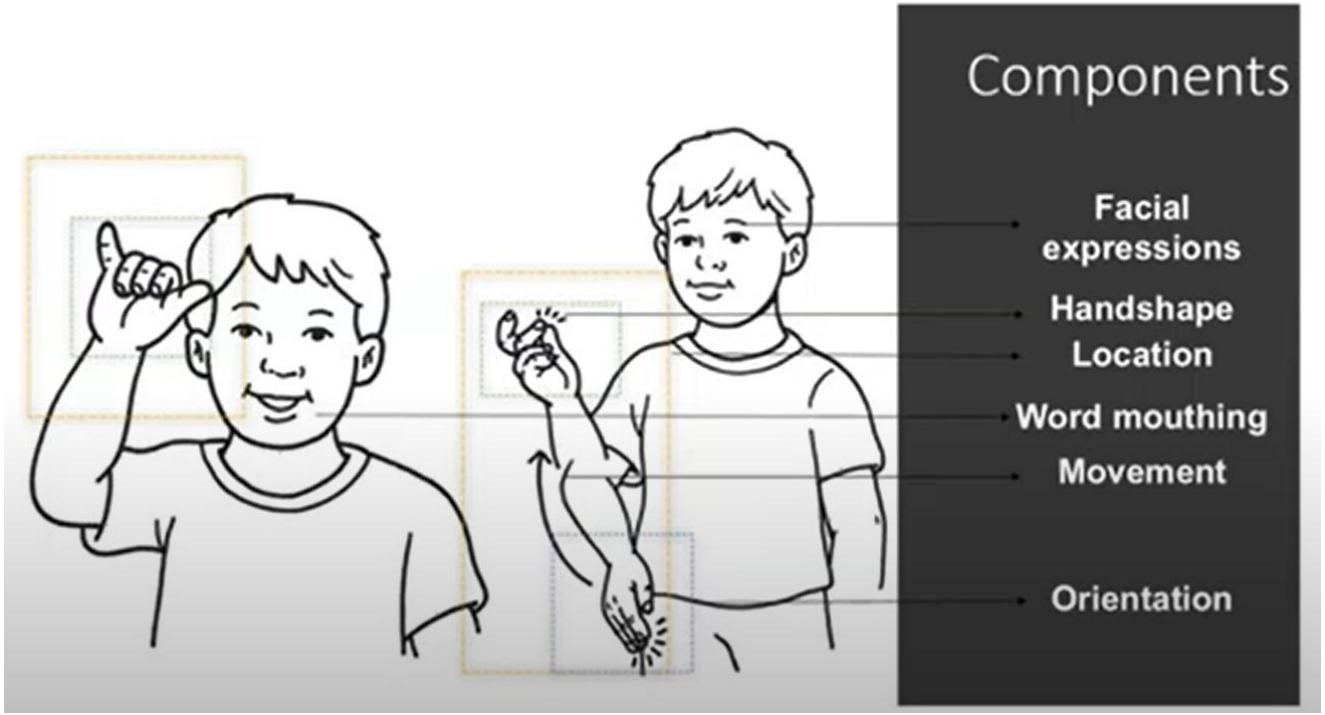
Ryan Wong[1], **Necati Cihan Camgoz**[2], **Richard Bowden**[1]
[1]University of Surrey, [2]Meta Reality Labs
{r.wong, r.bowden}@surrey.ac.uk, neccam@meta.com

Reviewed by Susang Kim

Contents

1.Introduction – Challenges of Sign Language processing

Sign language relies not only on hand movements but also on facial expressions, body posture, and spatial orientation. Capturing and understanding these multiple modalities simultaneously is complex.



Facial expressions convey both linguistic information and emotions. For instance, raising the eyebrows is essential for indicating general questions in most sign languages.

https://youtu.be/E5nKeEvoAK0
https://www.uib.no/en/hf/136874/complex-system-facial-expressions-sign-language

1.Introduction - Sign Language Recognition and Translation & Gloss

Gloss is a transcription system that converts sign language into written/spoken language by breaking signs into their smallest meaningful parts across multiple communication channels (hands, face, body).

Sign2gloss2text (S2G2T), recognizes the sign language video as gloss and translates the gloss into spoken language text.
Sign2text (S2T), which directly generates spoken language text from sign language video end to end.
Sign2(gloss+text) (S2(G+T)), multitasks by outputting glosses and text and can use external glosses as supervision signals.
Gloss2text (G2T), which can reflect the translation performance from gloss sequence to text.

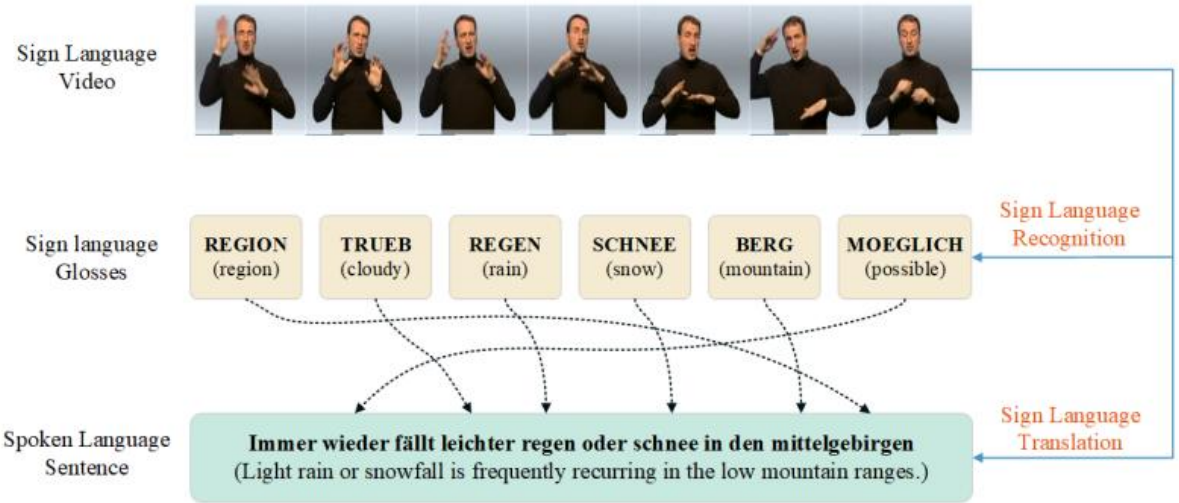


Figure 1. The difference between SLR and SLT.

Creating datasets with gloss annotations is both time-consuming and resource-intensive.

Hence, a recent trend is to shift towards gloss-free sign language translation, which is the primary focus of paper.

LIANG, Zeyu; LI, Huailing; CHAI, Jianping. Sign language translation: A survey of approaches and techniques. Electronics, 2023.

