APPENDIX

A Building *SignCorpus*

In this section, we explain the various stages of data processing in detail for preparing the large pretraining dataset, from curation of sources till storage of quality-checked final data.

A.1 Data curation

SL	Sources
American	LifePrint Signs, The Bible Recap, ASL Meredith, ASL Dictionary, AzEIN, Foundations of Literacy, IRSVideosASL, Kelly Keller, Centers for Disease Control and Prevention, Sign1News, ASL THAT, The Daily Moth, Deaf Newspaper, Start ASL
Australian	Start ASL, deaftvaussie, ASHPHYXIA, AuslanforFamilies, DeafSportRecQld, Can:Do Classroom, The Deaf Society, Access Plus WA Deaf, CommRepublic, Australian Government Development of Health, Conexu Foundation
British	Signature Deaf, dwpsign, Woodgreen Evangelical Church, Providence Baptist Chapel Bedford, The Deaf Academy, BSL learning with Mel, Deaf Business Academy, StorySign UK, National Deaf Children's Society, 57davison, SheffieldCCouncil, Clive Morgan, Commanding Hands, UK Parliament, Northamptonshire Healthcare NHS Foundation Trust, Royal Association for Deaf people, 2011Census, rdashcommunications, South Gloucestershire Council, NHS Greater Glasgow and Clyde, Tabernacle Deaf
Chinese	babysign.org, ChineseSignLanguage.org, PNNPTS, pei fen, Bridge Creative Ventures, Learn CHINESE channel - Largest Video Library, Lin Hao, Zhong Guo Shou Yu, HongKong Bible Conference
Greek	StarTvGreece, OpenTV.gr, newsontime.gr, Noimatizoume, Theodoros Martzos, Fotis Kokkos, HandsUp
Indian	ISH News, MBM Vadodara, NewzHook, NIOS, SignLibrary
Korean	KBS News, Korea Welfare University, MyHand MySign, Gyeonggi Deaf Broadcasting, DBN Korea Deaf Broadcasting, Human Korean, DEAFiN, HE N
Russian	GARAGEMCA, Videos in Sign Language, Victor Fabry, Ksenia B., Channel Right, DeafSign RUSSIA, City of Gestures, School of Sign Language, IRRTVdeaf, Irk VOG, Russian Bible Society, Anna Leo, The Arts Museum, ARK Media, Videos for Deaf, ParfenovDEAF, WeDeaf, Deafmos Media, VOGinfo, UslishMir, Gesture lang
Spanish	Visualfy, InformativosTvc, ASORNA, CNLSE Youtube, INSOR Colombia, Roberto De Angeliss, Direccion de Politicas Inclusivas IDM, Taida R. ASAE, LSE Facil, Laory82, MOVILSE, Aprender - Entre Rios
Turkish	Celal Uca, Isaret Adam, Isaret TV, Isaret Dili Egitimi, titivi, Bilal OzTurk, IEEF, TRT Haber, eren utku kayaarslan

Table 6: List of all YouTube channels from which we curate the SignCorpus dataset, along with their hyperlinks

In Table 7, we show the total number of channels, playlists and videos that we crawl for each of the 9 new sign languages we study in our work. We observe that videos for American sign language is relatively much easily available on the Internet compared to other sign languages, hence accounting for more than half of the total videos we scrape in total. We also open-source the precise playlist-wise statistics explaining the properties of the dataset through a spreadsheet[3].

Sign Language	Number of Channels	Number of Playlists	Number of videos
American	14	26	46,781
Australian	11	11	1,124
British	26	102	4,026
Chinese	11	19	9,551
Greek	7	8	3,810
Korean	8	13	7,253
Russian	21	51	9,246
Spanish	13	13	1,655
Turkish	9	11	974
Total	120	254	84,420

Table 7: Channel-wise and playlist-wise statistics of videos across 9 new languages that we collect from YouTube

We further study the distribution of all the content by looking at their domains to which each playlist belongs to. The consolidated domain-wise statistics is shown in Table 8, indicating the number of hours of content collected from each domain. We see that news content dominate the overall new dataset, accounting for more than 60% of the total data. The top-2 next domains are from tutorials and religious content, which account for more than 30% of the content.

