QUERYD: A VIDEO DATASET WITH HIGH-QUALITY TEXTUAL AND AUDIO NARRATIONS

Andreea-Maria Oncescu   João F. Henriques   Yang Liu   Andrew Zisserman   Samuel Albanie

Visual Geometry Group, University of Oxford, UK
http://www.robots.ox.ac.uk/~vgg/data/queryd

ABSTRACT

We introduce QuerYD, a new large-scale dataset for retrieval and event localisation in video. A unique feature of our dataset is the availability of two audio tracks for each video: the original audio, and a high-quality spoken description of the visual content. The dataset is based on YouDescribe, a volunteer project that assists visually-impaired people by attaching voiced narrations to existing YouTube videos. This ever-growing collection of videos contains highly detailed, temporally aligned audio and text annotations. The content descriptions are more relevant than dialogue, and more detailed than previous description attempts, which can be observed to contain many superficial or uninformative descriptions. To demonstrate the utility of the QuerYD dataset, we show that it can be used to train and benchmark strong models for retrieval and event localisation. All data, code and models will be made available, and we hope that QuerYD inspires further research on video understanding with written and spoken natural language.

Index Terms— Audio description, retrieval

1. INTRODUCTION

The development of new datasets has been instrumental in the immense progress of modern machine learning, working hand-in-hand with methodological improvements. New, larger datasets allow incremental gains in performance with no change in methodology [1], and prevent progress stalling from over-fitting to saturated benchmarks [2]. In this work, we propose a dataset that supports investigations into the relationship between video and natural language.

For this task, obtaining rich and diverse training data is essential, an endeavour that is made difficult by the fact that "describing a video" is a relatively under-constrained objective. Dataset designers that rely on crowdsourced annotations, usually through paid micro-work platforms such as Amazon Mechanical Turk, must specify elaborate guidelines and multi-stage verification mechanisms to attempt to remove low-effort solutions, given the monetary incentive [3].

We propose a new dataset for retrieval and localisation,

sourced from the YouDescribe community[1] that contributes audio descriptions for YouTube videos. The videos cover diverse domains and the descriptions are precisely localised in time. They exhibit richer vocabulary than existing publicly available text-video description datasets (sec. 4).

The aim of the annotators is to *communicate* the contents of videos to visually-impaired people, resulting in diverse and high quality audio descriptions. Annotators add spoken narrations to videos from YouTube, subject to the time-constraint of describing the action in real-time. The proposed dataset expands previous datasets across four axes: (1) *Modality*. In addition to the original audio track of each video, QuerYD contains a separate audio track containing spoken narrations describing the action. This modality is highly complementary to the standard audio, the content of which can be unrelated to the visual content (e.g. background music, character dialogue about off-screen events, and aspects of the action that do not produce sound). In contrast, spoken narrations have a one-to-one relation to the video's content, encoded as audio. (2) *Quantity*. QuerYD is large-scale, containing over 200 hours of video and 74 hours of audio descriptions. It also contains a higher density of descriptions than other datasets (as measured by words-per-second). (3) *Quality*. Since the descriptions in QuerYD are created by volunteers who aim for high-quality descriptions of videos for the visually-impaired, and narrations are rated by other users, there is an added incentive for quality when compared to micro-work platforms. We demonstrate this difference empirically, observing a larger vocabulary size, larger number of sentences per video, and sentence lengths, when compared to other datasets (which generally contain text captions instead of audio narrations). (4) *Scalability*. QuerYD is not a static dataset, since it is based on an ever-growing collection of audio descriptions which is continually evolving with new narrations added every day. By publishing our open-source toolkit, we will empower future researchers to obtain more up-to-date snapshots of the audio descriptions data, ensuring that it stays relevant as the data demand grows.

