1. Supplementary materials

In the supplementary material, we provide details and extra experiments that are not shown in the main paper.

2. Effects of the local and global temporal context perception for sign recognition

As depicted in Fig. 1 of the main paper, the temporal aggregation module in baseline framework incorporates both local and global temporal context perception. This baseline, which combines local-global temporal context perception, has been widely adopted in recent works [4, 6, 10–12, 19]. Specifically, the local temporal context perception captures adjacent frames to better identify isolated signs [3, 11], while the global temporal context perception emphasizes sequence correlation [4, 8, 13].

As mentioned in Sec.3.3 of the main paper, we conducted a simple ablation study on the widely-used RWTH-2014 dataset [9] to demonstrate the importance of incorporating both local and global temporal context perception in the temporal aggregation module. In practice, following the baseline (as detailed in Sec.3.1 of the main paper) we remain either 1D-TCN (CSLR-LocTAM) or BLSTM (CSLR-GloTAM) respectively in the temporal aggregation module, as detailed in Sec.4.1 of the main paper. Notably, our baseline, CSLR-LocTAM, and CSLR-GloTAM have the same structure with [6, 10, 11, 19]. Results in Tab. 1 indicate that coupling 1D-TCN and BLSTM in the temporal aggregation module is among the top-performing ones. Therefore, we confirm the necessity of incorporating both local and global temporal context perception in temporal aggregation module, as also highlighted in previous leading studies.

3. Ablation on language models

In this section, we employ three different language models for CTCA to evaluate the language model's importance. As shown in the Tab. 2, varying the language model represents similar WERs, but CTCA with the pre-trained BERT obtains the best result. This result shows the semantic correlation among glosses extracted by the pre-trained BERT is more effectively.

4. Experiments about chain depth variants of BLSTM in the general CSLR framework

To evaluate the chain depth influence mentioned in Sec.3.2 of the main paper, we also consider about whether to increase the number of layers of BLSTM. As shown in Table 3, with more BLSTM layers, the IIW, as well as the WER values, become worse. For these results we consider that previous works [6, 13] have presented that in general CSLR network the CTC loss provides a limited contribution to the learning of the spatial perception module(SPM), which the penalty for SPM is hard to conduct from the BLSTM, due to the chain rules of back-propagation. This makes the BLSTM is prone to over fit on the sequential order of sign actions [3, 6]. Therefore, increasing the number of BLSTM layers makes the situation worse.

Table 1. Ablation results (WER, %) of local and global temporal context perception for temporal aggregation module of the baseline on RWTH-2014 [9]. "Baseline" is followed the Fig 1 of the main paper, and the "CSLR-LocTAM" and "CSLR-GloTAM" has been introduced in Sec.4.1 of the main paper.

Methods	Dev		Test	
	del/ins	WER	del/ins	WER
Baseline	7.0/3.0	21.8	6.7/2.7	22.1
CSLR-LocTAM	9.4/3.1	24.2	9.2/2.9	23.8
CSLR-GloTAM	8.4/3.5	23.0	8.5/2.9	23.1

Table 2. Results of different language models on the RWTH-2014.

Methods	Dev(WER)	Test(WER)
3-layer BLSTM	19.9	20.5
6-layer Transformer	19.8	20.7
pre-trained BERT	**19.5**	20.1

5. Experiments about model ensemble

In Tab. 4, we conduct the ensemble by averaging outputs of global & local temporal perception branch. The ensemble model obtains a similar but slightly worse performance. For this phenomenon, we explain that the proposed CTCA focuses on exploring a desired temporal aggregation module, and we find that it should be a shallow architecture to allow more effective training of spatial perception module but also should be a deep one for a high temporal aggregation capability. Therefore, we conduct a cross-temporal context aggregation (CTCA) to transfer the local temporal context and the linguistic prior to the global perception module. This operation causes the global perception module is more strong than the local. Therefore, when averaging their probabilities, the final probability might be worse.

Table 3. Baseline CSLR with different numbers of BLSTM layers.

