Supplementary Material:
Aligning Subtitles in Sign Language Videos

Hannah Bull[1*] Triantafyllos Afouras[2*] Gül Varol[2,3]
Samuel Albanie[2,4] Liliane Momeni[2] Andrew Zisserman[2]
[1] LISN, Univ Paris-Saclay, CNRS, France
[2] Visual Geometry Group, University of Oxford, UK
[3] LIGM, École des Ponts, Univ Gustave Eiffel, CNRS, France
[4] Department of Engineering, University of Cambridge, UK
hannah.bull@lisn.upsaclay.fr; {afourast,gul,albanie,liliane,az}@robots.ox.ac.uk
https://www.robots.ox.ac.uk/~vgg/research/bslalign/

We provide supplementary details on our training and evaluation datasets (Sec. A), further implementation details (Sec. B), additional qualitative results (Sec. C), additional experiments (Sec. D), and a broader impact statement (Sec. E).

A. Dataset details

BSL-1K$_{aligned}$. The training set contains 7 cooking, 9 food-related travel, 1 environment-related travel and 3 lifestyle documentary shows. The test set contains 2 nature and 2 cooking shows. The 4 test episodes are chosen to evaluate the alignment model in different settings: seen/unseen signer and seen/unseen programme genre (which affects the number of out-of-vocabulary words) as shown in Tab. A.1.

The signing-aligned subtitles were annotated by one deaf native BSL signer and a random subset was verified by another deaf native BSL signer, taking around 200 hours for the 24 episodes. The instruction was to shift the start and end times of each subtitle to correspond to the signing using the open-source VIA tool [4]. The process was refined over several iterations, incorporating annotator feedback. A handful of subtitles were excluded due to annotation uncertainty.

BSL Corpus [10, 11]. For our task, we employ the *Free-Translation* annotation tier, which provides written English subtitles to accompany portions of the *Conversation* and *Interview* subsets of the corpus. In total, the annotations cover a total of 227 videos after cropping to include a single signer. Of these, 141 are sourced from the *Interview* subset and 86 videos are sourced from the *Conversation* subset. For consistency with prior work, we follow the train, validation and test partition employed by [1, 9]. However, since this partition does not fully span the dataset, we add any dataset instances that were not present in the partition to the training set.

BOBSL. The test set contains 36 videos, almost all of which are factual documentaries related to nature, science and the environment. There are also a handful of food-related shows.

	#vids.	#hours	#subs	#inst.	Vocab.	OOV
Train	20	14.4	13.8K	128.1K	8.6K	\
Test (total)	4	3.3	2.0K	18.6K	2.8K	726
signer$_{seen}$, genre$_{seen}$	1	0.7	648	6.1K	1.3K	188
signer$_{seen}$, genre$_{unseen}$	1	0.9	465	4.1K	1.0K	233
signer$_{unseen}$, genre$_{seen}$	1	0.7	506	5.6K	1.1K	99
signer$_{unseen}$, genre$_{unseen}$	1	1.0	360	2.8K	882	234

Table A.1: **BSL-1K$_{aligned}$:** The test set videos were chosen to evaluate performance on episodes with either signers or genre unseen during training.

B. Implementation details

Text embeddings. For the text embeddings, we use a pretrained BERT model from HuggingFace[1] with a standard architecture of 12-layers, 12-heads and 768 model size. The model is pretrained on BookCorpus[2] and English Wikipedia[3].

Positional encodings. For the input to the video encoder, we use 512-dimensional sinusoidal positional encodings as in [12]. The positional encodings are added to the video features before feeding to the Transformer.

[*] Equal contribution

[1] https://huggingface.co/bert-base-uncased

[2] https://yknzhu.wixsite.com/mbweb

[3] https://en.wikipedia.org

Output thresholding. The output of our model is a temporal sequence of predictions between 0 and 1. For the single-subtitle SAT model, we consider the start of the subtitle to be the first time when the prediction is above $\tau = 0.5$ and the end of the subtitle to be the last time when the prediction is above $\tau = 0.5$ in the search window. When we apply a global alignment step with DTW to correct for overlapping subtitles, we no longer use these thresholds, but rather the temporal sequence of predictions between 0 and 1 using the method described in the main paper.

