A. Implementation Details

We elaborate on our training and inference procedure in this section. Unless otherwise specified we use a batch size of 8, the Adam optimizer with a weight decay of 1e-3 and the cosine annealing scheduler.

Visual Encoder Pretraining. For general-domain pretraining on K-400 [3] and WLASL [4], our training procedure follows [5]. All of the five blocks in S3D backbone followed by a spatial average pooling layer and a linear classification layer are used in the general-domain pretraining. For within-domain pretraining on Sign2Gloss, we only use the first four blocks of the pretrained S3D and spatially pool the S3D features into size of $T/4 \times 832$ as inputs into our head network. The data augmentations for within-domain pretraining include temporally-consistent spatial random crop with range of [0.7-1.0] and frame-rate augmentations with range of [$\times$0.5-$\times$1.5]. All frames are spatially resized to 224$\times$224 as inputs. We train our visual encoder for 80 epochs with an initial learning rate of 1e-3. During inference, we use all frames and resize them to 224$\times$224 without cropping. For gloss sequence prediction, we adopt CTC beam search decoder with widths ranging from one to ten and choose the width of the best performance on dev set for test set evaluation.

CTC Decoder. Once the within-domain pretraining (Sign2Gloss) of our visual encoder is finished, we can use it for gloss prediction. Concretely, given a sign video $\mathcal{V} = (v_1, ..., v_T)$ with T frames, the visual encoder $\mathcal{E}$ predicts gloss probabilities $\mathcal{P} = (p_1, ..., p_{T/4})$ where p_t represents the distribution of gloss probability at step t. CTC decoder uses $\mathcal{P}$ as the input to estimate the most confident gloss sequence using beam search decoding algorithm. More details can be found in [2].

Translation Pretraining. For general-domain pretraining, we use the release of mBART-large-cc25[1] as initialization. For within-domain pretraining, i.e., Gloss2Text task, we train the translation network with the cross-entropy loss for 80 epochs with an initial learning rate of 1e-5. We also use dropout of 0.3 and label smoothing of 0.2 to prevent overfitting. For memory efficiency, we prune the mBART word embedding by preserving words in the target language, i.e., Chinese for CSL-Daily [6] and German for PHOENIX2014T [1]. The pruned word embedding is frozen during training. Following mBART, we use a language id symbol, i.e. 'zh_CN' or 'de_DE', as [EOS] of encoder inputs and [BOS] of decoder inputs for language identification.

Joint Training. Two independently pretrained networks are loaded as the initialization for joint training. Visual-language mapper (V-L Mapper) establishes a bridge between features of the visual modality and language modality. To reduce computational cost, we freeze the S3D backbone during joint training. For PHOENIX-2014T [1] dataset, we use gloss representations (see Figure 3 of the paper) as the input of V-L Mapper. In CSL-Daily [6] dataset, we observe that the gloss annotations almost contain all language information for generating the text, i.e., in many cases spoken text can be accurately predicted by simply reordering, or even copying, the gloss sequence. Therefore in CSL-Daily, the glosses' linguistic semantics are more helpful than their visual semantics for translation task (Sign2Text). As a consequence, we take the gloss probabilities (see Figure 3 of the paper) as the input of the V-L Mapper where the FC layer is initialized using the weights of the pretrained word embedding in the translation network. The other settings for the two datasets are identical. The learning rate is set as 1e-5 for the translation network and 1e-3 for trainable layers in the visual encoder and V-L mapper. We train the whole network under the joint supervision of the CTC loss and cross-entropy loss with a loss weight of 1.0 for 40 epochs. During evaluation, we use beam search decoding with a beam width of 4 and length penalty of 1.

[1] https://huggingface.co/facebook/mbart-large-cc25

B. More ablations

Loss weights. In Sign2Text joint training, we vary weights of CTC loss and Cross-Entropy loss to study their effects on translation performance. The results are shown in table 1. We find that our method is insensitive to the loss weights.

Temporal downsampling. We also temporally downsample input videos to 1/2 and 1/3 and use the downsampled videos to train Sign2Gloss and Sign2Text of PHOENIX-2014T. Comparison between different downsampling strides are shown in table 2. It can be seen that temporal downsampling greatly degrades performance in both Sign2Gloss and Sign2Text.

Tuning S3D layers in Sign2Text joint training We freeze the S3D backbone for efficient Sign2Text training. Table 3 examines freezing a subset of S3D blocks. Tuning S3D layers does not improve the performance, for which we conjecture three reasons. 1) the hyperparameters for joint training with tunable S3D layers need re-tuning. 2) Excessive number of trainable parameters lead to overfitting. 3) With a well-pretrained visual encoder, visual features are not the bottleneck for translation performance in Sign2Text join training.

C. Qualitative Analysis

We report some qualitative results on PHOENIX-2014T [1] and CSL-Daily [6] datasets. We first demonstrate the effectiveness of our Sign2Text end-to-end training. Then we reveal the current method's limitations to shed some light on future work.