---

[3] Spreadsheet: *SignCorpus* detailed stats

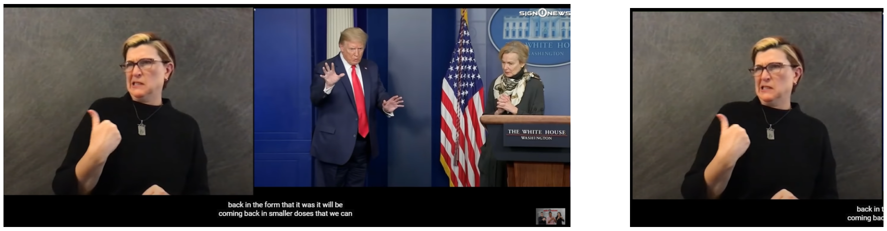
Domain	**Approximate Number of Hours**
News	3346
Sign Language Tutorials	849
Religious Information	703
Advice & Information	91
Conferences	69
Health	42
Dictionary	37
COVID Information	27
General Information	17
Poems	14
History	13
StoryTelling	11
Policies & Procedures	10
Crafts & Arts	8
Mental Health	7
Services	5
Communication	5
Tax Information	4
Teaching Bussines	4
Facts	4
Smart Energy	2
Prospectus Information	1
Prologue	1
Census	1
Fire Terminology	1
Employment & Career advice	1
Total	5274

Table 8: Overall domain-wise statistics along with corresponding number of hours in the *SignCorpus* dataset

A.2 Preprocessing of videos

A.2.1 Cropping the region of interest

In many news and other explanatory videos, we find that usually the signer is localized to a certain spatial-region in the video, while the remaining region of the frame (or even the full frame) would be dedicated to playing video-excerpts related to the content being described. In our case, since we are interested only with signing activity for pretraining purposes, we crop out only the region-of-interest where the signer would be predominantly present in the video by manually estimating the coordinates. We find that in almost all playlists, the cropping coordinates for signer remains consistent, although there are variations among different playlists in a channel.


SIGN1NEWS
THE WHITE HOUSE
WASHINGTON
back in the form that it was it will be
coming back in smaller doses that we can

(a) Video from a news reporting channel.

(b) After cropping the signer region.

Figure 3: Example demonstrating why spatial-cropping is essential

A.2.2 Removing ambiguous regions

Upon manually inspecting the videos, we find that some sign language videos recorded in-the-wild can sometimes have more than 1 person in a frame besides the signer, or there are also cases where the recorded videos were a conversation between 2 or more signers. Also, sometimes when the focus of the video is shifted from a signer to a different context (like playing a sample video in full-screen from an event in news channels), there could be no person some video regions. To remove all such ambiguous regions, we use an open-source person detector model and run it across all the videos we collect. It is to be noted that for computational efficiency, we only run it for 1 frame every 2 seconds. We then drop all the regions which have no person or more than 1 person, and dump the resulting temporally-cropped video as different videos since the frame-dropping removes the continuity of signing. As mentioned in Section 3.3, this results in a decrease of the overall content size by 6.5%.

A.3 Subtitles Extraction

In addition to releasing the mere poses, for some amount of the entire *SignCorpus*, we align the existing subtitles for the videos that we collected. Although we do not use the subtitles for pretraining or any other purpose in our current work, we believe that releasing this data as well would be beneficial to the research community to explore ways to exploit these additional signals in-addition to just pose inputs. For instance, these noisy subtitles could serve as pseudo-labels for CSLR training or pretraining.

