SignCLIP: Connecting Text and Sign Language by Contrastive Learning

Anonymous ACL submission

Abstract

We present SignCLIP, which re-purposes CLIP (Contrastive Language-Image Pretraining) to project spoken language text and sign language videos, two classes of natural languages of distinct modalities, into the same space. SignCLIP is an efficient method of learning useful visual representations for sign language processing from large-scale, multilingual video-text pairs, without directly optimizing for a specific task or sign language which is often of limited size.

We pretrain SignCLIP on *Spreadthesign*, a prominent sign language dictionary consisting of ∼500 thousand video clips in up to 44 sign languages, and evaluate it with various downstream datasets. SignCLIP discerns in-domain signing with notable text-to-video/video-to-text retrieval accuracy. It also performs competitively for out-of-domain downstream tasks such as isolated sign language recognition upon essential few-shot prompting or fine-tuning.

We analyze the latent space formed by the spoken language text and sign language poses, which provides additional linguistic insights. Our code and models are openly available[1].

1 Introduction

Sign(ed) languages are the primary communication means for ∼70 million deaf people worldwide[2]. They use the visual-gestural modality to convey meaning through manual articulations in combination with non-manual elements like the face and body (Sandler and Lillo-Martin, 2006). Sign language processing (SLP) (Bragg et al., 2019; Yin et al., 2021) is a subfield of natural language processing (NLP) that is intertwined with computer vision (CV) and sign language linguistics.

SLP spans the tasks of sign language recognition (Adaloglou et al., 2021), translation (De Coster et al., 2023), and production (Rastgoo et al., 2021).

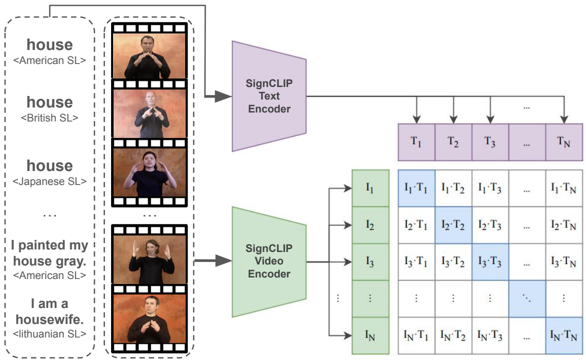

Figure 1: Illustration of SignCLIP, comprising a text encoder and a video encoder jointly trained on pairs of text and multilingual signing examples. Every sign is articulated in diverse languages and contexts with subtle differences in hand shape, movement, place of articulation, etc. The matrix part is taken from CLIP.

Typical datasets, equipped with spoken language text and sign language glosses in addition to videos, that support SLP research are RWTH-PHOENIX-Weather 2014T, in German Sign Language (DGS), introduced by Forster et al. (2014); Camgoz et al. (2018); and CSL-Daily, in Chinese Sign Language (CSL), introduced by Zhou et al. (2021). However, advances on a specific task/dataset/language are often limited and non-transferable to more generic and challenging settings (Müller et al., 2022, 2023) due to the small, domain-specific vocabulary (1,066 and 2,000 signs, respectively) and data size (11 and 23 hours, respectively). Recent sign language corpora of a thousand signing hours have emerged for relatively high-resource sign languages, e.g., BOBSL (Albanie et al., 2021) for British Sign Language (BSL) and YouTube-ASL (Uthus et al., 2024) for American Sign Language (ASL).

In the meanwhile, outside the world of SLP, there is great progress in deep pretrained models of different modalities, GPT (Achiam et al., 2023) for text, masked autoencoders (He et al., 2022) for images, and wav2vec 2.0 (Baevski et al., 2020) for speech,

[1]The link is hidden for anonymous review.
[2]https://wfdeaf.org/our-work/

to name a few. They are commonly pretrained on a huge amount of data (e.g., over 15 trillion tokens for Llama 3, Touvron et al. (2023)) with very little or weak supervision, but present striking multilingual and multi-task ability, for zero-shot prediction, fine-tuning, and representation learning.

Ideally, the visual aspect of sign languages, i.e., larger lexical similarity due to iconicity (Johnston and Schembri, 2007) (illustrated in Figure 1) and smaller grammatical variation (Sandler and Lillo-Martin, 2006) than spoken languages, makes transferring between different sign languages easier than the text of different spoken languages or even scripts. The latter faces unfair tokenization issues (Petrov et al., 2024) and out-of-vocabulary errors. At the same time, unlike discrete text tokens (see the comparison in Table 1), the dense continuous video signal is expensive to process computationally and seems daunting for the above-mentioned self-supervised training approaches.

Since SLP tasks and datasets usually involve both the text and visual/signed modalities, we take inspiration from OpenAI's CLIP model (Radford et al., 2021) but use contrastive learning to connect text with sign language videos instead of images. For video understanding, follow-up work such as VideoCLIP (Xu et al., 2021) mainly deals with tasks including action recognition (Zhu et al., 2020) and VideoQA (Xu et al., 2016; Yu et al., 2018). However, both CLIP, VideoCLIP, and other existing multimodal models understand visual content on a coarse-grained level and generic domain and do not address the intricacy of sign language. We show the lack of sign language understanding in contemporary AI models both intuitively (Figure 3 in Appendix A) and empirically (Table 2).

This work uses sign-language-specific data to train a CLIP-like model for SLP. We first validate the approach's feasibility on fingerspelling, a subsystem of sign language, by a model named FingerCLIP (§4), which correctly understands the fingerspelling of individual letters. We then curate the Spreadthesign[3] dictionary as a large-scale pretraining dataset consisting of ∼500 hours of signing, as well as diverse public downstream task datasets to run comprehensive pretraining, fine-tuning, and evaluation on full-fledged sign languages. By contrastive training on ∼500 thousand video-text pairs, we obtain a multimodal and multilingual model named SignCLIP (§5), illustrated by Figure 1. SignCLIP excels at various SLP tasks and datasets and presents a compelling latent space for signed video content aligned with spoken language text (§6).

[3]https://www.spreadthesign.com/. The use of the data is under a license granted by Spreadthesign.

2 Background: Sign Language Representation

Representation is a key challenge for SLP. Unlike spoken languages, sign languages have no widely adopted written form. As sign languages are conveyed through the visual-gestural modality, video recording is the most straightforward way to capture them. The final goal of SignCLIP is to represent a sign language video clip by an embedding that aligns with a ubiquitous text encoder like Sentence-BERT (Reimers and Gurevych, 2019). Cheng et al. (2023), a work related to ours, uses a combination of a domain-agnostic and a domain-aware video encoder to address video-to-text retrieval. Generally, end-to-end training on the raw videos is computationally costly, and various intermediate representations alleviate this issue.

VideoCLIP and Video Encoders Our work adapts VideoCLIP, which is pretrained by general instructional videos from the HowTo100M (Miech et al., 2019) dataset. We aim at replacing HowTo100M with domain-specific sign language videos, such as the dataset How2Sign (Duarte et al., 2021), albeit on a considerably smaller scale.

Videos are very dense temporally (frame rate) and spatially (video resolution). A 3D-CNN-based video encoder is often used to extract informative features with reduced dimensionalities for downstream tasks. VideoCLIP uses an S3D (Zhang et al., 2018) model pretrained on HowTo100M that produces one video token (i.e., a video embedding for the temporal window) per second. For SLP, it is possible to use a video encoder pretrained specifically on sign language videos. A prominent one is the I3D (Carreira and Zisserman, 2017) model pretrained on the BSL sign language recognition task (Varol et al., 2021) with the BSK-1K dataset (Albanie et al., 2020). A more recent approach to simultaneously address temporal and spatial complexity is the Video Swin Transformer proposed by Liu et al. (2022), and Prajwal et al. (2022) trains one such model for BSL fingerspelling recognition.

Pose Estimation A potentially more interpretable and universal way of extracting sign language-related features from videos is human pose estimation (Zheng et al., 2023), for example, using MediaPipe Holistic (Lugaresi et al., 2019; Grishchenko and Bazarevsky, 2020). Each video frame is converted into the location (X, Y, Z) of 543 full body keypoints in a 3D space. However, the immediate applicability of the pose estimation systems for SLP is questionable (Moryossef et al., 2021), and known issues such as the lack of accurate depth information are presented (Holmes et al., 2024). Generally, there is a trade-off between user-friendliness and accuracy for pose estimation tools and SLP researchers often prefer MediaPipe Holistic for the former (Selvaraj et al., 2022) over others like OpenPose (Cao et al., 2019) and AlphaPose (Fang et al., 2022). An even more universal alternative approach to track keypoints named *Tracking Everything Everywhere All at Once* is proposed by Wang et al. (2023). Sevilla et al. (2024) uses a similar CoTracker (Karaev et al., 2023) model to study sign language prosody. Such models produce smoother signals than traditional pose estimation.

Discrete Representation Non-standard written forms of sign language, including SignWriting (Sutton, 1990), HamNoSys (Prillwitz and Zienert, 1990), and glosses (Johnston, 2008), offer the possibility to incorporate sign language content into a text-based NLP pipeline (Jiang et al., 2023). However, a good segmentation (Moryossef et al., 2023) and transcription model to process raw video input is first required, which is not well-researched and is not part of this paper.