1.Introduction – Sign Language Recognition

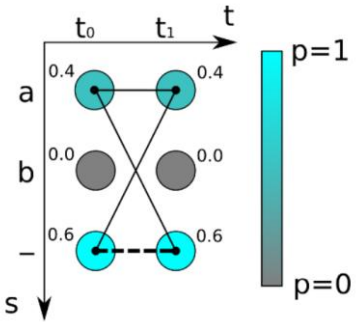
Isolated Sign Language Recognition (ISLR) : ISR focuses on recognizing individual signs. Each sign is treated as a separate unit without any connection to a preceding or following sign.
Continuous Sign Recognition (CSR) : CSR focuses on recognizing continuous sign sequences without pauses, requiring temporal understanding and sign segmentation.

Aspect	Isolated Sign Recognition (ISR)	Continuous Sign Recognition (CSR)
Input Type	Single, isolated signs	Continuous stream of signs
Complexity	Lower complexity	Higher complexity
Segmentation	Not required	Critical to separate signs
Applications	Learning tools, dictionaries	Real-time transcription, communication
Key Focus	Accurate classification of individual signs	Temporal modeling and context understanding

2.Related Works – CTC loss

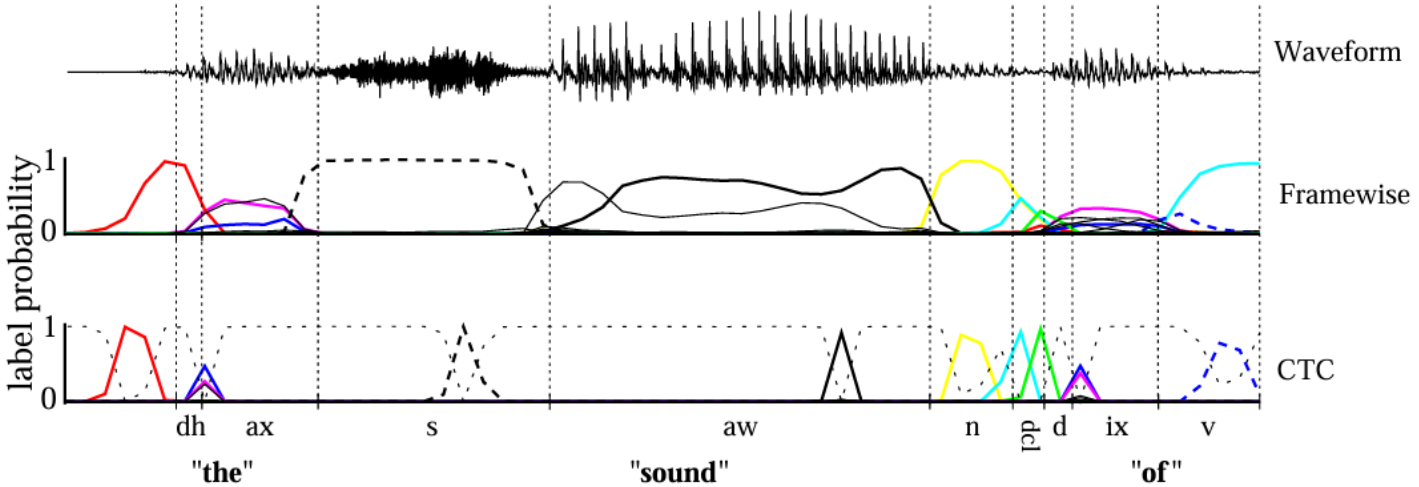
Connectionist Temporal Classification (CTC) is a method for handling unaligned sequence data, commonly used in speech and handwriting recognition. It works with RNNs to manage output sequences of different lengths from input sequences, preserving temporal order. CTC uses a 'blank' label to prevent merging consecutive identical outputs. (hhheellllloo ➔ hel-lo (peuso character) ➔ hello)

$$\mathcal{L}_{CTC} = -\log P(\mathbf{y}|\mathbf{X}), \quad P(\mathbf{y}|\mathbf{X}) = \sum_{\pi \in \mathcal{B}^{-1}(\mathbf{y})} P(\pi|\mathbf{X}), \quad P(\pi|\mathbf{X}) = \prod_{t=1}^{T} p(\pi_t|x_t).$$



Vocabulary = {a,b}
Add blank '-'
Probability = {a,b,'-'}
egg → -e-g-g-
apple → -a-p-p-l-e-

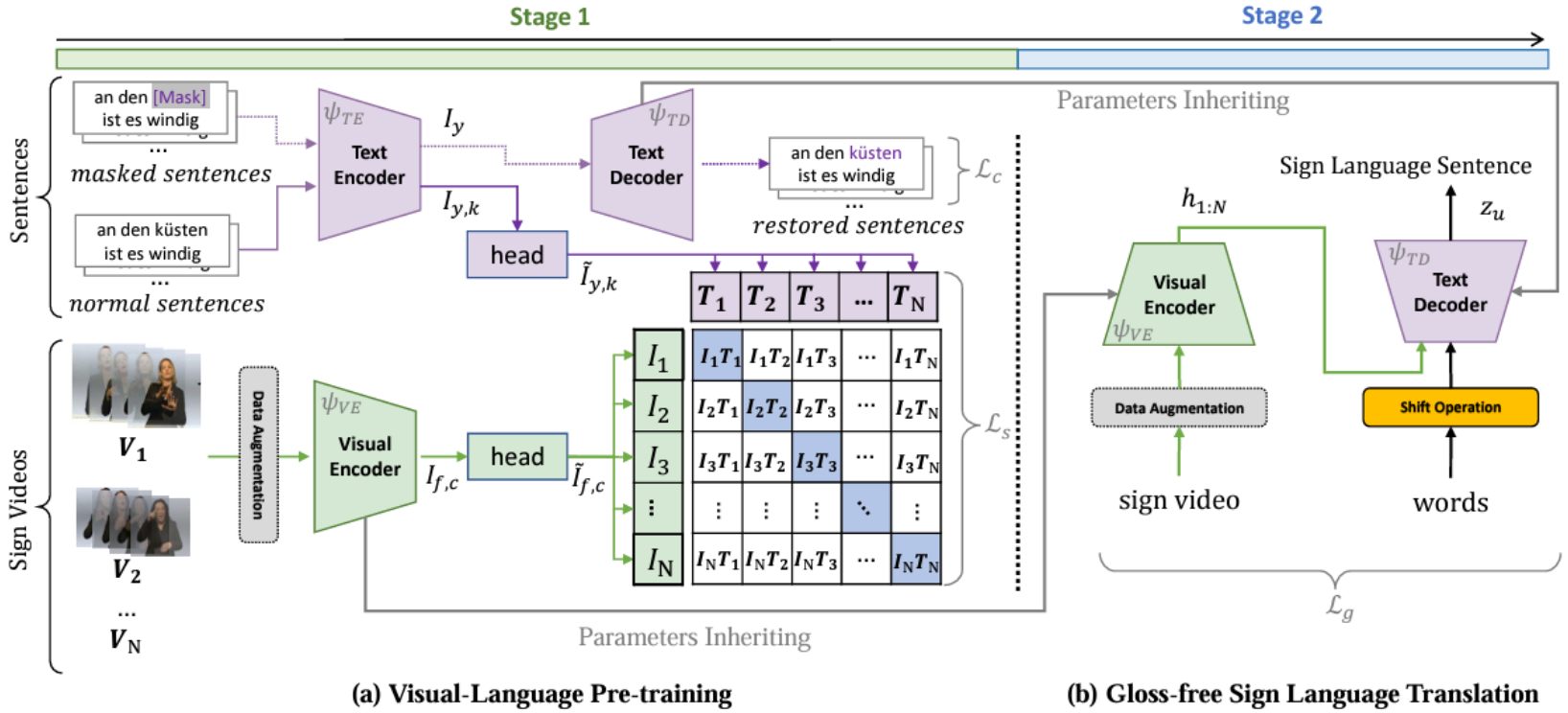
The goal of CTC is to maximize this probability for the correct output y. (alignment paths π)



