DIFFSIGN: A DIFFERENT APPROACH TO CONTINUOUS SIGN LANGUAGE RECOGNITION WITH DIFFUSION MODEL
SUPPLEMENTARY MATERIAL

Anonymous authors
Paper under double-blind review

1 MORE ABLATION STUDIES

1.1 STUDY ON DIFFERENT VISUAL ENCODERS

Since we use a video-to-text generation approach to accomplish the CSLR task, the ability to adapt to different types and depths of visual encoders is crucial for our method. To illustrate that our enhancements are not reliant on any specific encoder, we examine a range of different visual encoders. As shown in Table 1, using a deeper network does not significantly affect the final recognition accuracy. This result reinforces the idea that the improvements in recognition stem from our design and further validate the generality of our approach.

Table 1: Comparison with different visual encoders on the CSL-Daily dataset.

Visual encoder	Dev(%)	Test(%)
ResNet18	24.3	24.0
ResNet50	24.4	23.9
ResNet102	24.2	24.0
ResNet152	24.5	23.8
I3D	24.3	24.1
S3D	24.5	23.8
ResNet34	24.3	23.9

1.2 STUDY ON DATA AUGMENTATION

We apply three data augmentation techniques during training: random cropping, horizontal flipping, and random temporal scaling, in line with the methods described in Hu et al. (2023). In Table 2, we evaluate the effect of each strategy. The findings show that data augmentation greatly improves performance, especially with random cropping. We propose that the network might be utilizing shortcuts, like concentrating on the hands' exact positions in video frames. By incorporating random cropping, we promote the network's ability to learn more abstract, high-level features, which helps diminish its reliance on these shortcuts.

Table 2: Ablations for data augmentation strategies on the CSL-Daily dataset.

Crop	Flip	Temporal Scaling	Dev(%)	Test(%)
✘	✘	✘	31.5	30.9
✔	✘	✘	**26.4**	**25.9**
✘	✔	✘	27.7	27.1
✘	✘	✔	30.2	29.5
✔	✔	✘	25.9	25.2
✔	✔	✔	**24.3**	**23.9**

1.3 STUDY ON DIFFERENT SAMPLING ROUNDS

Table 3 compares the impact of our default configuration sampling for 50 rounds with different sampling rounds on the accuracy (WER) of generated gloss sequences. As mentioned in the main text, DDM used for generating images typically requires sampling hundreds of rounds, with the middle and later stages mainly focused on restoring low-level visual features such as image textures. These low-level features are redundant for generating gloss sequences. Therefore, setting the sampling rounds too high will result in the generated gloss features containing excessive redundant information, leading to an increase in WER. Conversely, if the sampling rounds are set too low, the generated gloss features will contain a lot of noise, which will also affect the accuracy of downstream recognition tasks. Therefore, we ultimately chose to sample 50 rounds for its best performance.

Table 3: Comparison with different sampling rounds on the CSL-Daily dataset.

Sampling rounds	Dev(%)	Test(%)
10	33.4	33.0
25	26.2	25.7
100	25.4	24.6
250	26.5	26.0
50	24.3	23.9

2 MORE VISUALIZATIONS

2.1 VISUALIZATION FOR ATTENTION MATRICES OBTAINED IN GLOSS-LEVEL FEATURE ENHANCEMENT

In the gloss-level feature enhancement strategy, we optimize visual features by applying gloss-level contrastive learning to the attention matrix between the visual embeddings and gloss tokens. This is done in order to alleviate the inherent ambiguity and uncertainty issues of visual features. As illustrated in Fig.1, the pre-GFE similarity matrix comprises numerous embeddings that are highly similar to multiple gloss tokens. However, following the application of GFE, each embedding is distinctly associated with a gloss token, and the gaps between adjacent sign language actions are also clarified. The application of gloss-level contrastive learning enables the precise localisation and identification of each sign action, facilitating the accurate recognition of those that were previously overlooked or misidentified.

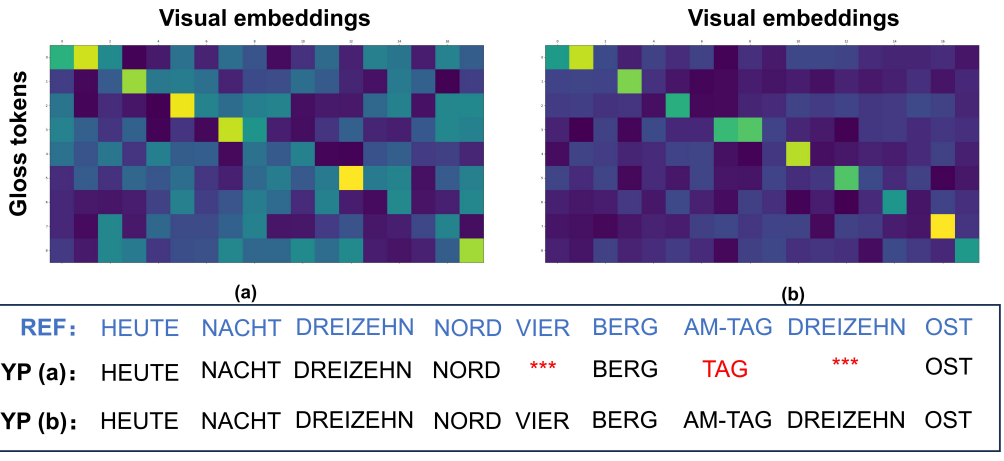

Figure 1: visualization for attention matrices obtained in gloss-level feature enhancement (GFE) at the last sampling round. (a) Before applying GFE. (b) After applying GFE.

3 ADDITIONAL DETAILS

3.1 THE DATA SCARCITY PROBLEM IN CSLR

Many previous methods Gan et al. (2023); Cheng et al. (2023) apply cross-modal contrastive learning at the sentence-level to improve the alignment globally. The data-hungry nature of contrastive learning makes it challenging to achieve high-quality contrastive learning in CSLR tasks with data scarcity issues. Applying contrastive learning at the gloss-level not only helps to achieve more fine-grained feature optimization but also serves as a means of data augmentation to address data scarcity issues. Compared to Gan et al. (2023); Cheng et al. (2023), which conducts contrastive learning at the sentence-level, our gloss-level feature representation enhancement with contrastive learning allows a single sign language video to generate multiple gloss-level samples. This significantly increases the number of samples and addresses the issue of data scarcity in the CSLR task. For a visual feature sequence of length N and a gloss feature sequence of length $N^{'}$, we can obtain $N \times N^{'}$ gloss-level sample pairs to accomplish sufficient contrastive learning.

REFERENCES

Yiting Cheng, Fangyun Wei, Jianmin Bao, Dong Chen, and Wenqiang Zhang. Cico: Domain-aware sign language retrieval via cross-lingual contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19016–19026, 2023.

Shiwei Gan, Yafeng Yin, Zhiwei Jiang, Kang Xia, Lei Xie, and Sanglu Lu. Contrastive learning for sign language recognition and translation. In *IJCAI*, pp. 763–772, 2023.

Lianyu Hu, Liqing Gao, Zekang Liu, and Wei Feng. Continuous sign language recognition with correlation network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2529–2539, 2023.