Recently, vector quantization (VQ) approaches (Van Den Oord et al., 2017) such as SignVQNet (Hwang et al., 2023) demonstrate the ability to convert the continuous signal of videos/poses to discrete tokens similar to spoken language sub-words (Sennrich et al., 2016), which might be a promising direction to pursue in future work.

Comparison In this work, we only empirically experiment with the video encoder and pose-based methods since there are not yet mature and open solutions for the others at the time of writing. Given a hypothetical 10-second, 30 FPS, 480p (640×480), RGB (3 channels) video of 12 consecutive signs, we compare the dimensionalities of the most common representations in Table 1. These approaches compress the raw videos to a sequence of video tokens compatible with a Transformer (Vaswani et al., 2017) or a pretrained language model for further training and processing (Gong et al., 2024).

Representation	Temporal	Spatial	Interpretable
Original video	10x30	640×480×3	–
S3D (HowTo100M)	10	512	no
I3D (BSL-1K)	10	1024	no
MediaPipe Holistic	10×30	543×3	yes
VQ (e.g., SignVQNet)	10	1024*	no
SignWriting/HamNoSys/gloss	12	1024*	yes

Table 1: Temporal and spatial dimensions of different sign language representations for a 10-second, 30 FPS, 480p (640×480), RGB (3 channels) video consisting of 12 signs. In the parentheses are the datasets on which the models are pretrained. For discrete representations, we assume the embedding size to be 1024, marked with an asterisk (*), but this can be chosen arbitrarily.

3 Model Architecture

We follow the setups in VideoCLIP and reuse their codebase[4]. The most essential model architecture with minor modifications to adapt our experiments is described here. We take pairs of video and text samples (v, t) as inputs, where for each video clip, $\boldsymbol{c_v}$ is a sequence of continuous video frames and is processed by a video encoder $f_{\theta_{ve}}$. This is then followed by a trainable MLP projection layer, $f_{\theta_{vp}}$, to project the embedding to the same dimension, $d = 768$, as the word embedding on the text side:

$$\boldsymbol{x_v} = f_{\theta_{vp}}(stopgrad(f_{\theta_{ve}}(\boldsymbol{c_v})) \tag{1}$$

The frozen video encoder $f_{\theta_{ve}}$ is by default a 3D-CNN network but can be replaced by any other visual backbone or a black-box pose estimation system as summarized in Table 1. Likewise, text token vectors $\boldsymbol{x_t}$ are acquired through embedding lookup from a frozen BERT model (Devlin et al., 2019). Then $\boldsymbol{x_v}$ and $\boldsymbol{x_t}$ are fed into two separate trainable Transformers, f_{θ_v} and f_{θ_t}, followed by average pooling over the sequence of the tokens to obtain the temporally aggregated embeddings:

$$z_v = Avg(f_{\theta_v(\boldsymbol{x_v})}), z_t = Avg(f_{\theta_t(\boldsymbol{x_t})}) \tag{2}$$

We optionally add two linear multimodal projection layers on top of $\boldsymbol{z_v}$ and $\boldsymbol{z_t}$, which are missing in VideoCLIP but present in CLIP (see Figure 3 of the CLIP paper). Finally, we employ the InfoNCE loss (Oord et al., 2018) as the contrastive objective to discern the relationship between the embedded N video-text pairs in each mini-batch and run contrastive training over the whole dataset.

[4] https://github.com/facebookresearch/fairseq/tree/main/examples/MMPT

4 FingerCLIP

As a proof of concept, we first apply this contrastive training approach to Fingerspelling (Battison, 1978; Wilcox, 1992; Brentari and Padden, 2001), a subsystem of sign languages heavily influenced by the surrounding spoken languages. For concepts that do not (yet) have associated signs (names of people, locations, organizations, etc.), sign language users borrow a word of a spoken language by spelling it letter-by-letter with predefined signs for that language's alphabet. Isolated fingerspelling recognition[5] can therefore be considered a toy task similar to the MNIST (Deng, 2012) handwritten digits classification task in CV. We name the model FingerCLIP, a mini-version of SignCLIP (§5).

Dataset We start with the RWTH German Fingerspelling Database (Dreuw et al., 2006), containing ∼1400 videos of 35 DGS fingerspelling and number signs, five of which contain inherent motion. We provide details and an illustration of the dataset in Appendix B. We split all examples randomly into training/validation/test sets at the ratio of 8:1:1.

Training Details We adhere to most implementation details outlined in VideoCLIP unless otherwise specified. The two trainable Transformers, f_{θ_v} and f_{θ_t}, are initialized with the pretrained *bert-base-uncased* weights. We train 25 epochs within two hours on a Tesla V100-SXM2-32GB GPU, validated by loss on the validation set. For contrastive training, we construct each batch as a collection of 35 different signs with corresponding text prompts "Fingerspell the letter <letter_name> in DGS.". By optimizing the InfoNCE loss, we move the embedding of a sign closer to the paired text, and further away from the remaining 34 negative examples.

We test different combinations of video encoders: (a) *S3D HowTo100M* video features[6]; (b) *I3D BSL-1K (M+D+A)* video features[7]; (c) *MediaPipe Holistic* pose estimation, and training strategies: (a) zero-shot VideoCLIP (no training); (b) fine-tuning VideoCLIP; (c) training from scratch.

MediaPipe Holistic runs offline on a low-end CPU device. Pose estimation is normalized to a consistent scale by setting the mean width of each person's shoulders to 1, and the mid-point to $(0, 0)$. The leg values are removed since they are irrelevant to signing. We further augment the data by randomly rotating, shearing, and scaling the poses.

Evaluation We view fingerspelling understanding as a text-to-video/video-to-text retrieval task. The candidates are ranked for both directions by a dot-product-based similarity score to each text/video query in the latent space. For the test text prompt of each sign, there is possibly more than one correct video (e.g., the same letter signed by different signers) in the test video pool, and they are all considered successful retrieval. We thus evaluate the text-to-video retrieval task by *precision@k*, i.e., what percent of the top k candidates are correct answers. Each test video query has only one correct text prompt out of the 35 possible prompts. We thus evaluate the video-text retrieval task by *recall@k*, i.e., the chance to include the only correct answer by taking the top k candidates. *precision@1* and *recall@1* can be interpreted as the retrieval accuracy. We also add the metric median retrieval rank, i.e., the median value of the rank of the first correct answer in the candidate lists. We present the experimental results in Table 2.

Discussion FingerCLIP distinguishes itself from the supervised baseline method (Dreuw et al., 2006) in that it is not directly optimized for classification. Instead, a contrastive objective ties positive pairs of text and videos by learning meaningful embeddings, which are then used for similarity-based retrieval. We find video-to-text retrieval reasonably more challenging than text-to-video retrieval[8] since text-to-video retrieval is also often of less value, as a trivial dictionary look-up does the job perfectly.

E1, zero-shot VideoCLIP, presenting random guess results, shows that a common video understanding network pretrained on HowTo100M does not necessarily address the nuance of sign language, even simply as fingerspelling. Therefore, dedicated training on sign language data is essential. Comparing *E1.1* and *E1.2*, neither is fine-tuning an existing VideoCLIP checkpoint helpful, so in the rest of the paper, models are trained from scratch.

In *E2*, *I3D BSL-1K* sign-language-specific features outperform *HowTo100M S3D* video features (*E1.2*), especially when downsampled to the same dimension as S3D (*E2.1*). *MediaPipe Holistic* pose estimation as a feature extractor (*E3*) works better

[5]Continuous fingerspelling recognition is more complex and is often solved by a CTC loss like speech recognition.

[6]https://github.com/antoine77340/S3D_HowTo100M

[7]https://www.robots.ox.ac.uk/~vgg/research/bslattend/

[8]Note that the relatively low *precision@5/10* values are uncomparable to the high *recall@5/10* values, since the former makes the task harder, while the latter simplifies the task.

Experiment		Text-to-video				Video-to-text			
		P@1↑	P@5↑	P@10↑	MedianR↓	R@1↑	R@5↑	R@10↑	MedianR↓
E0	Dreuw et al. (2006) (supervised HMM, appearance-based features)	–	–	–	–	0.64	–	–	–
Explore training strategy									
E1	VideoCLIP zero-shot (+ *S3D HowTo100M* video features)	0.03	0.02	0.03	22	0.02	0.14	0.28	18
E1.1	VideoCLIP fine-tuned (+ *S3D HowTo100M* video features)	0.40	0.36	0.30	2	0.31	0.75	0.89	2
E1.2	VideoCLIP trained from scratch (+ *S3D HowTo100M* video features)	0.54	0.35	0.28	**1**	0.28	0.69	0.87	3
Explore video-based (I3D) features									
E2	FingerCLIP trained from scratch (+ *I3D BSL-1K* video features)	0.63	0.47	0.37	**1**	0.37	0.78	0.91	2
E2.1	**E2** + feature dimension average pooled from 1024 to 512	0.74	0.56	**0.44**	**1**	0.47	0.82	0.94	2
Explore pose-based features									
E3	FingerCLIP trained from scratch (*MediaPipe Holistic* pose features)	0.89	0.67	0.42	**1**	0.68	0.97	**1.00**	**1**
E3.1	**E3** + dominant hand features only (26 times less keypoints)	**1.00**	0.72	0.42	**1**	0.82	0.99	**1.00**	**1**
E3.2	**E3.1** + 2D augmentation on pose features ($\sigma = 0.2$)	0.91	**0.74**	0.43	**1**	**0.93**	**1.00**	**1.00**	**1**

Table 2: FingerCLIP experimental results evaluated on the test set. *P@k* denotes *precision@k*, *R@k* denotes *recall@k*, and *MedianR* denotes the median retrieval rank. The best score of each column is in bold. *E0* is taken from Dreuw et al. (2006) as a baseline (*R@1* derived from the best error rate 35.7%).

than 3D-CNN-based video encoders, presumably because it is more universal than an I3D model pretrained on a particular dataset and sign language. For fingerspelling understanding, focusing on the dominant hand (*E3.1*) is beneficial[9], which drastically reduces the number of keypoints from 543 to 21. 2D data augmentation further improves the overall performance. Since *MediaPipe Holistic* as features perform the best and are more interpretable and operable for potential data normalization and augmentation, we decide to use it as the frozen video encoder $f_{\theta_{ve}}$ for the rest of the paper[10].

