Supplementary for "Gloss-free Sign Language Translation: Improving from Visual-Language Pretraining"

A. More Implementation details

GFSLT Model Table 1 presents detailed information on the GFSLT model structure and feature sizes for each module. The input sign video, which may have varying lengths, is padded to the longest length and loaded into a batch. After ResNet [2] processing without a fully connected (FC) layer, the resulting visual feature has a size of $B \times T \times 512$. Two temporal modules, each consisting of Conv1D-BN1D-RELU-MaxPooling1D, are used to capture the short-term dependencies in the sign video, yielding features of size $B \times T/4 \times 1024$. These features are then passed through an MLP and a Transformer Encoder to prepare for decoding. In the decoder, the text inputs are first padded to a uniform length of U and passed through a Word Embedding Layer to obtain features of size $B \times U \times 1024$. The Transformer Decoder takes the outputs of the Transformer Encoder and the Word Embedding to generate one word at a time, and an FC layer is used to obtain the final prediction word.

B. More Ablation Studies

B.1. Impact of Mask Rate.

We adopt a token masking strategy in our approach similar to that used in Bert [1]. Specifically, we randomly replace $\rho\%$ of the tokens in a sentence using the following criteria: (i) 80% of these tokens are replaced with the special [Mask] token, and (ii) 10% are replaced with any other token, while the remaining 10% of the tokens are kept intact. As shown in Table 2, our experiments reveal that the optimal BLEU-4 score is achieved with a masking rate of 15%, which is consistent with the rate used in Bert. Interestingly, we observe that increasing or decreasing the masking rate does not yield significant benefits. This result could be attributed to the fact that the proposed approach, VLP, places more emphasis on pre-training the Visual Encoder than the Text Decoder.

B.2. Impact of Loss weight

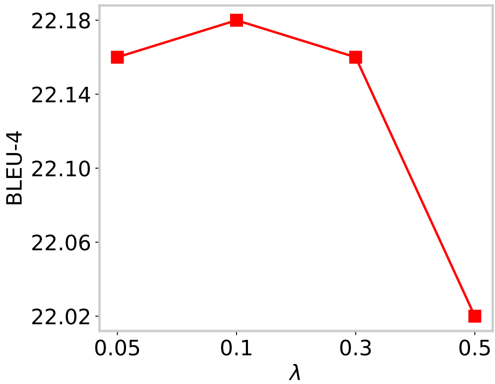

Figure 1: Impact of loss weight coefficient λ on network performance.

In fact, Text Decoder can be updated jointly or in stages. When updating jointly, the loss in the first stage consists of the following two parts:

$$\mathcal{L}_{total} = \mathcal{L}_s + \lambda \mathcal{L}_c \tag{1}$$

where λ is a scalar weight. In this experiment, we studied the effect of the loss weight coefficient λ on the pre-trained model. As illustrated in Figure 1, the influence of λ on the model performance is relatively minor, with performance fluctuations staying around ± 0.1. However, as λ increases, the model's performance begins to decline, indicating that it is not always beneficial to amplify the influence of the Text Decoder on VLP. As a result, we set λ to 0.1 in this paper.

B.3. Investigation VLP on CSL-Daily

We also conducted VLP and strong data augmentation ablation experiments on CSL-Daily. As shown in Table 3, the translation performance improved with VLP, and adding strong data augmentation in Stage 1 further helped. However, the model performance decreased when strong data augmentation was added only in Stage 2 without VLP. The best result was achieved when strong data augmentation was added to both stages. This finding is consistent with the results of our experiments on Phoenix14T.

Module	Stride	Kernel	Output Size
Sign Input	-	-	$B \times T \times 224 \times 224 \times 3$
Resnet wo/ fc	-	-	$B \times T \times 512$
Conv1D-BN1D-RELU	1	5	$B \times T \times 1024$
MaxPooling1D	2	2	$B \times T/2 \times 1024$
Conv1D-BN1D-RELU	1	5	$B \times T/2 \times 1024$
MaxPooling1D	2	2	$B \times T/4 \times 1024$
Linear-BN1D-RELU	-	-	$B \times T/4 \times 1024$
Transformer Encoder	-	-	$B \times T/4 \times 1024$
Text Input	-	-	$B \times U$
Word Embedding	-	-	$B \times U \times 1024$
Transformer Decoder	-	-	$B \times U \times 1024$
FC	-	-	$B \times U \times C$

Table 1: Detailed Gloss-Free SLT(GFSLT) Framework. B means batch size. T means the lengths of the longest input sign video in the batch. U means the lengths of the longest input text in the batch.

mask rate (ρ)	**Dev**				**Test**			
	BLEU1	**BLEU-2**	**BLEU-3**	**BLEU-4**	**BLEU-1**	**BLEU-2**	**BLEU-3**	**BLEU-4**
10%	43.30	33.05	26.04	22.03	43.54	32.90	25.61	20.84
15%	44.08	33.56	**26.74**	**22.12**	43.71	**33.18**	**26.11**	**21.44**
20%	**44.15**	**33.72**	26.35	22.07	**43.85**	33.08	25.97	21.32

Table 2: Effect of mask rate for network performance. The gray box represents the mask rate we finally adopted in this paper.

VLP	**Aug-S1**	**Aug-S2**	**Dev**				**Test**			
			BLEU-1	**BLEU-2**	**BLEU-3**	**BLEU-4**	**BLEU-1**	**BLEU-2**	**BLEU-3**	**BLEU-4**
✗	✗	✗	37.60	23.30	14.89	9.92	37.69	23.28	14.93	9.88
✓	✗	✗	37.38	23.26	14.91	9.97	37.84	23.60	15.23	10.29
✓	✓	✗	38.34	24.13	15.56	10.32	38.31	23.80	15.33	10.27
✗	✗	✓	34.36	21.00	13.50	9.14	34.07	20.77	13.40	9.03
✓	✓	✓	**39.20**	**25.02**	**16.35**	**11.07**	**39.37**	**24.93**	**16.26**	**11.00**

Table 3: Effect of VLP and data augmentation strategies on CSL-Daily dataset. VLP: Visual-Language Pre-training, Aug-S1: strong data augmentation employed during stage 1 for sign video, Aug-S2: strong data augmentation employed during stage 2 for sign video.

References

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 1

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1