Natural Language-Assisted Sign Language Recognition Supplementary Material

A. More Implementation Details

Bidirectional Lateral Connections. We apply bidirectional lateral connections [1] to the first four S3D blocks for video-video, keypoint-keypoint, and video-keypoint information exchange. For video-keypoint connections (dashed lines in Figure 4 in the main paper), since the input spatial resolutions of the video and keypoint encoder are 224×224 and 112×112, respectively, we use 2D convolution (Figure 6a) and transposed convolution layers (Figure 6b) with a stride of 2 and a kernel size of 3×3 to match the spatial resolutions. For video-video and keypoint-keypoint connections (dotted dashed lines in Figure 4 in the main paper), due to the input length difference, we use 1D convolution (Figure 6c) and transposed convolution layers (Figure 6d) with a stride of 2 and a kernel size of 3 to match the temporal resolutions. Figure 6 shows the bidirectional lateral connections.

Keypoint Illustration. We show the keypoints used in our VKNet in Figure 7. The keypoints are estimated by HRNet [9] trained on COCO-WholeBody [3]. We use a subset of keypoints including 11 upper body keypoints, 10 mouth keypoints, and 42 hand keypoints.

B. More Experiments

Head Choices for Inter-Modality Mixup. By default, we apply our inter-modality mixup on all head networks. To validate the effectiveness of this setting, we further conduct experiments on only applying it on partial heads. We categorize the head networks into three groups: video heads with input features $(\boldsymbol{f}^V_{64}, \boldsymbol{f}^V_{32})$, keypoint heads with input features $(\boldsymbol{f}^K_{64}, \boldsymbol{f}^K_{32})$, and joint heads with input features $(\boldsymbol{f}_{64}, \boldsymbol{f}_{32}, \boldsymbol{f})$. See Figure 4 in the main paper for their definitions. Table 10 shows that applying the inter-modality mixup on either one group of heads outperforms the baseline, and our default setting, applying the inter-modality mixup on all heads, achieves the best performance.

Keypoint Selection. We utilize HRNet [9] trained on COCO-WholeBody [3] to estimate 63 keypoints (11 for upper body, 42 for hands, and 10 for mouth) per frame. As shown in Table 11, we validate the effectiveness of each keypoint group by training several single-stream keypoint encoders. Only using upper body keypoints yields the lowest top-1 accuracy (21.37%). Employing hand keypoints significantly improves the top-1 accuracy by 27.17%. This result is also consistent to the fact that sign languages mainly convey information by signers' hand movement. Finally, the mouth keypoints also have a positive effect since signers usually mouth the words during signing.

Bidirectional Lateral Connections. Within the VKNet, we apply bidirectional lateral connections [1] for video-video, keypoint-keypoint, and video-keypoint information

Video	Keypoint	Joint	Per-instance		Per-class	
			Top-1	Top-5	Top-1	Top-5
			59.56	90.10	56.77	89.33
✓			60.42	91.07	57.62	90.37
	✓		60.08	90.62	57.27	89.76
		✓	59.83	90.72	56.88	90.11
✓	✓		60.56	91.24	57.87	90.37
✓	✓	✓	**61.05**	**91.45**	**58.05**	**90.70**

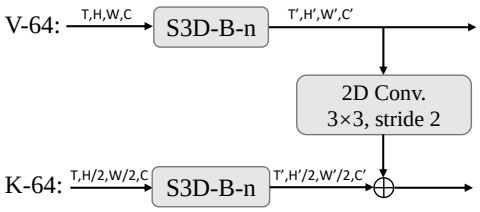
Table 10. Ablation studies on applying inter-modality mixup on different types of head networks.

