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

Yiting Cheng¹, Fangyun Wei²*, Jianmin Bao², Dong Chen², Wenqiang Zhang¹*
¹School of Computer Science, Fudan University, ²Microsoft Research Asia
{ytcheng18, wqzhang}@fudan.edu.cn, {fawe, jianbao, doch}@microsoft.com

Abstract

This work focuses on sign language retrieval—a recently proposed task for sign language understanding. Sign language retrieval consists of two sub-tasks: text-to-sign-video (T2V) retrieval and sign-video-to-text (V2T) retrieval. Different from traditional video-text retrieval, sign language videos, not only contain visual signals but also carry abundant semantic meanings by themselves due to the fact that sign languages are also natural languages. Considering this character, we formulate sign language retrieval as a cross-lingual retrieval problem as well as a video-text retrieval task. Concretely, we take into account the linguistic properties of both sign languages and natural languages, and simultaneously identify the fine-grained cross-lingual (i.e., sign-to-word) mappings while contrasting the texts and the sign videos in a joint embedding space. This process is termed as cross-lingual contrastive learning. Another challenge is raised by the data scarcity issue—sign language datasets are orders of magnitude smaller in scale than that of speech recognition. We alleviate this issue by adopting a domain-agnostic sign encoder pre-trained on large-scale sign videos into the target domain via pseudo-labeling. Our framework, termed as domain-aware sign language retrieval via Cross-lingual Contrastive learning or CiCo for short, outperforms the pioneering method by large margins on various datasets, e.g., +22.4 T2V and +28.0 V2T R@1 improvements on How2Sign dataset, and +13.7 T2V and +17.1 V2T R@1 improvements on PHOENIX-2014T dataset. Code and models are available at: https://github.com/FangyunWei/SLRT.

1. Introduction

Sign languages are the primary means of communication used by people who are deaf or hard of hearing. Sign language understanding [1, 10, 12–15, 18, 32, 33, 61, 73] is significant for overcoming the communication barrier between the hard-of-hearing and non-signers. Sign language recognition and translation (SLRT) has been extensively studied, with the goal of recognizing the arbitrary semantic meanings conveyed by sign languages. However, the lack of available data significantly limits the capability of SLRT. In this paper, we focus on developing a framework for a recently proposed sign language retrieval task [18]. Unlike SLRT, sign language retrieval focuses on retrieving the meanings that signers express from a closed-set, which can significantly reduce error rates in realistic deployment.

Sign language retrieval is both similar to and distinct from the traditional video-text retrieval. On the one hand, like video-text retrieval, sign language retrieval is also composed of two sub-tasks, i.e., text-to-sign-video (T2V) retrieval and sign-video-to-text (V2T) retrieval. Given a free-form written query and a large collection of sign language videos, the objective of T2V is to find the video that best matches the written query (Figure 1a). In contrast, the goal of V2T is to identify the most relevant text description given a query of sign language video (Figure 1b).

On the other hand, different from the video-text retrieval, sign languages, like most natural languages, have their own grammars and linguistic properties. Therefore, sign language videos not only contain visual signals, but also carry

*Corresponding author.
Sign language retrieval is extremely challenging due to the following reasons: (1) Sign languages are completely separate and distinct from natural languages since they have unique linguistic rules, word formation, and word order. The transcription between sign languages and natural languages is complicated, for instance, the word order is typically not preserved between sign languages and natural languages. It is necessary to automatically identify the sign-to-word mapping from the cross-lingual retrieval perspective; (2) In contrast to the text-video retrieval datasets [45, 48] which contain millions of training samples, sign language datasets are orders of magnitude smaller in scale—for example, there are only 30K video-text pairs in How2Sign [19] training set; (3) Sign languages convey information through the handshape, facial expression, and body movement, which requires models to distinguish fine-grained gestures and actions; (4) Sign language videos typically contain hundreds of frames. It is necessary to build efficient algorithms to lower the training cost and fit the long videos as well as the intermediate representations into limited GPU memory.

