

Supplementary Material: Sign Language Video Retrieval with Free-Form Textual Queries

Amanda Duarte^{1,2}

Samuel Albanie³

Xavier Giró-i-Nieto^{1,4}

Gül Varol⁵

¹*Universitat Politècnica de Catalunya, Spain* ²*Barcelona Supercomputing Center, Spain*

³*Department of Engineering, University of Cambridge, UK*

⁴*Institut de Robòtica i Informàtica Industrial, CSIC-UPC, Spain*

⁵*LIGM, École des Ponts, Univ Gustave Eiffel, CNRS, France*

https://imatge-upc.github.io/sl_retrieval/

This appendix provides additional qualitative analyses (Sec. A), implementation details (Sec. B), dataset details (Sec. C), additional experiments demonstrating the sensitivity of our model to different initialisations (Sec. D), and an experiment demonstrating challenges of using text-based retrieval via sign language translation (Sec. E).

A. Qualitative Analysis

Supplemental webpage. We qualitatively illustrate, in our project page, (https://imatge-upc.github.io/sl_retrieval/app-qualitative/index.html), the retrieval results using the best model on the How2Sign dataset (SR+CM combination from Tab. 6). For each query, we show the top three ranked videos as well as their corresponding topic category (see [9] for more details of video topic categories), signer ID and sentences (note that these are not used during retrieval, and are provided for visualisation purposes).

The top ten rows of the webpage show cases in which our model is able to correctly retrieve the video corresponding to the textual query. The middle five rows of the webpage show cases where the correct video is not retrieved successfully. For these failures, we nevertheless observe that the retrieval model makes reasonable mistakes (for instance, in the majority of cases, at least one of the top three ranked videos share the same topic category of the GT video). In the bottom five rows, we show examples of failure cases of our model.

Combination of cross modal and sign recognition. We noticed that our strongest retrieval model combines similarities from the cross modal embeddings and the sign recognition model (SR+CM combination from Tab. 6). In Fig. A.2, we illustrate two example queries for which the use of the sign recognition model substantially improves the performance of the cross modal embeddings.

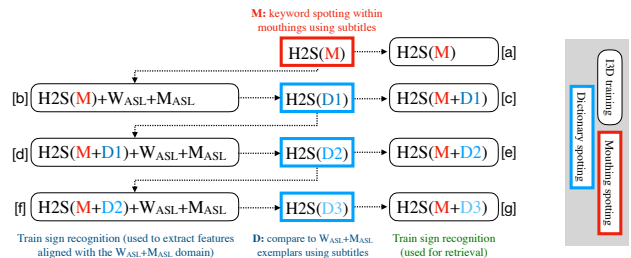


Figure A.1. **Pipeline sketch of SPOT-ALIGN iterations:** [a] We use a Mouthing-based sign spotting to obtain an initial set of automatic sign-level annotations on the How2Sign (H2S) dataset which we call here H2S(M). [b] Using the automatic annotations obtained, we jointly train on the continuous signing examples from H2S(M) and the dictionary-style signing videos from WLASL and MSASL, in order to obtain a feature space aligned between the two domains. A Dictionary-based sign spotting approach is then used to obtain a new set of sign spottings (D1) by re-querying How2Sign videos with lexicon exemplars. The process is then iterated with the new spottings, as described in the main paper.

B. Implementation Details

In this section, we provide a detailed sketch of the SPOT-ALIGN pipeline (Sec. B.1), as well as the additional implementation details for the sign video embedding (Sec. B.2), text embedding (Sec. B.3) and cross modal retrieval training (Sec. B.4).