In this sub-section, we briefly explain how we consolidate subtitles for videos that have captions. Since all the data that we collect are from YouTube, it is possible for videos to have subtitles along with it, which either transcribes the signer or the audio explanation in background, usually all 3 being in-sync. We categorize the collected videos into four types based on the type of captions available:

1. *Manually-uploaded captions* – Subtitles that are uploaded by the content creators.
2. *Auto-generated captions* – Subtitles that are automatically created by YouTube after trascribing the audio using their speech-to-text models.
3. *Embedded captions* – Subtitles that are hard-coded/burnt into the video frames, hence not available in digital format.
4. *No captions* – Videos that do not have any subtitles.[4]

Table 9 shows the statistics for each of the 9 new sign languages in *SignCorpus*, along different caption-types mentioned above. It can be seen that around 32% of data have subtitles (including both auto-generated and manually uploaded), around 6% of data have subtitles burned-into the video itself, and the remaining 62% of data do not have any captions (including embedded captions for languages for which good quality OCR is not available online).

Sign Language	No Captions	Uploaded Captions	Embedded Captions	Auto-generated Captions
American	660.37	202.84	50.67	2
Australian	54	13.86	7.91	0
British	524.79	26.73	27.96	197.28
Chinese	133.19	181.02	0	0.89
Greek	427.26	0.72	0	56.88
Korean	82.26	0.99	94.46	253.32
Russian	237.54	63.02	18.27	115.54
Spanish	106.14	12.5	0	47.84
Turkish	37.11	0	0	13.56
Total	2262.66	501.68	199.27	687.31

Table 9: Caption-wise statistics (in hours) for each language in *SignCorpus*

A.3.1 Digital subtitles

Upon manual analysis, we find that the auto-generated subtitles are of relatively lesser quality than that of manually-uploaded subtitles, which is not surprising as the outputs of speech-to-text is not accurate enough. We also note that the alignment of the subtitles with the actual signing is not perfect, because usually a sentence is broken into 2 or more subtitles based on the spoken audio in the background, which differs from the order in which signing is performed. Hence we recommend any users of this data to use the subtitles by considering them at the sentence-level instead of chunk-level, although this might not be possible for auto-generated captions as they do not have any punctuations (due to streaming ASR).

---

[4] It is to be noted that although the videos could have audio in-addition to signing activity, we do not attempt to transcribe them using speech-to-text systems in our current work.

A.3.2 Embedded subtitles

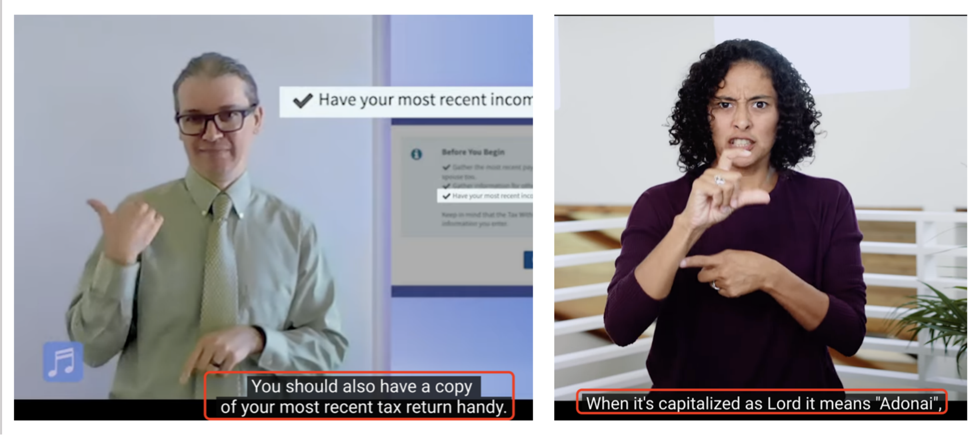

Have your most recent incom
You should also have a copy
of your most recent tax return handy.
When it's capitalized as Lord it means "Adonai",