5 SignCLIP Pretraining

To fully realize the power of contrastive learning, we train and evaluate SignCLIP on larger and more diverse sign language datasets. For efficient experimenting, we start exploring datasets consisting of relatively short-duration sign language video examples, e.g., for the task of isolated sign language recognition (ISLR) instead of machine translation. In Table 3, we summarize recent large-scale sign language datasets in this context, focusing on ASL, one of the highest-resourced languages in SLP.

5.1 Spreadthesign Pretraining Dataset

Spreadthesign is used as the pretraining dataset for its large-scale and multilingual nature[11]. In this work, we limit the text translations to English only to avoid a cartesian product number of data points, for the pretraining to be economical and sign-language-focused. After filtering English text, our dataset consists of 18,423 concepts in 41 sign languages, resulting in 456,913 video-text pairs with a total duration of ~500 hours. The data distribution is presented in Appendix C in detail. We split all examples randomly into training, validation, and test sets at the ratio of 98:1:1. A caveat of this dataset is that there is normally only one signing example per text concept per sign language, which means the pretraining is prone to overfitting the exact signs by particular signers. We still believe that the diverse text-signing pairs will guide the model to learn a useful visual representation.

We add the Spreadthesign data scale and that of CLIP and VideoCLIP to Table 3 for comparison. CLIP was trained on 400 million text-image pairs collected from the Internet (500K queries and up to 20K pairs per query); VideoCLIP was pretrained on 1.2 million videos from HowTo100M (each lasts ~6.5 minutes with ~110 clip-text pairs). We additionally add ImageNet (Deng et al., 2009), which has a relatively close scale to Spreadthesign.

5.2 Training and Evaluation Details

Most implementation details and evaluation protocols in FingerCLIP (§4) are reused for SignCLIP. The text prompts now consist of the text content prepended with a spoken and a sign language tag, inspired by multilingual machine translation in Johnson et al. (2017). For example, the text prompt for signing the phrase "Hello, can I help you?" in ASL is "<en> <ase> Hello, can I help you?", tagged by the ISO 639-3 language code. Fitting most examples, we limit the context length of the

[9] Note that the DGS finger alphabet is one-handed.

[10] An S3D/I3D model fine-tuned end-to-end may overcome some limitations of pose estimation and yield superior performance for specific tasks. In this work, we trade pursuing state-of-the-art numbers for the universality, interpretability, and cheap computation of pose estimation.

[11] https://www.spreadthesign.com/en.us/about/statistics/. The data used in this work was crawled in 2023 and might differ slightly from the official statistics.

Dataset	Language	Type/Task	#examples	#signs/#classes	#signers
RWTH German Fingerspelling (Dreuw et al., 2006)	DGS	Isolated Fingerspelling	1,400	35	20
ChicagoFSWild (Shi et al., 2018)	ASL	Continuous fingerspelling	7,304	–	160
ChicagoFSWild+ (Shi et al., 2019)	ASL	Continuous fingerspelling	55,232	–	260
Google – American Sign Language Fingerspelling Recognition	ASL	Continuous fingerspelling	67,208	–	100
Google – Isolated Sign Language Recognition (i.e., *asl-signs*) ✓	ASL	ISLR	94,477	250	21
WLASL (Li et al., 2020)	ASL	ISLR	21,083	2,000	100
ASL Citizen (Desai et al., 2023) ✓	ASL	ISLR	83,399	2,731	52
Sem-Lex (Kezar et al., 2023) ✓	ASL	ISLR	91,148	3,149	41
How2Sign (Duarte et al., 2021)	ASL	Continuous signing	35,000	16,000	11
ASL-LEX (Sehyr et al., 2021)	ASL	Dictionary (phonological)	2,723	2,723	(unknown)
Spreadthesign (SignCLIP, our filtered version)	Multilingual	Dictionary	456,913	18,423*	(unknown)
ImageNet (Deng et al., 2009)	–	Image classification	1,431,167	1,000	–
HowTo100M (Miech et al., 2019) (VideoCLIP)	–	Video understanding	136,000,000	–	–
CLIP (Radford et al., 2021)	–	Contrastive learning	400,000,000	–	–

Table 3: Summarization of datasets consisting of relatively short-duration video examples, compared with Spreadthesign and common CV datasets. SignCLIP has been tested with the datasets marked with a checkmark. *asl-signs* is part of PopSign ASL (Starner et al., 2024), full data not released yet. *#signs/#classes* for Spreadthesign is marked with an asterisk (*) since the signs of a concept across different sign languages are barely classified as one sign.

video Transformer f_{θ_v} to 256 tokens, equivalent to a 10-second, 25 FPS video clip; and that of the text Transformer f_{θ_t} to be 64. Both Transformers are initialized with the pretrained *bert-base-cased* weights and trained on an NVIDIA A100-SXM4-80GB GPU to maximally afford a batch size of 448[12]. We also measure the training efficiency by the number of parameters and the training time.

For evaluation, we include the same video-to-text metrics used in FingerCLIP and omit text-to-video for simplicity, as the latter correlates with and is more trivial than the former. We additionally test the models with three ASL ISLR datasets in a zero-shot way to evaluate out-of-domain generalization. We perform the following text preprocessing to mitigate the effect of shifted text distribution when building text prompts from the raw gloss labels: (1) lowercasing the glosses; (2) removing the gloss index for different variants of a sign; (3) filtering their test sets by known text labels in Spreadthesign, which reduces the total number of signs/classes.

Starting from a baseline setup that resembles *E3* in FingerCLIP, we increase the video Transformer layers from 6 to 12 and add linear multimodal projection layers after temporal pooling. For pose data, we always simplify the face by using the contour keypoints only, resulting in 203 keypoints. We further experiment with the following modifications:

Pose Data Preprocessing Before the regular normalization, $D_{shoulders} = 1, mid\text{–}point = (0,0)$, as performed in all experiments, we further try (1) removing redundant keypoints and repositioning the wrist to the hand model's prediction (*E6*); (2) standardizing the keypoints by subtracting the mean pose values of all examples from Spreadthesign and dividing by the standard deviation (*E6.1*); (3) anonymizing by removing the appearance from the first frame then adding the mean pose (*E6.2*).

Pose Data Augmentation After data normalization, we also employ data augmentation (Boháček and Hrúz, 2022) at training time to improve the models' robustness, including (1) randomly flipping the poses horizontally; (2) 2D spatial augmentation as done in FingerCLIP; (3) temporal augmentation of the signing speed by linear interpolation between frames; (4) Gaussian noise on keypoints.

5.3 Experimental Results and Discussion

We present the experimental results in Table 4. In this scenario, an accurate retrieval is harder than FingerCLIP because there are 3,939 unique text prompts in the test set of 4,531 examples. The in-domain results, *E6.1* at the top (attained with increased video layers and keypoint standardization), are impressive given the challenging nature. We attribute it to (1) the hypothesized multilingual transfer effect thanks to sign language iconicity; and (2) the broader supervision signal of contrastive learning than fixed labels, coming from example phrases consisting of individual signs (Figure 1).

On the other hand, zero-shot performance on out-of-domain data is deficient. We posit that to reach noticeable performance on out-of-domain data, few-shot learning or fine-tuning (§6.1) is essential given the current scale of pretraining. Nev-

[12] For reference, CLIP was trained with batch size 32,768 and VideoCLIP was trained with batch size 512.