B.1. SPOT-ALIGN iterations

In Fig. A.1 we provide a detailed sketch of the SPOT-ALIGN framework and how we obtain our Mouthing and Dictionary-based annotations for the How2Sign dataset. On the left ([b], [d], [f]), we show the iterations for the joint training between How2Sign, WLASL and MSASL, which is used for Dictionary-based sign spotting. On the right ([a], [c], [e], [g]), the training is only performed on the



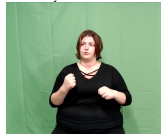
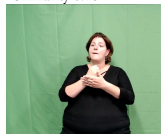



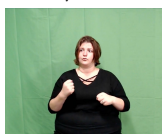
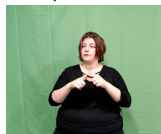
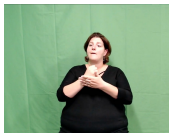
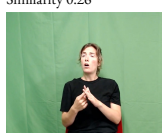
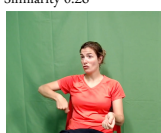
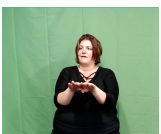
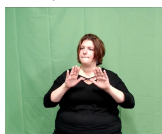
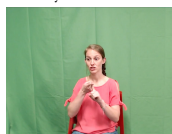
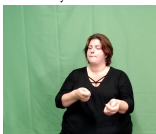

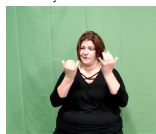
Text query	Sign video retrieval			
<p>"Then bring your feet together and by this time you should be able to have built up enough strength to do a full push up." (GT rank: 1)</p> <p>"So another example of shape we want to show you are in the teacup and we would take a look at that coming up in this series." (GT rank: 7)</p>	Combination			
<p>"Then bring your feet together and by this time you should be able to have built up enough strength to do a full push up." (GT rank: 16)</p> <p>"So another example of shape we want to show you are in the teacup and we would take a look at that coming up in this series." (GT rank: 112)</p>	Cross-Modal			
<p>"Then bring your feet together and by this time you should be able to have built up enough strength to do a full push up." (GT rank: 1) SR words: ['to', 'your', 'time', 'enough', 'full', 'have', 'push']</p> <p>"So another example of shape we want to show you are in the teacup and we would take a look at that coming up in this series." (GT rank: 3) SR words: ['show', 'in', 'that', 'and', 'you', 'up']</p>	Sign Recognition			
<div style="display: flex; justify-content: space-between;"> <div style="width: 24%;"> <p>Similarity 0.49</p>  <p>"Then bring your feet together and by this time you should be able to..."</p> </div> <div style="width: 24%;"> <p>Similarity 0.44</p>  <p>"A proper cardiovascular program should incorporate various aspects of training..."</p> </div> <div style="width: 24%;"> <p>Similarity 0.42</p>  <p>"Then when you get strong, then you can start picking up your feet."</p> </div> </div> <div style="display: flex; justify-content: space-between; margin-top: 10px;"> <div style="width: 24%;"> <p>Similarity 0.46</p>  <p>"So some other shapes when you are collecting pink luster..."</p> </div> <div style="width: 24%;"> <p>Similarity 0.46</p>  <p>"So, if we're looking at this house, for example, when you first walk in, you're going to see this vignette to your left."</p> </div> <div style="width: 24%;"> <p>Similarity 0.45</p>  <p>"So I'm shuffling this deck at the start of this segment because..."</p> </div> </div>	<div style="display: flex; justify-content: space-between;"> <div style="width: 24%;"> <p>Similarity 0.29</p>  <p>"A proper cardiovascular program should incorporate various aspects of training..."</p> </div> <div style="width: 24%;"> <p>Similarity 0.28</p>  <p>"Then when you get strong, then you can start picking up your feet."</p> </div> <div style="width: 24%;"> <p>Similarity 0.26</p>  <p>"Today we're going to work on stretching and strengthening the lower body."</p> </div> </div> <div style="display: flex; justify-content: space-between; margin-top: 10px;"> <div style="width: 24%;"> <p>Similarity 0.27</p>  <p>"So some other shapes when you are collecting pink luster..."</p> </div> <div style="width: 24%;"> <p>Similarity 0.26</p>  <p>"So, we're just going to start right in the arch, light, feathery strokes..."</p> </div> <div style="width: 24%;"> <p>Similarity 0.26</p>  <p>"Alright, now this next shot I am showing you is kind of illegal in pool halls..."</p> </div> </div>	<div style="display: flex; justify-content: space-between;"> <div style="width: 24%;"> <p>Similarity 0.28</p>  <p>"Then bring your feet together and by this time you should be able to..."</p> </div> <div style="width: 24%;"> <p>Similarity 0.22</p>  <p>"So you would push this lever and you'll pull it up into a riding position."</p> </div> <div style="width: 24%;"> <p>Similarity 0.21</p>  <p>"Here we're going to cover the initial contact, getting off first, the straight blast or the chain punch..."</p> </div> </div> <div style="display: flex; justify-content: space-between; margin-top: 10px;"> <div style="width: 24%;"> <p>Similarity 0.23</p>  <p>"If I make the reach cast like this, it pulls the fly back, so as I am making my reach cast stop..."</p> </div> <div style="width: 24%;"> <p>Similarity 0.22</p>  <p>"Now that we have our seasoned chicken wings and our seasoned flour we need to get those together so we are going to..."</p> </div> <div style="width: 24%;"> <p>Similarity 0.21</p>  <p>"So another example of shape we want to show you are in the teacup and we would take a look at that coming up in this series."</p> </div> </div>		