In this work, we concentrate on resolving the challenges listed above:

- We consider the linguistic rules (e.g., word order) of both sign languages and natural languages. We formulate sign language retrieval as a cross-lingual retrieval task as well as a video-text retrieval problem. While contrasting the sign videos and the texts in a joint embedding space as achieved in most vision-language pre-training frameworks [5, 44, 57], we simultaneously identify the fine-grained cross-lingual (sign-to-word) mappings between two types of languages via our proposed cross-lingual contrastive learning as shown in Figure 2.

- Data scarcity typically brings in the over-fitting issue. To alleviate this issue, we adopt transfer learning and adapt a recently released domain-agnostic sign encoder [61] pre-trained on large-scale sign-videos to the target domain. Although this encoder is capable of distinguishing the fine-grained signs, direct transferring may be sub-optimal due to the unavoidable domain gap between the pre-training dataset and sign language retrieval datasets. To tackle this problem, we further fine-tune a domain-aware sign encoder on pseudo-labeled data from target datasets. The final sign encoder is composed of the well-optimized domain-aware sign encoder and the powerful domain-agnostic sign encoder.

- In order to effectively model long videos, we decouple our framework into two disjoint parts: (1) a sign encoder which adopts a sliding window on sign-videos to pre-extract their vision features; (2) a cross-lingual contrastive learning module which encodes the extracted vision features and their corresponding texts in a joint embedding space.

Our framework, called domain-aware sign language retrieval via Cross-lingual Contrastive learning or CiCo for short, outperforms the pioneer SPOT-ALIGN [18] by large margins on various datasets, achieving 56.6 (+22.4) T2V and 51.6 (+28.0) V2T R@1 accuracy (improvement) on How2Sign [19] dataset, and 69.5 (+13.7) T2V and 70.2 (+17.1) V2T R@1 accuracy (improvement) on PHOENIX-2014T [8] dataset. With its simplicity and strong performance, we hope our approach can serve as a solid baseline for future research.

2. Related Work

Sign Language Understanding. Sign language understanding aims at interpreting the semantic information con-
veyed within sign videos. Researchers have explored such capability on various tasks including sign language recognition [13,20,21,30,38,58,73], sign spotting [1,18,47,61], sign language translation [8,9,12,37,72] and our focused sign language retrieval [18].

One of the fundamental tasks of sign language understanding is sign language recognition (SLR), which aims to transcribe a sign video into a gloss sequence. Previous works on SLR focus on designing carefully engineered features [20,21,58] or modeling temporal dependencies [55,56]. Recently, the success of 3D convolutional neural networks in action related tasks [41,60,63] is transferred to SLR. In particular, the I3D [11] architecture has proven to be effective for this task [1,30,36,38,61]. In this work, we also adopt this network architecture in our sign encoder.

Sign spotting is a particular variant of sign language recognition. It aims to localize all instances of a given sign within an untrimmed video. Recent works tackle this task with auxiliary cue of subtitles, introducing automatic annotation systems by using mouthing [1], dictionaries [47] and attention maps of Transformer [61]. SPOT-ALIGN [18] extends existing spotting methods [1,47] with an iterative training schema, which alternates between repeated sign spotting and model fine-tuning. In this work, pseudo-labeling is served as our sign spotting approach to localize isolated signs in untrimmed videos from target sign language retrieval datasets. Compared with above efforts, our approach only employs a pre-trained sign encoder without utilizing additional auxiliary cue, which is proved to be simple yet efficient.

Early works of sign language retrieval primarily investigate query-by-example searching [4,70], which queries individual instances with given sign examples. Our work focuses on free-form textual retrieval—a recently introduced task by SPOT-ALIGN [18]. It symbolizes the real-world scenario of searching sign videos with natural languages. The pioneer SPOT-ALIGN [18] purely formulates sign language retrieval as a video-text retrieval task, where the cross-modal alignment is modeled upon overall global embeddings of sign videos and texts. However, the linguistic properties of sign languages are ignored. In contrast, we formulate sign language retrieval as a joint task of text-video retrieval and cross-lingual retrieval.