Training details. We use the Adam optimiser with a batch size of 64. We train with a learning rate of 10^{-5} at the word-pretraining stage, and of 5×10^{-6} at finetuning with subtitles. At the word pretraining stage, the model is trained over 5 epochs. In one epoch of word pretraining, there are approximately 700K sign instances (including sign spotting both with mouthings and dictionaries). At this point the word alignment model obtains a frame-level accuracy of 30.38% and F1@50 of 40.75% on the 1630 sign instances of the test set episodes. During full-sentence finetuning, the model is trained over 80 epochs.

C. Additional qualitative analysis

Effect of global alignment with DTW. In Fig. A.1, we present results before and after the global alignment with DTW on a long timeline. We observe that the single-subtitle Transformer model produces overlapping regions between consecutive subtitles which are resolved after the global DTW stage. Consequently, we see that the overall duration of subtitles decreases after DTW (see Fig. A.2). During the DTW stage, we order subtitles by their predicted order, not by the original order of S_{audio}. Indeed, in BSL-1K$_{aligned}$, 1.6% of subtitles in S_{gt} do not respect the original order of S_{audio}. On the test set, 1.6% of subtitles in S_{pred} switch position with respect to S_{audio}.

Results on BSL-1K$_{aligned}$. Fig. A.4 demonstrates qualitative results on BSL Corpus.

Results on BSL Corpus. Fig. A.4 demonstrates qualitative results on BSL Corpus.

Results on BOBSL. Fig. A.5 demonstrates qualitative results on BOBSL.

D. Additional experiments

We analyse performance on each test set episode and perform ablations to evaluate the influence of our data augmentations and the encoding choice for the subtitle text.

Performance on unseen signers/genres. Tab. A.2 shows the SAT model results by test set episode. Our model tends to result in larger improvements over the S^{+}_{audio} baseline for signers seen in the training episodes, but still outperforms the S^{+}_{audio} baseline for unseen signers in unseen genres. More training data would be needed to better generalise to unseen signers.

Test episode						
signer	genre	Method	frame-acc	F1@.10	F1@.25	F1@.50
seen	*seen*	S^{+}_{audio}	45.48	66.92	55.02	31.84
		SAT	**60.23**	**77.74**	**68.47**	**49.00**
seen	*unseen*	S^{+}_{audio}	64.31	74.84	64.73	34.19
		SAT	**72.56**	**81.29**	**74.19**	**52.47**
unseen	*seen*	S^{+}_{audio}	56.30	**80.79**	69.70	44.95
		SAT	**63.68**	80.32	**72.40**	**52.82**
unseen	*unseen*	S^{+}_{audio}	71.84	63.29	53.16	33.76
		SAT	**74.93**	**69.76**	**59.92**	**34.32**

Table A.2: **Performance breakdown by test episode:** Our model improves upon the S^{+}_{audio} baseline for all the combinations of seen/unseen for signer and genre. The improvements however are greater in the test episodes where the signer has been seen during training.

Text encoding choice. We experiment with word2vec [8] encodings for subtitle words instead of BERT as used in the main paper experiments. We use the pretrained word2vec model from [7], forming sentence embeddings by max pooling the encodings of all words over the channel dimension. In Tab. A.3, we see that this results in lower performance compared to using the BERT encodings. We hypothesize that this is due to word2vec using a limited vocabulary, ignoring word order, and lacking the large-scale pretraining of the BERT model.

Method	frame-acc	F1@.10	F1@.25	F1@.50
word2vec	67.16	74.59	64.96	42.06
BERT	**68.72**	**77.80**	**69.29**	**48.15**

Table A.3: **Text encoding:** We experiment with word2vec encodings instead of BERT to embed words in the subtitle.