Loss weights		Dev					Test				
CTC	translation	R	B1	B2	B3	B4	R	B1	B2	B3	B4
0.5	1	52.95	54.03	41.32	33.23	27.70	**53.17**	**54.49**	**42.01**	**33.90**	28.33
1	1	53.10	53.95	41.12	33.14	27.61	52.65	53.97	41.75	33.84	**28.39**
2	1	53.15	53.90	41.03	32.94	27.35	52.80	54.08	41.48	33.34	27.70
1	0.5	52.91	54.09	41.46	**33.45**	27.91	53.12	54.19	41.60	33.67	28.27
1	2	**53.59**	**54.35**	**41.50**	33.44	**27.92**	52.73	53.68	41.31	33.42	28.01

Table 1. Ablation study of varied weights on the CTC loss and translation loss on **PHOENIX Sign2Text** training.

Downsample rate	Sign2Gloss		Sign2Text									
	Dev	Test	Dev					Test				
	WER	WER	R	B1	B2	B3	B4	R	B1	B2	B3	B4
1	**21.90**	**22.45**	53.10	53.95	41.12	33.14	27.61	**52.64**	**53.97**	**41.75**	**33.84**	**28.39**
1/2	29.54	30.73	50.74	51.33	38.51	30.58	25.24	50.78	52.01	39.14	31.02	25.59
1/3	40.97	40.43	46.65	47.62	34.16	26.23	21.07	46.45	47.66	34.72	26.92	21.89

Table 2. Ablation study of different temporal downsampling rate temporal of input videos on PHOENIX dataset.

Frozen blocks	Dev					Test				
	R	B1	B2	B3	B4	R	B1	B2	B3	B4
1	**53.28**	53.75	41.14	33.31	27.95	52.35	53.57	41.21	33.23	27.79
1-2	52.67	53.42	40.76	32.91	27.51	52.39	53.57	41.20	33.26	27.82
1-3	53.03	**54.24**	**41.71**	**33.69**	**28.16**	52.61	53.89	41.48	33.51	28.03
1-4 (default)	53.10	53.95	41.12	33.14	27.61	**52.65**	**53.97**	**41.75**	**33.84**	**28.39**

Table 3. Freezing different S3D blocks on **PHOENIX Sign2Text**.

Effectiveness of Joint Sign2Text Training. To illustrate that our end-to-end Sign2Text training can utilize rich visual information from sign videos and semantic knowledge from text transcriptions to produce translation of high quality, we compare it with our Gloss2Text model and Sign2Gloss2Text pipeline. Table 4 shows the ground-truth glosses (Gloss), ground-truth text references (Text), gloss predictions from visual encoder (Sign2Gloss), and translation results from three approaches, namely Gloss2Text, Sign2Gloss2Text and Sign2Text.

In Example (a) and (c), we can see that when Sign2Gloss model predicts wrong glosses, the two-stage pipeline (Sign2Gloss2Text) will be influenced, resulting in the wrong translation texts. For instance, in Example (c), the sign of 'Selfish' is wrongly predicted as 'Happiness' by Sign2Gloss model, which further misleads Sign2Gloss2Text pipeline to generate the wrong translation 'We cannot do happy things'. Nevertheless, our end-to-end Sign2Text model mitigates the error propagation issue, e.g., correctly generating 'We cannot do selfish things'. Moreover, Example (b) and (d) demonstrate that our end-to-end Sign2Text model outperforms Gloss2Text translation model by correctly predicting the words which are not included in the gloss annotations. For example, in Example (b), our Sign2Text model manages to translate some subtle words such as 'dominate' and 'half', and in Example (d), unlike Gloss2Text model that translates according to the gloss's literal meaning, our Sign2Text model translates the sign of 'Sugar' into 'drink'. This suggests that our Sign2Text model is capable of leveraging visual information from sign videos and supplement the knowledge that the discrete glosses can not fully capture.

Limitations. Here we present some failure cases in Table 5. We observe that the our method has some difficulty in identifying numbers and location entities due to their low frequency in the training corpus. Also, when dealing with long inputs, the translation results may either leave out some information or be not fluent. We look forward to overcoming the difficulty by handling long-tailed sign distribution in future works. There are several limitations of our work. First, our current approach relies on continuous sentence-level gloss annotations. Although some existing SLT datasets provide gloss annotations, they are expensive

to obtain. We hope to lift the need for manual gloss labels in the future. Second, both CSL-Daily and PHOENIX-2014T are recorded under constrained conditions with limited vocabulary and number of signers. All of the signings are interpreted from spoken language. While our model achieves good results in these two benchmarks, it is worth further studying its performance in wild scenarios with large vocabulary, diverse signers and conversational signings. As with broader social impact, it should be noted that sign language modelling may result in increased surveillance of deaf communities.

References

[1] Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. Neural sign language translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[2] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, 2006.

[3] Will Kay, João Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset. *CoRR*, 2017.

[4] Dongxu Li, Cristian Rodriguez, Xin Yu, and Hongdong Li. 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*, 2020.

