CiCo: Domain-Aware Sign Language Retrieval via Cross-Lingual Contrastive Learning Supplementary Material

Yiting Cheng[1], Fangyun Wei[2]*, Jianmin Bao[2], Dong Chen[2], Wenqiang Zhang[1]*
[1]School of Computer Science, Fudan University, [2]Microsoft Research Asia
{ytcheng18, wqzhang}@fudan.edu.cn, {fawe, jianbao, doch}@microsoft.com

Model	T2V			V2T		
	R@1	R@5	R@10	R@1	R@5	R@10
CLIP [3]	35.4	53.4	60.9	30.7	49.1	57.1
CiCo	**56.6**	**69.9**	**74.7**	**51.6**	**64.8**	**70.1**

Table 1. Comparison between Cico and CLIP.

Strategy	T2V			V2T		
	R@1	R@5	R@10	R@1	R@5	R@10
Max	21.1	38.0	46.4	17.8	34.9	42.9
Softmax	32.6	50.3	58.2	29.0	46.6	54.0
Mean	**56.6**	**69.9**	**74.7**	**51.6**	**64.8**	**70.1**

Table 2. Study on different strategies of global similarity calculation in cross-lingual contrastive learning.

Stride	T2V			V2T		
	R@1	R@5	R@10	R@1	R@5	R@10
1	**56.6**	**69.9**	**74.7**	**51.6**	**64.8**	**70.1**
2	44.8	60.5	68.1	39.7	55.5	63.0
4	24.3	42.3	49.8	14.4	30.2	37.4
8	23.6	40.8	49.1	15.3	31.5	39.6

Table 3. Study on different sliding window strides used in sign encoder.

A. More Experiments

CiCo vs CLIP. We compare our approach CiCo with CLIP [3], which is one of the most representative vision-language models. CLIP can be easily generalized to sign language retrieval by replacing our cross-lingual contrastive learning with CLIP. The other settings including sign encoder and text augmentation still remain unchanged. As shown in Table 1, CiCo surpasses CLIP by +21.2 T2V and +20.9 V2T R@1 scores. The reason is that CLIP contrasts the overall features of two modalities, while our cross-lingual contrastive learning concentrates on identifying the fine-grained sign-to-word mappings during modeling global similarities of texts and sign videos.

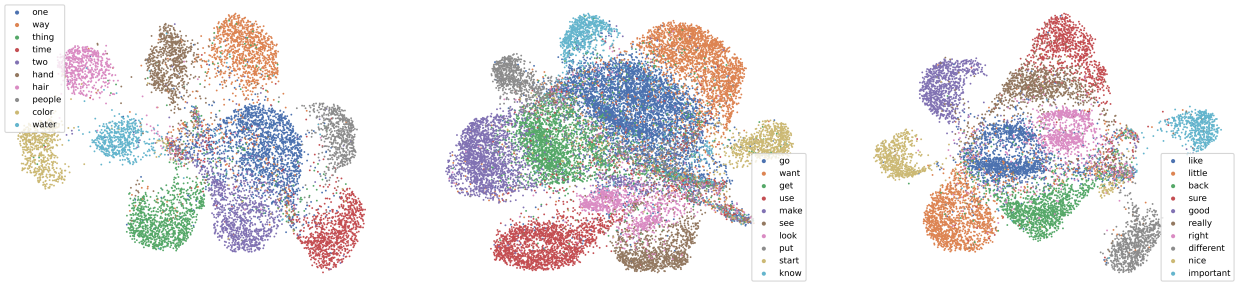
Different Strategies of Global Similarity Calculation in Cross-Lingual Contrastive Learning. As described in Section 3.3 and illustrated in Figure 3b, we adopt "Mean" strategy which averages sign-to-text similarities and word-to-video similarities to obtain the global video-to-text similarity and text-to-video similarity, respectively. In Section 4.3 of the main paper, we study different strategies to identify the fine-grained sign-to-word mappings, now we investigate different ways of global similarity calculation. Table 2 shows the results of two variants termed "Max" and "Softmax" besides the default "Mean" strategy. "Max" assigns global similarity with the maximum score of sign-to-text similarities (or word-to-video similarities). "Softmax" stands for a combination of Softmax, multiplication and sum (refer to Section 4.3 for details). The default "Mean" strategy achieves the best result.

