

Supplementary Materials of “SEDS: Semantically Enhanced Dual-Stream Encoder for Sign Language Retrieval”

Anonymous Authors

A INTRODUCTION

This supplementary material provides further details and additional experiments not included in the main paper. It begins by introducing the process of keypoints extraction and selection (B.1), and the architecture of Pose encoder (B.2). Then, we compare the efficiency between different modules in the SEDS framework to demonstrate the significant difference in computational requirements between the two modal encoders (C). After that, we present ablation studies on the effect of Signbert Initialization (D.1), number of cross gloss attention layers (D.2), using keypoints from different parts (D.3). Furthermore, we study the influence of fine-tuning hyperparameters on performance, such as batch size (D.4) and learning rate (D.5). Finally, we show more visualization of Top-3 TVR results (E.1) and the fine-grained similarity matrices (E.2). All ablation studies and qualitative experiments are conducted on the How2Sign [3] dataset.

B ADDITIONAL IMPLEMENTATION DETAILS

B.1 Keypoints Extraction and Selection

For Pose estimation, we choose RTMPose[5] trained on COCO-WholeBody[7] as the estimator. We adopt the implementation by MMPose[2] and extract 133 2D keypoints. As mentioned in the main paper, we select 49 keypoints and divide them into three groups: 7 keypoints for the body, 21 keypoints for the left hand, and 21 keypoints for the right hand as Pose input. In the next section, we investigate the impact of leveraging keypoints from different parts, such as the mouth and face, as Pose input. We visualize these keypoints from different parts in Fig. 1.

B.2 Pose Encoder Architecture

We leverage the RGB encoder and Pose encoder to extract RGB and Pose features, respectively. The RGB Encoder is a pre-trained I3D[1] network. Therefore, for the sake of simplicity, we do not go into details here. We only detail the structure of the Pose Encoder that consists of GCN-hand, GCN-body and 1D Conv Fusion module. As shown in Tab. 1, where ST-GCN[9] layer i means the i -th ST-GCN layer in each module.

As depicted in Fig. 2, the 2D coordinates produced by the estimator serve as input. We utilize GCN-hand and GCN-body to respectively extract features related to the hand and body. Subsequently, these features are combined through the 1D Conv Fusion module, yielding the final Pose features.

C EFFICIENCY COMPARISON ON DIFFERENT MODULES

In order to provide a more intuitive demonstration of the efficiency of different modules in the SEDS framework, we present the parameters, FLOPs and input size of these modules in Tab. 2. From the table, it is obvious that there is not much difference in the number



Figure 1: The coordinates of each group in our approach. They represent the body, left hand, right hand, mouth, and face respectively.

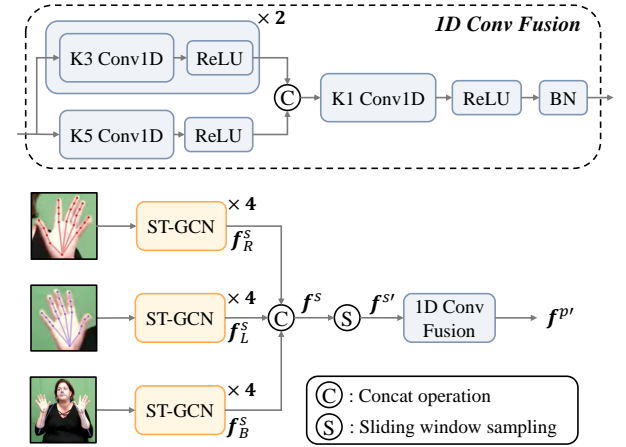


Figure 2: The architecture of Pose Encoder.

Structure	GCN-hand	GCN-body
ST-GCN layer 1	128	128
ST-GCN layer 2	128	128
ST-GCN layer 3	256	256
ST-GCN layer 4	512	512
Concat	1536	
1D Conv Fusion	1536	

Table 1: Output dimension of different layers in Pose Encoder.

of parameters between each module. However, for the consumption of FLOPs, the RGB encoder is more than two hundred times that of other modules, primarily due to extensive convolution operations by multiple negative samples. In comparison to RGB encoders, Pose encoders and other modules require significantly less computational resources, allowing for end-to-end training. Furthermore, the input size of each module indicates that the density of RGB visual data inevitably consumes a significant amount of memory.