**Vision-Language Models.** Learning general-purpose representations for visual and textual modalities is a longstanding topic [7,35]. The idea has been investigated decades ago [49]. Recently, the prominent success of CLIP [51], ALIGN [28] and ALBEF [39] has demonstrated the capability of learning joint cross-modal representations with large-scale web data using simple image-text contrastive learning. Similar idea has also been explored in video-text retrieval, which is the closest area of our work.

The common practice in image-text retrieval [23,25,29,51] and video-text retrieval [5,24,42–44,57,68] is to encode the images/videos and texts into the overall representations. It is reasonable since the visual signals of typical image-text and video-text retrieval datasets mainly describe the certain objects or describable events. In contrast, sign videos, as the carriers of sign languages, convey abundant semantics by themselves. There exist fine-grained mappings between sign videos and natural languages. We find that identifying such cross-lingual (sign-to-word) mappings significantly boosts the performance of sign language retrieval.

**Cross-Lingual Information Retrieval.** In this work, we also formulate sign-language retrieval as a cross-lingual task. In natural language processing community, cross-lingual information retrieval (CLIR) [34,59,65] refers to the task of retrieving documents between different languages. One of the most common approaches of CLIR is to learn sentence-level embedding alignment by mapping pre-acquired monolingual embeddings [46,50,69] of different languages into a shared space. Recently, researchers have exploited fine-grained word-level mappings with self-training [2,3,59]. They find that mappings can be learned by initiating with a seed dictionary and alternating between alignment modeling and dictionary mining. Our work also exploits fine-grained word-level mappings, called sign-to-word mappings, in the context of sign language retrieval. Identifying sign-to-word mappings is challenging since sign languages are expressed in visual modality. The proposed cross-lingual contrastive learning tackles this challenge by exploiting the fine-grained cross-modal interactions.

### 3. Methodology

In this section, we first formulate the task of sign language retrieval in Section 3.1. Next, we introduce our framework which is composed of two parts: 1) a sign encoder which extracts discriminative and domain-aligned features of sign videos (Section 3.2); 2) a cross-lingual contrastive learning framework, which contrasts sign-video-text pairs while concurrently identifying the fine-grained sign-to-word mappings (Section 3.3). At last, we explore text augmentations in Section 3.4.

#### 3.1. Task Formulation

Let $\mathcal{V}$ and $\mathcal{T}$ denote a set of sign videos and their corresponding texts (transcriptions), respectively. Sign language retrieval consists of two tasks namely text-to-sign-video retrieval (T2V), and sign-video-to-text retrieval (V2T), respectively. The objective of T2V is to find the sign video $v \in \mathcal{V}$ whose signing content best matches the text query. In contrast, the reverse task V2T requires the model to identify the most relevant text (transcription) $t \in \mathcal{T}$ given a query of sign video. We resolve sign language retrieval by learning a joint embedding space of sign videos and texts.
3.2. Sign Encoder

Process Sign Videos with Sliding Window. Sign videos from sign language retrieval datasets typically contain hundreds of frames. To efficiently train our model and lower the usage of GPU memory, given a sign video, we adopt a sliding window manner to extract a discriminative and domain-aligned feature per clip. Cross-lingual contrastive learning takes \( N \) sign-video-text pairs as inputs and contrasts paired data in a shared embedding space while implicitly identifying the fine-grained sign-to-word mappings during training.
encoder, on target datasets through pseudo-labeling. The final sign encoder is composed of the well-optimized domain-agnostic sign encoder and the powerful domain-agnostic sign encoder, as illustrated in Figure 3a.