Figure A.2. **Qualitative results:** We show two samples where text-based retrieval using sign recognition (SR) helps retrieval when combined with cross modal embeddings (CM). Top, middle and bottom rows show the retrieval results for the same query using the average of the similarities from SR and CM (Combination), Cross Modal and Sign Recognition models, respectively.

How2Sign dataset, which provides sign video embeddings for retrieval.

B.2. Sign recognition and sign video embedding

Sign recognition training. As explained in Sec. 3.4 of the main paper, we train a sign recognition model, a 3D convolutional neural network instantiated with an I3D [6] architecture pretrained on BOBSL [2]. We finetune this model on the How2Sign dataset using our automatic sign spotting annotations. In the final setting with mouthing (M) and dictionary (D_3) spottings from a vocabulary of 1887 signs, we have 206K training video clips, each corresponding to a single sign. Since the spottings represent a point in time, rather than a segment with beginning-end times, we determine a fixed window for each video clip. For mouthing annotations, this window is defined as 15 frames before the annotation time and 4 frames after ($[-15, 4]$). For dictionary annotations, the window is similarly set to $[-3, 22]$. During training, we randomly sample 16 consecutive frames from this window, such that the RGB video input to the network becomes of dimension $16 \times 3 \times 224 \times 224$. We apply a similar spatial cropping randomly from 256×256 resolution. We further employ augmentations such as colour jittering, resizing and horizontal flipping.

We perform a total of 25 epochs on the training data, starting with a learning rate of $1e-2$, reduced by a factor of 10 at epoch 20. We optimise using SGD with momentum (with a value of 0.9) and a minibatch of size 4.

At test time, for recognition, we apply a sliding window averaging in time, and center cropping in space. At test time, for text-based retrieval, we obtain the predicted class per 16-frame sliding window (with a stride of 1 frame), and record the corresponding word out of the 1887-vocabulary if the probability is above the 0.5 threshold. The resulting set of words are merged in case of repetitions, and are compared against the queried text to obtain an intersection over union (IOU) score, used as the similarity.

Sign video embedding. As noted in Sec. 3.2 of the main paper, we employ the I3D recognition model (described above) to instantiate our sign video embedding. More specifically, we use the outputs corresponding to the spatio-temporally pooled vector before the last (classification) layer. This produces a 1024-dimensional real-valued vector for each 16 consecutive RGB frames. We extract these features densely with a stride 1 from How2Sign sign language sentences to obtain the sequence of sign video embeddings.

B.3. Text embedding

We consider several text embeddings in this work. When conducting experiments on the How2Sign dataset, we explore the following English language embeddings:

GPT [15] is a 768-dimensional embedding that uses a

Transformer decoder which is trained on the BookCorpus [20] dataset.