Figure 4: Sample frames showing videos containing embedded captions

For extracting captions that are hard-coded in the video, we employ an OCR-based approach to digitize the text content as briefly shown in Figure 5. We use a text detection model to localize the subtitle box, as shown in Figure 4 to perform validation if the subtitles are present in the region of interest. Then we perform text recognition on the detected subtitle region to extract text. We achieve both using an open-source library called EasyOCR[5], which supports Russian, Korean, and English (for American, Australian and British SLs). It is to be noted that we do not run the recognition model for all the frames; rather, for each frame f_{t+k} which has text detected, we check if there is any change in the subtitle by performing a structural similarity comparison (SSIM) with the previously detect keyframe f_t. We perform OCR for a frame f_{t+n} only when the SSIM value of that region against that of f_t exceeds a set threshold. We also then perform fuzzy matching of the extracted content with the previous text content to check if there is any real change (in-case SSIM-based check fails), only upon which we consider it as a next valid subtitle.

To evaluate the quality of OCR-extracted subtitles, we manually subtitle a set of 10 videos from a British playlist, and compare it against the corresponding auto-extracted subtitles (after passing the sentences through a language model called Gramformer to attempt to correct spelling mistakes from OCR). We find that on average, the WER score was around 13.2, thus indicating that although the extracted text-data is not very accurate, it can still serve as useful signals during multi-modal pretraining.

A.4 Processing the data

As explained in our paper, the *SignCorpus* dataset is based on pose-modality. In this section, we explain the details of how pose extraction works, how we perform quality checking, and final data storage format.

---

[5] `https://github.com/JaidedAI/EasyOCR`

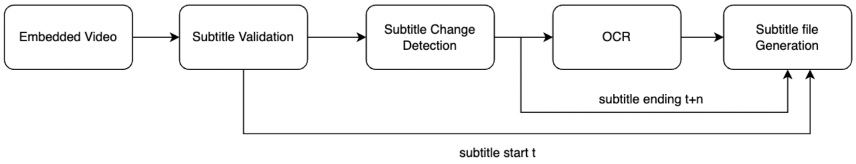
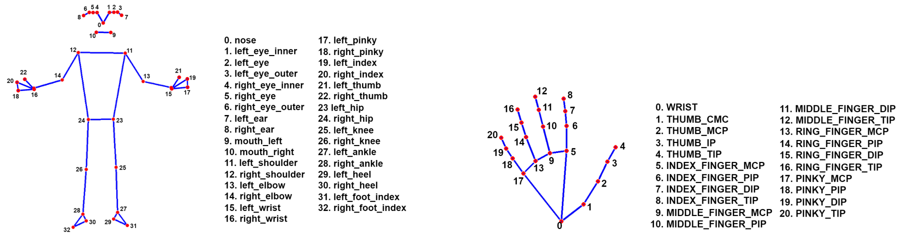


Figure 5: Subtitle Extraction Pipeline

A.4.1 Pose Extraction

As mentioned in Section 3.2, we use the MediaPipe platform to perform pose extraction to obtain the whole body keypoints as well as hand keypoints. For each frame (with a person), we save 75 keypoints – 33 pose landmarks from the body (as shown in Figure 6[a]) and 21 additional landmarks for each hand (as shown in Figure 6[b]).



(a) 33 body keypoints

(b) 21 hand keypoints

Figure 6: Pose landmarks extracted using MediaPipe

For sign language, since only the upper body (including hands and face) is predominantly used to communicate, we ignore the keypoints from the lower body when we train SLR models, hence using a total of 59 keypoints as shown in Figure 7.

A.4.2 Quality Checking

As mentioned in Section 3.3, we perform various quality checks to ensure we store only the clean data finally. Table 10 shows the various quality checks that we perform on the pose videos of each sign language, also indicating the average fraction of total frames for which the QC was successful.