Amount of training data. By increasing the amount of training data, we improve performance of our model on the test set. Tab. A.4 shows our results when training on random subsets of 25%, 50% and 75% of the videos in our training data. For subset selection, we randomly sample 4 times, and report the average performance across 4 trainings, as well as the standard deviation.

#training videos	frame-acc	F1@.10	F1@.25	F1@.50
5	$66.62^{\pm 0.16}$	$75.55^{\pm 0.86}$	$66.04^{\pm 1.09}$	$43.24^{\pm 0.81}$
10	$67.40^{\pm 0.28}$	$75.74^{\pm 0.25}$	$66.60^{\pm 0.25}$	$45.41^{\pm 0.88}$
15	$67.71^{\pm 0.23}$	$75.24^{\pm 0.43}$	$66.29^{\pm 0.84}$	$46.16^{\pm 0.66}$
20	**68.72**	**77.80**	**69.29**	**48.15**

Table A.4: **Amount of training data:** We train with a subset of our videos, using 5, 10, or 15 episodes instead of the total 20 used in the paper. We observe increased performance as we increase the training size.

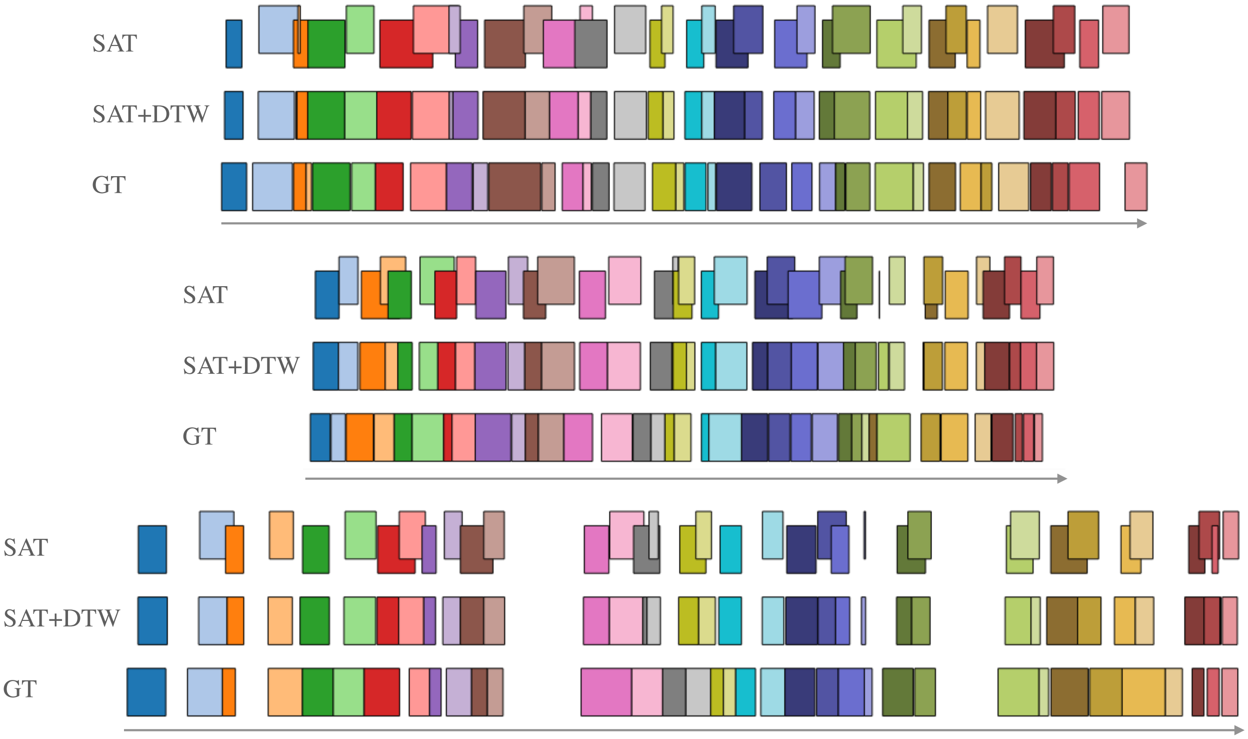

Figure A.1: **DTW:** Our SAT model predicts the locations of subtitles independently of each other, and thus there can be overlaps in subtitle localisations. Using a global alignment step with DTW, we resolve these overlaps and improve performance.