GRAVES, Alex, et al. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. ICML 2006.
https://github.com/hccho2/CTC-Loss/blob/master/CTCLoss.pdf

2.Related Works – GFSLT-VLP (ICCV 2023)

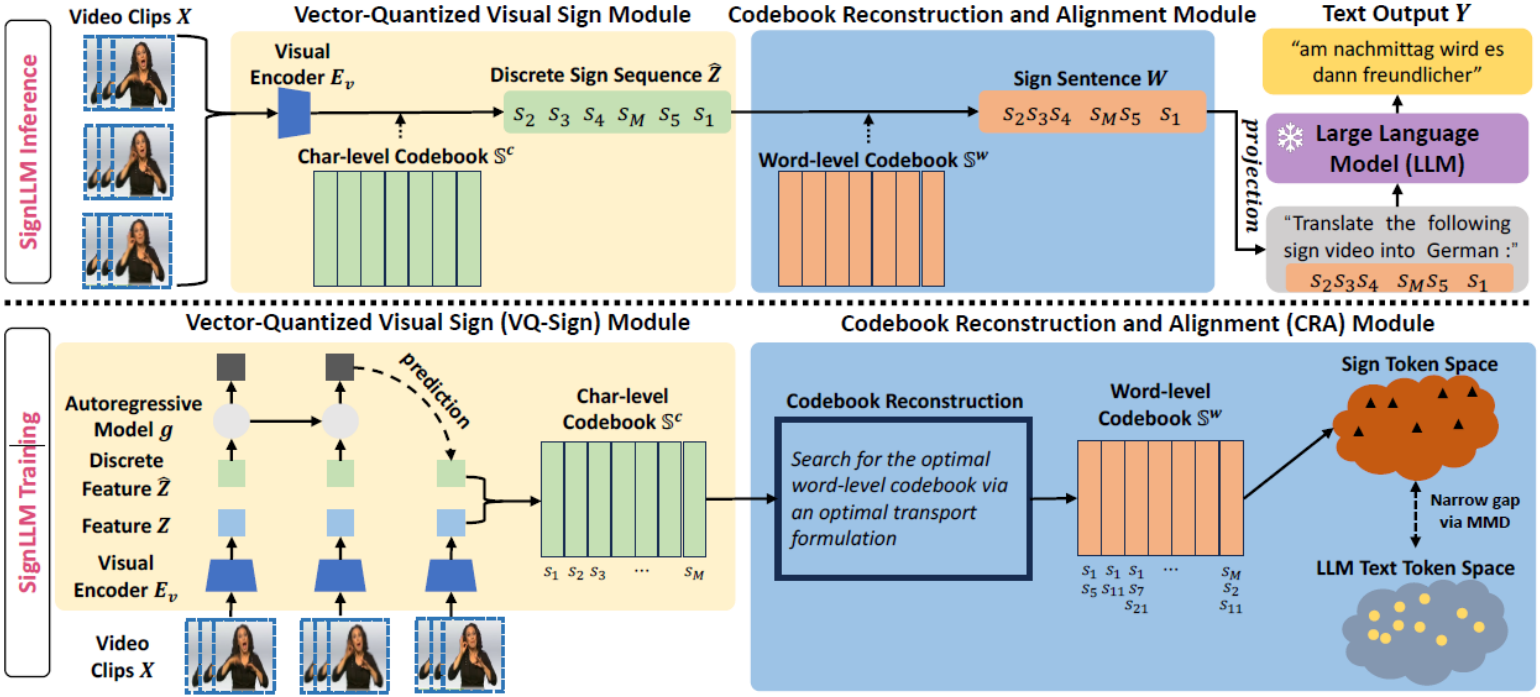
A novel pre-training paradigm that incorporates masked self-supervised learning together with contrastive language-image pre-training to facilitate the gloss-free SLT task.



ZHOU, Benjia, et al. Gloss-free sign language translation: Improving from visual-language pretraining. ICCV 2023.

2.Related Works – SignLLM (CVPR 2024)

(1) The first to harness the power of off-the-shelf and frozen LLMs for SLT.
(2) To make the input sign video compatible with LLMs, our SignLLM framework incorporates two designs: a VQSign module to quantize the sign video into a sequence of discrete character-level sign tokens
(3) CRA module that transforms the character-level sign tokens to word-level sign tokens.



GONG, Jia, et al. Llms are good sign language translators. CVPR 2024.

3.Method – Overview of Sign2GPT

Sign2GPT leverages pretrained large vision and language models to enhance sign language translation performance. It introduces a novel pretraining strategy with two main components: (1) an algorithm for automatically generating pseudo-glosses, (2) a prototype-driven method for pretraining the sign encoder using these pseudo-glosses. This approach removes the need for manual gloss annotations and ensures glosses do not require sign-order formatting.