SL	Total Frames	Body	Wrist	Positional	Left Hand	Right Hand	Overall Pass
ASL	89442498	0.95	0.95	0.97	0.74	0.8	0.96
Auslan	7304151	0.98	0.97	0.98	0.79	0.8	0.95
BSL	75702498	0.97	0.97	0.96	0.66	0.64	0.87
CSL	30477634	0.97	0.97	0.98	0.78	0.72	0.97
Greek	48588262	0.99	0.99	0.99	0.7	0.58	0.98
KSL	53411325	0.97	0.97	1.0	0.61	0.67	0.99
Russian	39331567	0.98	0.98	0.98	0.76	0.78	0.95
Spanish	14374405	0.99	0.99	0.98	0.64	0.62	0.97
Turkish	4693355	0.92	0.91	0.99	0.7	0.73	0.97

Table 10: Fractions of Data passing through each of the Pose-based Quality Checks (QC) that we perform for each new language in *SignCorpus*

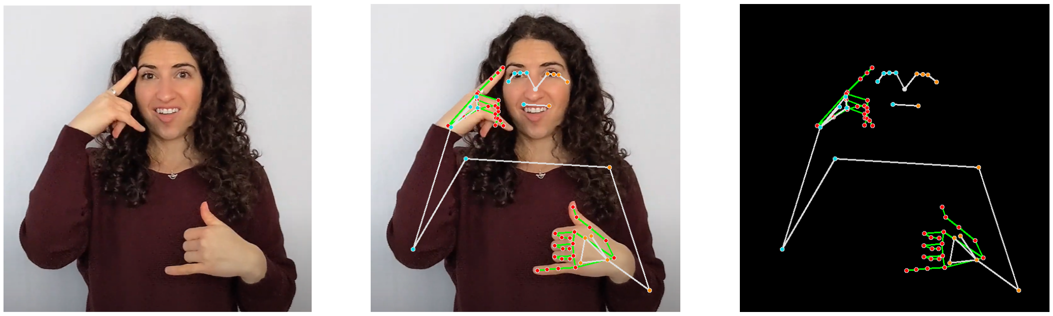


Figure 7: (a) Original Image (b)Keypoints along with image (c) Keypoints

We briefly explain each quality check below:

- **Body**: Fraction of frames which contains 70% of the first 23 pose landmarks corresponding to the upper-body (see Figure 6[a]).
- **Wrist**: Fraction of frames containing either of the left or right wrist.
- **Positional**: Fraction of frames which are positionally correct. For instance, we check if the y coordinate of shoulder lies between nose and hip of the person; if x coordinate of nose between x coordinates of right and left shoulder, etc.
- **Left Hand**: The conditional fraction of frames containing all 21 left hand landmarks out of those frames in which left hand wrist is present.
- **Right Hand**: The conditional fraction of frames containing all 21 right hand landmarks out of those frames in which right hand wrist is present.
- **Overall pass**: Fraction of data passing through all the sanity checks, after filtering the data using different thresholds for each QC.

As mentioned in Section 3.3, we set different QC threshold for different types of video content for each sign language, as shown in Table 11. These thresholds have been obtained after observing many videos manually, which we found to be usually consistent for a given channel. Hence we decide to set the thresholds for each QC based on the video category to ease the process of filtering the content per language. The threshold for Positional check (shown in Table 10) is set to 85% irrespective of the category of the video as we always allow only till 15% of the bad frames (where signer's orientation is not front facing) in a video.

Sign Language	Body	Wrist	Left Hand	Right Hand
American	33.02/51.48/-	16.37/44.87/-	17.33/23.32/-	22.87/27.48/-
Australian	-/59.06/59.06	-/56.79/59.79	-/38.70/38.70	-/41.04/41.04
British	35/38/37	23/15/37	20/30/20	20/30/20
Chinese	0/54.53/52.91	0/58.86/53.17	0/24.16/24.16	0/25.32/25.32
Greek	-/90/55	-/90.94/52	-/14.9/14.05	-/14.9/10.49
Korean	50/79.04/63	50/75.84/63	10/34/33.79	10/34/33.79
Russian	12.26/82.70/54.95	12.26/82.57/54.73	0/29.0/20.63	0/29.72/29.84
Spanish	-/71.97/62.94	-/54.56/54.52	-/15.18/31.2	-/15.0/ 32.8
Turkish	-/53.15/53.15	-/57.98/57.98	-/17.51/17.51	-/14.82/14.82

Table 11: Quality Check thresholds for each category (isolated/continuous/multiple-isolated) across various languages in *SignCorpus*.