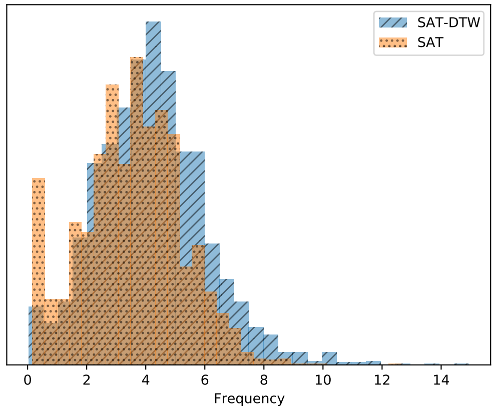

Figure A.2: **Duration before and after DTW**: The median duration of S_{gt} is 3.3s. Before DTW, the median duration of our predicted subtitles is 4.1s, but after DTW the median duration is reduced back down to 3.5s by resolving conflicts in overlapping subtitles.

Window size	frame-acc	F1@.10	F1@.25	F1@.50
8 sec	66.98	73.12	64.66	44.13
12 sec	68.63	75.52	67.56	47.29
16 sec	68.51	76.18	68.63	48.10
20 sec	**68.72**	**77.80**	**69.29**	**48.15**

Table A.5: **Search window size** T**:** We vary T between 50 and 125 frames (corresponding to 8- and 20-second inputs, respectively). Larger windows tend to perform better, possibly due to increased contextual information and the fact that the difference between S_{audio} and the aligned subtitle S_{gt} can be in the order of 10s.

Size of the search window T**.** In Tab. A.5, we report the performance against different choices for input duration T We conclude that larger search windows generally improve performance, at the cost of computational complexity. This might be due to increased supervision, since with larger windows the training sees more negative examples, as well as due to better coverage at test time. A too short window size inhibits recovery of the correct location, if the correct location falls outside of the window boundaries.

Sensitivity analysis. During inference, we predict the location of a subtitle within a 20 second search window surrounding the location of S^{+}_{audio}. In order to analyse the sensitivity of the choice of search window, we shift the window by 1s, 3s and 5s at inference time. Tab. A.6 shows that the choice of window within a margin of a few seconds does