**Pseudo-Labeling on Target Datasets.** Now we describe the details of pseudo-labeling. Given a sign video from a target dataset, we adopt a sliding window with the stride of 1 and the window size of 16 to generate a set of temporally overlapped clips. For each clip, we first utilize the pre-trained domain-agnostic sign encoder to produce its prediction. Then we binarize the prediction with a pre-defined threshold $\lambda$ to generate the corresponding pseudo label. The invalid samples, whose maximum score is lower than $\lambda$, are filtered. We repeat the above process for all sign videos and eventually build a pseudo-labeled set. Our domain-aware sign encoder, which is initialized by the domain-agnostic sign encoder, is fine-tuned on the pseudo-labeled set via a standard cross-entropy loss for classification training.

**Feature Extraction with Sign Encoder.** So far, we acquire a domain-aware sign encoder approximately aligned in target domain. Nevertheless, its capability is restricted by the unavoidable noises in pseudo-labels and the limited amount of pseudo-labeled samples. Recall that we already have a powerful domain-agnostic sign encoder pre-trained on large-scale dataset in hand, inspiring us to make use of both domain-agnostic sign encoder $h_\xi(\cdot)$ and domain-aware sign encoder $h_\theta(\cdot)$ to extract discriminative and domain-aligned features. Our final sign encoder $H(\cdot)$, as shown in Figure 3a, is a weighted combination of $h_\xi(\cdot)$ and $h_\theta(\cdot)$ with a trade-off hyper-parameter $\alpha$. As described above, $H(\cdot)$ encodes sign videos in a sliding window manner. For simplicity, we use $H(v)$ to denote feature extraction on sign video $v$, which is formulated as:

$$H(v) = \alpha h_\xi(v) + (1 - \alpha) h_\theta(v).$$

(1)

### 3.3. Cross-Lingual Contrastive Learning

The objective of cross-lingual contrastive learning (CLCL) is to learn a joint embedding space of sign videos and texts while concurrently identifying the fine-grained sign-to-word mappings during training. An overview is shown in Figure 3b. CLCL takes a mini-batch $\{(v_n, t_n)\}_{n=1}^N$ containing $N$ sign-video-text pairs as input, and contrasts paired data in a shared embedding space for sign language retrieval.

**Sign Features and Word Features.** Given a sign video $v \in \{v_n\}_{n=1}^N$, we first adopt our sign encoder $H(\cdot)$ described in Section 3.2 to extract its intermediate feature. Note $H(\cdot)$ encodes sign videos in a sliding window manner, and thus there are no interactions among different clips. To facilitate information exchange, we further append a 12-layer Transformer [62] $F(\cdot)$ onto $H(\cdot)$ to extract sign features $S$ of sign video $v$, which is formulated as $S = F(H(v)) \in \mathbb{R}^{M \times D}$, where $M$ denotes the number of clips, and $D$ is the hidden dimension. Given a text $t \in \{t_n\}_{n=1}^N$, we convert $t$ into a lower-cased byte pair encoding (BPE) representation [53], which is subsequently fed into another 12-layer Transformer $G(\cdot)$ to generate the word features $W = G(t) \in \mathbb{R}^{L \times D}$, where $L$ represents word number.

Since CLIP [51] shows excellent transfer capability in various downstream tasks [17, 22, 27, 54, 66, 67], we initialize $F(\cdot)$ and $G(\cdot)$ with CLIP’s image encoder (ViT-B) and text encoder, respectively, to ease the learning. Though CLIP’s vision encoder takes image patches as inputs, we experimentally find that it generalizes well in our scenario where input data is in a different modality.