Upper Body	Hand	Mouth	#Keypoints	Per-instance		Per-class	
				Top-1	Top-5	Top-1	Top-5
✓			11	21.37	50.66	19.78	49.00
✓	✓		53	48.54	81.45	45.52	79.94
	✓	✓	52	48.64	81.83	45.64	80.36
✓	✓	✓	63	**49.10**	**82.00**	**46.18**	**80.71**

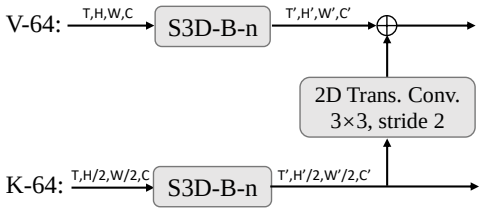
Table 11. Ablation study on keypoint selection.

V-V	K-K	V-K	Per-instance		Per-class	
			Top-1	Top-5	Top-1	Top-5
			56.85	86.87	53.34	85.60
✓			57.12	87.11	54.21	85.94
✓	✓		57.16	87.56	54.03	86.54
✓	✓	✓	**57.19**	**88.29**	**54.35**	**87.49**

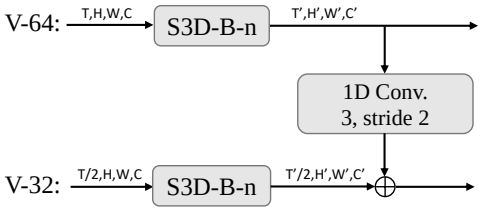
Table 12. Ablation studies on different types of bidirectional lateral connections. (V-V: video-video; K-K: keypoint-keypoint; V-K: video-keypoint.)



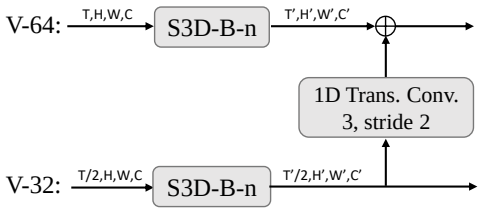
(a) Lateral connection from the 64-frame video encoder to the 64-frame keypoint encoder. We use a 2D convolution layer with a kernel size of 3×3 and a stride of 2 to match the spatial resolutions. The same holds true for the 32-frame input.



(b) Lateral connection from the 64-frame keypoint encoder to the 64-frame video encoder. We use a 2D transposed convolution layer with a kernel size of 3×3 and a stride of 2 to match the spatial resolutions. The same holds true for the 32-frame input.



(c) Lateral connection from the 64-frame video encoder to the 32-frame one. We use a 1D convolution layer with a kernel size of 3 and a stride of 2 to match the temporal resolutions. The same holds true for the keypoint encoder.



(d) Lateral connection from the 32-frame video encoder to the 64-frame one. We use a 1D transposed convolution layer with a kernel size of 3 and a stride of 2 to match the temporal resolutions. The same holds true for the keypoint encoder.

Figure 6. Illustration of the lateral connections. Note that we split bidirectional lateral connections into unidirectional ones for better illustration.