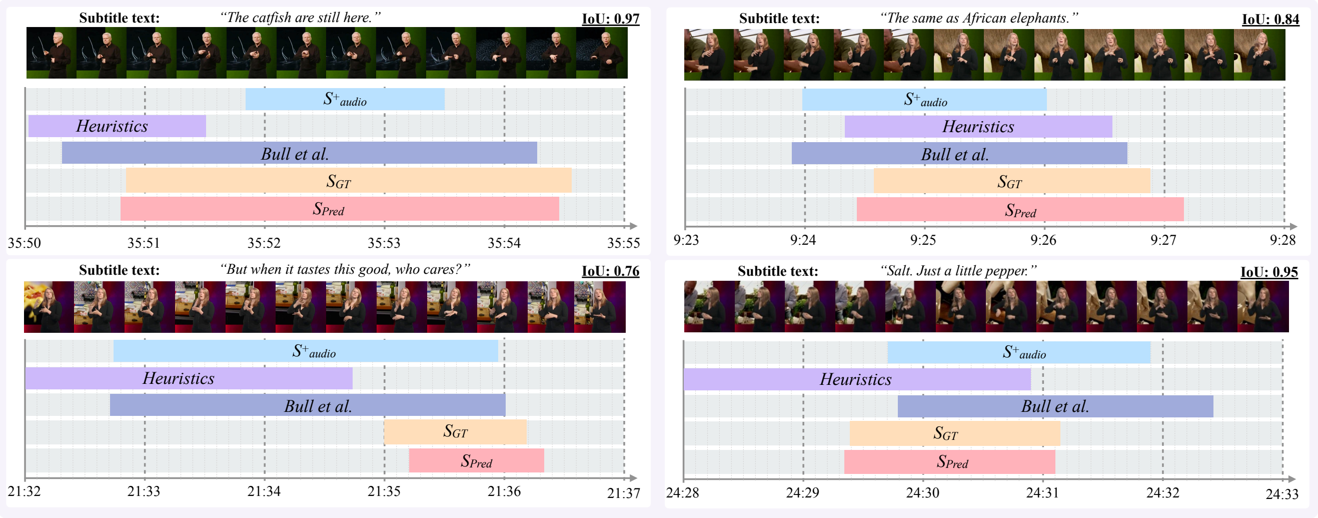

Figure A.3: **Qualitative results on BSL-1K$_{aligned}$:** This figure shows short time windows of 5s with shifted audio-aligned subtitles (S$^{+}_{audio}$), ground truth signing-aligned subtitles (S$_{gt}$) and our predicted signing-aligned subtitles (S$_{pred}$). In practice, we input 20 seconds of video during training and testing as our search window.

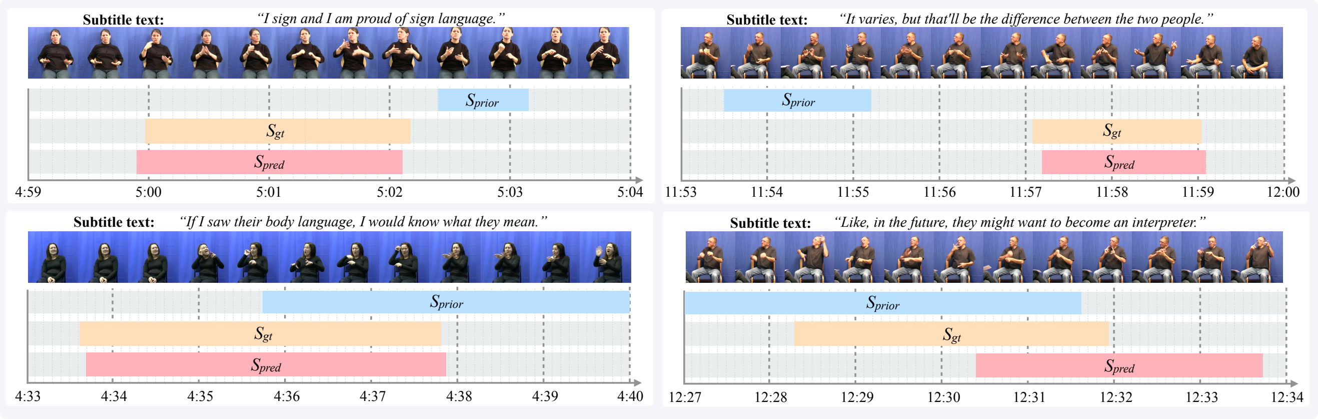

Figure A.4: **Qualitative results on BSL Corpus:** This figure shows short time windows of 5s and 7s with shifted and rescaled subtitles (S$_{prior}$), ground truth aligned subtitles (S$_{gt}$) and our predicted subtitles (S$_{pred}$). In practice, we input 20 seconds of video during training and testing for our search window. The shifted and rescaled subtitles (S$_{prior}$) are created using a random shift with standard deviation of 3.5s and a random change in length of standard deviation 1.5s.

not have a large impact on the results.

However, if we keep the position of the search window constant and change the position of the prior estimate S$^{+}_{audio}$, then this has a significant effect on results. Tab. A.7 shows the results of an experiment where we shift the prior estimate S$^{+}_{audio}$ by 1s, 3s and 5s at inference time. The performance degrades when the model is given a worse prior as input, i.e., shifting S$^{+}_{audio}$.