Methods	SPM-IIW	TAM-IIW	Dev	Test
1-layer(2048)	6.4E-5	8.3E-5	21.5	22.5
2-layers(1024)	5.8E-5	7.4E-5	21.3	22.2
7-layers(512)	7.4E-4	1.6E-4	26.1	27.2
28-layers(256)	3.8E-4	6.1E-4	97.6	97.9

Table 4. Ensemble of local & global branches on the RWTH-2014.

Methods	Dev(WER)↓	Test(WER)↓
CTCA	19.5	20.1
Ensemble	20.4	22.0

6. Complement the implementation details

In this section we will introduce the main implementation details. Unless otherwise specified we use the ResNet18 as initialization for the spatial perception module, and the batch size and weight decay are set to 4 and $1e-4$, respectively. In addition, we train our CTCA for 100 epochs with the Adam optimizer [7] and an initial learning rate of $1e-4$, the learning rate is decayed (0.1) at 30 and 60 epochs. Specifically, in the CSL-Daily [18] benchmark, the initial learning rate and weight decay are set to $5e-5$ and $1e-6$, respectively. And the learning rate is decayed (0.5) at 30 and 60 epochs.

Furthermore, for the gloss feature embedding we exploit pre-trained German and Chinese BERT model [2, 5] as the BERT initialization for RWTH benchmarks [1, 9] and CSL-Daily benchmark, respectively.

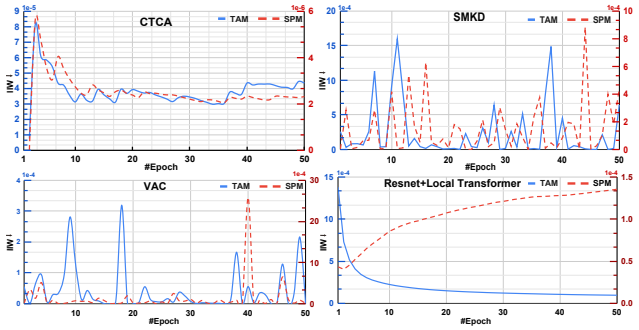

Figure 1. IIW trends of SPM & TAM of CTCA, SMKD, VAC and local transformer. Similar to Fig. 2 of the main paper, the left y-axis (blue) presents IIW values of SPM, and the right (red) y-axis contains IIW values of TAM.

7. Qualitative complementary experiments

The IIW trends and analysis. In this section we extensively study the limitations and desirable properties of Temporal Aggregation Module (TAM) via IIW. As shown in Fig. 1, IIW trends of SPM & TAM of two SOTA works VAC and SMKD are not consistent and keep high IIW values during the training process, inferring that SPM has low generalization and is not well-trained. In addition. we also use IIW to analyze the local-global-mixed architecture, local transformer in [16, 19], which the resnet18 is the SPM, the local transformer is the TAM. In Fig. 1, the local-global-mixed architecture's TAM can be effectively optimized with

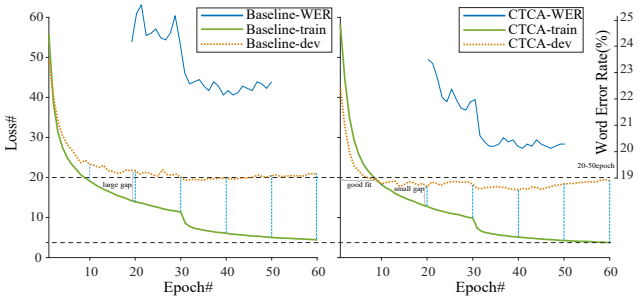

Figure 2. The visualization of training loss curve, test loss curve and the WER(%) values of the baseline and the CTCA on the RWTH-2014 test set.