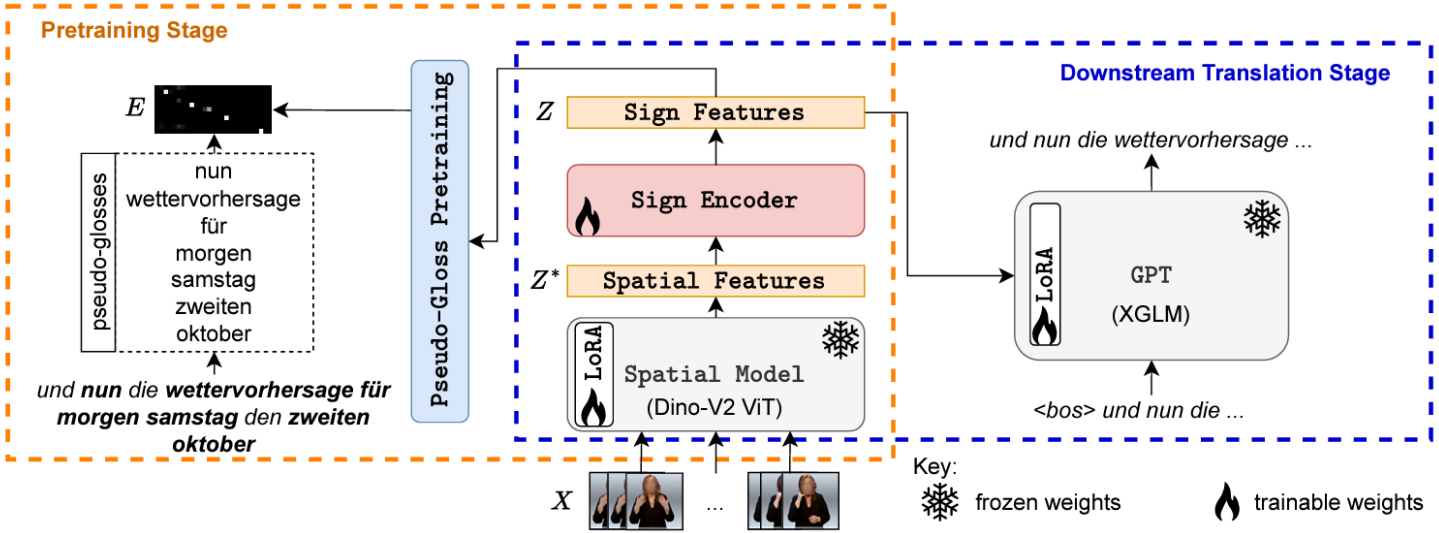
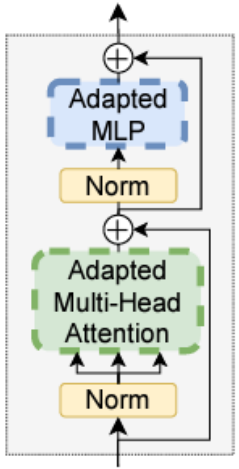


Figure 1: Overview of Sign2GPT, which consists of a pretraining stage that makes use of pseudo-glosses and downstream translation that leverages a frozen GPT model.

ZHOU, Hao, et al. Improving sign language translation with monolingual data by sign back-translation. CVPR 2021.

3.Mehotds – Model Architecture (Spatial Backbone)

Extract spatial features Z^* (linear transformation, batch normalization to feed it into sign encoder)
Input video frames $X = \{x_0, x_0, x_0, \ldots, x_{T^*}\}$ with T^*frmes, dimension is C represented as $Z^* \in \mathbb{R}^{T^* \times C}$
Fine-tuning Dino-V2 is essential for adapting it to the unique characteristics of SLT datasets.
LoRA is applied to the top encoder layers, targeting FC layers in the MLP and Multi-Head Attention.



Adapted Spatial Layer

Dino-V2 Vision Transformer(ViT-S/14)
- Fast and memory-efficient attention. / Sequence packing.
- Efficient stochastic depth. / Fully-Sharded Data Parallel.
- Model distillation.

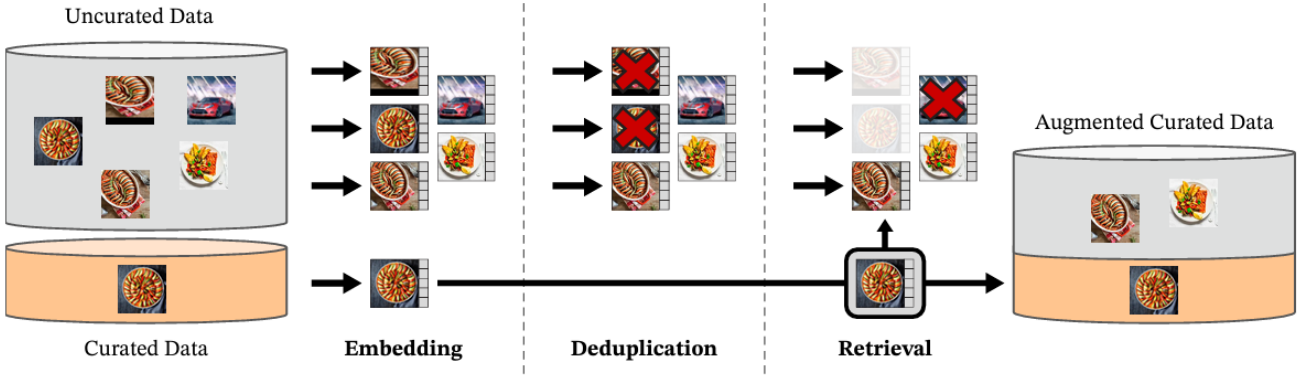


Figure 3: **Overview of our data processing pipeline.** Images from curated and uncurated data sources are first mapped to embeddings. Uncurated images are then deduplicated before being matched to curated images. The resulting combination augments the initial dataset through a self-supervised retrieval system.