A.4.3 Final dataset storage

We store our final datasets after quality-checking in a hierarchical data format called HDF5, as shown in Figure 8. We choose this format because HDF5 files are self-describing and allows storing rich metadata along with the main data. We create individual HDF5 files for each playlist processed in this work, across different channels and sign languages. Inside each HDF5, we store the data in 3 different groups – *keypoints*, *confidence_scores* and *metadata*. The *keypoints* group contains the pose landmarks extracted for all the videos in the playlist, and this is the only data that we use in our work for pretraining. The *confidence_scores* group contains a fractional value for each keypoint in all videos, indicating the level of visibility of that specific keypoint. The *metadata* group contains all the miscellaneous details about the playlist, like video URLs, duration, video title, etc.

B Creation of label-aligned data for ISLR and finger-spelling

B.1 Rules for vocabulary normalization and unification

As explained in Section 5.1, for all the existing 11 ISLR datasets we study in this work, we label-align all the vocabularies to a common form, precisely a normalized standard of English glosses. The following list specifies the instructions that we create and comply-with to standardize the vocabulary:

1. Use capital letters only.

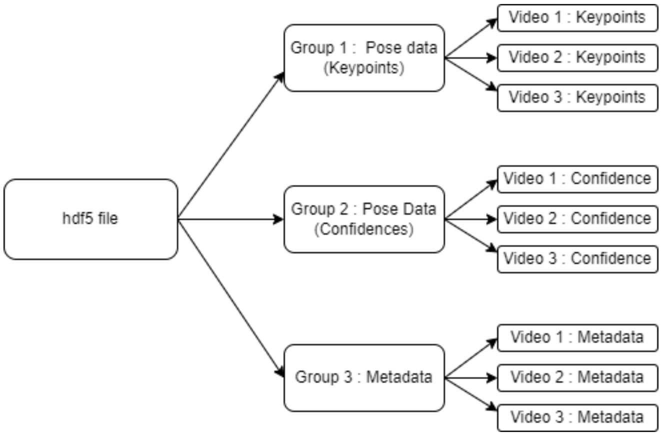


Figure 8: HDF5 file schema used for *SignCorpus* data storage

2. Do not allow single letter words, since it is reserved only for fingerspelling alphabet (for e.g., use "ME" instead of "I" for the first-person nominative pronoun)
3. **Lemmatization:**
    - Avoid plural as much as possible (unless dataset has both singular & plural)
    - Avoid "-ing" forms of verbs.
4. Use ’_’ for spaces and hyphens. No other punctuations marks were used.
5. Use ’+’ for finger-spelt words (D+O+G), and use only A-Z (irrespective of any language)
6. Standardize with a single word if multiple synonyms are present (e.g., "small" vs "little", "large" vs "big")
7. Unify all the different ways to sign a gloss (like "YOU (1)", "YOU (2)", . . . –> "YOU")
8. Prefer American spelling over British spelling (e.g. COLOR vs COLOUR, REALIZE vs REALISE)
9. Remove words in brackets, and include it along with the word if relevant. (e.g. "Grandma (Paternal)" –> PATERNAL_GRANDMOTHER")

B.2 Finger-spelling

As explained in Section 5.2, we curate data for finger-spelling for 7 different sign languages from different sources, as listed in Table 12 with hyperlinks. Except Greek, all the other sign languages predominantly use the English-based roman alphabet for fingerspelling. In total, we find 59 unique characters in *MultiSign-FS* which we directly use for the multilingual experiments explained in Section 6.