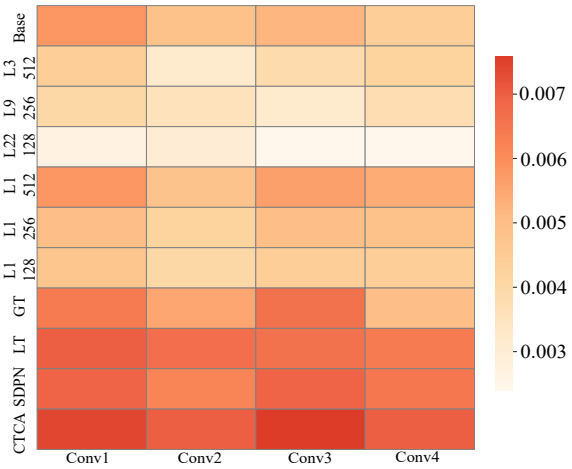

Figure 3. Statistics of layer-wised gradients of different method's spatial perception module. "Base", "GT", "LT", "SDPN" and "CTCA" , denote the Baseline, the CSLR-GloTAM, the CSLR-LocTAM, the SPM-shared dual-path network, and the CTCA, respectively. Lα denotes the α chain depth of temporal aggregation module. "512" is the channel number.

decreasing IIW values while its SPM suffers from insufficiently trained, *i.e.*, the IIW value gradually increases.

Above all, these analysis also infers that TAM should be shallow but has high local & global aggregation capability. To this end, we propose CTCA that contains a dual-path network and the cross-context knowledge distillation loss function. The dual-path network consists of a shared SPM, and two parallel branches for GloTAM and LocTAM. The LocTAM is removed during inference. The proposed loss function can transfer the local temporal context and the linguistic prior to the global perception module. With our contributions, IIW trends of SPM & TAM are towards the desired style [15], that is, they keep similar trends and go down, thus relieving the generalization problem mentioned in Fig 2 of the main paper.

The train-test losses gap. Besides counting the information stored in weights (IIW) [15] in the main paper

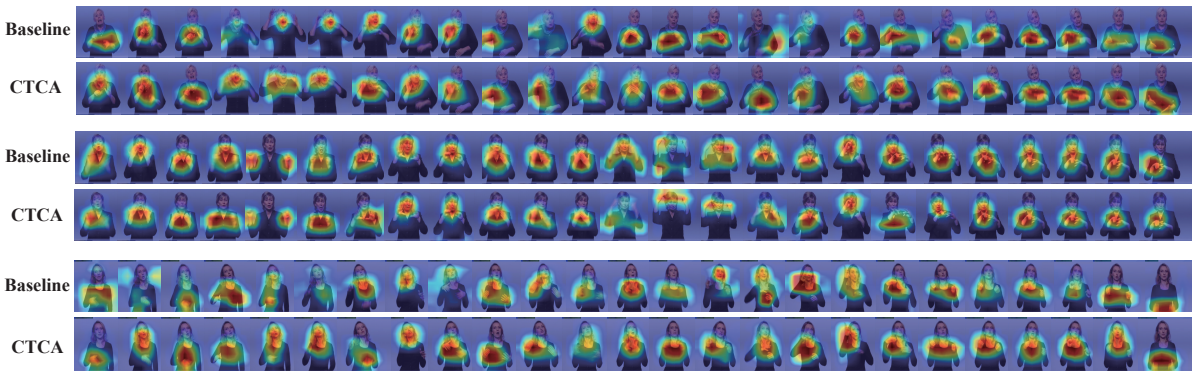

Figure 4. Ablation on GradCAM [14] visualization. Samples come from the test set of the RWTH-2014. Colors of images change from blue to yellow and to red, meaning the model will pay higher attention to the regions.