**Cross-Lingual Similarity.** There exists inherent sign-to-word mappings between sign languages and natural languages. To incorporate this prior knowledge into learning, we introduce cross-lingual similarity—an indicator to identify sign-to-word mappings between $i$-th sign video $v_i$ and $j$-th text $t_j$. Concretely, given sign features $S_i \in \mathbb{R}^{M \times D}$ of $v_i$, and word features $W_j \in \mathbb{R}^{L \times D}$ of $t_j$, we calculate a cross-lingual similarity matrix $E_{i,j} = S_i^T W_j \in \mathbb{R}^{M \times L}$. Each element in $E_{i,j}$ represents the similarity of a sign clip in $v_i$ and a word in $t_j$.

**Cross-Lingual Contrastive Learning.** Directly apply supervisions on token-wise similarity matrix $E_{i,j}$ is infeasible due to the absence of fine-grained sign-to-word annotations. Inspired by the recent progress of vision-language contrastive learning [23, 25, 29, 39, 40, 51], we turn to contrast the global representations of sign videos and texts. The underlying idea is to calculate a global similarity $z$ of $v_i$ and $t_j$ based on $E_{i,j} \in \mathbb{R}^{M \times L}$.

**Sign-Video-to-Text Contrast.** We first introduce sign-video-to-text contrast as shown in Figure 3b. To be specific, we utilize a Softmax operation to each row of $E_{i,j}$, and multiply the resulting matrix with $E_{i,j}$ to generate a sign-to-word similarity matrix $E'_{i,j} \in \mathbb{R}^{M \times T}$, where each row represents the re-weighted similarities between a sign clip in $v_i$ and all words in $t_j$. After that, we adopt a row-wise addition operation on $E'_{i,j}$ to yield the sign-to-text similarity vector $e_{i,j} \in \mathbb{R}^M$, where each element denotes a similarity of a sign clip in $v_i$ and whole text $t_j$. At last, we average all elements in $e_{i,j}$ to produce the global similarity $z$ of sign video $v_i$ and text $t_j$.

In the same way, we can calculate the similarities for both positive pairs $\{(v_i, t_j)\}_{i=1}^N$ and negative pairs $\{(v_i, t_j)\}_{i,j=1,i \neq j}$ in a mini-batch, yielding a video-to-text similarity matrix $Z_{vT} \in \mathbb{R}^{N \times N}$, where $Z_{vT}$ denotes the global similarity of $v_i$ and $t_j$. Following CLIP [51], we adopt InfoNCE loss [26] to pull the embeddings of matched image-text pairs together while pushing those of non-matched pairs apart, which is formulated as
Our model is also evaluated on Datasets.

4. Experiment

Contrastive learning is insensitive to word order. To take this into account, we provide our approach.

We define the augmented text similarity matrix $Z_{T2V}$.

**Loss Function.** The overall loss for cross-lingual contrastive learning is a weighted sum of $L_{V2T}$ and $L_{T2V}$ with a trade-off hyper-parameter $\beta$:

$$\mathcal{L} = \beta L_{V2T} + (1 - \beta) L_{T2V}. \quad (3)$$

3.4. Text Augmentation

Considering that the datasets of sign language retrieval are typically small-scale, we explore text augmentations to improve the generalization of our approach. EDA [64] introduces three simple yet efficient data augmentations in text classification task: random delete randomly removes words in a sentence; synonym replacement randomly selects words from a sentence that are not stop words and replaces them with synonyms; random swap randomly chooses two words in a sentence and swaps their positions. The first two augmentations have been proven effective in text classification task. However, we experimentally find that our focused sign language retrieval is sensitive to random delete and synonym replacement augmentations. To guarantee that the augmented texts preserve the original semantic meanings, we only adopt the random swap augmentation in our approach. We suppose there are two reasons: 1) the word order of sign languages and natural languages are constitutionally distinct, and reordering does not affect semantic meanings; 2) the proposed cross-lingual contrastive learning is insensitive to word order.

4. Experiment

4.1. Datasets and Implementation Details

Datasets. We primarily focus on How2Sign [19] dataset. Our model is also evaluated on PHOENIX-2014T [8] and CSL-Daily [71], which are primarily used for sign language recognition and translation in previous works.