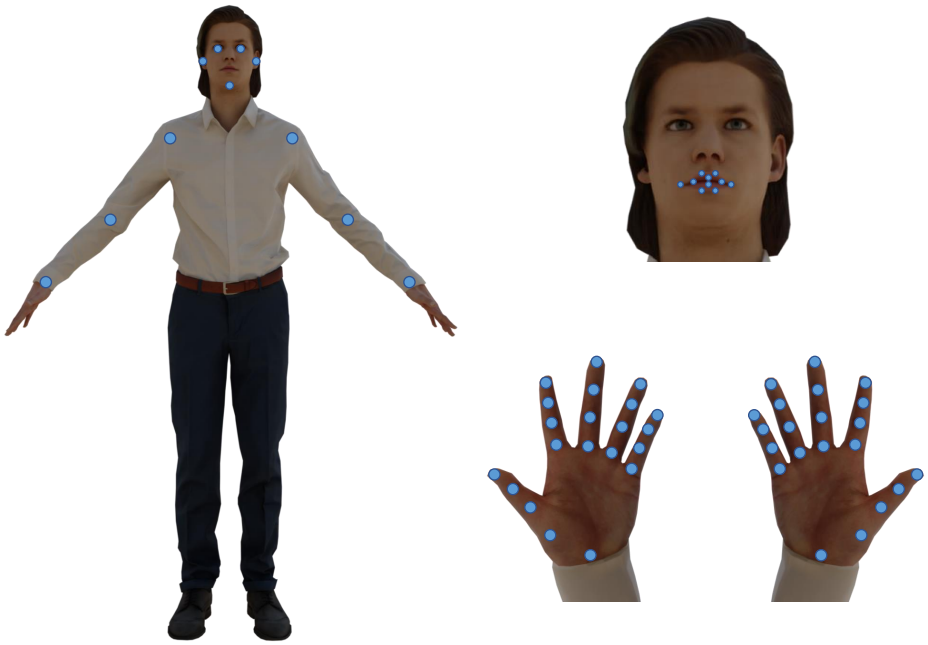


Figure 7. Illustration of the keypoints (11 upper body keypoints, 10 mouth keypoints, and 42 hand keypoints) used in our VKNet.

exchange. See Figure 4 in the main paper for their illustration. As shown in Table 12, each type of bidirectional lateral connections has a positive effect on model performance, and our default setting, using all of the three types of the lateral connections, can achieve the best performance.

VKNet vs. SlowFast. Our VKNet consists of two sub-networks, VKNet-64 and VKNet-32, to jointly model video-keypoint pairs with different temporal receptive fields. The results in Table 4 in the main paper suggest that modeling different video-keypoint pairs with varied temporal receptive fields improves the model generalization capability. One network that is related to our VKNet is Slow-

Method	Per-instance		Per-class	
	Top-1	Top-5	Top-1	Top-5
SlowFast	56.81	87.60	53.69	86.68
VKNet	**57.19**	**88.29**	**54.35**	**87.49**

Table 13. Comparison between SlowFast and our VKNet.

Method	Per-instance		Per-class	
	Top-1	Top-5	Top-1	Top-5
Contrastive Learning	59.90	91.28	57.23	90.59
Inter-Modality Mixup	**61.05**	**91.45**	**58.05**	**90.70**

Table 14. Comparison between contrastive learning and our inter-modality mixup.

Method	Per-instance		Per-class	
	Top-1	Top-5	Top-1	Top-5
Word2vec [5]	60.63	91.14	57.53	90.42
GloVe [7]	60.81	90.90	57.73	90.27
FastText [6]	**61.05**	**91.45**	**58.05**	**90.70**
BERT [4]	60.11	90.83	57.15	90.05

Table 15. Comparison among different word representation learning methods.

Fast [2], which consists of two streams taking RGB videos with low/high frame rate as inputs while having a fixed temporal receptive field. For a fair comparison between SlowFast and our VKNet, we replace the "temporal crop" operation in Figure 3 in the main paper with "temporal sampling", *i.e.*, sampling a 32-frame pair from the 64-frame one with a stride of 2 frames, to mimic the SlowFast. As shown in Table 13, our VKNet can consistently outperform SlowFast on all of the four metrics, showing that VKNet is a stronger backbone for sign language recognition.

Inter-Modality Mixup vs. Contrastive Learning. Our inter-modality mixup blends vision and language features to better maximize the separability of signs. Its effectiveness is shown in Table 6 in the main paper. One work that is related to our inter-modality mixup is CLIP [8], which jointly trains an image encoder and a text encoder with a contrastive loss by maximizing the cosine similarity of positive image-text pairs while minimizing the similarity of negative pairs. Following the practice in CLIP, we replace our inter-modality mixup loss $\mathcal{L}_{IMM}$ with a contrastive loss between the vision feature $\boldsymbol{f}$ and gloss features $\bar{\boldsymbol{E}}$. As shown in Table 14, our inter-modality mixup can consistently outperform the contrastive learning method on all of the four metrics. The results demonstrate that our inter-modality mixup is a more effective approach to exploit semantic information contained in glosses.