In summary, our contributions are: (1) we propose QuerYD, a new dataset for video retrieval and localisation;
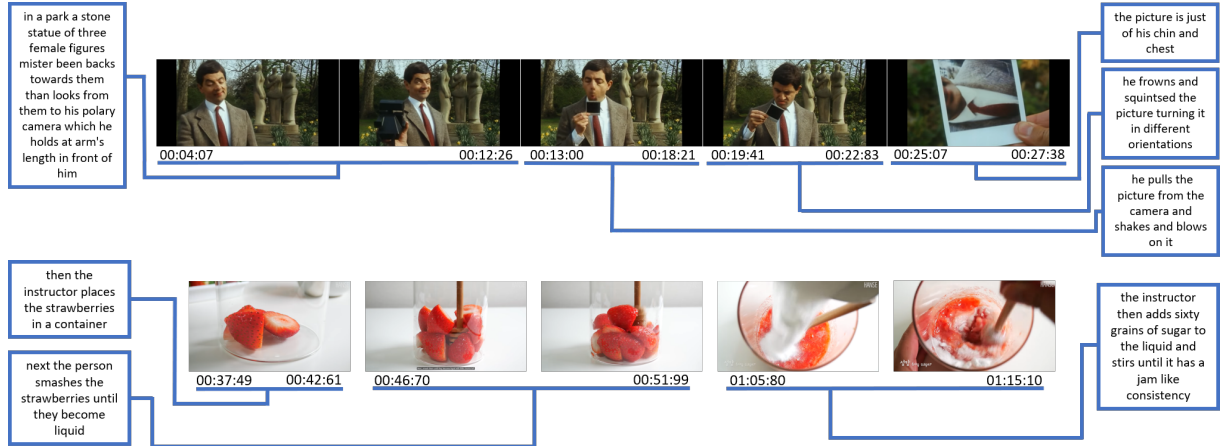
---

[1] http://youdescribe.ski.org

Fig. 1: Qualitative examples from the QuerYD dataset. Audio descriptions for video segments were provided and transcribed by voluntary annotators. Some frames are individually described while other annotations take place over a sequence of frames. The time localisation of the audio descriptions is presented in the format *minutes:seconds:miliseconds*.

(2) we provide an analysis of the data, demonstrating that it possesses a richer diversity of text descriptions than prior work; (3) we provide baseline performances with existing state-of-the-art models for text-video understanding tasks.

2. RELATED WORK

Researchers in the computer vision and natural language processing communities have been striving to bridge the gap between videos and natural language. We have seen significant progress in many tasks, such as text-to-video retrieval, text-based action or event localisation, and video captioning. This success has resulted both from advances in deep learning as well as from the availability of video description datasets.

Video-Text Retrieval Benchmarks. In the past, video-text retrieval datasets have addressed controlled settings [4, 5] or a single domain, such as cooking [6, 7], and movies [8, 9, 10]. Some new video datasets [11, 12, 13, 14, 15] have been collected from YouTube and Flicker, which are open-domain and realistic. However, most existing datasets are manually annotated. Obtaining such annotations is time-consuming and expensive and therefore difficult to do at large scale. Some recent works obtained text annotations by automatically transcribing narrations instead, and collected relatively large-scale datasets, such as CrossTask[16], YouCook2 [17], HowTo100M [18]. Unfortunately, these datasets are all sourced from a *narrow* domain: instructional videos related to pre-defined tasks. Moreover, collecting annotations from narration introduces some incoherence (noise) in the video-text pairs, due to the narrator talking about things unrelated to the video, or describing something before or after it happens.

In this work, we explore a different source for text annotations – Audio Description (AD). AD provides descriptions of the visual content in a video which allows visually-impaired people to understand the video. Compared to manual anno-

tations and narration transcriptions from instructional videos, ADs describe precisely what is shown on the screen and are always temporally well-aligned to the video. Related to our work, LSMDC [8] also considers AD as a source of supervision. However, as opposed to our work, LSMDC purely focuses on movies and short clips (3s on average) and, due to copyright retrictions, the material is not easily accessible. Our dataset contains more videos with longer duration, covers a wider range of domains and source videos are directly available from YouTube.

Event Localisation Benchmarks. Event localisation is a task that aims to retrieve a specific temporal segment from a video given a natural language text description. It requires the event localisation methods to determine not only what occurs in a video but also when. Some datasets have been collected for this task in the past few years, including Charades-STA [19], DiDemo [13], and the Activity-Net [14] dataset. However, the average length of the video in these datasets is less than 180s and the number of events per video is less than 4, which makes the event localisation less challenging. In contrast, our QuerYD dataset consists of longer videos and more distinct moments from each unedited video footage paired with descriptions that can uniquely localise the moment. It is also worth noting that instead of using manual annotations which can suffer from ambiguity in the event definition and disagreement of the start and end times, AD provide more fine-grained and precise temporal segment annotations.

3. THE QuerYD DATASET

In this section, we first give a high-level overview of the QuerYD dataset. Afterwards, we provide qualitative examples and quantitative analyses of the videos and descriptions that comprise the dataset, compared to others in the literature.