Metrics	RGB Encoder	Pose Encoder	RGB Interaction	Pose Interaction	Text Encoder	CGAF Module
Parameters	17.8M	9.66M	57.7M	57.8M	25.2M	17.4
FLOPs	28551.3G	132.1G	59.7G	60.1G	12.9G	18.1G
Input Size	(16×64, 3, 224, 224)	(16×64, 49, 2)	(16, 64, 1024)	(16, 64, 1536)	(16, 32, 512)	(16, 64×2, 512)

Table 2: The efficiency comparison on different modules in our SEDS framework. We set the batch size to 16 and assume the RGB encoder for online feature extraction. Where 64 is the length of video clips, 32 is the length of text words, 49 is the total number of keypoints, 1024 is the output size of the RGB encoder and 1536 is the output size of the Pose encoder.

Signbert Initialization	Structure	T2V			V2T		
		R@1↑	R@5↑	R@10↑	R@1↑	R@5↑	R@10↑
✗	Pose	48.4	64.5	71.3	42.9	59.6	66.6
✓	Pose	55.9	69.6	74.9	50.9	64.7	69.3
✗	SEDS	57.5	72.4	77.5	53.1	67.0	71.9
✓	SEDS	62.5	75.1	80.1	57.9	70.4	74.9

Table 3: The ablation study on How2Sign to investigate the influence of Signbert Initialization.

Parts	T2V			V2T		
	R@1↑	R@5↑	R@10↑	R@1↑	R@5↑	R@10↑
hand	58.7	72.4	77.7	54.1	67.3	73.3
hand+body+face	62.0	74.8	79.7	56.8	69.3	74.2
hand+body+mouth	61.9	74.9	79.6	57.0	69.9	74.4
hand+body+face+mouth	62.3	75.0	79.4	57.6	69.6	74.8
hand+body	62.5	75.1	80.1	57.9	70.4	74.9

Table 4: The ablation study on How2Sign to investigate the influence of using keypoints from different parts to input Pose Encoder.

layers	T2V			V2T		
	R@1↑	R@5↑	R@10↑	R@1↑	R@5↑	R@10↑
0	60.2	73.3	78.1	55.2	68.5	73.3
1	61.6	74.8	79.0	56.4	69.3	74.2
2	62.5	75.1	80.1	57.9	70.4	74.9
3	62.2	74.9	80.0	57.6	70.2	74.4
4	61.8	74.2	79.1	56.8	69.9	74.0

Table 5: The ablation study on How2Sign to investigate the influence of the number of cross gloss attention layers.

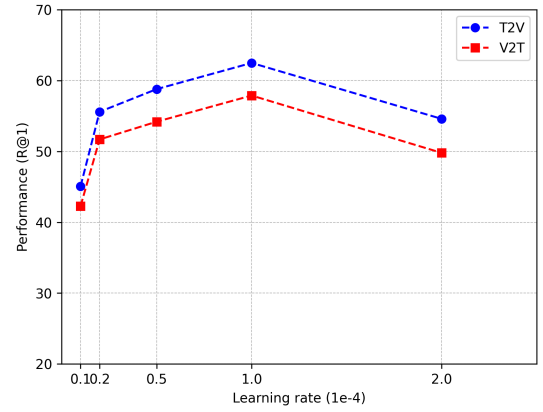
batchsize	T2V			V2T		
	R@1↑	R@5↑	R@10↑	R@1↑	R@5↑	R@10↑
16	42.6	58.9	65.5	40.7	55.8	61.7
32	49.7	65.2	71.1	47.0	62.1	67.8
64	57.4	71.7	76.6	53.0	66.1	71.1
128	62.5	75.1	80.1	57.9	70.4	74.9

Table 6: The ablation study on How2Sign to investigate the influence of the batch size.