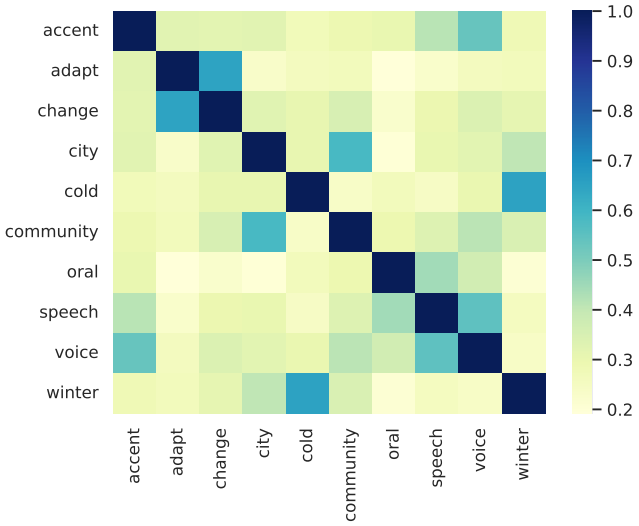


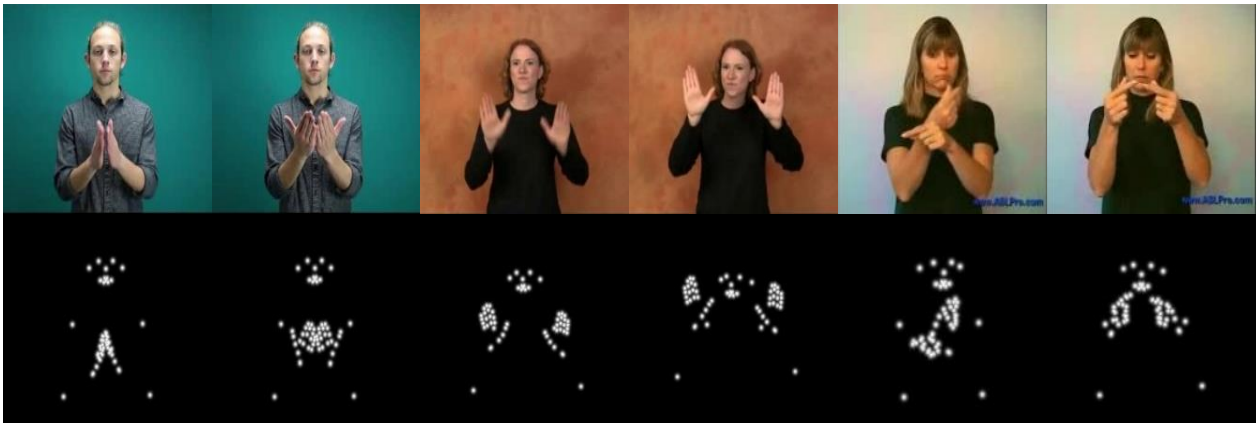
Figure 8. Visualization of gloss feature similarities. We adopt fastText to extract gloss features.

Word Representation Learning Methods. We adopt fastText [6] as our default gloss feature extractor. Here we investigate other alternatives as shown in Table 15. Word2vec [5] and GloVe [7] are two classical word representation learning methods which are widely-adopted in NLP community. They perform comparably to each other that GloVe achieves better results on the top-1 accuracy while word2vec is superior regarding to the top-5 accuracy. As an improvement of word2vec, fastText leads to better results on all of the four metrics. Finally, we also utilize an advanced model, BERT-base [4], to extract word representations by averaging the outputs of the last layer. However, it performs worse than all the other methods since it is not dedicated to word representation learning.

