A. Overview

In this supplementary material, we provide details and extra experiment that are not shown in the main paper. Firstly, we provide more visualization results about the similarity matrix (Sect. B). Then we introduce the details of the proposed GSBA method (Sect. C). At last, we give the extra experiment to explore the influence of the receptive field (Sect. D).

B. Visualization of the Similarity Matrix

In Sect.3.4, we give an instance to explore the property of the features after sharing the weights. As shown in Fig. 3, we also visualize the similarity matrices in different cases: without WS, with WS, and with WS & GSBA. We can find that only with the WS method, the GCF of a key frame will have a strong correlation with the LVFs of the frames nearby it. Meanwhile, with the gloss segmentation, the GCF of a key frame will focus on more LVFs nearby it. More visualization results on the training set and the development set are given in Fig. 4. And we find that all of them show a similar characteristic with the instance we gave in Sect.3.4.

C. Gloss Segment Boundary Assignment

The pseudo-code of GSBA is described in Algorithm 1. It consists of two functions: LOCATE and EXPAND.

For the LOCATE function, it is used to locate the key frame of the current class $c_i$. We scan the predict probability distribution $\tilde{Y}_g$ to get the location $t$.

For the EXPAND function, it is used to expand the key frame. We treat the located key frame as an anchor frame. We first set an expanded radius $d$ to limit the maximum expansion distance and a direction $s \in \{-1, 1\}$ to determine the expand direction. Then, we expand the frame from $t + s$ to $t + s + d$ frame. If the cosine similarity between the GCF of the current expanding frame and the weight vector of $c_i$ is the smallest among the classes $c_j \in I$, we then annotate this frame with the label $c_i$. Otherwise, we will stop the expansion process.

---

Algorithm 1 Gloss Segment Boundary Assignment

**Input:** video’s GCF sequence $G$, ground truth sign gloss sequence $l$, predict probability distribution $\tilde{Y}_g$, classifier weight vectors $W$, expanded radius $d$.

**Output:** pseudo gloss segment labels $y^{seg}$

1. function LOCATE($l_i$, pos)
2. for $t \leftarrow pos; t \leq |G|$ do
3. $c_t \leftarrow \arg\min\{\tilde{y}_t\}$
4. if $c_t == l_i$ then
5. pos $\leftarrow t$
6. return $t$, pos
7. else
8. break
9. end if
10. end for
11. end function

12. function EXPAND($t$, $l_i$, $s$)
13. for $j \leftarrow 1; j \leq d$ do
14. $c_{t+js} \leftarrow \arg\min\{\langle g_t, w_{c_j} \rangle\}_{c_j \in I}$
15. if $c_{t+js} == l_i$ then
16. $y^{seg} \leftarrow \{t + js, l_i\}$
17. else
18. break
19. end if
20. end for
21. end function

22. for $i \leftarrow 1; i \leq |G|$ do
23. $y^{seg} \leftarrow \{i, blank\}$
24. end for

25. pos $= 1$
26. for $i \leftarrow 1; i \leq |l|$ do
27. $t, pos \leftarrow$ LOCATE($l_i$, pos)
28. $y^{seg} \leftarrow \{t, l_i\}$
29. EXPAND($t$, $l_i$, $-1$)
30. EXPAND($t$, $l_i$, $1$)
31. end for
32. return $y^{seg}$
Based on the two functions mentioned above, we first initialize each frame’s label as the blank class. Then we use the LOCATE function to locate the key frames for each class $l_i$ in the ground truth sign gloss sequence $l_i$. After that, we use the EXPAND function to expand the key frames and update the pseudo gloss segment labels $Y^{seg}_i$. Then we smooth the $Y^{seg}_i$, and get the smoothed labels $\tilde{Y}^{seg}_i$ as:

$$\tilde{y}_{ij} = \begin{cases} 1 - \varepsilon & \text{if } j = y_{ij}^{seg} \\ \varepsilon \frac{|T|}{|T|+1} & \text{otherwise}, \end{cases}$$

where $\varepsilon$ is the label smoothing rate. More visualizations of the pseudo gloss segment label produced by GSBA are shown in Fig. 5.

Note that, we active the GSBA after epoch 20 to avoid the unreliable segment proposal at the initial optimization stage. And we enlarge the expanded radius $d$ after the training of the contextual module tends to steady to introduce more spatial-temporal information.

**D. Details on Temporal Receptive Field**

We define $r$ as the temporal receptive field (TRF) of the visual module. As $r$ is relevant to a temporal window in the contextual module, and suitable size of the visual module’s TRF will better match the followed contextual module. As shown in Fig. 1 and Fig. 2, we visualize the self-similarity matrices of the visual module with different TRFs and find that the self-similarity matrix with larger TRF tends to be more diagonal. The blank and non-blank features will be hard to distinguish if the TRF is small, and this will increase the difficulty of aligning features from the two modules. Moreover, we compare the performances with different TRF of the visual module as shown in Table 1. We observe that small $r$ will results in performance deterioration due to the loss of temporal information. Besides, the overuse of the pooling operation will also harm the performance. Among the selected TRFs, the optimal structure is $C_5 - P_2 - C_5$. So we set the visual module as $C_5 - P_2 - C_5$ by default.

<table>
<thead>
<tr>
<th>Visual Module</th>
<th>TRF</th>
<th>Dev (%)</th>
<th>Test (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$C_5$</td>
<td>5</td>
<td>22.5</td>
<td>23.3</td>
</tr>
<tr>
<td>$C_5 - C_5$</td>
<td>9</td>
<td>22.4</td>
<td>23.2</td>
</tr>
<tr>
<td>$C_5 - P_2 - C_5$</td>
<td>14</td>
<td><strong>21.2</strong></td>
<td><strong>21.4</strong></td>
</tr>
<tr>
<td>$C_5 - P_2 - C_5 - P_2$</td>
<td>16</td>
<td>22.1</td>
<td>22.7</td>
</tr>
</tbody>
</table>

Table 1. Ablation studies on the TRF of the visual module on the PHOENIX14 dataset (only trained in the synchronous training stage).

Figure 1. The heatmaps of LVFs’ self similarity matrix with receptive fields 5.

Figure 2. The heatmaps of LVFs’ self similarity matrix with receptive fields 9.
Figure 3. The heatmap of the LVFs’ and GCFs’ self-similarity matrices and the similarity matrix between the LVFs and the GCFs (the darker color represents the higher similarity). From top to bottom are the results that network training without WS, with WS and with WS & GSBA.
Figure 4. The heatmap of the LVFs’ and GCFs’ self-similarity matrices and the similarity matrix between the LVFs and the GCFs in different examples (the darker color represents the higher similarity).
(a) Example1.  
(b) Example2.  
(c) Example3.  
(d) Example4.  
(e) Example5.  
(f) Example6.  
(g) Example7.  
(h) Example8.  
(i) Example9.

Figure 5. From top to bottom are the spike phenomenon and the pseudo gloss segment labels produced by GSBA with $d = 1, 2, 3$ (different colors represent different classes).