How2Sign is a large-scale continuous American sign language (ASL) dataset consisting of a parallel corpus of about 80 hours of sign videos with subtitle annotations. It covers a wide range of instructional videos corresponding to various categories. There are 31164, 1740 and 2356 sign-video-text pairs in training, validation and test sets, respectively. Following SPOT-ALIGN [18], we remove the invalid pairs where the subtitle alignment is detected to exceed the video duration, remaining 31085, 1739 and 2348 available pairs in training, validation and test sets, respectively. The resolution of sign videos is 1280×720, we crop the human bodies of signers with Faster R-CNN [52] to generate valid videos.

PHOENIX-2014T is a German sign language (Deutsche Gebärdensprache, DGS) dataset collected in the domain of weather forecast from TV broadcast, consisting of 7096, 519 and 642 video-text pairs in training, validation and test sets, respectively.

CSL-Daily is a recently released Chinese sign language (CSL) dataset. The topic of CSL-Daily revolves around people’s daily lives, including 18401, 1077, 1176 parallel samples in training, validation and test set, respectively.

**Evaluation Metric.** Following previous works [18, 43, 51, 57], retrieval performance is evaluated by recall at rank K (R@K, higher is better) and median rank (MedR, lower is better). We evaluate our approach on both text-to-sign-video (T2V) retrieval and sign-video-to-text (V2T) retrieval tasks. We report R@1, R@5, and R@10 in all experiments, and additionally report MedR when comparing with state-of-the-art approaches.

**Implementation Details.** The sign encoder takes videos of resolution of 256×256 as input. The domain-agnostic sign encoder is a I3D [11] network pre-trained on BSL-1K [61]. In pseudo label generation, we set the threshold $\lambda$ to 0.6 to filter samples with low-confidence. Non-maximum suppression (NMS) with a temporal window of 24 frames is utilized to remove the duplicates among the pseudo-labeled samples. A collection of approximate 64K pseudo-labeled samples covering a vocabulary of 1220 words is eventually generated. The domain-aware sign encoder is initialized with the domain-agnostic one and fine-tuned with a learning rate of $1 \times 10^{-2}$ and batch size of 4 for 15 epochs. In the training of cross-lingual contrastive learning, the vision transformer and text transformer are initialized by the image encoder and text encoder in CLIP (ViT-B/32) [51]. The maximum length of sign clip features and text features are set to 64 and 32, respectively. The model is fine-tuned with Adam optimizer [31] with batch size of 512. The initial learning rate is set to $1 \times 10^{-5}$, which is decreased with a cosine schedule following the CLIP [51]. We set $\alpha = 0.8$ in Eq. 1 and $\beta = 0.5$ in Eq. 3. The languages of PHOENIX-2014T and CSL-Daily are German and Chinese.
Table 1. Comparison with the different variants of the pioneer SPOT-ALIGN (SA) [18] on How2Sign [19] dataset.

<table>
<thead>
<tr>
<th>Model</th>
<th>T2V R@1↑ R@5↑ R@10↑ MedR↓</th>
<th>V2T R@1↑ R@5↑ R@10↑ MedR↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>SA-SR [18]</td>
<td>18.9 32.1 36.5 62.0</td>
<td>11.6 27.4 32.5 69.0</td>
</tr>
<tr>
<td>SA-CM [18]</td>
<td>24.3 40.7 46.5 16.0</td>
<td>17.9 40.1 46.9 14.0</td>
</tr>
<tr>
<td>SA-COMB [18]</td>
<td>34.2 48.0 52.6 8.0</td>
<td>23.6 47.0 53.0 7.5</td>
</tr>
<tr>
<td>Ours</td>
<td>56.6 69.9 74.7 1.0</td>
<td>51.6 64.8 70.1 1.0</td>
</tr>
</tbody>
</table>

Table 2. Comparison with the different variants of the pioneer SPOT-ALIGN (SA) [18] on PHOENIX-2014T [8] dataset.