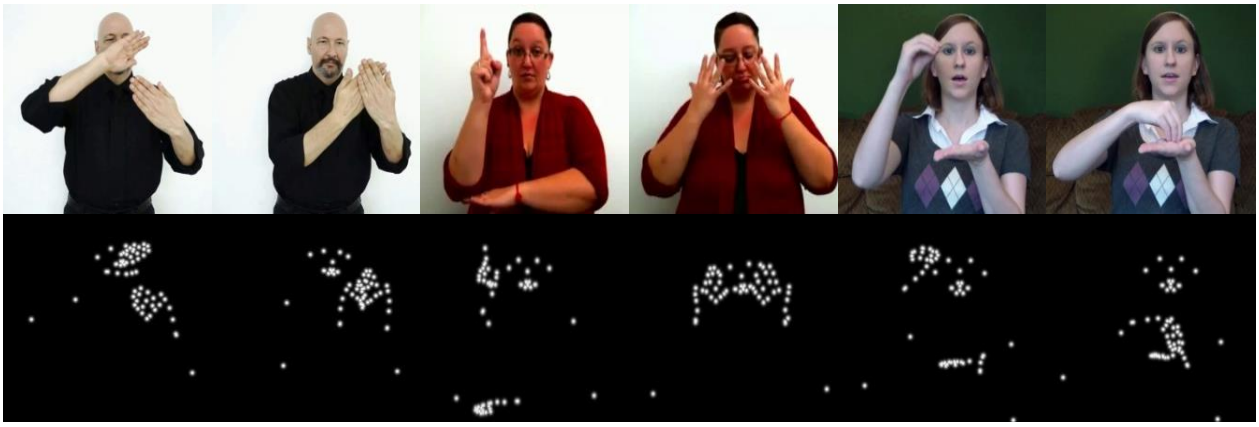
C. Visualization

Gloss Feature Similarity. The gloss feature similarities play a key role in our language-aware label smoothing. We select several glosses from the vocabulary and visualize the cosine similarities between their gloss features as a heatmap in Figure 8. We can see that the similarity matrix can roughly reflect the semantic similarities between glosses. For example, the pairs: ("adapt", "change"), ("city", "community"), ("cold", "winter"), ("speech", "oral"), and ("accent", "voice"), have high similarities, which are consistent to human understanding.

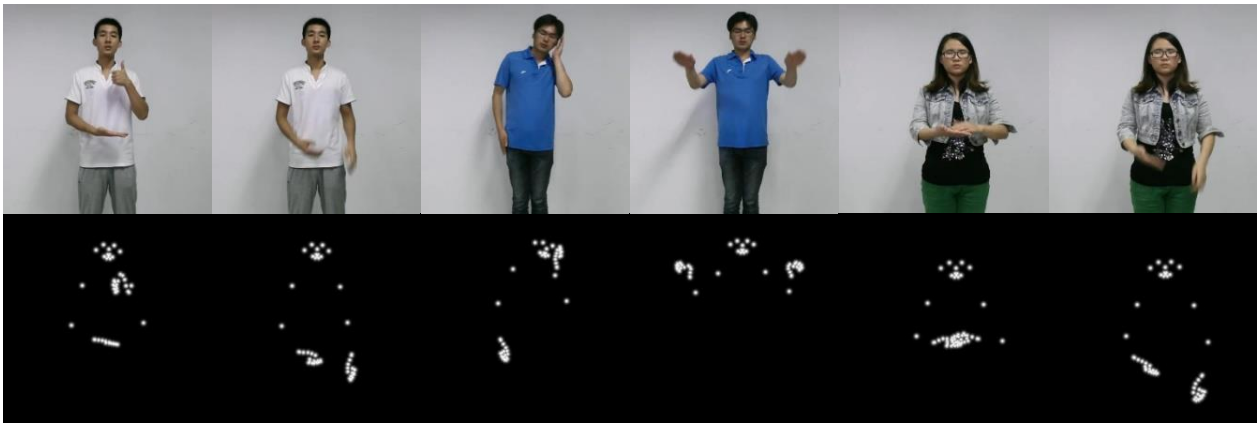
Keypoint Heatmaps. As shown in Figure 9, we visualize the keypoint heatmaps extracted by HRNet [9] by randomly selecting six frames of three signers from the test sets of WLASL2000, MSASL1000, and NMFs-CSL, respectively. We can clearly see that the heatmaps are robust to signer appearances, background variations, hand positions, and palm



(a) WLASL2000.



(b) MSASL1000.