Dataset overview. The QuerYD dataset comprises 207

Partition	Train	Validation	Test
Untrimmed videos	1,815	388	390
Total descriptions	22,008	4,716	4,717
Localised descriptions	9,113	1,952	1,954

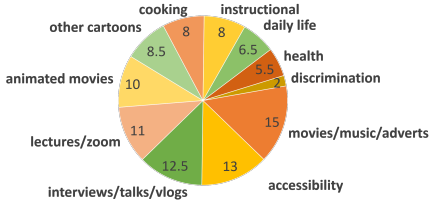**Table 1**: The partitions of the proposed QuerYD dataset.

Fig. 2: Detailed classification of the QuerYD video content by category, and associated proportion (%).

Fig. 3: A comparison of the linguistic complexity and diversity of QuerYD with popular text-video datasets. Note that QuerYD has a higher proportion of unique nouns and verbs.

Dataset	#Videos	Video Length	#Clips	Clip Length	#Sentence per video	Sent Length per clip	Vocabulary				
							Total	Noun	Verb	Adj	Adverb
DiDeMo	9605	25s	37185	6.3s	5.86 ± 0.34	7.50 ± 3.23	7865	3475	1316	841	339
ACT	20,000	180s	54926	36.18s	3.56 ± 1.67	13.58 ± 6.44	12413	5218	2162	1590	534
QuerYD	2593	278s	13019	7.72s	12.25 ± 15.27	19.91 ± 22.89	28515	8825	3551	3128	907

Table 2: Comparison of the event localisation datasets.

==hours of video accompanied by 74 hours of Audio Description (AD) transcriptions, resulting in 31,441 descriptions.== The videos, which are sourced from YouTube, cover a diverse range of visual content shown in Fig. 2. Of the total transcriptions, 13,019 are localised within video content (with precise start and end times), and the remaining 18,422 descriptions are *coarsely* localised (they are assigned a single time in a video, but their temporal extent is not annotated). The training, validation and test partitions of the dataset, obtained by uniform sampling of videos, are shown in Tab. 1.

Dataset analysis. We analysed several aspects of the QuerYD dataset: the diversity of vocabulary (Fig. 3), linguistic complexity (Tab. 2), as well as the video duration and the distribution of localised segments (Tab. 2).

We compare the QuerYD dataset with other datasets containing annotated videos such as LSMDC[8], VaTeX [15], YouCook2 [17]. Because the QuerYD videos are chosen by voluntary annotators from YouTube, the topics vary greatly with a vocabulary more varied than that of more specific datasets, as can be seen in Fig. 3 and Table 2. The vocabulary size is calculated by first assigning words to their corresponding part-of-speech and then lemmatising the tokens using the Spacy library [20]. As well as having more varied video topics, the QuerYD dataset has a wider range of video lengths as demonstrated in Tab. 2, with 9 videos longer than 40 minutes. The average video time is 278 seconds for untrimmed video, while the average of the video segments is of 7.72 seconds. This results in better temporal localisation since the audio description provided describes only the current scene.

Qualitative examples. In this section, we explore some qualitative examples from QuerYD. Examples of videos and captions are given in Fig. 1 together with the timestamps of the corresponding audio descriptions. In the top picture there are 4 rich descriptions covering 27 seconds of one video. These descriptions have a high content of verbs and nouns, reflecting the statistics in Fig. 3. The bottom image is another example of detailed descriptions of a short video with
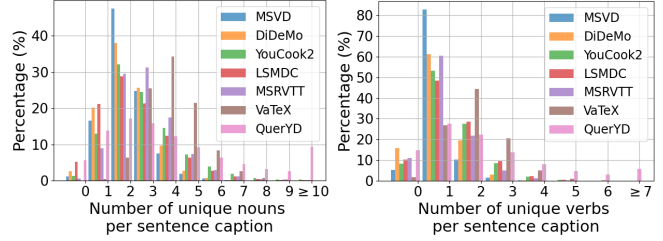
extremely fine-grained localisation.

Data collection. QuerYD is gathered from user-contributed descriptions provided by the YouDescribe community, who contribute audio descriptions to videos hosted on YouTube to assist the visually-impaired [21]. A portion of these audio descriptions is further accompanied by user-provided transcriptions. To handle cases in which such a transcription is not provided, we use the Google Speech-to-Text API [22] to transcribe audio descriptions. Unlike speech transcription from speech "in the wild", audio descriptions are recorded in a clean environment and contain only the voice of the speaker, leading to high quality transcriptions. The YouDescribe privacy policy [21] ensures that all captions are public with the consent of annotators. Lastly, we also group the localised descriptions with their corresponding clips for the localisation task and eliminate empty audio descriptions.