Sliding Window Stride in Sign Encoder. Our sign encoder adopts a sliding window manner to extract features of continuous sign videos. The default sliding window stride is set as 1. We vary the stride and show the results in Table 3. Setting stride as 1 yields the best performance.

Fine-Tuning Hyper-Parameters. Recall that in the training of cross-lingual contrastive learning, our vision transformer and text transformer are initialized by the image encoder and text encoder in CLIP (ViT-B/32) [3]. Here we study the fine-tuning hyper-parameters, *i.e.*, learning rate in Figure 2a and batch size in Figure 2b. A learning rate of 1e-5 yields best result. The increase of batch size sustainably promotes the performance. In our experiment, we set the batch size to 512 due to the limited GPU memory.

Other Hyper-Parameters. There are four remaining hyper-parameters in CiCo: 1) α defined in Eq.(1) controls the weights of features extracted by domain-agnostic sign

*Corresponding author.



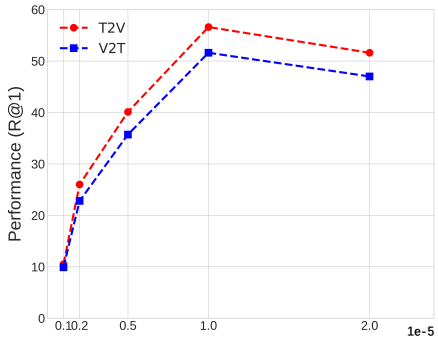
(a) Top-10 Nouns. (b) Top-10 Verbs. (c) Top-10 Adjective/Adverbs.

Figure 1. Feature visualization of sign video clips. We map features extracted by our sign encoder to 2D space with UMAP [2].

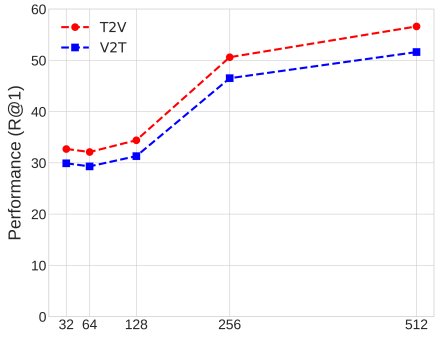
encoder and domain-aware sign encoder; 2) β defined in Eq.(3) controls the weights of sign-video-to-text contrast and text-to-sign-video contrast; 3) the temperature σ of row-wise and column-wise Softmax; 4) the maximum length of sign clip feature L. The studies are shown in Table 4, Table 5, Table 6 and Table 7, respectively.

B. Qualitative Results

Visualization of the Identified Sign-to-Word Mappings. Recall that in cross-lingual contrastive learning, we implicitly identify the sign-to-word mappings by calculating the fine-grained cross-lingual similarities (see Figure 3b of the main paper). Once the model is well optimized, we could infer the input texts and sign videos to produce a cross-lingual similarity matrix, which approximately reflects the sign-to-word mappings. For each word, we could identify its corresponding sign which has the maximal activation value. After that, the sign-to-word mapping is established. In Figure 1, we utilize UMAP [2] to visualize the



(a) Learning rate.



(b) Batch size.

Figure 2. Study on fine-tinning hyper-parameters in contrastive learning.

α	T2V			V2T		
	R@1	R@5	R@10	R@1	R@5	R@10
0.2	53.0	67.5	72.5	47.6	62.7	67.2
0.4	55.4	68.7	74.0	49.9	62.5	68.6
0.6	55.1	68.5	73.4	49.6	63.9	68.9
0.8	**56.6**	**69.9**	**74.7**	**51.6**	**64.8**	**70.1**

Table 4. Study of α defined in Eq.(1).