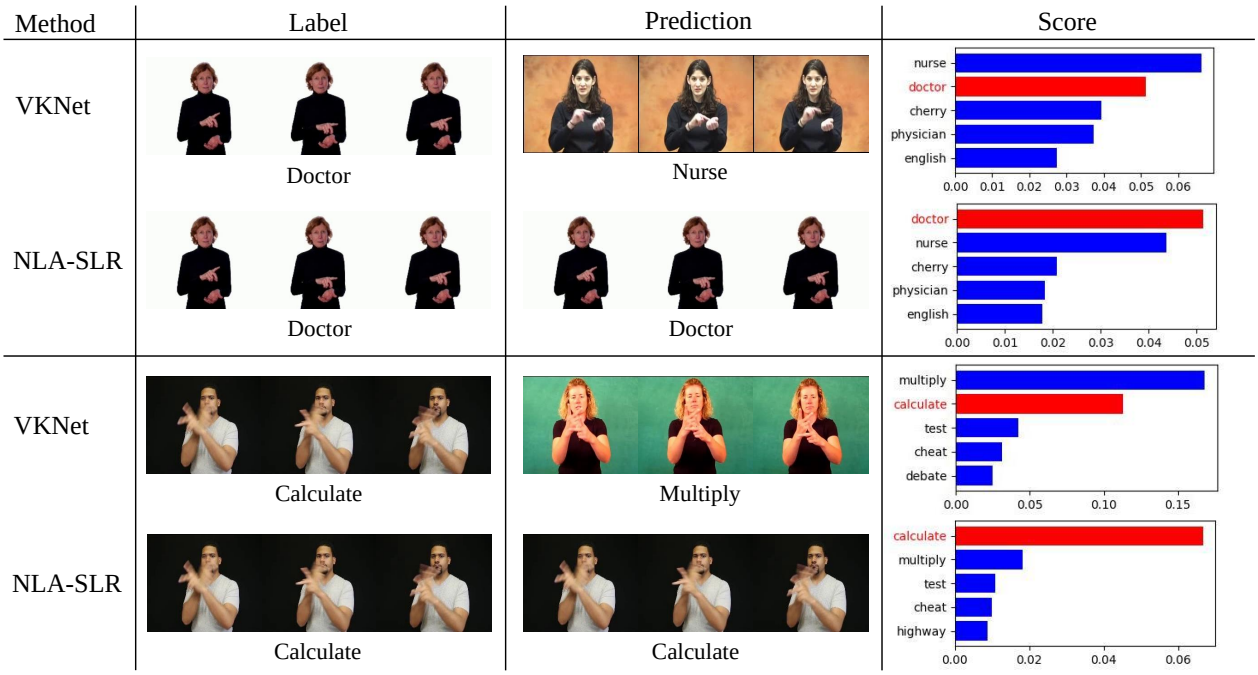
(c) NMFs-CSL.

Figure 9. Visualizations for the randomly selected frames and their corresponding keypoint heatmaps estimated by HRNet.