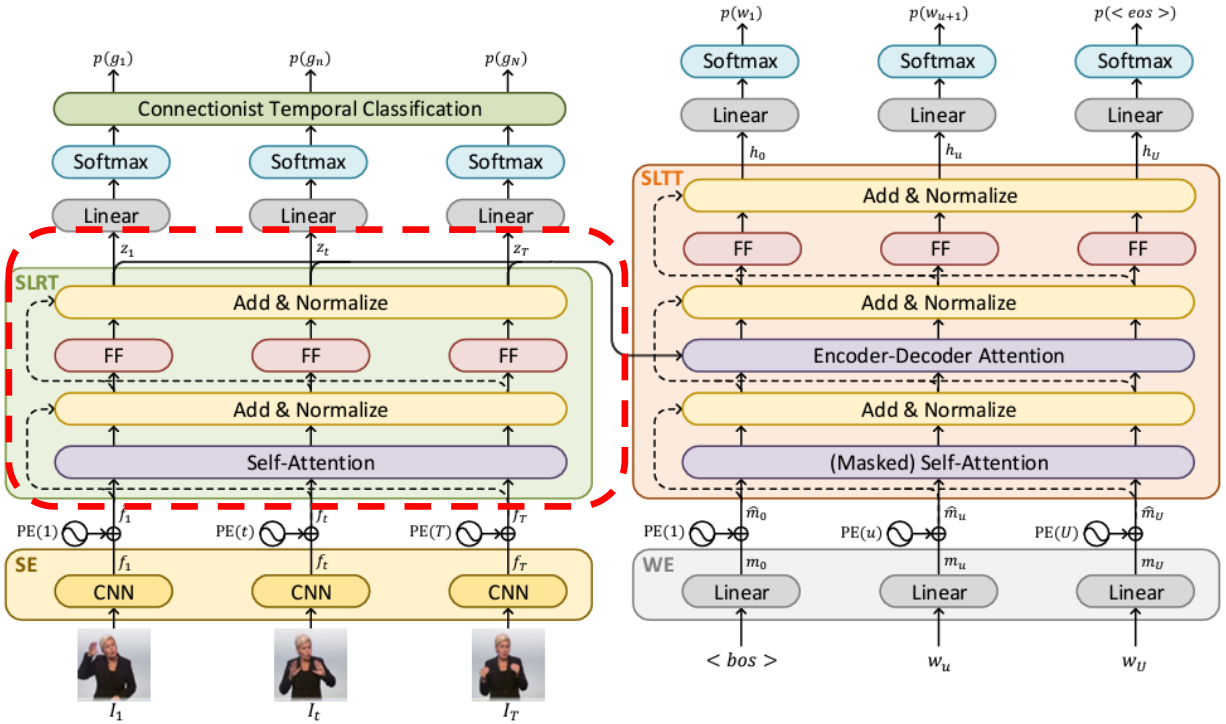
OQUAB, Maxime, et al. Dinov2: Learning robust visual features without supervision. TMLR 2024.

3.Mehotds – Model Architecture (Sign Encoder)

A spatio-temporal transformer model inspired by prior sign language translation approaches.

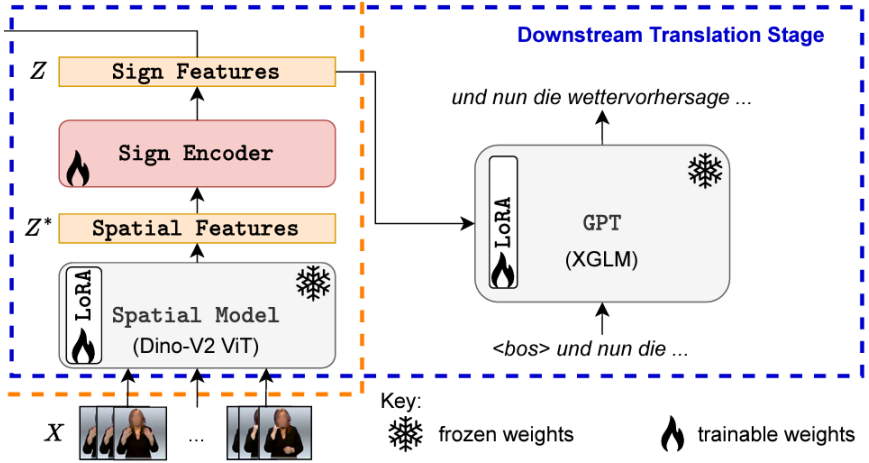
Employ temporal down sampling after specific layers within our encoder. ($T^*$to $\frac{T^*}{2}$) 3 kernel and 2 stride.

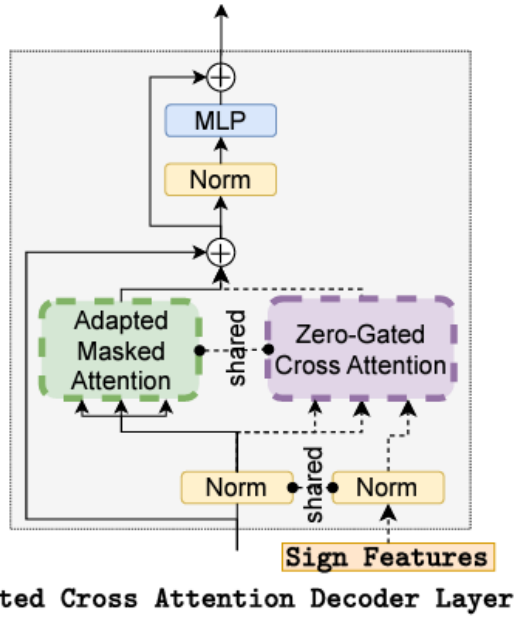
Use local self-attention with a window size of seven, a technique proven to be highly effective in SLT tasks.



CAMGOZ, Necati Cihan, et al. Sign language transformers: Joint end-to-end sign language recognition and translation. CVPR 2020.

3.Mehotds – Model Architecture (Language Decoder)





Gated Cross Attention Decoder Layer

Adapt the XGLM(1.7B) model

This enhancement shares weights from the pretrained masked multi-head attention and integrates a separate
LoRA for masked multi-head attention
(Adapted Masked Attention) and
cross-attention (Zero-Gated Cross Attention).

$$\text{GatedAttention}(Q,K,V) = \left(\boldsymbol{g} \times \texttt{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)\right)V$$

K and V represent the inputs from the sign features
Q originates from the textual features.
g is a learnable gate parameter for each attention head which is clamped between 0 and 1 and initialized to zero to preserve linguistic knowledge at the start of training.

ZHOU, Hao, et al. Improving sign language translation with monolingual data by sign back-translation. CVPR 2021.

3.Method – Pre-training Stage

Video-to-text training for SLT
Freeze the pretrained VLM and use the adapters for sign language domain transfer.
Our model's primary focus on capturing and utilizing sign language features and then leverages the language model's linguistic ability to adapt the features to spoken language translation.