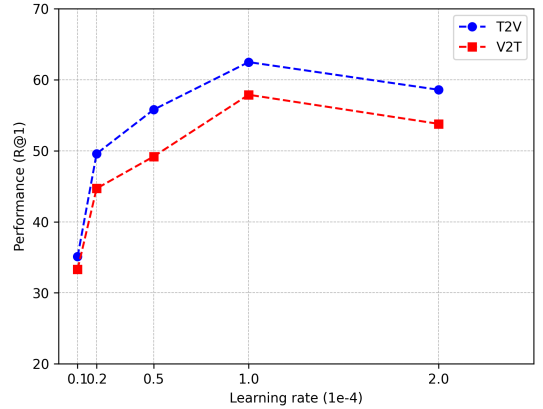
D MORE EXPERIMENTS

D.1 Ablation Study on Signbert Initialization

In our framework, the GCN network module for hand keypoints in Pose encoder is initialized with the Signbert[4], which understands the hand Pose through the masked modeling strategy and the Hand-model-aware decoder based on MANO[8]. We compare a randomly initialized baseline with one initialized by Signbert in



(a) Learning rate for Pose encoder



(b) Learning rate for CGAF module

Figure 3: The ablation study on How2Sign to investigate the influence of the learning rate for different modules.

Tab. 3. Both Pose encoder and SEDS show significant performance improvements due to the initialization of Signbert.

D.2 Ablation Study on Cross Gloss Attention Layers

In the SEDS framework, the Cross Gloss Attention Fusion (CGAF) module aggregates the local sequence features with similar semantics from inter-modal and intra-modal. We utilize two cross gloss attention layers in CGAF module for fusion of two modalities. In Tab. 5, we investigate the impact of different layer numbers. The performance of model is optimal at two layers, exhibiting a gradual increase at few layers, and a slight decrease after two layers. This

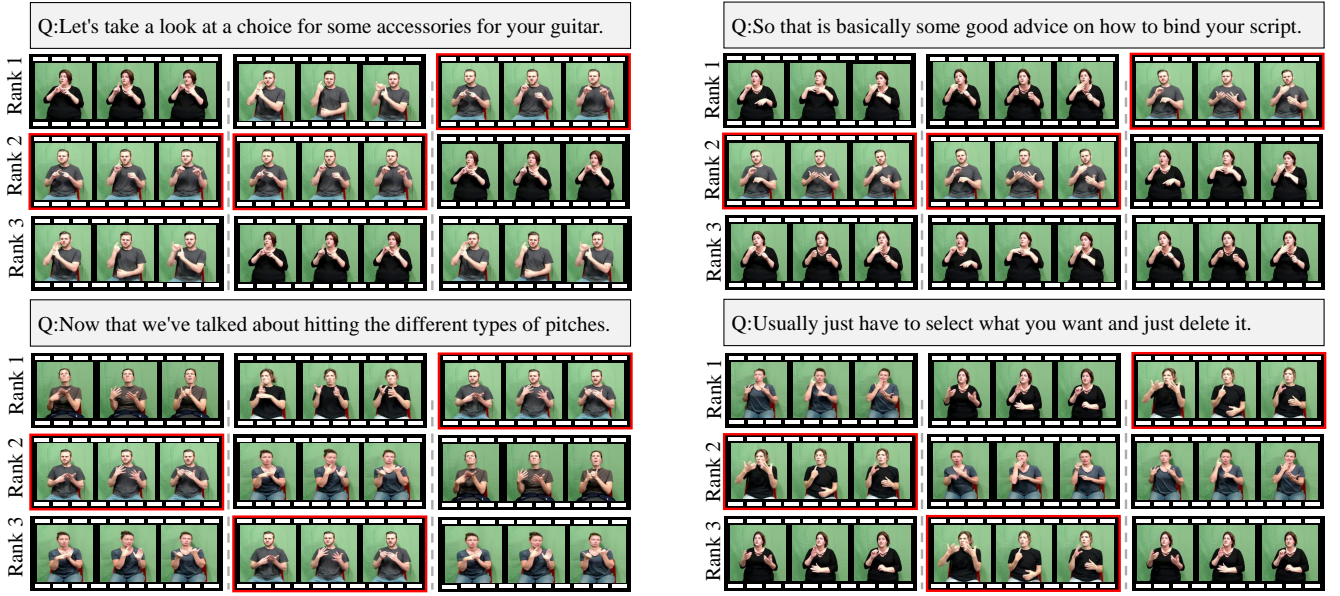


Figure 4: More Top-3 TVR results on How2Sign. Left: by Pose encoder. Middle: by RGB encoder. Right: by our framework. Red: correct videos. Q: query annotation.

indicates that only when there are sufficient layers, Pose features and RGB features are interacted and fused more fully. However, the excessive number of layers results in the two features losing their discrimination in the fusion process.