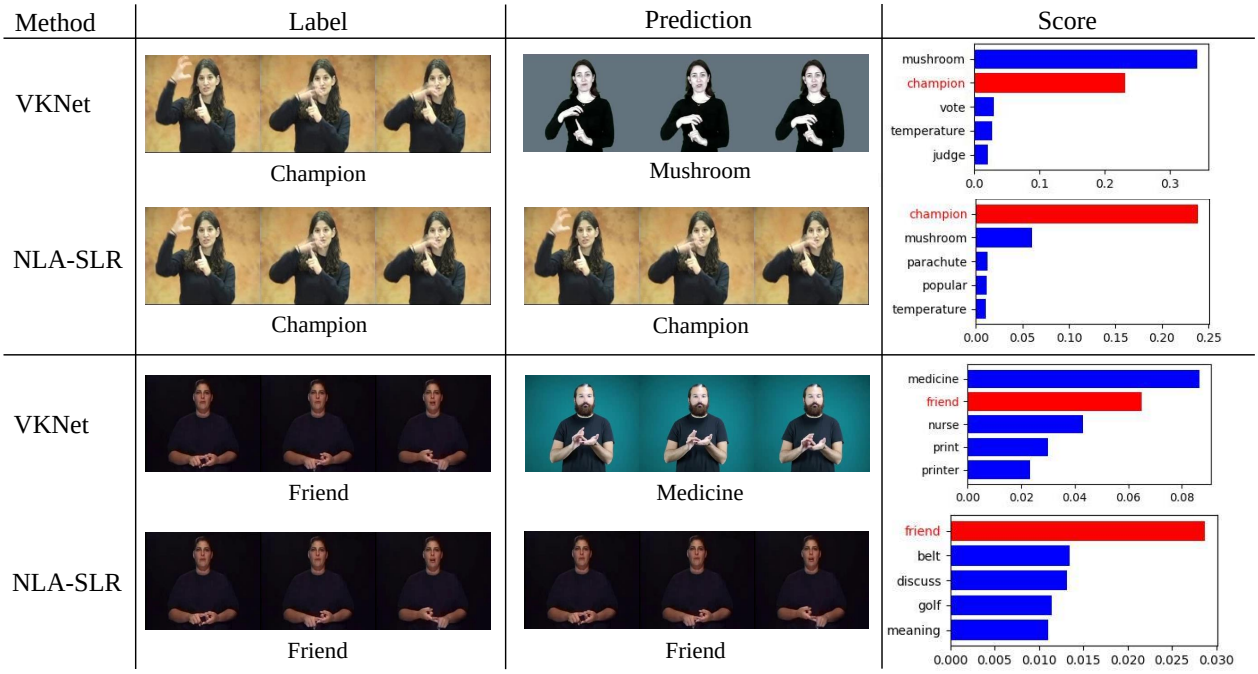
orientations.

D. Qualitative Results

As shown in Figure 10, we conduct qualitative analysis for our NLA-SLR. We find that compared with VKNet (baseline), our NLA-SLR can well classify visually indistinguishable signs (VISigns) with either similar or distinct meanings. As shown in Figure 10a, our NLA-SLR can successfully distinguish ("doctor", "nurse") and ("calculate", "multiply"), which are VISigns with similar semantic meanings, whereas the baseline, VKNet, fails to classify them. Besides, as shown in Figure 10b, our NLA-SLR can also recognize VISigns with distinct semantic meanings:



(a) VISigns with *similar* semantic meanings.



(b) VISigns with *distinct* semantic meanings.

Figure 10. Qualitative results on WLASL2000. (Here for NLA-SLR, we do not use intra-modality mixup for a fair comparison. The ground-truth gloss is highlighted in red.)

("champion", "mushroom") and ("friend", "medicine"). We owe these success to the two proposed techniques: language-aware label smoothing and inter-modality mixup.

E. Social Impact and Limitation

Sign language is the primary communication method among the deaf community. Thus, research on sign language recognition can help bridge the communication gap

between the normal-hearing and hearing-impaired people.

The proposed method is data-driven. Thus, the model performance may be affected by the biases in the training data. Besides, our backbone relies on pre-extracted keypoints; inaccurate keypoint estimation may hurt the model performance. We believe that stronger keypoint estimators may further improve sign language recognition in the future.

References

[1] Haodong Duan, Yue Zhao, Kai Chen, Dahua Lin, and Bo Dai. Revisiting skeleton-based action recognition. In *CVPR*, pages 2969–2978, 2022. 1

[2] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *ICCV*, pages 6202–6211, 2019. 3

[3] Sheng Jin, Lumin Xu, Jin Xu, Can Wang, Wentao Liu, Chen Qian, Wanli Ouyang, and Ping Luo. Whole-body human pose estimation in the wild. In *ECCV*, pages 196–214, 2020. 1

[4] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, pages 4171–4186, 2019. 3

[5] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013. 3

[6] Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. Advances in pre-training distributed word representations. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, 2018. 3

[7] Jeffrey Pennington, Richard Socher, and Christopher D Manning. GloVe: Global vectors for word representation. In *EMNLP*, pages 1532–1543, 2014. 3

[8] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763, 2021. 3

[9] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *CVPR*, pages 5693–5703, 2019. 1, 3