Experiment		Video-to-text (In-domain) R@1↑	R@5↑	R@10↑	MedianR↓	Video-to-text (out-of-domain) AS MedianR↓	AC MedianR↓	SL MedianR↓	Efficiency #Params↓	Time↓
E4	Baseline	0.33	0.64	0.77	3/3939	103/213	253/1625	455/1967	**175M**	29h
Initial architectural changes										
E5	**E4** + six more video layers	0.37	0.68	0.80	**2**	104	192	382	217M	28h
E5.1	**E5** + multimodal projection layer	0.38	0.69	0.80	**2**	104	216	418	218M	15h
Pose data preprocessing										
E6	**E5** + keypoint reduction & reposition	0.37	0.68	0.80	**2**	105	230	665	217M	32h
E6.1	**E6** + keypoint standardization	**0.40**	**0.71**	**0.83**	**2**	**99**	273	551	217M	**14h**
E6.2	**E6.1** + pose anonymization	0.37	0.68	0.79	**2**	101	251	577	217M	40h
Pose data augmentation										
E7	**E5** + pose random flipping ($p = 0.2$)	0.36	0.67	0.79	**2**	105	200	435	217M	29h
E7.1	**E5** + spatial 2D augmentation ($\sigma = 0.2$)	0.35	0.65	0.78	3	102	219	377	217M	39h
E7.2	**E5** + temporal augmentation ($\sigma = 0.2$)	0.39	0.69	0.80	**2**	104	187	372	217M	62h
E7.3	**E5** + Gaussian noise ($\sigma = 0.001$)	0.37	0.68	0.80	**2**	104	198	364	217M	29h
Test-time-only normalization										
E8*	**E7.2** + flipping to right-handed	0.39	0.69	0.80	**2**	103	**187**	**359**	–	–
E8.1*	**E8** + pose anonymization (zero-shot-only)	–	–	–	–	101	214	380	–	–

Table 4: SignCLIP experimental results evaluated on the test set. *R@k* denotes *recall@k*, and *MedianR* denotes the median retrieval rank as well as the total number of unique signs/classes. *AS* = *asl-signs*, *AC* = *ASL Citizen*, *SL* = *Sem-Lex*. Experiments marked with an asterisk (*) are test-time only. The best score of each column is in bold.

ertheless, starting from *E7*, we experiment with a few data augmentation techniques to increase data variation since the previous standardization benefits in-domain results but hurts out-of-domain results. We then attempt test-time-only normalization to shift the test distribution of the poses closer to training. As a result, we manage to gradually fight against overfitting and improve the overall zero-shot performance by temporal augmentation (*E7.2*) and test-time pose flipping if the right hand is not present (*E8*). The former provides robustness to signing speed change, and the latter is helpful for unseen test examples signed by left-handed signers.

6 Downstream Tasks and Analysis

In this section, we evaluate SignCLIP on several downstream tasks and datasets, and we discuss the ideas for further enhancement and evaluation in §8.

6.1 Isolated Sign Language Recognition

We provide a comprehensive evaluation for the three ASL ISLR datasets in Table 5. For zero-shot prediction, we follow the optimal setups in *E8*/*E8.1* but skip any text preprocessing on raw glosses for a fairer comparison. For few-shot learning, we randomly sample 10 example videos for each class from training data and use a nonparametric k-nearest neighbors (KNN) algorithm to infer test labels based on pose embedding similarity. We additionally train a supervised logistic regression classifier by *scikit-learn* on top of the embedded poses, also known as linear probe (Alain and Bengio, 2016). We also train SignCLIP models from scratch or fine-tune them from the *E7.2* checkpoint with the three datasets for 25 epochs with batch size 256. Outliers longer than 256 frames are removed.

Data	Model	R@1↑	R@5↑	R@10↑	MedianR↓
AS	SignCLIP (E8.1) zero-shot	0.01	0.03	0.05	118/250
	SignCLIP (E8.1) 10-shot	0.01	0.03	0.06	109
	SignCLIP (E8.1) + linear	0.12	0.30	0.41	17
	SignCLIP train from scratch	0.74	0.91	**0.94**	**1**
	SignCLIP fine-tuned	0.74	0.91	**0.94**	**1**
	SignCLIP fine-tuned + linear	**0.78**	**0.92**	**0.94**	**1**
	SOTA PopSign ASL	0.84*	–	–	–
	SOTA Kaggle competition	0.89*	–	–	–
AC	SignCLIP (E8) zero-shot	0.00	0.00	0.00	1296/2731
	SignCLIP (E8) 10-shot	0.05	0.16	0.23	56
	SignCLIP (E8) + linear	0.23	0.47	0.57	7
	SignCLIP train from scratch	0.39	0.71	0.80	2
	SignCLIP fine-tuned	0.46	0.77	0.84	2
	SignCLIP fine-tuned + linear	**0.60**	**0.84**	**0.89**	**1**
	SOTA ASL Citizen	0.63	0.86	0.91	–
SL	SignCLIP (E8) zero-shot	0.01	0.02	0.03	853/3837
	SignCLIP (E8) 10-shot	0.02	0.06	0.10	235
	SignCLIP (E8) + linear	0.15	0.29	0.36	38
	SignCLIP train from scratch	0.14	0.30	0.38	32
	SignCLIP fine-tuned	0.16	0.34	0.42	22
	SignCLIP fine-tuned + linear	**0.30**	**0.48**	**0.55**	**6**
	SOTA Sem-Lex	0.67*	–	–	–
	SOTA + auxiliary phonology	0.69*	–	–	–

Table 5: Comprehensive evaluations of SignCLIP on the test set of three ASL ISLR datasets. *AS* = *asl-signs*, *AC* = *ASL Citizen*, *SL* = *Sem-Lex*. The best score SignCLIP achieves on each dataset is in bold. SOTA numbers marked with an asterisk (*) are not directly comparable to ours. *SOTA PopSign ASL* and *SOTA Kaggle* are trained with *asl-signs* but tested on a private test set; *SOTA Sem-Lex* is tested with a reduced test set of 2,731 classes that are aligned with *ASL Citizen*/*ASL-LEX*.

In general, proper zero-shot prediction is hindered by distribution shift on both modalities: (1) for *asl-signs* dataset, zero-shot prediction is flawed because we only have pose data normalized in an unknown way from Kaggle instead of the raw videos[13]; (2) *ASL Citizen* uses all upper case glosses while *Sem-Lex* uses snake case glosses, both unseen during training. For *asl-signs*, we find pose anonymization eases the domain shift issue in zero-shot (*E8.1*) and other settings. Besides, pose-based few-shot KNN greatly improves the deficient zero-shot results, bypassing the influence of out-of-domain glosses. Finally, fine-tuning SignCLIP with a pretrained checkpoint, compared to training from scratch, yields closer results to the state-of-the-art (SOTA) reported in the three papers.

6.2 Sign Language Identification

Sign language identification (Gebre et al., 2013) can be achieved by simply ranking text prompts of different sign languages without actual content, e.g., "<en> <ase>" for ASL. We evaluate the best checkpoint *E6.1* for in-domain test data and obtain 0.99 *recall@1*, which solves the task perfectly without any direct supervision. However, the identification task is ill-defined in that the model can learn to identify signers for particular sign languages.

6.3 Latent Space Exploration

We make SignCLIP accessible via a RESTful API and perform analysis in a Colab notebook[14].

Distributional Hypothesis for Sign Language We revisit the distributional hypothesis (Lenci and Sahlgren, 2023) in a sign language context based on the pose/sign embeddings instead of text/word (Mikolov et al., 2013). As illustrated by Figure 2, unlike a word with a discrete, unique text token, each sign has multiple realizations scattered in a continuous space. By aligning pose representation to text with contrastive learning, we have realized distributional semantics in sign language, reflected by the cluster center of each sign, while maintaining individual variance.

What Is the Most Iconic Sign Crosslingually? Iconicity is one of the key motivations for training a multilingual SignCLIP (Figure 1), and now we ask this linguistic question back to the model. We rank a sampled subset of 302 signs from *Spreadthesign* based on the variance of the pose embeddings across 20+ different sign languages. As a result, the sign for "scorpion" with a universal hook hand shape ranks at the top, and the motivational example "house" sign also ranks high (51/302). On the opposite, the signs for numbers tend to rank low due to the diverse signing styles across languages. We append the full rank in Appendix D.

[13]The authors of PopSign ASL promise to release the full data, then we can do a fairer and more meaningful evaluation.

[14]The link is hidden for anonymous review.

Figure 2: *King – Man + Woman = Queen* analogy revisited. 14 video examples of each sign are randomly sampled from the ASL Citizen dataset, embedded by a fine-tuned SignCLIP pose encoder, and then visualized by *t-SNE (perplexity=15)* with different shapes and colors. Cluster centers are represented with a big symbol.

7 Conclusion: Where Are We for SLP?

This work involves SLP, an interdisciplinary field that suffers from low-resourceness compared to mainstream NLP and CV. To overcome the non-generalizability of a specific dataset/task/language, we adapt (Video-)CLIP and propose SignCLIP. SignCLIP is trained on *Spreadthesign*, a multilingual, generic sign language dictionary consisting of ∼500 hours of signing in 41 sign languages, and is evaluated extensively for various purposes.

SignCLIP demonstrates excellent in-domain performance but falls short of immediate zero-shot prediction on downstream ISLR tasks. This finding is consistent with previous CV studies (Li et al., 2017) before CLIP reached its scale. Similar models trained on smaller datasets close to an ImageNet scale performed much worse than supervised baselines on common benchmarks. However, a supervised approach usually requires meticulous, task-specific data collection. This is more demanding for SLP, a niche domain lacking human experts.

As a middle ground between full zero-shot prediction and full supervision, few-shot learning or fine-tuning is essential to tackle domain shift and is more realistic given the present methodology and data scale. The cross-lingual transfer effect is prospective for sign language due to iconicity and smaller grammatical discrepancies. With no surprise, under a sensible data condition, the universal pretraining paradigm that is transforming NLP and CV is also a promising research direction for SLP.

8 Limitations

The main limitation of this work is the data scale, which interestingly is the primary breakthrough of the original CLIP work. Apart from the inherent data scarcity issue in SLP research, the concrete limitations of SignCLIP come in a few aspects.