GPT-2-xl [16] is a 1600-dimensional embedding (employing 1558M parameters, also in a Transformer architecture [19]) that is trained on the WebText corpus (containing millions of pages of web text).

Albert-XL [10] is a 2048-dimensional embedding that builds on BERT [8] to increase its efficiency. It is trained with a loss that models inter-sentence coherence on the BookCorpus [20] and Wikipedia [8] datasets.

W2V [14] is a 300-dimensional word embedding, trained on the Google News corpus (we use the `GoogleNews-vectors-negative300.bin.gz` model from <https://code.google.com/archive/p/word2vec/>).

GroVLE [4]. This is a 300-dimensional embedding that aims to be vision-centric: it is adapted from Word2Vec [14]

For experiments on the PHOENIX2014T dataset, we use a German language model:

German GPT-2 [7] (based on the original GPT-2 architecture of [15]) is a 768-dimensional embedding. The model is trained on the OSCAR [17] corpus, together with a blend of smaller German language data. We use the parameters made available at <https://huggingface.co/dbmdz/german-gpt2>.

B.4. Cross modal retrieval

The dimensionality of the shared embedding space (denoted by the variable C in Sec. 3.2) used in this work is 512. The margin hyperparameter, m , introduced in Eqn. 1, is set to 0.2, following [13]. All cross modal embeddings are trained for 40 epochs using the RAdam optimiser [12] with a learning rate of 0.001, a weight decay of $1E-5$ and a batch size of 128. For each experiment, the epoch achieving the highest geometric mean of $R@1$, $R@5$ and $R@10$ on the validation set was used to select the final model for test set evaluation. The NetVLAD [3] layer employed in the text encoder uses 20 clusters. Sign video embeddings (which form the input to the video encoder, ϕ_v described in Sec. 3.2) are extracted densely (i.e. with a temporal stride 1).

C. Dataset Details

To construct training, validation and test retrieval partitions from the How2Sign dataset, we select video segments with their corresponding manually aligned subtitles (released by the authors of [9]). This provides an initial pool of 31,164 training, 1,740 validation and 2,356 test videos with corresponding translations. After initial inspection, we found that while most annotations were produced to a high quality, a small number of the manually aligned subtitles were invalid (i.e. exhibited no temporal overlap with

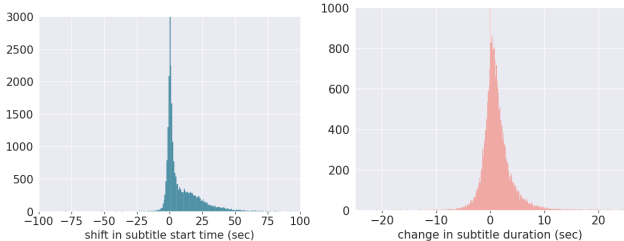


Figure A.3. We plot two histograms to illustrate the difference between the sign-aligned and speech-aligned subtitles. On the left, we show the distribution of $t_{beg}^{sign} - t_{beg}^{speech}$, i.e. the shift between the beginning subtitle times between the sign- and speech-aligned versions. On the right, we similarly plot the distribution of changes in durations. The peaks at the zero-bin are at 7,000 and 5,000 for left and right plots, respectively, which are truncated for better visibility.

Table A.1. **Sensitivity to initialisation:** We investigate the effects of different initialisation for our sign video embedding. We experiment with random and WLASL initialisation. $D_{1,BSL1K,ft(H_M WM)}$ means obtaining D_1 by pretraining the [b] model (see Fig. A.1) on BSL-1K [18] and finetuning jointly on H2S mouthing annotations and WLASL/MSASL exemplars.