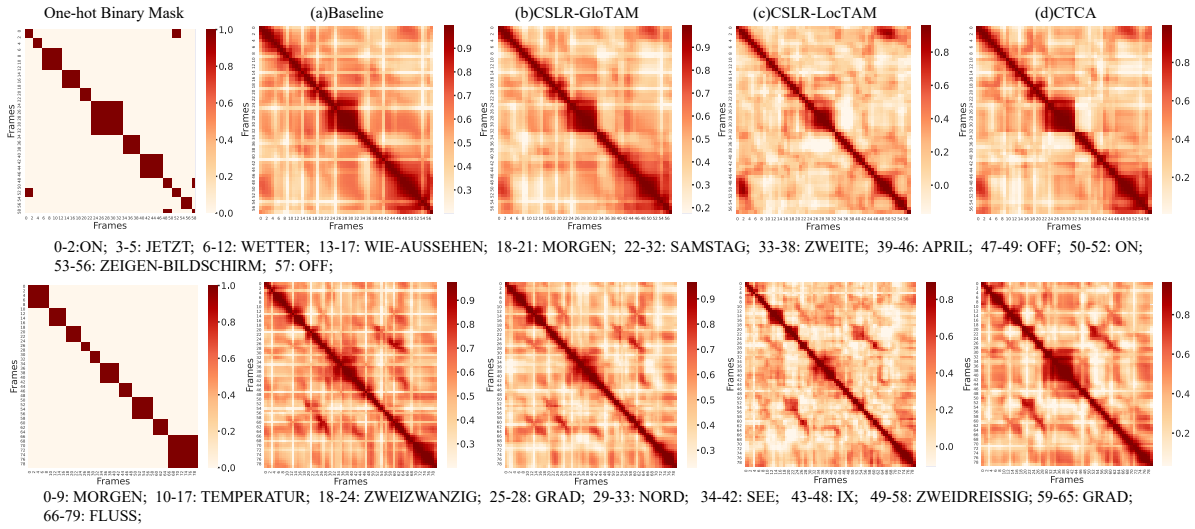

Figure 5. Two test samples are chosen for self-similarity matrices heatmaps visualization of temporal aggregation module's representations (the reder color represents the higher similarity).

Fig 1, we also employ the complementary train-test losses gap of the model to verify the model generalization ability in Fig. 2. In Fig. 2, the CTCA has better model generalization ability and performance than the baseline, which has lower WER values and a lower gap between the training loss curve and the test loss curve, especially before the 10-th epoch the CTCA achieves a good fitting.

The influence of temporal aggregation module chain depth on the spatial perception module. Motivated CAM methods [14,17] propose that if the magnitude of gradient of the layer is larger, it provides more current category information. In this section, we visualize the statistics of layer-wised gradients of SPM to indicate the influence of chain depth of TAM to the SPM, complementary. In practice, we visualize the summation gradient of each convolution layer at the last block in SPM(owning informative category information) for all RWTH-2014 training samples. We adopt the model of fixed epoch *i.e.*, 35-th.

In Fig. 3, under the same experiments setting with Sec.3.2, results of distinct chain depths TAMs are shown. As the TAM chain depth increases the magnitude of gradients of SPM decreases, which indicates that the SPM with large chain depth TAM is unable to pay more attention to recognize and locate the current signs leading to sub-optimization. In addition, the CSLR-GloTAM and the CSLR-LocTAM can be regarded as extreme cases of TAM reducing the chain depth. Fig. 3 also shows that the magnitude of gradients of the CSLR-GloTAM, CSLR-LocTAM, and SPM-shared dual-path network (SDPN) is large than the baseline, which also indicates that reducing TAM chain depth is a benefit to the SPM discriminating signs.

T-SNE visualization of spatial perception module representations. To explore whether the spatial perception module representations can be enhanced by the SDPN and

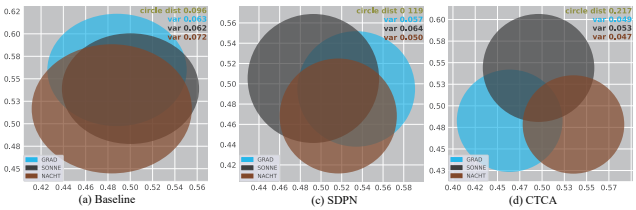

Figure 6. T-SNE visualization of distinct methods' spatial perception module representations. Each circle is plotted by the intra-glosses mean (circle center) and intra-glosses variance (circle radius) of SPM representations, which covers all samples of corresponding glosses. The "circle dist" and "var" are the distance of three circle centers and the variances of three circles, respectively.

CTCA, we also provide T-SNE visualization of SPM representations of top-3 high-frequency glosses from the RWTH-2014 training set. If the "circle dist" is large indicating less coincidence of inter-glosses representations distribution, and if the "var' is small denoting the distribution of the intra-glosses representation is compact. As shown in Fig. 6, the "circle dist" and "var" of Fig. 6 (b) (SDPN) and Fig. 6 (c) (CTCA) are larger and smaller than Fig. 6 (a) (Baseline), respectively. Especially three circles in Fig. 6 (c) are the smallest and highly distinguishable. These results enumerate the architecture of SDBN is benefit to SPM discriminating signs, and demonstrate that the cross-context knowledge distillation loss can enhance the SPM representations power effectively.