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

[6] Hao Zhou, Wengang Zhou, Weizhen Qi, Junfu Pu, and Houqiang 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.

Example (a)	
Gloss (GT)	HEUTE / NACHT / NORD / REGION / ELF / GRAD / ALPEN / EINS / GRAD (Today / Night / North / Region / Eleven / Degrees / **Alps** / One / Degree)
Sign2Gloss	HEUTE / NACHT / NORD / ELF / GRAD / REGION / EINS / GRAD (Today / Night / North / Eleven / Degrees / **Region** / One / Degrees)
Text (GT)	Heute Nacht elf Grad im Norden und ein Grad an den Alpen. (Tonight eleven degrees in the north and one degree in the **Alps**.)
Sign2Gloss2Text	Heute Nacht elf Grad an der Nordsee ein Grad im Breisgau. (Tonight eleven degrees on the north sea one degree in **Breisgau**.)
Sign2Text	Heute Nacht elf Grad im Norden und ein Grad an den Alpen. (Tonight eleven degrees in the north and one degree in the **Alps**.)
Example (b)	
Gloss (GT)	MORGEN / IX / NORDOST / WOLKE / ÜBERWIEGEND / SCHAUER (Tomorrow / IX / Northeast / Clouds / Presently / Showers)
Sign2Gloss	MORGEN / NORDOST / WOLKE / ÜBERWIEGEND / SCHAUER (Tomorrow / Northeast / Clouds / Presently / Showers)
Text (GT)	Morgen überwiegen in der Nordosthälfte die Wolken und es fallen einzelne Schauer. (Tomorrow the clouds will **dominate** in the northeast **half** and there will be a few showers.)
Sign2Gloss2Text	Morgen im Norden und Nordosten dichte Wolken mit einzelnen Schauern. (Tomorrow in the north and northeast dense clouds with a few showers.)
Sign2Text	Morgen überwiegen in der Nordosthälfte die Wolken und es fallen einige Schauer. (Tomorrow the clouds will **dominate** in the northeast **half** and some showers will fall.)
Example (c)	
Gloss (GT)	我们 / 做 / 人 / 自私 / 不行 (We / Do / People / **Selfish** / Cannot)
Sign2Gloss	我们 / 做 / 幸福 / 不行 (We / Do / **Happiness** / Cannot)
Text (GT)	我们做人不能自私。 (We cannot be **selfish**.)
Gloss2Text	我们做人不能自私自利。 (We cannot be **selfish**.)
Sign2Gloss2Text	我们不能做幸福的事情。 (We cannot do **happy** things.)
Sign2Text	我们不能做自私自利的事情。 (We cannot do **selfish** things.)
Example (d)	
Gloss (GT)	今天 / 菜 / 咸 / 我 / 想 / 喝 / 糖 (Today / Food / Salty / I / Want to / Drink / **Sugar**)
Text (GT)	今天的菜好咸，我想喝饮料。 (The food is quite salty today and I want to drink a **drink**.)
Gloss2Text	今天菜很咸,我想喝糖。 (The food is quite salty today and I want to drink some **sugar**.)
Sign2Text	今天的菜真咸,我想喝饮料。 (The food is quite salty today and I want to drink a **drink**.)

Table 4. Qualitative analysis on PHOENIX-2014T (Example (a) and (b)) and CSL-Daily (Example (c) and (d)) datasets.

Example (a)	
Text (GT)	Auch am Sonntag ist es neben Wolkenfeldern teilweise freundlich und trocken bei minus fünf bis plus zwei Grad. (On sunday, too, it is partly friendly and dry in addition to cloud fields, with minus five to plus **two** degrees.)
Sign2Text	Am Sonntag teils wolkig teils freundlich und trocken minus fünf bis plus drei Grad. (On sunday partly cloudy partly friendly and dry minus five to plus **three** degrees.)
Example (b)	
Text (GT)	Also wir haben morgen HöchstTemperaturen in Mitteleuropa von achtzehn bis sechsundzwanzig Grad. (So tomorrow we have maximum temperatures of **eighteen** to twenty-six degrees in central Europe.)
Sign2Text	Morgen Temperaturen von fünfzehn Grad in Mitteleuropa bis sechsundzwanzig Grad in der Lausitz. (Tomorrow temperatures from **fifteen** degrees in central Europe to twenty-six degrees in **lausitz**.)
Example (c)	
Text (GT)	他今年四岁。 (He is **four** years old.)
Sign2Text	他今年三岁。 (He is **three** years old.)
Example (d)	
Text (GT)	我们不但要从成功中总结经验，还要从失败中吸取教训。 (We must not only draw lessons from our success, but also learn from our failures.)
Sign2Text	我们要善于吸取失败的教训。 (We must be good at learning from our failures.)

Table 5. Failure cases from PHOENIX-2014T (Example (a) and (b)) and CSL-Daily (Example (c) and (d)) datasets. Our method has difficulty in recognizing numbers and location entities due to their low frequency in the training corpus.