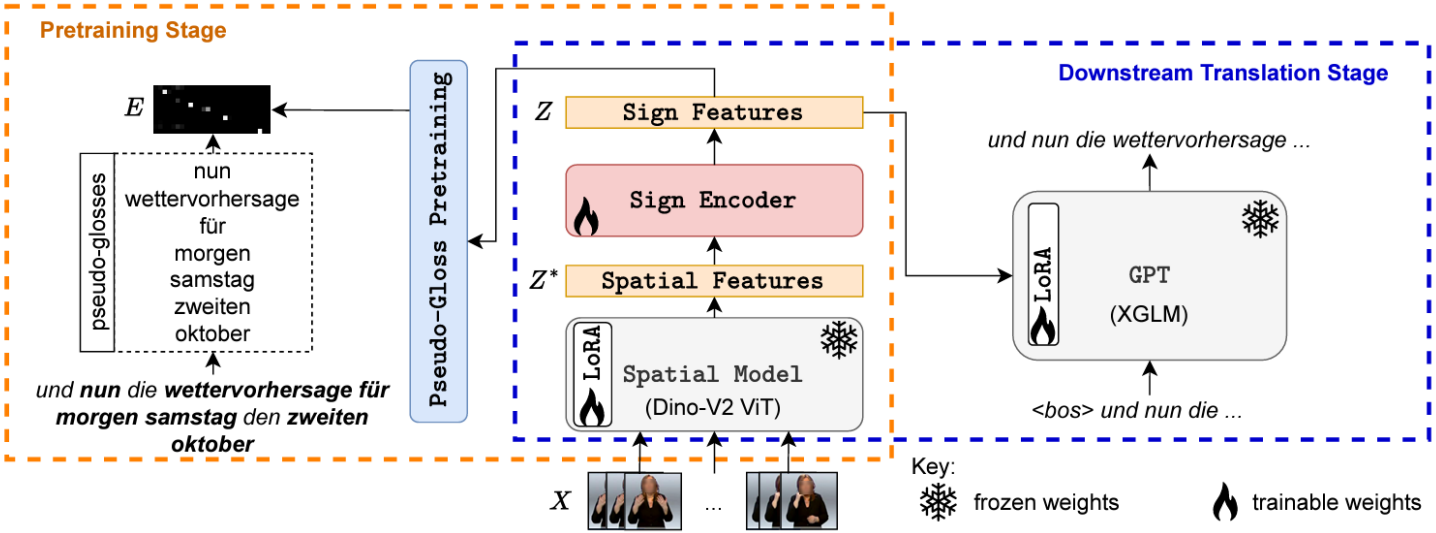


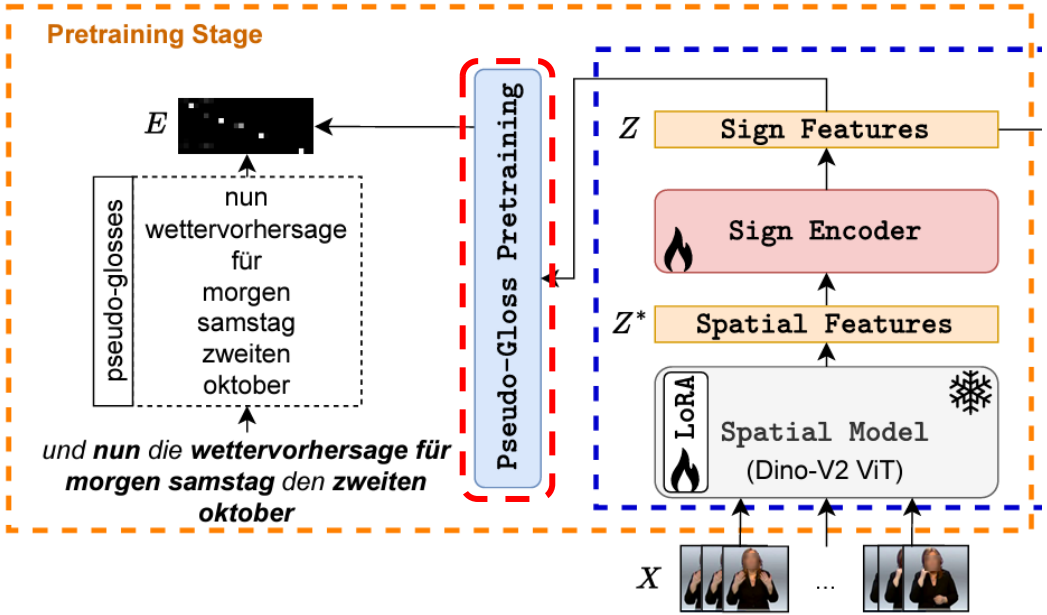
Figure 1: Overview of Sign2GPT, which consists of a pretraining stage that makes use of pseudo-glosses and downstream translation that leverages a frozen GPT model.

ZHOU, Hao, et al. Improving sign language translation with monolingual data by sign back-translation. CVPR 2021.

3.Method – Pre-training Stage

Pseudo-gloss generation : Spoken language -> Pseudo-gloss (using spaCy library)
German-Phoenix14T(lemmatization), Chinese-CSL-Daily(word segmentation)

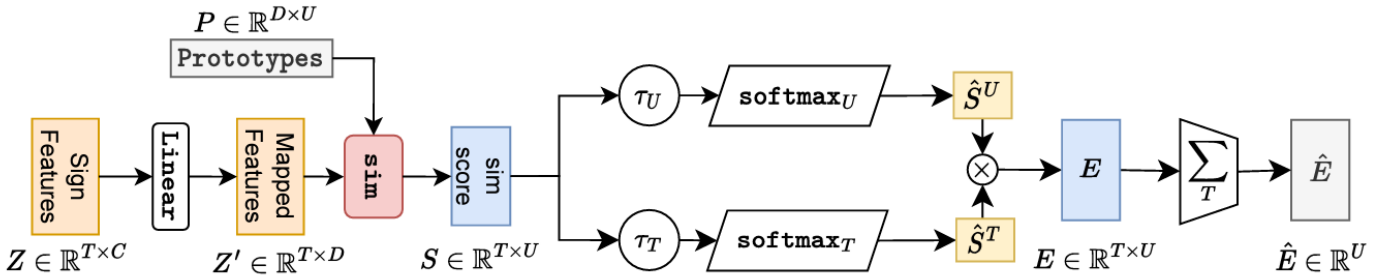
➔ It's important to highlight that our pseudo-glosses are in spoken language order,
unlike manually annotated glosses which are in sign order. (CTC loss are not suitable)



3.Method – Pre-training Stage

Pseudo-gloss pretraining : Generated prototypes for each pseudo-gloss.
Note that during pretraining the learned sign representations are temporally invariant, therefore we also add sinusoidal positional encoding before FC_m for the translation task.

Generated prototypes for each pseudo-gloss.



$$Z = \{z_0, z_1, ..., z_i, ..., z_T\} \qquad s_i = \mathtt{sim}(z_i', P) = \frac{z_i' \cdot P}{\|\hat{z}_i\| \|P\|}$$

4.Experiments – Dataset & Evaluation Protocol

Table 1. Summary of public available video-based sign language benchmarks popular for computer vision research. (SignDict: the corpus has isolated or segmented sign videos as a dictionary. Continuous: the corpus is composed of videos of continuous sign sentences and gloss-level annotations. Translation: the corpus has spoken language translation annotations.)