SL	**Sources**	**Alphabet**
American	WLASL, MSASL, ASLLVD, SpreadTheSign	A–Z
Argentine	YouTube (1, 2, 3, 4, 5, 6, 7, 8)	A–Z, Ñ
Chinese	DEVISIGN, SpreadTheSign, YouTube (1, 2, 3, 4)	A–Z, CH, NH, SH, ZH
German	SpreadTheSign, Gebärden lernen, YouTube (1, 2, 3, 4, 5)	A–Z
Greek	SpreadTheSign, YouTube (1, 2, 3, 4, 5, 6)	A,B,Γ,Δ,E,Z,H,Θ,I,K,Λ,M,N,Ξ,O,Π,P,Σ,T,Υ,Φ,X,Ψ,Ω
Indian	SpreadTheSign, YouTube (1, 2, 3, 4, 5, 6, 7, 8)	A–Z
Turkish	SpreadTheSign, YouTube (1, 2, 3, 4, 5)	A–Z, Ç, Ö, Ü, Ğ, İ, Ş

Table 12: List of all sign languages for which we curate fingerspelling data, along with their sources and alphabet list

C Experimental settings

C.1 Model implementation and Training setup

We implement the model using the PyTorch machine-learning framework, and use Lightning library for all experiments. We implement the backbone network by adapting the decoupling GCN code implemented in the original work that extend it to SL-GCN[6]. For implementing the pretraining strategy, we adapt the original code for Dense Predive Coding by extending it to sign language using the above mentioned backbone. All the code and configuration files will be made public via the 👐`OpenHands` library. We perform the same normalization and augmentations during training as explained in our previous work.

C.2 Hyperparameters

In this subsection, we briefly mention the settings that we follow for all the experiments we perform after performing a hyperparamter search. All the training and fine-tuning experiments are run on a virtual machine with Nvidia A30 GPU.

For pretraining we used a learning rate of 10^{-4}, a batch size of 128, with each epoch taking around 24 minutes (where each epoch denotes one cycle through all the videos throughout the dataset sampling one window from each video to train).

For ISLR fine-tuning, with CosineAnnealing as the learning rate scheduler, we use a learning rate of 10^{-3} with batch size 32 and during the last few epochs, we increase the batch size to 64 (without changing the LR). For *MultiSign-ISLR* , we model was finetuned for around 200 epochs, with each epoch taking around 30 minutes.

For finger-spelling, the baseline model is a lightweight SL-GCN with 2 layers of decoupling GCN. Owing to the small size of the model, the average training time for each language was around 20minutes, with a batch size of 16 and learning rate 10^{-3}. For the multilingual fine-tuning experiments, we use a LR of 10^{-4} and batch size of 32, for which the total training time ranged from 1.5-2 hours since it is a relatively smaller dataset compared to ISLR.