Shift window	frame-acc	F1@.10	F1@.25	F1@.50
0s	**68.72**	**77.80**	**69.29**	48.15
1s	68.53	76.99	69.23	47.69
3s	68.53	76.99	68.32	47.90
5s	68.32	76.58	68.42	**48.50**

Table A.6: **Shifting search window:** We shift the search window at inference time by 1s, 3s and 5s. This does not have a major impact on results.

Sampling the prior estimate. We consider an alternative choice of prior where we randomly sample S$_{audio}$ during training from a Gaussian distribution with sample mean (3.2s) and standard deviation (3.6s) of the difference between the start of S$_{gt}$ and S$_{audio}$. This choice seems equally valid in comparison to our original prior, which shifts S$_{audio}$ by the estimated mean of 3.2s. We

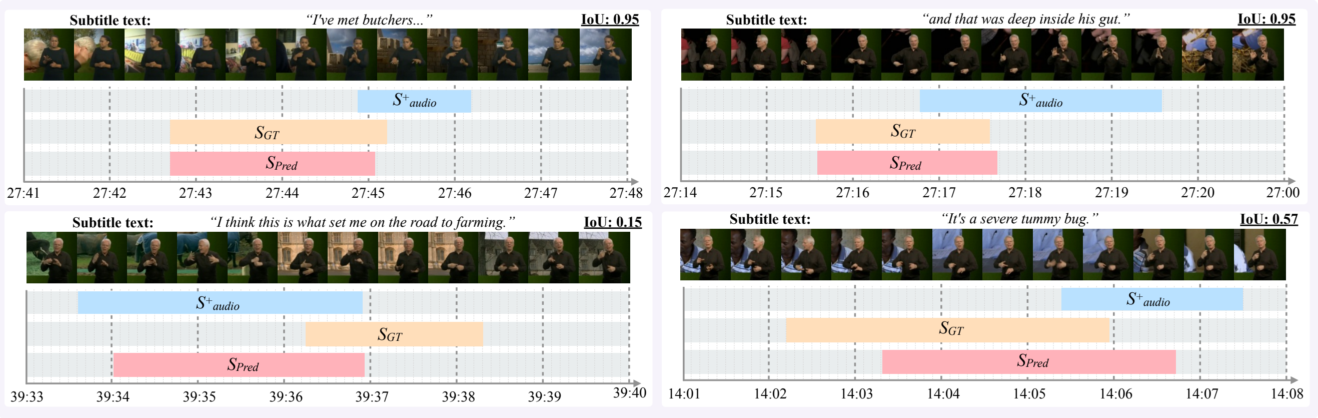

Figure A.5: **Qualitative results on BOBSL:** This figure shows short time windows of 7s with shifted audio-aligned subtitles (S^+_{audio}), ground truth signing-aligned subtitles (S_{gt}) and our predicted signing-aligned subtitles (S_{pred}).

Shift prior	frame-acc	F1@.10	F1@.25	F1@.50
0s	**68.72**	**77.80**	**69.29**	**48.15**
1s	68.26	75.77	67.36	45.67
3s	58.69	58.08	47.80	28.18
5s	46.11	35.49	26.21	12.52

Table A.7: **Shifting prior estimate S^+_{audio}:** By shifting the location of the prior S^+_{audio} at inference time by respectively 1s, 3s and 5s, the performance degrades.

No. heads	frame-acc	F1@.10	F1@.25	F1@.50
1	66.00	75.35	66.13	44.08
2	**68.72**	**77.80**	**69.29**	**48.15**
4	67.99	75.50	67.60	46.97

Table A.8: **Number of attention heads:** We choose 2-head attention for our final model.

obtain similar results, i.e. a slightly higher frame accuracy (69.15 vs 68.72), but slightly lower F1 scores ({F1@.10, F1@.25, F1@.50}={75.42 vs 77.80, 67.61 vs 69.29, 47.59 vs 48.15}).