Sign-Vid-Emb	Init [a][c]	#tr. ann.	Acc. top-5	Sign Recognition				Cross-modal retrieval			
				R@1↑	R@5↑	R@10↑	MedR↓	R@1↑	R@5↑	R@10↑	MedR↓
M	BOBSL	9K	27.0	0.6	2.3	4.4	1174.5	16.4 _{1.2}	31.1 _{0.8}	38.2 _{0.8}	32.7 _{3.1}
$M+D_{1,BSL1K,ft(H_M WM)}$	BOBSL	38K	77.8	10.2	21.2	26.5	136.3	20.6 _{1.1}	36.7 _{0.6}	43.3 _{0.9}	22.0 _{2.6}
M	BSL-1K	9K	25.1	0.6	2.4	4.4	1174.5	18.0 _{0.7}	32.4 _{0.6}	39.3 _{0.7}	27.8 _{1.6}
$M+D_{1,BSL1K,ft(H_M WM)}$	BSL-1K	38K	77.7	10.4	22.6	27.6	131.5	20.8 _{0.8}	36.9 _{0.9}	43.5 _{0.8}	20.5 _{0.5}
M	WLASL	9K	23.5	1.1	2.9	4.2	1175.5	11.3 _{0.5}	23.0 _{0.5}	29.5 _{0.6}	67.3 _{7.2}
$M+D_{1,WLASL,ft(H_M WM)}$	WLASL	60K	72.7	9.3	19.9	24.5	208.8	17.1 _{0.6}	31.5 _{0.6}	38.3 _{0.4}	32.8 _{2.5}
M	random	9K	6.5	0.0	0.0	0.0	1174.5	0.6 _{0.1}	2.0 _{0.2}	3.3 _{0.5}	530.7 _{16.3}
$M+D_{1,random,ft(H_M WM)}$	random	136K	28.0	0.0	0.5	0.8	1175.0	2.4 _{0.2}	7.2 _{0.2}	10.2 _{0.0}	221.3 _{6.4}

the video). We excluded these invalid subtitles from our retrieval benchmark, producing final splits of: 31,075 training, 1,739 validation and 2,348 test videos.

In Fig. A.3, we visualise the difference between the timings of the original subtitles versus the manually aligned subtitles. We note that the signing is on average behind the speech, constituting a misalignment when using the original subtitle timings. This misalignment explains the performance drop we demonstrated in Tab. 5 of the main paper when experimenting with the original subtitles instead of the manually aligned ones.

D. Sensitivity to Initialisation

We provide in Tab. A.1 comparisons for training with annotations from Mouthing (M) and the first iteration of Dictionary (D1) spottings ([a] and [c] in Fig. A.1) from four different initialisations: I3D weights pretrained on BOBSL [2], BSL-1K [18], WLASL [11], or randomly initialised. Note that all BOBSL, BSL-1K and WLASL models are also initialised from Kinetics. Here, we rerun the Dictionary-based sign spotting to obtain different sets of D_1 annotations by initialising from WLASL-pretrained and

random weights (instead of BSL-1K model from [18] in the rest of the experiments). While random initialisation significantly hurts performance, the WLASL-pretrained model performs slightly worse than [2], demonstrating that our method can work provided a reasonable initialisation. Assuming access to WLASL is realistic since we use it in step [b].

The performances of BOBSL versus BSL-1K pretraining are similar in Tab. A.1. Our preliminary results also suggest that similar trends from Tab. 1 hold when pretraining all rows on BSL-1K instead of BOBSL. We therefore report all our models ([a, c, e, g]) with BOBSL pretraining since this dataset [2] has recently become available (unlike the BSL-1K source data [1], which is not public). However, we clarify that the Dictionary spottings were obtained with models ([b, d, f]) pretrained on BSL-1K.

Furthermore, we investigate whether the domain alignment between WLASL+MSASL exemplars and How2Sign is beneficial by comparing $M+D_{1,BSL1K,ft(H_M WM)}$ and $M+D_{1,BSL1K}$. The latter consists of 112K spottings (as opposed to 38K); however, the top-5 recognition accuracy drops to 60.0% (from 77.7%) suggesting the poor quality of feature alignment between the two domains in the absence of joint finetuning.