---

[6] `https://github.com/jackyjsy/CVPR21Chal-SLR`

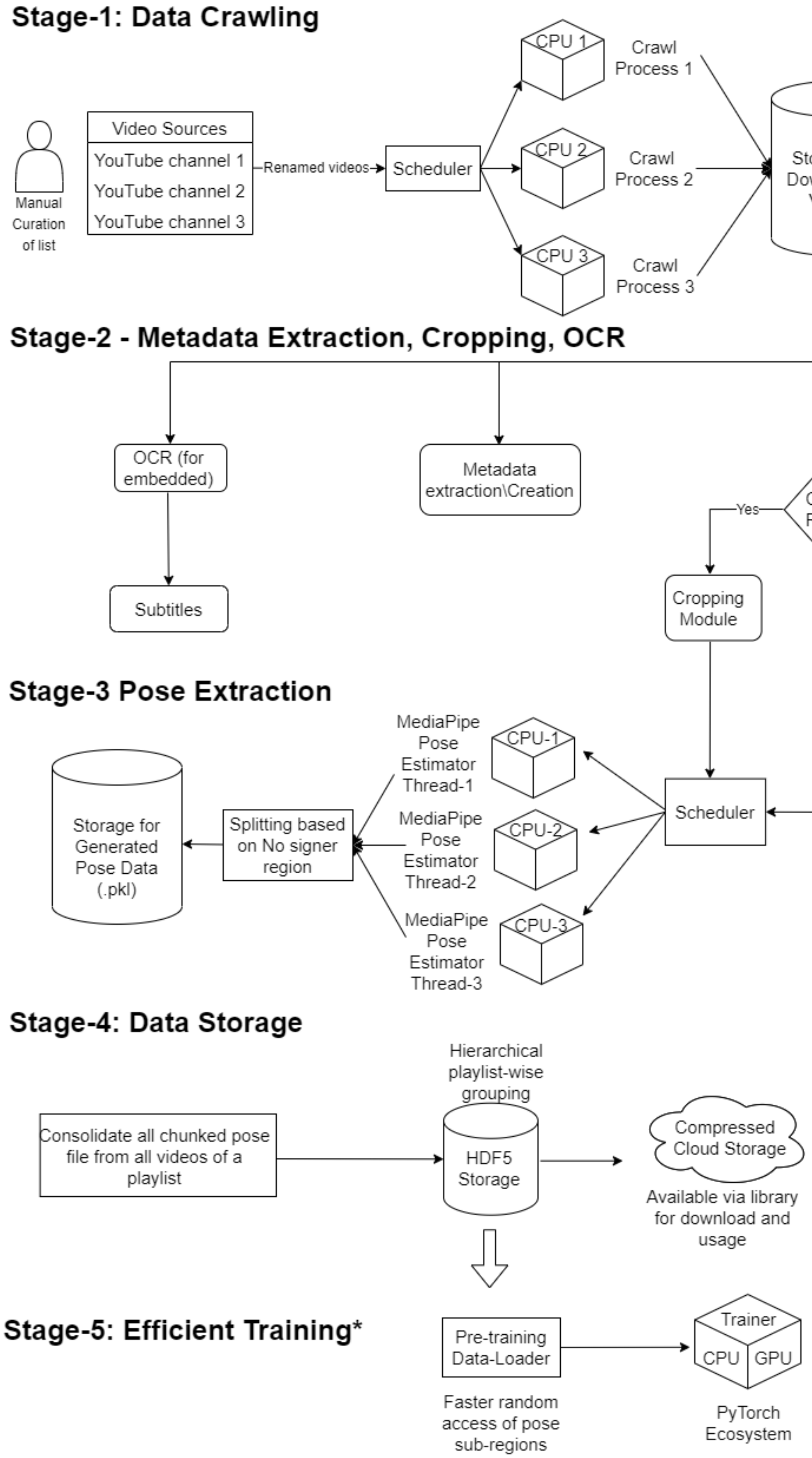

Stage-1: Data Crawling
Manual Curation of list
Video Sources
YouTube channel 1
YouTube channel 2
YouTube channel 3
Renamed videos
Scheduler
CPU 1
CPU 2
CPU 3
Crawl Process 1
Crawl Process 2
Crawl Process 3
Storage for Downloaded Videos
Stage-2 - Metadata Extraction, Cropping, OCR
OCR (for embedded)
Subtitles
Metadata extraction\Creation
Cropping Required
Yes
No
Cropping Module
Stage-3 Pose Extraction
Scheduler
CPU-1
CPU-2
CPU-3
MediaPipe Pose Estimator Thread-1
MediaPipe Pose Estimator Thread-2
MediaPipe Pose Estimator Thread-3
Splitting based on No signer region
Storage for Generated Pose Data (.pkl)
Stage-4: Data Storage
Consolidate all chunked pose file from all videos of a playlist
Hierarchical playlist-wise grouping
HDF5 Storage
Compressed Cloud Storage
Available via library for download and usage
Stage-5: Efficient Training*
Pre-training Data-Loader
Faster random access of pose sub-regions
Trainer
CPU
GPU
PyTorch Ecosystem

Figure 9: The full pipeline built for processing and preparing the *SignCorpus* dataset