<table>
<thead>
<tr>
<th>Model</th>
<th>T2V R@1↑ R@5↑ R@10↑ MedR↓</th>
<th>V2T R@1↑ R@5↑ R@10↑ MedR↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>Translation [10]</td>
<td>30.2 53.1 63.4 4.5</td>
<td>28.8 52.0 60.8 56.1</td>
</tr>
<tr>
<td>SA-CM [18]</td>
<td>48.6 76.5 84.6 2.0</td>
<td>50.3 78.4 84.4 1.0</td>
</tr>
<tr>
<td>SA-COMB [18]</td>
<td>55.8 79.6 87.2 1.0</td>
<td>53.1 79.4 86.1 1.0</td>
</tr>
<tr>
<td>Ours</td>
<td>69.5 86.6 92.1 1.0</td>
<td>70.2 88.0 92.8 1.0</td>
</tr>
</tbody>
</table>

Table 3. We additionally provide a baseline for CSL-Daily [71] dataset.

4.2. Comparison with State-of-the-art Methods

We compare our method with different variants of the pioneer, called SPOT-ALIGN [18], on How2Sign and PHOENIX-2014T. We also provide the results on CSL-Daily as a baseline.

Table 1 and Table 2 show the comparisons between our approach and SPOT-ALIGN [18] on How2Sign and PHOENIX-2014T, respectively. SPOT-ALIGN builds the final combination (COMB) model by integrating its primary cross-modal (CM) model with an auxiliary retrieval model (sign recognition (SR) model for How2Sign and Translation [10] for PHOENIX-2014T). Our method outperforms the COMB model, which achieves best results in SPOT-ALIGN, by large margins, achieving +22.4 T2V and +28.0 V2T R@1 improvements on How2Sign, +13.7 T2V and +17.1 V2T R@1 improvements on PHOENIX-2014T. It is worth mentioning that the SPOT-ALIGN conducts three rounds of sign spotting [1,47] and encoder training. In contrast, we simplify the training of sign encoder and only perform a single round of training on pseudo-labeled data. We also provide a baseline on CSL-Daily dataset as shown in Table 3, demonstrating that our model can be generalized to various sign languages.

4.3. Ablation Study

We conduct all ablation studies on the most challenging How2Sign dataset.

The Effectiveness of Each Proposed Component. Table 4 shows the ablation of each component. We first build a baseline where the sign encoder is the domain-agnostic one and contrastive learning is trained with the standard contrastive loss [51]. Then we add the proposed sign encoder (SE), cross-lingual contrastive learning (CLCL) and text augmentation (TA) to the baseline step by step. SE encodes domain-relevant and discriminative features, yielding +1.8 T2V and +4.5 V2T R@1 gains. The proposed CLCL significantly boosts the performance by +20.7 T2V and +19.1 V2T R@1 improvements, demonstrating that identifying fine-grained sign-to-word mappings is essential for sign language retrieval. The introduction of TA further promotes the retrieval task, achieving 56.6 R@1 on V2T and 51.6 R@1 on T2V, respectively.

Various Sign Encoders. As described in Section 3.2, our sign encoder is composed of a domain-agnostic sign encoder (Single-Ag) [61] and a domain-aware sign encoder (Single-Aw). We also study different ways to integrate the features extracted by Single-Ag and Single-Aw, including average and weighted sum.