First, to be not dataset/task/language-specific while possibly large-scale, we choose *Spreadthesign* as our pretraining dataset (§5.1), which is indeed highly multilingual and untied to any SLP task. However, we are unable to release the dataset publicly, and future researchers who are interested in it have to first resolve or purchase the license from *Spreadthesign*. Fortunately, it is possible to augment or replace *Spreadthesign* with recently released public datasets of comparable or larger size, which should increase data density and variation sign-wise, but with potentially decreased diversity and balance among different sign languages.

Secondly, training large models on video data is very costly. The principle of this work is to finish every training process in less than three days on a single Nvidia A100 GPU, and we managed this goal (Table 4) by using pose-based models, a moderate context length of 256 frames, and limiting spoken language to English only. To scale up, multiple GPU training can be exploited, and architectural modifications that reduce the sequence length must be employed, for which we refer to several techniques discussed in §2.

As a result of the above-mentioned optimizations, we will be able to remove the limitation of the relatively short context, and then train and evaluate for longer-range tasks such as machine translation and sign language production. This will make SignCLIP more versatile and generalizable, as well as support other types of deep pretrained networks for SLP, such as large language models.

Acknowledgments

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Nikolas Adaloglou, Theocharis Chatzis, Ilias Papastratis, Andreas Stergioulas, Georgios Th Papadopoulos, Vassia Zacharopoulou, George J Xydopoulos, Klimnis Atzakas, Dimitris Papazachariou, and Petros Daras. 2021. A comprehensive study on deep learning-based methods for sign language recognition. *IEEE Transactions on Multimedia*, 24:1750–1762.

Guillaume Alain and Yoshua Bengio. 2016. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*.

Samuel Albanie, Gül Varol, Liliane Momeni, Triantafyllos Afouras, Joon Son Chung, Neil Fox, and Andrew Zisserman. 2020. BSL-1K: Scaling up co-articulated sign language recognition using mouthing cues. In *European Conference on Computer Vision (ECCV)*.

Samuel Albanie, Gül Varol, Liliane Momeni, Hannah Bull, Triantafyllos Afouras, Himel Chowdhury, Neil Fox, Bencie Woll, Rob Cooper, Andrew McParland, and Andrew Zisserman. 2021. BOBSL: BBC-Oxford British Sign Language Dataset.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Robbin Battison. 1978. *Lexical borrowing in American sign language.* ERIC, Linstok Press, Inc., Silver Spring, Maryland 20901.

Matyáš Boháček and Marek Hrúz. 2022. Sign pose-based transformer for word-level sign language recognition. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 182–191.

Danielle Bragg, Oscar Koller, Mary Bellard, Larwan Berke, Patrick Boudreault, Annelies Braffort, Naomi Caselli, Matt Huenerfauth, Hernisa Kacorri, Tessa Verhoef, et al. 2019. Sign language recognition, generation, and translation: An interdisciplinary perspective. In *The 21st International ACM SIGACCESS Conference on Computers and Accessibility*, pages 16–31.

Diane Brentari and Carol Padden. 2001. A language with multiple origins: Native and foreign vocabulary in american sign language. *Foreign vocabulary in sign language: A cross-linguistic investigation of word formation*, pages 87–119.

Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. 2018. Neural sign language translation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. 2019. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Joao Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6299–6308.

Yiting Cheng, Fangyun Wei, Jianmin Bao, Dong Chen, and Wenqiang Zhang. 2023. Cico: Domain-aware sign language retrieval via cross-lingual contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19016–19026.

Mathieu De Coster, Dimitar Shterionov, Mieke Van Herreweghe, and Joni Dambre. 2023. Machine translation from signed to spoken languages: State of the art and challenges. *Universal Access in the Information Society*, pages 1–27.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

Li Deng. 2012. The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE signal processing magazine*, 29(6):141–142.

Aashaka Desai, Lauren Berger, Fyodor O Minakov, Vanessa Milan, Chinmay Singh, Kriston Pumphrey, Richard E Ladner, Hal Daumé III, Alex X Lu, Naomi Caselli, and Danielle Bragg. 2023. Asl citizen: A community-sourced dataset for advancing isolated sign language recognition. *arXiv preprint arXiv:2304.05934*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Philippe Dreuw, Thomas Deselaers, Daniel Keysers, and Hermann Ney. 2006. Modeling image variability in appearance-based gesture recognition. In *ECCV workshop on statistical methods in multi-image and video processing*, pages 7–18.

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)*.

Hao-Shu Fang, Jiefeng Li, Hongyang Tang, Chao Xu, Haoyi Zhu, Yuliang Xiu, Yong-Lu Li, and Cewu Lu. 2022. Alphapose: Whole-body regional multi-person pose estimation and tracking in real-time. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Jens Forster, Christoph Schmidt, Oscar Koller, Martin Bellgardt, and Hermann Ney. 2014. Extensions of the sign language recognition and translation corpus RWTH-PHOENIX-weather. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1911–1916, Reykjavik, Iceland. European Language Resources Association (ELRA).

Binyam Gebrekidan Gebre, Peter Wittenburg, and Tom Heskes. 2013. Automatic sign language identification. In *2013 IEEE International Conference on Image Processing*, pages 2626–2630. IEEE.

Jia Gong, Lin Geng Foo, Yixuan He, Hossein Rahmani, and Jun Liu. 2024. Llms are good sign language translators. *arXiv preprint arXiv:2404.00925*.

Ivan Grishchenko and Valentin Bazarevsky. 2020. Mediapipe holistic.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 16000–16009.

Ruth M Holmes, Ellen Rushe, and Anthony Ventresque. 2024. The key points: Using feature importance to identify shortcomings in sign language recognition models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15970–15975.

Eui Jun Hwang, Huije Lee, and Jong C Park. 2023. Autoregressive sign language production: A gloss-free approach with discrete representations. *arXiv preprint arXiv:2309.12179*.

Zifan Jiang, Amit Moryossef, Mathias Müller, and Sarah Ebling. 2023. Machine translation between spoken languages and signed languages represented in SignWriting. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1706–1724, Dubrovnik, Croatia. Association for Computational Linguistics.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Trevor Johnston and Adam Schembri. 2007. *Australian Sign Language (Auslan): An Introduction to Sign Language Linguistics*. Cambridge University Press, Cambridge, UK.

Trevor Alexander Johnston. 2008. From archive to corpus: transcription and annotation in the creation of signed language corpora. In *Pacific Asia Conference on Language, Information and Computation*.

Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. 2023. Cotracker: It is better to track together. *arXiv:2307.07635*.

Lee Kezar, Jesse Thomason, Naomi Caselli, Zed Sehyr, and Elana Pontecorvo. 2023. The sem-lex benchmark: Modeling asl signs and their phonemes. In *Proceedings of the 25th International ACM SIGACCESS Conference on Computers and Accessibility*, pages 1–10.

Alessandro Lenci and Magnus Sahlgren. 2023. *Distributional semantics*. Cambridge University Press.

Ang Li, Allan Jabri, Armand Joulin, and Laurens Van Der Maaten. 2017. Learning visual n-grams from web data. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4183–4192.

Dongxu Li, Cristian Rodriguez, Xin Yu, and Hongdong Li. 2020. Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. In *The IEEE Winter Conference on Applications of Computer Vision*, pages 1459–1469.

Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. 2022. Video swin transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 3202–3211.

Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, et al. 2019. Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*.

Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. 2019. HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips. In *International Conference on Computer Vision (ICCV)*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.

Amit Moryossef, Zifan Jiang, Mathias Müller, Sarah Ebling, and Yoav Goldberg. 2023. Linguistically motivated sign language segmentation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12703–12724, Singapore. Association for Computational Linguistics.

Amit Moryossef, Ioannis Tsochantaridis, Joe Dinn, Necati Cihan Camgoz, Richard Bowden, Tao Jiang, Annette Rios, Mathias Muller, and Sarah Ebling. 2021. Evaluating the immediate applicability of pose estimation for sign language recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 3434–3440.

Mathias Müller, Malihe Alikhani, Eleftherios Avramidis, Richard Bowden, Annelies Braffort, Necati Cihan Camgöz, Sarah Ebling, Cristina España-Bonet, Anne Göhring, Roman Grundkiewicz, Mert Inan, Zifan Jiang, Oscar Koller, Amit Moryossef, Annette Rios, Dimitar Shterionov, Sandra Sidler-Miserez, Katja Tissi, and Davy Van Landuyt. 2023. Findings of the second WMT shared task on sign language translation (WMT-SLT23). In *Proceedings of the Eighth Conference on Machine Translation*, pages 68–94, Singapore. Association for Computational Linguistics.

Mathias Müller, Sarah Ebling, Eleftherios Avramidis, Alessia Battisti, Michèle Berger, Richard Bowden, Annelies Braffort, Necati Cihan Camgöz, Cristina España-bonet, Roman Grundkiewicz, Zifan Jiang, Oscar Koller, Amit Moryossef, Regula Perrollaz, Sabine Reinhard, Annette Rios, Dimitar Shterionov, Sandra Sidler-miserez, and Katja Tissi. 2022. Findings of the first WMT shared task on sign language translation (WMT-SLT22). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 744–772, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Aleksandar Petrov, Emanuele La Malfa, Philip Torr, and Adel Bibi. 2024. Language model tokenizers introduce unfairness between languages. *Advances in Neural Information Processing Systems*, 36.

K R Prajwal, Hannah Bull, Liliane Momeni, Samuel Albanie, Gül Varol, and Andrew Zisserman. 2022. Weakly-supervised fingerspelling recognition in british sign language videos. In *British Machine Vision Conference*.