D.3 Ablation Study on Keypoints from Different Parts

Tab. 4 illustrates the performance changes of the SEDS framework when different combinations of keypoints are used for the Pose Encoder. We observe a notable decrease in model performance when only hand keypoints were introduced. This is because body keypoints not only provide a comprehensive description of overall human motion but also enhance the robustness of the model, reducing the dependency of the RGB model on specific individuals and backgrounds. Additionally, we find that there is a slight decrease in model performance when keypoints from face or mouth[6] are introduced alone. This is due to the RGB modality inherently containing rich facial expression information. So the keypoints from the face or mouth alone disrupt the existing facial expression information. Using both face and mouth keypoints produces similar results to using hand and body keypoints, as the information they provide is largely redundant with existing facial expression data from the RGB modality. Due to the feature extraction of excess keypoints bringing extra computational complexity, we only use the combination of hand and body keypoints as input in Pose Encoder.

D.4 Ablation Study on Training Batch Size

As shown in Tab. 6, the continuous increase of batch size steadily improve the retrieval performance. In contrastive learning, batch size directly decides the number of negative samples in a batch during the training process. Due to GPU memory limitations, we set the batch size to 128.

D.5 Ablation Study on Learning Rate for Different Modules

In Fig. 3, we study the different learning rates of two modules. The performance of both modules reach their peak at a learning rate of $1e-4$, but the trend of change is different. The most complex part of the Pose encoder is the GCN module for both hands, which is initialized by Signbert. In contrast to the CGAF module, the Pose encoder demonstrates superior performance at low learning rates due to the influence of prior knowledge. However, excessive learning rates result in a more obvious decline in performance, particularly in the Pose encoder, which is attributed to the deterioration of prior knowledge. And the performance of randomly initialized CGAF modules exhibit a relatively minor decline.

E MORE QUALITATIVE RESULTS

E.1 More Top-3 TVR Results.

We show more Top-3 TVR results in Fig. 4. The different emphases of the Pose encoder and RGB encoder are clearly displayed. Concurrently, our framework SEDS integrates the benefits of both and retrieves accurate samples in all examples.

E.2 More Visualization of the Fine-grained Similarity Matrices

As shown in Fig. 5, we show more visualization results of the fine-grained similarity matrices. These examples intuitively illustrate the positive impact of our proposed Pose-RGB fine-grained matching objective. It ensures a balanced distribution of the Pose-Text and RGB-Text fine-grained similarity matrices during the fusion process, thereby enhancing the representation ability of the fusion features.