E. Text-based Retrieval Attempt through Sign Language Translation

As mentioned in Sec. 1 of the main paper, a text-based retrieval solution using sign language translation on videos is not a viable option due to unsatisfactory video-to-text translation performance of current state-of-the-art models [5] on open-vocabulary domains. Here, we provide a brief justification by training the encoder-decoder Transformer model of [5] on the How2Sign dataset using the same sign video embeddings as in our cross-modal retrieval setting, i.e. the densely extracted I3D features. We keep all the hyperparameters identical to the publicly available setting of [5] and obtain poor BLEU scores of 1.74 and 17.08 for BLEU-4 and BLEU-1, respectively. We provide in our project page (https://imatge-upc.github.io/sl_retrieval/app-translation/index.html) the ground truth (left) and the predicted (right) sentences on the validation set and observe that the predictions tend to be generic sentences that do not correspond to the input sign language video, with the exception that sometimes the model predicts one word right out of the entire sentence.

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 recogni-

- tion using mouthing cues. In *European Conference on Computer Vision (ECCV)*. Springer, 2020. 4
- [2] Samuel Albanie, Gül Varol, Liliane Momeni, Hannah Bull, Triantafyllos Afouras, Himel Chowdhury, Neil Fox, Bencie Woll, Rob Cooper, Andrew McParland, et al. Bbc-oxford british sign language dataset. *arXiv preprint arXiv:2111.03635*, 2021. 3, 4
- [3] Relja Arandjelović, Petr Gronát, Akihiko Torii, Tomás Pa-jdla, and Josef Sivic. Netvlad: Cnn architecture for weakly supervised place recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40:1437–1451, 2018. 3
- [4] Andrea Burns, Reuben Tan, Kate Saenko, Stan Sclaroff, and Bryan A. Plummer. Language Features Matter: Effective language representations for vision-language tasks. In *The IEEE International Conference on Computer Vision (ICCV)*, 2019. 3
- [5] Necati Cihan Camgoz, Oscar Koller, Simon Hadfield, and Richard Bowden. Sign language transformers: Joint end-to-end sign language recognition and translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10023–10033, 2020. 4
- [6] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 3
- [7] Branden Chan, Stefan Schweter, and Timo Möller. German’s next language model. *ArXiv*, abs/2010.10906, 2020. 3
- [8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, 2019. 3
- [9] Amanda Duarte, Shruti Palaskar, Lucas Ventura, Deepti Ghadiyaram, Kenneth DeHaan, Florian Metze, Jordi Torres, and Xavier Giro-i Nieto. How2sign: a large-scale multi-modal dataset for continuous american sign language. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 1, 3
- [10] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations. *ArXiv*, abs/1909.11942, 2020. 3
- [11] Dongxu Li, Cristian Rodriguez, Xin Yu, and Hongdong Li. Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1459–1469, 2020. 4
- [12] Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han. On the variance of the adaptive learning rate and beyond. *International Conference on Learning Representations*, 2019. 3
- [13] Antoine Miech, Ivan Laptev, and Josef Sivic. Learning a text-video embedding from incomplete and heterogeneous data. *arXiv preprint arXiv:1804.02516*, 2018. 3
- [14] Tomas Mikolov, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *ICLR*, 2013. 3
- [15] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. 2018. 3
- [16] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. In *OpenAI blog*, 2019. 3
- [17] Pedro Ortiz Suarez, Benoît Sagot, and Laurent Romary. Asynchronous pipeline for processing huge corpora on medium to low resource infrastructures. In *7th Workshop on the Challenges in the Management of Large Corpora (CMLC-7)*, 2019. 3
- [18] Gül Varol, Liliane Momeni, Samuel Albanie, Triantafyllos Afouras, and Andrew Zisserman. Read and attend: Temporal localisation in sign language videos. In *CVPR*, 2021. 4
- [19] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 3
- [20] Yukun Zhu, Ryan Kiros, Richard S. Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 19–27, 2015. 3