Siegmund Prillwitz and Heiko Zienert. 1990. Hamburg notation system for sign language: Development of a sign writing with computer application. In *Current trends in European Sign Language Research. Proceedings of the 3rd European Congress on Sign Language Research*, pages 355–379.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Razieh Rastgoo, Kourosh Kiani, Sergio Escalera, and Mohammad Sabokrou. 2021. Sign language production: A review. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 3451–3461.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Wendy Sandler and Diane Lillo-Martin. 2006. *Sign language and linguistic universals*. Cambridge University Press.

Zed Sevcikova Sehyr, Naomi Caselli, Ariel M Cohen-Goldberg, and Karen Emmorey. 2021. The asl-lex 2.0 project: A database of lexical and phonological properties for 2,723 signs in american sign language. *The Journal of Deaf Studies and Deaf Education*, 26(2):263–277.

Prem Selvaraj, Gokul Nc, Pratyush Kumar, and Mitesh Khapra. 2022. OpenHands: Making sign language recognition accessible with pose-based pretrained models across languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2114–2133, Dublin, Ireland. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Antonio F. G. Sevilla, José María Lahoz-Bengoechea, and Alberto Diaz. 2024. Automated extraction of prosodic structure from unannotated sign language video. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1808–1816, Torino, Italia. ELRA and ICCL.

Bowen Shi, Aurora Martinez Del Rio, Jonathan Keane, Jonathan Michaux, Diane Brentari, Greg Shakhnarovich, and Karen Livescu. 2018. American sign language fingerspelling recognition in the wild. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 145–152. IEEE.

Bowen Shi, Aurora Martinez Del Rio, Jonathan Keane, Diane Brentari, Greg Shakhnarovich, and Karen Livescu. 2019. Fingerspelling recognition in the wild with iterative visual attention. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5400–5409.

Thad Starner, Sean Forbes, Matthew So, David Martin, Rohit Sridhar, Gururaj Deshpande, Sam Sepah, Sahir Shahryar, Khushi Bhardwaj, Tyler Kwok, et al. 2024. Popsign asl v1. 0: An isolated american sign language dataset collected via smartphones. *Advances in Neural Information Processing Systems*, 36.

Valerie Sutton. 1990. *Lessons in sign writing*. SignWriting.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Dave Uthus, Garrett Tanzer, and Manfred Georg. 2024. Youtube-asl: A large-scale, open-domain american sign language-english parallel corpus. *Advances in Neural Information Processing Systems*, 36.

Aaron Van Den Oord, Oriol Vinyals, et al. 2017. Neural discrete representation learning. *Advances in neural information processing systems*, 30.

Gül Varol, Liliane Momeni, Samuel Albanie, Triantafyllos Afouras, and Andrew Zisserman. 2021. Read and attend: Temporal localisation in sign language videos. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Qianqian Wang, Yen-Yu Chang, Ruojin Cai, Zhengqi Li, Bharath Hariharan, Aleksander Holynski, and Noah Snavely. 2023. Tracking everything everywhere all at once. In *International Conference on Computer Vision*.

Sherman Wilcox. 1992. *The phonetics of fingerspelling*, volume 4. John Benjamins Publishing.

Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. 2021. VideoCLIP: Contrastive pre-training for zero-shot video-text understanding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6787–6800, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 5288–5296.

Kayo Yin, Amit Moryossef, Julie Hochgesang, Yoav Goldberg, and Malihe Alikhani. 2021. Including signed languages in natural language processing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7347–7360, Online. Association for Computational Linguistics.

Youngjae Yu, Jongseok Kim, and Gunhee Kim. 2018. A joint sequence fusion model for video question answering and retrieval. In *Proceedings of the European conference on computer vision (ECCV)*, pages 471–487.

Da Zhang, Xiyang Dai, Xin Wang, and Yuan-Fang Wang. 2018. S3d: Single shot multi-span detector via fully 3d convolutional network. In *Proceedings of the British Machine Vision Conference (BMVC)*.

Ce Zheng, Wenhan Wu, Chen Chen, Taojiannan Yang, Sijie Zhu, Ju Shen, Nasser Kehtarnavaz, and Mubarak Shah. 2023. Deep learning-based human pose estimation: A survey. *ACM Comput. Surv.*

Hao Zhou, Wengang Zhou, Weizhen Qi, Junfu Pu, and Houqiang Li. 2021. Improving sign language translation with monolingual data by sign back-translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1316–1325.

Yi Zhu, Xinyu Li, Chunhui Liu, Mohammadreza Zolfaghari, Yuanjun Xiong, Chongruo Wu, Zhi Zhang, Joseph Tighe, R Manmatha, and Mu Li. 2020. A comprehensive study of deep video action recognition. *arXiv preprint arXiv:2012.06567*.

A An Intuitive Test on ChatGPT's Sign Language Understanding Ability

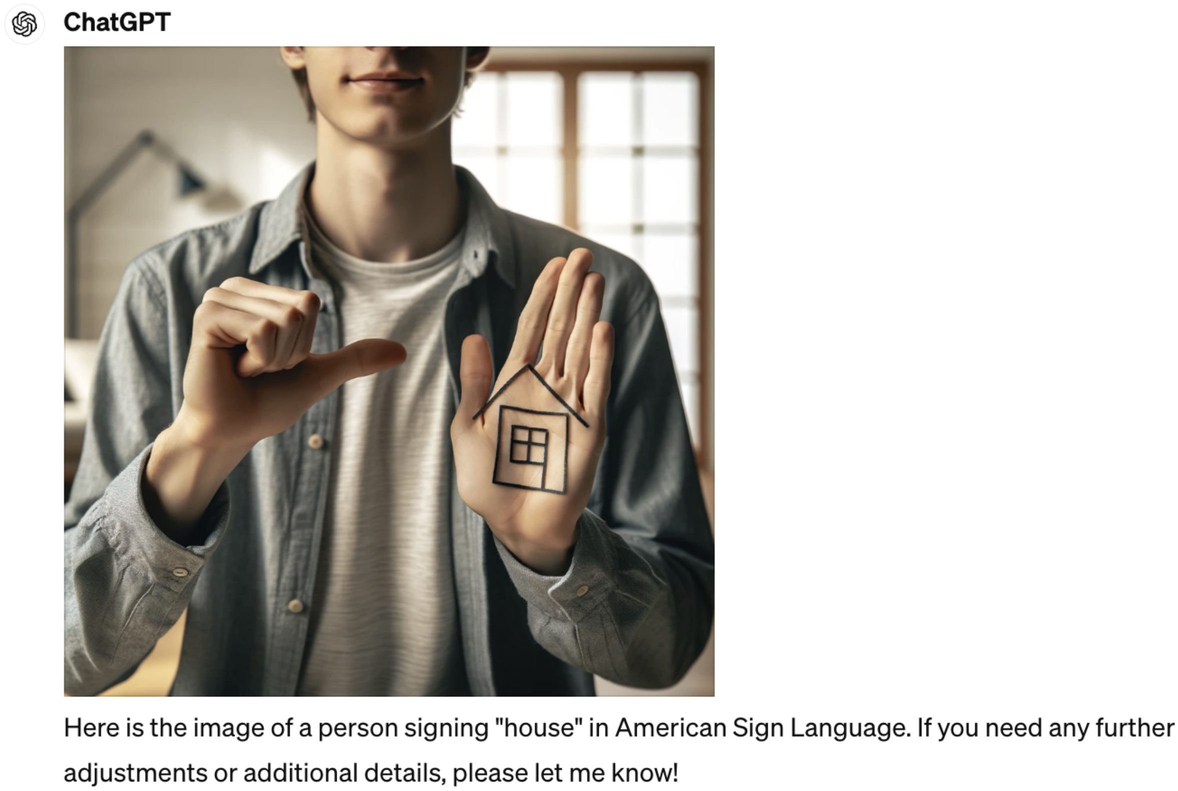

Figure 3: Screenshot of prompting ChatGPT 4o to sign "house" in ASL, which lacks sign language knowledge and tries to sketch a picture of a house on the open palm, tested in June 2024.

B Introduction of the RWTH German Fingerspelling Database

RWTH German Fingerspelling Database contains ∼1400 videos of 20 different signers. Each video was recorded by one webcam named *cam1*, recording the dominant hands only with a resolution of 320x240 at 25 FPS; and one camcorder named *cam2*, recording the whole body with a resolution of 352x288 at 25 FPS. We exclude all *cam1* videos for pose-based models since we assume that the pose estimation system expects whole-body input.