Dataset	Language	Attribute				Statitics			Source
		SignDict	Continuous	Translation	Resolution	#Signs	#Videos (avg. signs)	#Signers	
DEVISIGN [48]	CSL	✓			-	2,000	24,000 (1)	8	Lab
ICSL [52]	CSL	✓			1280×720	500	125,000 (1)	50	Lab
MSASL [20]	ASL	✓			-	1,000	25,513 (1)	222	Web
WLASL [24]	ASL	✓			-	2,000	21,083 (1)	119	Web
BSL-1K [1]	BSL	✓			-	1,064	273,000 (1)	40	TV
INCLUDE [40]	ISL	✓			1920×1080	263	4,287 (1)	7	Lab
PHOENIX-2014 [23]	DGS		✓		210×260	1,081	6,841 (11)	9	TV
CCSL [17]	CSL	✓	✓		1280×720	178	25,000 (4)	50	Lab
SIGNUM [47]	DGS	✓	✓	✓ (German)	776×578	455	15,075 (7)	25	Lab
PHOENIX-2014T [10]	DGS		✓	✓ (German)	210×260	1,066	8,257 (9)	9	TV
CSL-Daily (ours)	CSL	✓	✓	✓ (Chinese)	1920×1080	2,000	20,654 (7)	10	Lab

BLEU (Bilingual Evaluation Understudy) : Measures the quality of machine-translated text by comparing it to one or more reference translations. Focuses on precision rather than recall. (n-gram precision)

ROUGE–L (Recall-Oriented Understudy for Gisting Evaluation) : Measures the overlap between the generated text and reference texts using recall, precision, and F1-score. (–L means Longest Common Subsequence)

4.Experiments – Training Settings(Pretraining)

Batch size 8
2 A100 GPUs
Subsampling every second frame
Spatial adapters applied to the top three layers of the spatial model.(Due to memory constraints)
Sign encoder : 4 layer transformer with hidden dimension 512.(8 attention heads, intermediate size of 2048)

Training
Bfloat16 / Flash attention v2
Adam optimizer with learning rate of 3x10^4, weight decay of 0.001.
100 epochs with gradient clipping of 1.0.
A one-cycle cosine learning rate scheduler with warmup for initial 5 epochs.
Data augmentation (jitter, random resized cropping from 256x256 to 224x224 pixels, rotation, horizontal flips.

During evaluation, center cropping to 224x224 pixels.

Prototype (τ_U) and time temperature (τ_T) set to 0.1

For the CSL-Daily dataset, the number of pseudo-glosses significantly exceeds the vocabulary size.
(based on Chinese)

4.Experiments – Training Settings(Downstream Translation)

Table 1: Training setting with (a) the number of pseudo-glosses during pretraining and (b) parameter counts during downstream translation.

Dataset	# Vocab	# p-glosses
Phoenix14T	2,887	2,533
CSL-Daily	2,343	7,918

(a) Vocabulary versus pseudo-glosses per dataset

Component	# Params	# Trainable
Spatial	22,328,448	271,872
Sign Encoder	12,613,632	12,613,632
Decoder	1,736,710,528	3,803,520
Total	1,771,652,608	16,689,024

(b) Number of model parameters during translation

Cross-entropy loss with label smoothing set to 0.1.
LoRA rank and alpha values are both set to 4.
During inference we employ a beam search with a width of 4.

4.Experiments - Results on Phoenix14T

Table 2: Comparison of test set results on Phoenix14T. We present our gloss-free results for three experimental settings: (1) Without pseudo-gloss pretraining (**Sign2GPT**), (2) with pseudo-gloss pretraining (**Sign2GPT(w/PGP)**), and (3) extracted features Z from the frozen spatial and sign encoder model that has been trained with pseudo-gloss pretraining (**Sign(Z)2GPT(w/PGP)**).

Method	Test Set				
	BLEU1	BLEU2	BLEU3	BLEU4	ROUGE
Gloss-based					
SL-Transformer (Camgoz et al., 2020b)	46.61	33.73	26.19	21.32	–
BN-TIN-Transf.+BT (Zhou et al., 2021)	50.80	37.75	29.72	24.32	49.54
MMTLB (Chen et al., 2022a)	53.97	41.75	33.84	28.39	52.65
SLTU_{NET} (Zhang et al., 2023a)	52.92	41.76	33.99	28.47	52.11
TwoStream-SLT (Chen et al., 2022b)	54.90	42.43	34.46	28.95	53.48
Gloss-free					
NSLT (Camgoz et al., 2018)	29.86	17.52	11.96	9.00	30.70
TSPNet (Li et al., 2020b)	36.10	23.12	16.88	13.41	34.96
CSGCR (Zhao et al., 2021)	36.71	25.40	18.86	15.18	38.85
GASLT (Yin et al., 2023)	39.07	26.74	21.86	15.74	39.86
GFSLT (Zhou et al., 2023)	41.39	31.00	24.20	19.66	40.93
GFSLT-VLP (Zhou et al., 2023)	43.71	33.18	26.11	21.44	42.49
Sign2GPT	45.43	32.03	24.23	19.42	45.23
Sign2GPT(w/PGP)	**49.54**	**35.96**	**28.83**	**22.52**	**48.90**
Sign(Z)2GPT(w/PGP)	47.06	33.61	25.85	20.93	47.11

BLEU-n represents the weighted average translation precision up to n-grams. Typically, uniform weights are used, meaning the weights for 1-grams to n-grams are all set to 1/n.

BLEU-1, BLEU-2, BLEU-3, and BLEU-4 scores represent **n-gram precision at different levels of granularity** in a machine translation or text generation task.
BLEU-1~4
(unigram ~ N-gram)

4.Experiments - Results on the CSL-Daily

Table 3: Comparison of test set results on the CSL-Daily. We present gloss-free results for three experimental settings: (1) Without pseudo-gloss pretraining (**Sign2GPT**), (2) with pseudo-gloss pretraining (**Sign2GPT(w/PGP)**), and (3) extracted features Z from the frozen spatial and sign encoder model that has been trained with pseudo-gloss pretraining (**Sign(Z)2GPT(w/PGP)**).

Method	Test Set				
	BLEU1	BLEU2	BLEU3	BLEU4	ROUGE
Gloss-based					
SL-Transformer (Camgoz et al., 2020b)	37.38	24.36	16.55	11.79	36.74
BN-TIN-Transf.+BT (Zhou et al., 2021)	51.42	37.26	27.76	21.34	49.31
MMTLB (Chen et al., 2022a)	53.31	40.41	30.87	23.92	53.25
SLTU_{NET} (Zhang et al., 2023a)	54.98	41.44	31.84	25.01	54.08
TwoStream-SLT (Chen et al., 2022b)	55.44	42.59	32.87	25.79	55.72
Gloss-free					
GASLT (Yin et al., 2023)	19.90	9.94	5.98	4.07	20.35
NSLT (Camgoz et al., 2018)	34.16	19.57	11.84	7.56	34.54
GFSLT (Zhou et al., 2023)	37.69	23.28	14.93	9.88	35.16
GFSLT-VLP (Zhou et al., 2023)	39.37	24.93	16.26	11.00	36.44
Sign2GPT	34.80	24.00	17.27	12.96	41.12
Sign2GPT(w/PGP)	**41.75**	**28.73**	**20.60**	**15.40**	**42.36**
Sign(Z)2GPT(w/PGP)	32.73	20.52	13.75	9.73	33.39

4.Experiments – Qualitative Results

Leverage the output values of E, with dimensions.
Notice that our pretraining has automatic localization capabilities irrespective of the order of the pseudo-gloss when using the output E.