GradCAM visualization of spatial perception module rep- resentations. Moreover, we also adopt the Grad-CAM [14] to show spatial activations of signs to demonstrate the SPM generalization ability. In practice, feature maps with shape of 7×7 from the last layer at the stage 4 of ResNet18 will be employed to compute spatial activations. As shown in Fig. 4, we obvious that the SPM of CTCA is able to locate regions of signs occurrence more precisely than the baseline, which shows the strong generalization ability of CTCA's SPM.

The visualization of the self-similarity matrices heatmaps. To further illustrate the local-global temporal reception context of CTCA qualitatively, we follow the SMKD [6] to visualize the self-similarity matrices heatmaps of 1D-TCN branch and BLSTM branch representations. In the Fig. 5, we choose some videos in the RWTH-2014 [9] test set and can see that the self-similarity matrix heatmap of CTCA (d) is able to regard as the fusion of 1D-TCN (b) and BLSTM (c), which means the BLSTM branch perceiving local-global temporal context.

References

[1] Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. Neural sign language translation. In *CVPR*, 2018. 2

[2] Branden Chan, Stefan Schweter, and Timo Möller. German's next language model. In *ICCL*, 2020. 2

[3] Ka Leong Cheng, Zhaoyang Yang, Qifeng Chen, and Yu-Wing Tai. Fully convolutional networks for continuous sign language recognition. In *ECCV*, 2020. 1

[4] Runpeng Cui, Hu Liu, and Changshui Zhang. Recurrent convolutional neural networks for continuous sign language recognition by staged optimization. In *CVPR*, 2017. 1

[5] Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. Revisiting pre-trained models for Chinese natural language processing. In *EMNLP*, 2020. 2

[6] Aiming Hao, Yuecong Min, and Xilin Chen. Self-mutual distillation learning for continuous sign language recognition. In *ICCV*, 2021. 1, 4

[7] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 2

[8] Oscar Koller, Necati Cihan Camgoz, Hermann Ney, and Richard Bowden. Weakly supervised learning with multi-stream cnn-lstm-hmms to discover sequential parallelism in sign language videos. *PAMI*, 42(9), 2019. 1

[9] Oscar Koller, Jens Forster, and Hermann Ney. Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers. *Computer Vision and Image Understanding*, 141, 2015. 1, 2, 4

[10] Zekang Liu Lianyu Hu, Liqing Gao and Wei Feng. Temporal lift pooling for continuous sign language recognition. In *ECCV*. Springer, 2022. 1

[11] Yuecong Min, Aiming Hao, Xiujuan Chai, and Xilin Chen. Visual alignment constraint for continuous sign language recognition. In *ICCV*, 2021. 1

[12] Junfu Pu, Wengang Zhou, Hezhen Hu, and Houqiang Li. Boosting continuous sign language recognition via cross modality augmentation. In *ACMMM*, 2020. 1

[13] Junfu Pu, Wengang Zhou, and Houqiang Li. Iterative alignment network for continuous sign language recognition. In *CVPR*, 2019. 1

[14] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, 2017. 3, 4

[15] Zifeng Wang, Shao-Lun Huang, Ercan Engin Kuruoglu, Jimeng Sun, Xi Chen, and Yefeng Zheng. Pac-bayes information bottleneck. In *ICLR*, 2022. 2

[16] Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V. Le. Qanet: Combining local convolution with global self-attention for reading comprehension. In *ICLR*, 2018. 2

[17] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *CVPR*, pages 2921–2929, 2016. 3

[18] Hao Zhou, Wengang Zhou, Weizhen Qi, Junfu Pu, and Houqiang Li. Improving sign language translation with monolingual data by sign back-translation. In *CVPR*, 2021. 2

[19] Ronglai Zuo and Brian Mak. C2slr: Consistency-enhanced continuous sign language recognition. In *CVPR*, 2022. 1, 2