Number of attention heads. In Tab. A.8, we ablate 1, 2 and 4 attention heads. We conclude that the model with 2-head attention performs best.

E. Broader impact

The World Federation of the Deaf states that there are 70 million Deaf individuals world-wide using more than 200 sign languages.[4] Unfortunately, many technologies for spoken and written languages do not exist for signed languages. We hope that our work contributes towards addressing this imbalance by providing inclusive technologies for signed languages for several applications, discussed next.

One direct application of our method is an assistive subtitling tool for signing vloggers to align their subtitles (this technology is currently only available for spoken and written languages). A second application is to create bilingual written-signed corpora aligned at a sentence or phrase-like level. Such corpora can be used in contextual or concordance dictionaries, useful for translation or for language learning [5]. Moreover, they can be used as training data for translation between signing and written text. For context, note that machine translation—which can now be performed to an acceptable level in many written languages to enable cross-lingual access to content—remains far from human performance for sign languages [6]. To enable progress for this task (and others that have been highlighted as important by members of Deaf communities), a key stumbling block is the availability of larger annotated datasets [2]. Our work aims to take steps towards addressing this challenge, since automatic subtitle alignment represents an important pre-processing step that has been performed manually for existing translation datasets, e.g. [3]. However, scaling manual annotation to larger datasets is prohibitively expensive (as noted in the submission, aligning one hour of video takes approximately 10-15 hours of annotation time).

We note that there are also potential risks associated with our contributions. First, there is a chance with any computational advances in sign language modelling that it leads to increased surveillance of Deaf communities (and of content moderation more generally). Second, we note that our training data, obtained from public broadcast footage, may not be demographically representative of the population as a whole, and therefore is susceptible to bias. Additionally, the videos contain BSL interpreted from English, not original BSL content. Subtitle alignment may work less effectively for individuals who are not well-represented in the training data.

[4] http://wfdeaf.org/our-work/

References

[1] Samuel Albanie, Gül Varol, Liliane Momeni, Triantafyllos Afouras, Joon Son Chung, Neil Fox, and Andrew Zisserman. BSL-1K: Scaling up co-articulated sign language recognition using mouthing cues. In *Proc. ECCV*, 2020. 1

[2] Danielle Bragg, Oscar Koller, et al. Sign language recognition, generation, and translation: An interdisciplinary perspective. In *ACM SIGACCESS*, 2019. 5

[3] Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. Neural sign language translation. In *CVPR*, 2018. 5

[4] Abhishek Dutta and Andrew Zisserman. The via annotation software for images, audio and video. In *Proc. ACMM*, volume 27 of *MM 19*, New York, USA, Oct 2019. ACM, ACM. to appear in Proceedings of the 27th ACM International Conference on Multimedia (MM 19). 1

[5] Marion Kaczmarek and Michael Filhol. Use cases for a sign language concordancer. In *Proceedings of the LREC2020 9th Workshop on the Representation and Processing of Sign Languages*, pages 113–116, 2020. 5

[6] Oscar Koller. Quantitative survey of the state of the art in sign language recognition. *arXiv:2008.09918*, 2020. 5

[7] Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. End-to-end learning of visual representations from uncurated instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9879–9889, 2020. 2

[8] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. *arXiv:1310.4546*, 2013. 2

[9] Katrin Renz, Nicolaj Stache, Samuel Albanie, and Gül Varol. Sign segmentation with temporal convolutional networks. In *International Conference on Acoustics, Speech, and Signal Processing*, 2021. 1

[10] Adam Schembri, Jordan Fenlon, Ramas Rentelis, and Kearsy Cormier. British Sign Language Corpus Project: A corpus of digital video data and annotations of British Sign Language 2008-2017 (Third Edition), 2017. 1

[11] Adam Schembri, Jordan Fenlon, Ramas Rentelis, Sally Reynolds, and Kearsy Cormier. Building the British Sign Language Corpus. *Language Documentation & Conservation*, 7:136–154, 2013. 1

[12] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 1