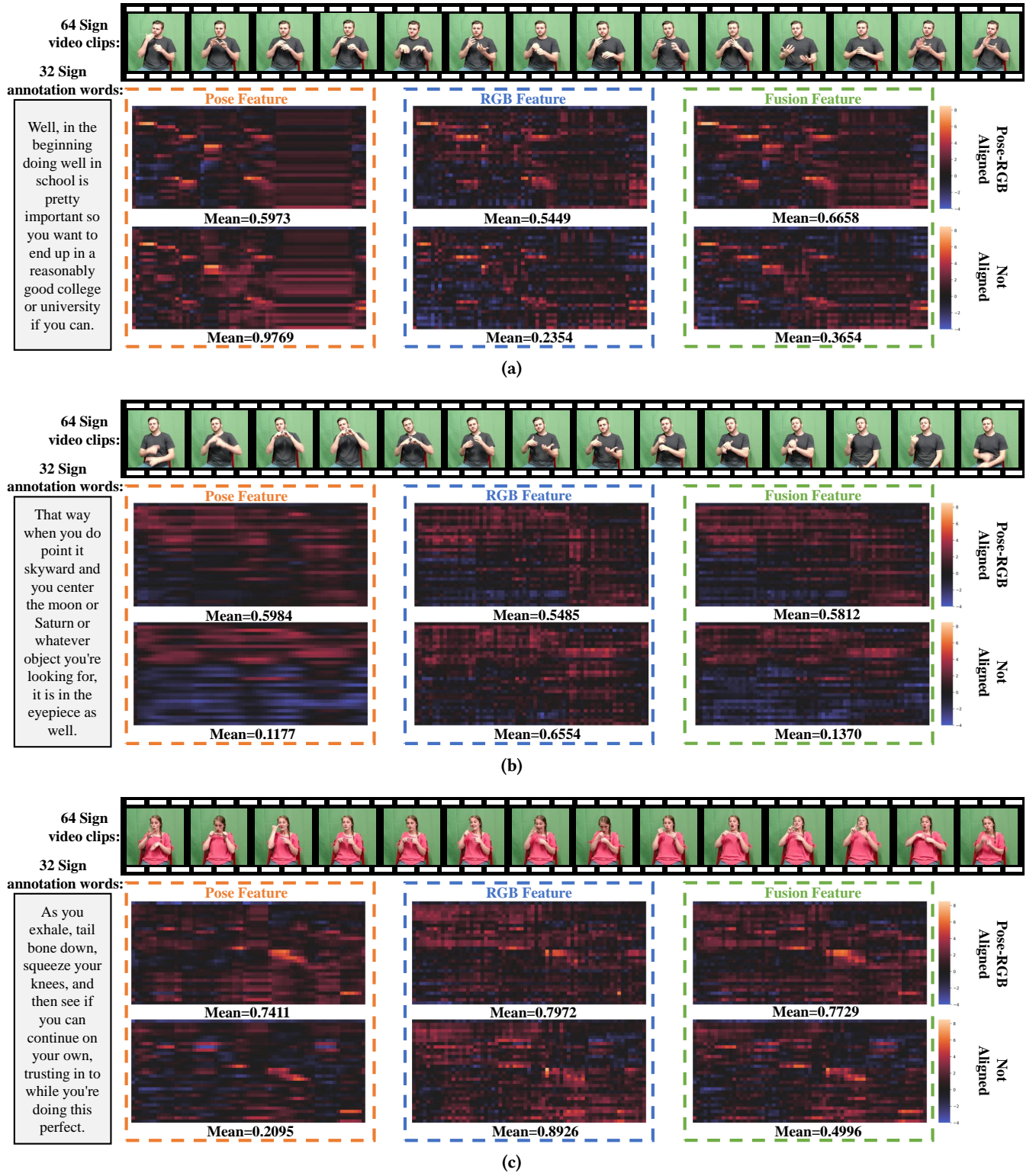


Figure 5: More visualization results of the fine-grained similarity matrices. We sample 64 clips in the video and 32 words in the text to extract features, calculate similarity, and display them in a heatmap format.

REFERENCES

[1] João Carreira and Andrew Zisserman. 2017. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. 4724–4733.

[2] MMPose Contributors. 2020. OpenMMLab Pose Estimation Toolbox and Benchmark. <https://github.com/open-mmlab/mmpose>.

[3] Amanda Duarte, Shruti Palaskar, Lucas Ventura, Deepti Ghadiyaram, Kenneth DeHaan, Florian Metze, Jordi Torres, and Xavier Giro-i Nieto. 2021. How2Sign: a large-scale multimodal dataset for continuous american sign language. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. 2735–2744.

[4] Hezhen Hu, Weichao Zhao, Wengang Zhou, Yuechen Wang, and Houqiang Li. 2021. SignBERT: Pre-Training of Hand-Model-Aware Representation for Sign Language Recognition. In *International Conference on Computer Vision (ICCV)*. 11067–11076.

[5] Tao Jiang, Peng Lu, Li Zhang, Ningsheng Ma, Rui Han, Chengqi Lyu, Yining Li, and Kai Chen. 2023. RTMPose: Real-Time Multi-Person Pose Estimation based on MMPose. *arXiv preprint arXiv:2303.07399*.

[6] Peiqi Jiao, Yuecong Min, Yanan Li, Xiaotao Wang, Lei Lei, and Xilin Chen. 2023. CoSign: Exploring Co-occurrence Signals in Skeleton-based Continuous Sign Language Recognition. In *International Conference on Computer Vision (ICCV)*. 20619–20629.

[7] Sheng Jin, Lumin Xu, Jin Xu, Can Wang, Wentao Liu, Chen Qian, Wanli Ouyang, and Ping Luo. 2020. Whole-Body Human Pose Estimation in the Wild. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

[8] Javier Romero, Dimitrios Tzionas, and Michael J. Black. 2017. Embodied hands: modeling and capturing hands and bodies together. In *ACM Transactions on Graphics (ToG)*. 1–17.

[9] Bing Yu, Haoteng Yin, and Zhanxing Zhu. 2018. Spatio-Temporal Graph Convolutional Networks: A Deep Learning Framework for Traffic Forecasting. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI’18)*. AAAI Press, 3634–3640.