Variants of Cross-Lingual Contrastive Learning. As described in Section 3.3 and illustrated in Figure 3b, through the use of a combination of softmax, multiplication, sum, and average operations, we convert the fine-grained cross-lingual (sign-to-word) similarity to the coarse-grained sign-video-to-text similarity to enable contrastive learning. We
refer to this process as “Softmax”. Here we study two variants termed “Mean” and “Max”. The “Mean” and “Max” strategies simply replace the combination of softmax, multiplication and sum operations with a simple mean operation and a max operation, respectively. The results are shown in Table 6. We observe that the “Mean” strategy performs worst since it merely evaluates the overall similarity of a text and a sign video and ignores the fine-grained sign-to-word mappings during training. The “Max” strategy identifies hard sign-to-word mappings, i.e., one sign is associated with the most similar word, and vice versa. Nevertheless, since we do not have the ground-truth of sign-to-word mappings, it is challenging for models to identify the confident one-to-one mappings in a weakly supervised manner, as discussed in previous works on multiple instance learning [47]. In contrast, our default “Softmax” strategy localizes the soft sign-to-word mappings and achieves the best results.

### Study of Text Augmentations

In Table 7, we study three text augmentations described in Section 3.4: random delete (RD), synonym replacement (SR) and random swap (RS). We observe a slight performance drop when introducing RD and SR into training. Sign language retrieval is not only a video-text retrieval problem, but also a cross-lingual retrieval challenge, deletion and replacement may break the intrinsic sign-to-word mappings. In contrast, RS augmentation preserves the semantics of texts and we observe a +2.6 T2V and a +1.6 V2T R@1 improvements over the counterpart without any text augmentations.

### The Effects of CLIP Initialization

In our framework, the vision and text Transformer are initialized with the CLIP’s vision encoder and text encoder. Note that the input modality of our vision Transformer and that of CLIP’s vision encoder are different. Ours takes a feature sequence as input while the input of CLIP’s image encoder is a set of image patches. In Table 8, we compare a randomly initialized baseline with the one initialized by CLIP. We find that CLIP-initialization significantly improves the performance, yielding +11.4 T2V and +11.0 V2T R@1 improvements though the input originates from a completely distinct modality.

### Study on Text Transformer

The text Transformer in our model is initialized with the CLIP’s text encoder. Considering the generalization capability of the pre-trained CLIP, we attempt to freeze the shallow layers of our text transformer. As shown in Table 9, the performance gradually decreases as the number of the frozen layers increases. CLIP is trained on large-scale image-text pairs, however, the domain gap between image-text data and sign-video-text data is non-negligible. Though CLIP shows promising transfer capacity, it is optimal to fine-tune the whole model on target datasets.

### 5. Conclusion

In this paper, we propose a novel framework named Cross-lingual Contrastive learning (CiCo) for recently introduced sign language retrieval task. We formulate sign language retrieval as a cross-lingual retrieval task as well as a video-text retrieval problem. CiCo models the fine-grained cross-lingual mappings between sign videos and texts via the proposed cross-lingual contrastive learning. We also introduce a sign encoder, which is composed of a domain-agnostic encoder and a domain-aware one, to extract discriminative and domain-aligned features. Our Cico outperforms the pioneer SPOT-ALIGN by large margins on How2Sign and PHOENIX-2014T benchmarks. We also provide a baseline on CSL-Daily. We hope our approach could serve as a solid baseline for future research.

### Acknowledgement

This work was partially supported by National Natural Science Foundation of China (No. 62072112), National Key R&D Program of China (No. 2020AAA0108301), and Scientific and Technological Innovation Action Plan of Shanghai Science and Technology Committee (No. 22511101502, 22511102202).
References


[28] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In International Conference on Machine Learning, pages 4904–4916. PMLR, 2021. 3


[38] Dongxu Li, Xin Yu, Chenchen Xu, Lars Petersson, and Hongdong Li. Transferring cross-domain knowledge for video sign language recognition. In IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 6205–6214, 2020. 3


[51] 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 super-


[70] Shilin Zhang and Bo Zhang. Using revised string edit distance to sign language video retrieval. In International Conference on Computational Intelligence and Natural Computing, volume 1, pages 45–49. IEEE, 2010. 3


[73] Ronglai Zuo, Fangyun Wei, and Brian Mak. Natural language-assisted sign language recognition. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023. 1, 3