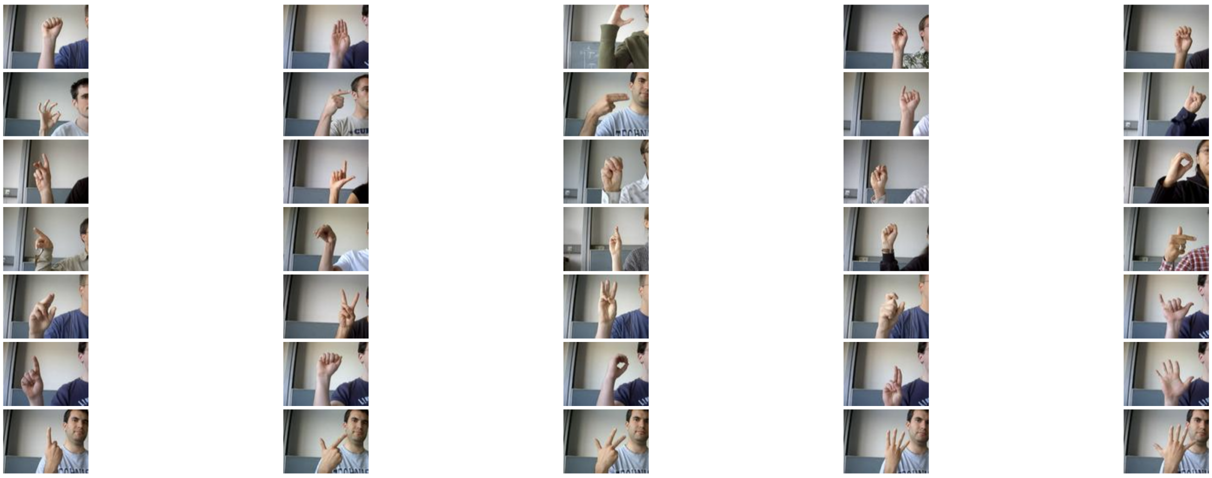

Figure 4: Examples of the German finger-alphabet taken from the RWTH gesture database recorded with the webcam showing the letters A-Z, Ä, Ö, Ü, SCH, and the numbers 1 to 5. Note that J, Z, Ä, Ö, and Ü are dynamic gestures. Figure taken from https://www-i6.informatik.rwth-aachen.de/aslr/fingerspelling.php.

C Extended Spreadthesign Data Analysis

Following the data statistics presented in §5.1, Figure 5 illustrates the distribution of the video examples in 41 sign languages we use from Spreadthesign.

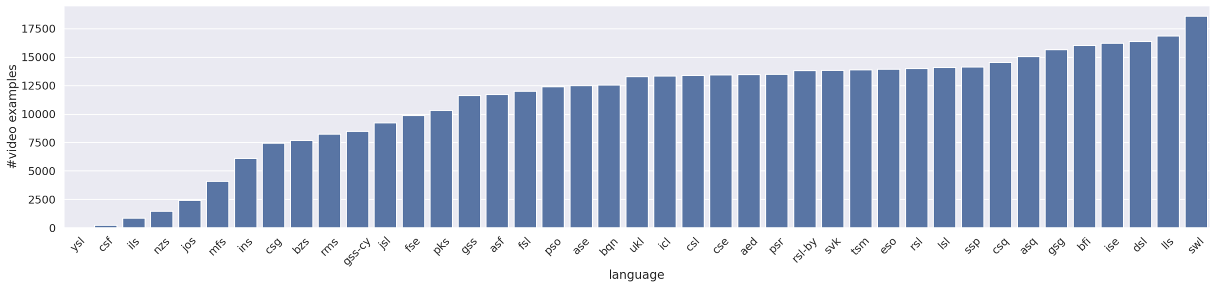

Figure 5: Sign language distribution of video examples in Spreadthesign, using the ISO 639-3 language codes.

Figure 6 illustrates the distribution of pose/video length in Spreadthesign, depending on which we decide the pretraining context length to be 256.

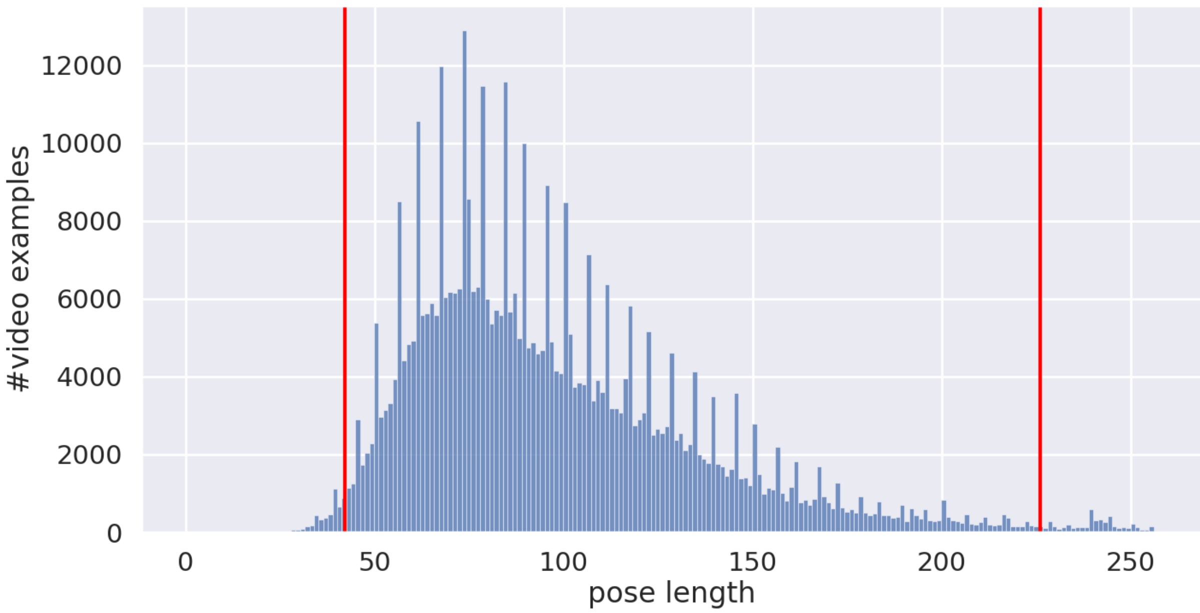

Figure 6: Pose length distribution of video examples in Spreadthesign. The two red vertical lines denote the 1st and 99th percentile of the number of frames.

Figure 7 illustrates the distribution of the number of video examples for 18,423 cross-lingual concepts in Spreadthesign. Most concept has only one video example, or one video example per sign language, and very few concepts have more than one video example for one specific sign language.

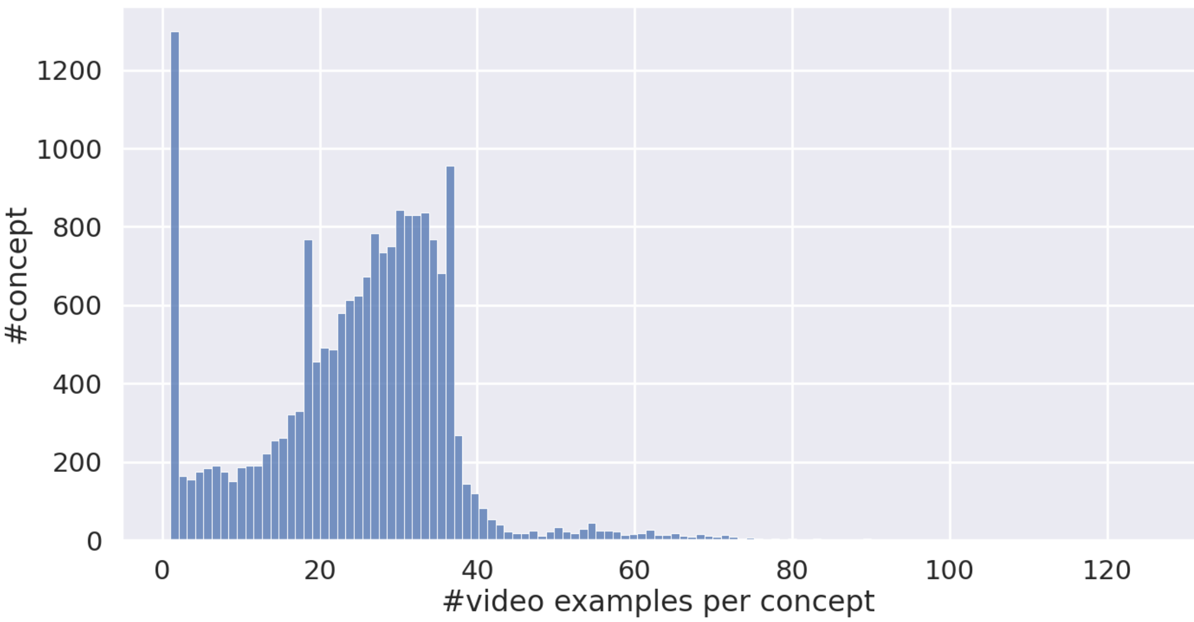

Figure 7: Concept distribution of video examples in Spreadthesign.