β	T2V			V2T		
	R@1	R@5	R@10	R@1	R@5	R@10
0.0	39.0	56.3	63.1	26.6	49.9	57.8
0.2	44.8	62.1	68.1	39.8	55.5	62.5
0.4	45.8	62.4	68.7	40.6	57.7	64.1
0.5	**56.6**	**69.9**	**74.7**	**51.6**	**64.8**	**70.1**
0.6	54.9	69.6	74.5	49.6	63.5	68.6
0.8	54.1	68.7	73.3	48.3	62.1	67.8
1.0	52.5	67.1	72.1	48.8	62.8	67.4

Table 5. Study of β defined in Eq.(3).

σ	T2V			V2T		
	R@1	R@5	R@10	R@1	R@5	R@10
7e-04	41.3	58.9	65.5	38.2	54.4	61.3
7e-03	42.6	59.6	65.5	39.5	54.7	61.9
7e-02	**56.6**	**69.9**	**74.7**	**51.6**	**64.8**	**70.1**
7e-01	31.9	49.9	57.8	28.6	45.8	53.9

Table 6. Study of the temperature σ used in row-wise and column-wise Softmax.

L	T2V			V2T		
	R@1	R@5	R@10	R@1	R@5	R@10
4	17.3	31.2	38.6	14.3	26.8	34.1
8	38.2	55.4	62.3	34.1	50.2	56.4
16	50.9	66.6	72.0	45.9	60.3	66.7
32	53.6	67.3	73.5	48.9	61.7	67.8
64	**56.6**	**69.9**	**74.7**	**51.6**	**64.8**	**70.1**

Table 7. Study of maximum length of sign clip feature L.

features of the identified sign video clips for top-10 nouns, verbs and adjectives/adverbs within the How2Sign [1] vocabulary. The features of sign video clips associated with the same word form a compact cluster, demonstrating that our approach could identify the sign-to-word mappings during training.

More Examples of Sign-to-Word Mappings. We visualize a collection of signs associated with the words {"Big", "Different", "Hard", "Understand", "Vegetable", "Vehicle", "Water", "Baby"} in Figure 3. The mappings are automatically identified by our CiCo.

References

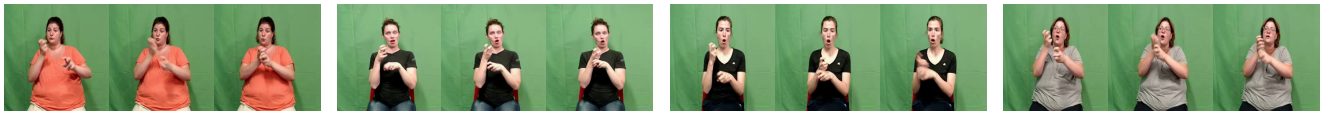
[1] Amanda Duarte, Shruti Palaskar, Lucas Ventura, Deepti Ghadiyaram, Kenneth DeHaan, Florian Metze, Jordi Torres, and Xavier Giro-i Nieto. How2Sign: A Large-scale Multimodal Dataset for Continuous American Sign Language. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 3, 4

[2] Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. Umap: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29), 2018. 2

[3] 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 *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 1

(a) “Big”.

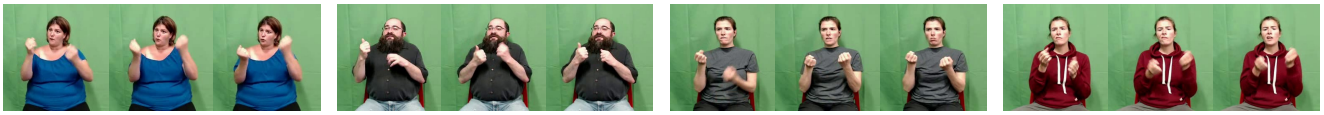
(b) “Different”.



(c) “Hard”.

(d) “Understand”.

(e) “Vegetable”.



(f) “Vehicle”.

(g) “Water”.

(h) “Baby”.

Figure 3. More examples of cross-lingual (sign-to-word) mappings identified by our approach on How2Sign [1] dataset.