Figure 4: Visualizations of the localization capabilities of our pretraining stage. We visualize only the pseudo-glosses from the target sentence (y-axis) over time (x-axis), with whiter regions indicating a higher probability of the pseudo-gloss occurring during the time segment. We also display the localized gloss (under the video frames) based on a threshold of 0.2 on E.

4.Experiments – ablation study

Table 4: Ablation of results on the Phoenix14T dataset showing different architecture changes with no pseudo-gloss pretraining (**Sign2GPT**) and with pseudo-gloss pretraining (**Sign2GPT(w/PGP)**).

Architecture	**BLEU4**
Sign2GPT	
Spatial Adapters + Local Attention + Downsampling	**19.55**
· No Spatial Adapters	16.38
· No Local Attention (+ Global Attention)	18.56
· No Downsampling on Sign Encoder	19.30
Sign2GPT(w/PGP)	
· No positional	21.68
· Learnable positional (zero init)	21.89
· Learnable positional (random init)	21.16
· Sinusoidal positional	**23.20**

Spatial Backbone	**Test Set**	
	BLEU4	**ROUGE**
ResNet18	12.28	38.31
DinoV2 (ViT-S/14)	**12.96**	**41.12**

Phoenix14T	**Precision**	**Recall**	**F1-Score**
no sign encoder	0.50	0.25	0.33
sign encoder	**0.52**	**0.39**	**0.44**

4.Experiments – ablation study

All words as tokens vs the selected pseudo-glosses (Phoenix14T)

Tokens	Precision	Recall	F1-Score
all words	**0.55**	0.28	0.37
pseudo-glosses	0.52	**0.39**	**0.44**

Dataset	Precision	Recall	F1-Score
Phoenix14T	0.52	0.39	0.44
CSL-Daily	0.38	0.34	0.36

Table 11: Ablation of XGLM backbones on Phoenix14T using our Sign2GPT architecture.

	Backbone	Test Set	
		BLEU4	ROUGE
	GFSLT-VLP (mBART) (Zhou et al., 2023)	21.44	42.49
Model Size	Sign2GPT(w/ PGP) (XGLM-564M)	22.29	48.21
	Sign2GPT(w/ PGP) (XGLM-1.7B)	**22.52**	**48.90**

4.Experiments – Examples of translation results on the Phoenix14T dataset.

Hypothesis:	im übrigen land scheint häufig die sonne und es gibt nur wenig schauer . **(in the rest of the country the sun often shines and there are only a few showers .)**
Pseudo-glosses:	übrig, gebiet, sonne, nur, locker, wolke
Reference:	in den übrigen gebieten viel sonne und nur ein paar lockere wolken . **(lots of sun in the remaining areas and only a few loose clouds .)**
Hypothesis:	am tag zwölf grad an der ostsee und bis zu zwanzig grad im süden . (**twelve degrees a day on the baltic sea and up to twenty degrees in the south .**)
Pseudo-glosses:	tag, zwölf, grad, ostsee, zwanzig, grad, niederbayer
Reference:	am tag zwölf grad an der ostsee und bis zwanzig grad in niederbayern . (**on the day twelve degrees on the baltic sea and up to twenty degrees in lower bavaria .**)
Hypothesis:	ich wünsche ihnen noch einen schönen abend und machen sie es gut . **(i wish you a nice evening and do well .)**
Pseudo-glosses:	ich, wünschen, ihnen, schön, abend, machen, sie, es, gut
Reference:	ich wünsche ihnen einen schönen abend und machen sie es gut . (**i wish you a nice evening and do well .)**
Hypothesis:	der wind weht schwach bis mäßig aus süd bis südost . (**the wind blows weakly to moderately from the south to southeast .)**
Pseudo-glosses:	dazu, wehen, schwach, wind, südost, süd
Reference:	dazu weht ein schwacher bis mäßiger wind aus südost bis süd . (**in addition a weak to moderate wind blows from the southeast to the south .)**

4.Experiments – Examples of translation results on the CSL-Daily dataset

Hypothesis:	这个地方不离饭店，走几步就到饭店的门口。**(This place is not far from the hotel, just a few steps to the door of the hotel.)**
Pseudo-glosses:	可以/ 这里/ 不/ 远/ 有/ 饭馆/ 走/ 几/ 分钟/ 就/ 到
Reference:	可以，离这里不远有一个饭馆，走几分钟就到了。**(Okay, there is a restaurant not far from here, it can be reached in a few minutes' walk.)**
Hypothesis:	公司很远，他为什么不打车呢？ **(The company is far away, why doesn't he take a taxi?)**
Pseudo-glosses:	公司/离家/ 很/ 远/ 他/ 为什么/ 不/ 打车
Reference:	公司离家很远，他为什么不打车？ **(The company is far from home, why doesn't he take a taxi?)**
Hypothesis:	我不去爬山，我有事情要去做。**(I'm not going to climb mountains, I have things to do.)**
Pseudo-glosses:	我/ 不/ 去/ 爬山/ 我/ 有事
Reference:	我不去爬山，我有事。**(I'm not going to climb the mountain, I have something to do.)**
Hypothesis:	我喜欢下雪。**(I like snow.)**
Pseudo-glosses:	我/ 喜欢/ 冬天/ 下雪/ 太/ 美
Reference:	我喜欢冬天，下雪太美了。**(I like winter, the snow is so beautiful.)**

5.Conclusion & Limitation

(+) Novel approach to address the challenging problem of Sign Translation in a gloss-free setting.

(+) Novel pretraining strategy that learns from pseudo-glosses which are generated automatically to learn word-level sign features, thereby allowing our sign encoder to be effectively pretrained without the use of manually annotated glosses.

(+) Sign2GPT architecture presents a promising direction for the exploration of fusing visual features to spoken language models for sign language recognition and translation tasks.

(+) Sign2GPT, demonstrates significant performance improvements over existing state-of-the-art techniques on the Phoenix14T and CSL-Daily datasets.

(-) The method is mostly based on existing models such as GPT, Dino-V2 and LoRA, so there is not much novelty from the architectural standpoint.

(-) Both large-scale vision and language model, while the title only mentions the language one.

(-) Utilization of pretrained model is not new. (LoRA)

https://openreview.net/forum?id=LqaEEs3UxU

Thanks

Any Questions?

You can send mail to
Susang Kim(healess1@gmail.com)