D Full Rank of the Signs for Iconicity Study

```
1 ('scorpion', 180.44417)
2 ('reduction', 181.40424)
3 ('illustration', 182.58504)
4 ('envelope', 182.63124)
5 ('observe', 183.66785)
6 ('gold', 184.15184)
7 ('telephone', 184.36679)
8 ('email', 185.0535)
9 ('center', 185.35873)
10 ('knife', 185.70123)
11 ('fake', 185.7079)
12 ('frozen', 186.34874)
13 ('high', 186.40546)
14 ('consume', 186.42805)
15 ('crocodile', 186.82196)
16 ('toothbrush', 186.83427)
17 ('Youtube', 186.87538)
18 ('earache', 186.9703)
19 ('Mexico', 186.98878)
20 ('ear', 187.11588)
21 ('Remind', 187.29926)
22 ('Notepad', 187.38866)
23 ('Put', 187.41467)
24 ('potato', 187.4263)
25 ('conceit', 187.43402)
26 ('thong', 187.492)
27 ('sauce', 187.50092)
28 ('obsessed', 187.57285)
29 ('drum', 187.8526)
30 ('Cuba', 188.00418)
31 ('generation', 188.16415)
32 ('grief', 188.4431)
33 ('guillotine', 188.59848)
34 ('to', 188.62285)
35 ('bind', 188.63608)
36 ('umbrella', 188.81358)
37 ('omit', 189.20493)
38 ('Superman', 189.20782)
39 ('advice', 189.48647)
40 ('Refuse', 189.52502)
41 ('speed', 189.6827)
42 ('diamond', 189.76154)
43 ('cute', 189.82486)
44 ('headache', 190.24924)
45 ('jealousy', 191.17343)
46 ('flag', 191.18753)
47 ('banana', 191.20346)
48 ('Wait!', 191.36217)
49 ('yet', 191.75407)
50 ('theft', 191.75734)
51 ('house', 191.79712)
52 ('percentage', 191.80008)
53 ('eye', 191.84584)
54 ('understanding', 191.95715)
55 ('badly', 192.02959)
56 ('skin', 192.04294)
57 ('dvd', 192.05695)
58 ('until', 192.08847)
59 ('Denmark', 192.28465)
60 ('flower', 192.2852)
61 ('sew', 192.4188)
62 ('arrest', 192.55325)
63 ('previous', 192.55621)
64 ('neighbor', 192.59973)
65 ('spoon', 192.78313)
66 ('bell', 192.82666)
67 ('click', 192.86957)
68 ('vaccination', 192.90399)
```

```
69  ('leading', 193.04103)
70  ('total', 193.0793)
71  ('internet', 193.26022)
72  ('sandwich', 193.27394)
73  ('erect', 193.31744)
74  ('signature', 193.40796)
75  ('six', 193.43289)
76  ('strange', 193.51538)
77  ('boy', 193.53194)
78  ('Farewell!', 193.86534)
79  ('cafe', 193.86969)
80  ('twice', 193.91452)
81  ('four', 193.96509)
82  ('silver', 193.97311)
83  ('China', 194.03874)
84  ('certify', 194.06538)
85  ('soup', 194.08856)
86  ('rape', 194.09244)
87  ('necessary', 194.1475)
88  ('curious', 194.27072)
89  ('tall', 194.3718)
90  ('battle', 194.45622)
91  ('promise', 194.51105)
92  ('deadline', 194.61055)
93  ('certificate', 194.736)
94  ('genuine', 194.79256)
95  ('blood', 194.82959)
96  ('ban', 194.84705)
97  ('uncle', 194.85689)
98  ('quarantine', 194.90247)
99  ('salad', 195.1925)
100 ('ant', 195.31888)
101 ('thief', 195.48566)
102 ('fall', 195.60696)
103 ('childlike', 195.6437)
104 ('Japan', 195.68546)
105 ('run', 195.70013)
106 ('booking', 195.75223)
107 ('homesick', 195.86086)
108 ('advanced', 195.86685)
109 ('Where?', 195.92314)
110 ('bridge', 195.98723)
111 ('beside', 196.01958)
112 ('cup', 196.03467)
113 ('spaghetti', 196.04333)
114 ('dizziness', 196.06311)
115 ('mixer', 196.10735)
116 ('Assessment', 196.14081)
117 ('amnesia', 196.15091)
118 ('gigantic', 196.21085)
119 ('priest', 196.2539)
120 ('sorry', 196.31093)
121 ('investment', 196.35571)
122 ('believe', 196.41522)
123 ('hang', 196.4303)
124 ('three', 196.43767)
125 ('hearing', 196.51305)
126 ('principal', 196.63132)
127 ('punctual', 196.70624)
128 ('adult', 196.91183)
129 ('thin', 196.98502)
130 ('word', 197.0534)
131 ('arm', 197.11966)
132 ('censorship', 197.13493)
133 ('several', 197.31216)
134 ('bewilder', 197.43254)
135 ('reply', 197.46024)
136 ('serious', 197.57898)
137 ('sewing', 197.62848)
138 ('do', 197.69405)
```

```
139 ('together', 197.83093)
140 ('hairgrowth', 197.85425)
141 ('bull', 197.98575)
142 ('honeymoon', 198.02545)
143 ('ball', 198.05637)
144 ('ancient', 198.14444)
145 ('selfish', 198.15906)
146 ('arise', 198.17198)
147 ('wedding', 198.21584)
148 ('hour', 198.27957)
149 ('granddaughter', 198.29315)
150 ('circle', 198.32275)
151 ('couch', 198.38385)
152 ('scientist', 198.44269)
153 ('important', 198.52599)
154 ('helicopter', 198.61218)
155 ('born', 198.69475)
156 ('trousers', 198.7119)
157 ('acceptable', 198.82736)
158 ('lamp', 198.86002)
159 ('appetite', 198.86995)
160 ('association', 198.87001)
161 ('leave', 198.95593)
162 ('dyslexia', 198.98013)
163 ('twin', 199.01114)
164 ('force', 199.04541)
165 ('insist', 199.06236)
166 ('vase', 199.08466)
167 ('easter', 199.23846)
168 ('plate', 199.27231)
169 ('best', 199.32513)
170 ('heal', 199.61261)
171 ('petrol', 199.672)
172 ('cleaner', 199.69077)
173 ('pepper', 199.92517)
174 ('economic', 199.97253)
175 ('yoghurt', 200.06439)
176 ('brother', 200.0689)
177 ('unpleasant', 200.13562)
178 ('grapes', 200.14981)
179 ('buy', 200.29669)
180 ('2', 200.31093)
181 ('frog', 200.41377)
182 ('committee', 200.56615)
183 ('complain', 200.57298)
184 ('40', 200.58508)
185 ('faultless', 200.59167)
186 ('letter', 200.63252)
187 ('angel', 200.65413)
188 ('corruption', 200.66101)
189 ('director', 200.67601)
190 ('export', 200.99376)
191 ('acne', 201.08725)
192 ('participate', 201.14157)
193 ('injury', 201.19518)
194 ('offline', 201.215)
195 ('hurt', 201.2529)
196 ('shy', 201.31795)
197 ('kilometer', 201.32117)
198 ('inauguration', 201.33997)
199 ('tale', 201.4152)
200 ('very', 201.45908)
201 ('law', 201.47714)
202 ('diploma', 201.56801)
203 ('music', 201.57074)
204 ('war', 201.63304)
205 ('school', 201.65495)
206 ('horse', 201.746)
207 ('heartburn', 201.86896)
208 ('21', 201.90666)
```

```
('surname', 201.94667)
('addicted', 201.98883)
('supper', 202.08438)
('fun', 202.09558)
('terrorist', 202.22504)
('nanny', 202.32727)
('departure', 202.34787)
('600', 202.38922)
('pet', 202.46806)
('thousand', 202.5422)
('ice', 202.56726)
('menu', 202.57507)
('revise', 202.69331)
('haired', 202.70969)
('feeling', 202.82637)
('divorced', 202.85403)
('person', 203.0279)
('dawn', 203.36467)
('anxious', 203.46112)
('autism', 203.48038)
('discussion', 203.56613)
('adoption', 203.58824)
('truth', 203.6054)
('enemy', 203.6127)
('midnight', 203.63316)
('psychology', 203.70671)
('possible', 203.8542)
('pale', 203.85995)
('cucumber', 203.87328)
('favourite', 203.95981)
('rice', 204.09598)
('bedroom', 204.11554)
('sea', 204.18881)
('shock', 204.216)
('Admitted', 204.22227)
('anxiety', 204.22739)
('ten', 204.28094)
('international', 204.30515)
('curly', 204.32364)
('alarm', 204.42891)
('corn', 204.57687)
('upset', 204.593)
('morning', 204.65657)
('spinach', 204.6566)
('celebrate', 204.74715)
('confused', 204.82571)
('25', 204.87206)
('achievement', 204.94815)
('climate', 205.02107)
('communication', 205.26228)
('delegate', 205.26901)
('doorbell', 205.41678)
('anaesthesia', 205.66046)
('world', 205.86342)
('television', 206.1195)
('information', 206.20271)
('style', 206.28078)
('chip', 206.31616)
('anonymous', 206.41803)
('fashioned', 206.46509)
('development', 206.46605)
('scarf', 206.58081)
('uploading', 206.93182)
('900', 206.93893)
('500', 207.1253)
('camera', 207.15207)
('homeless', 207.25655)
('automatic', 207.29578)
('1000000', 207.40567)
('chef', 207.72531)
```

```
279 ('50', 207.73314)
280 ('infant', 207.88846)
281 ('actress', 207.97646)
282 ('nurse', 208.75182)
283 ('800', 208.8017)
284 ('slow', 208.93741)
285 ('clinic', 208.93753)
286 ('apartment', 209.07938)
287 ('employed', 209.48071)
288 ('electrician', 209.54414)
289 ('painter', 209.57893)
290 ('desert', 209.70918)
291 ('Audiologist', 209.97559)
292 ('engine', 210.53745)
293 ('barber', 210.76971)
294 ('bathroom', 210.77377)
295 ('diabetic', 211.66092)
296 ('7', 211.76964)
297 ('27', 211.83792)
298 ('depressed', 212.88554)
299 ('employee', 213.17842)
300 ('8', 213.59476)
301 ('farmer', 216.97792)
302 ('29', 217.70363)
```