4. VIDEO UNDERSTANDING TASKS

In this section we demonstrate the application of QuerYD to two video understanding tasks: paragraph-level video retrieval and clip localisation. We consider three models:

(1) The *E2EWS* (End-to-end Weakly Supervised) model proposed by [23] is a cross-modal retrieval model that is trained using weak supervision from a large-scale corpus (100 million) of instructional videos (using speech content as the supervisory signal). The model employs a S3D [24] video feature extractor and a lightweight text encoder. We use the video and text encoders without any form of fine-tuning on QuerYD, providing a calibration of task difficulty.

(2) The *MoEE* (Mixture of Embedded Experts) model proposed by [25] comprises a multi-modal video model in combination with a system of context gates that learn to fuse together different pretrained "experts" (inspired by the classical Mixture of Experts model [26]) to form a robust cross-modal text-video embedding.

	Text $\Longrightarrow$ Video					Video $\Longrightarrow$ Text				
Method	R@1↑	R@5↑	R@10↑	MdR↓	MnR↓	R@1↑	R@5↑	R@10↑	MdR↓	MnR↓
E2EWS [18]	13.5±0.0	27.5±0.0	34.5±0.0	35.0±0.0	72.5±0.0	12.4±0.0	23.8±0.0	30.8±0.0	33.0±0.0	73.4±0.0
MoEE [25]	11.6±1.3	30.2±3.0	43.2±3.1	14.2±1.6	42.7±2.6	13.0±3.1	30.9±2.0	43.0±2.4	14.5±1.8	42.6±1.5
CE [28]	13.9±0.8	37.6±1.2	48.3±1.4	11.3±0.6	35.1±1.6	13.7±0.7	35.2±2.7	46.9±3.2	12.3±1.5	35.8±2.4

Table 3: Comparison of text-video retrieval methods trained with paragraph-level information on the QuerYD dataset. ↑ indicates that higher is better, ↓ that lower is better.

	Text $\Longrightarrow$ Video					Video $\Longrightarrow$ Text				
Method	R@1↑	R@5↑	R@10↑	MdR↓	MnR↓	R@1↑	R@5↑	R@10↑	MdR↓	MnR↓
E2EWS [18]	3.4±0.0	9.8±0.0	14.4±0.0	274.0±0.0	523.8±0.0	4.2±0.0	10.4±0.0	13.9±0.0	268.5±0.0	522.0±0.0

Table 4: Basedline for text-video retrieval methods on the QuerYD dataset. ↑ indicates that higher is better, ↓ that lower is better.

(3) The *CE* (Collaborative Experts) model similarly learns a cross-modal embedding by fusing together a collection of pretrained experts to form a video encoder. It uses a relation network [27] sub-architecture to combine together different modalities, and represents the state-of-the-art on several retrieval benchmarks. Except where otherwise noted, the MoEE and CE models adopt the same four pretrained experts for scene classification, action recognition, ambient sound classification and image classification described in [28] (described in detail in the supplemental material).

Text-video retrieval. We first consider the task of paragraph-level video-retrieval with natural language queries (e.g. as studied in [29, 28]). We report standard retrieval evaluation metrics, including median rank (MdR, lower is better), mean rank (MnR, lower is better) and recall at rank K (R@k, higher is better). We report the mean and standard deviation for three runs with different random seeds. We first compare the three different baseline models on our dataset and then investigate the importance of the use of different modalities.

Baselines. Our results, reported in Tab. 3, show that the CE model [28] trained on our QuerYD dataset performs best, outperforming the MoEE [25] and E2EWS [23] models (the latter being trained on a corpus of 100 million instructional videos, but used without any form of finetuning). Additionally, we propose a baseline for retrieval where the full QuerYD dataset is used as test set. The results when running the E2EWS [23] model without finetuning on the full QuerYD dataset are shown in Table 4.

The importance of different modalities for QuerYD retrieval. In Tab. 5, we assess the importance of different pretrained experts for the retrieval performance of the CE model [28] on QuerYD. We observe that the addition of each expert improves the performance, although to different extents: both object and action classification experts bring large gains in performance, while the ambient sound expert produces mixed results and has a smaller impact, indicating that this is a weaker cue, but one that still could benefit retrieval.

Clip localisation. We next study clip localisation. Corpus-level clip localisation without effective proposals is an extremely challenging task (e.g. [30] reports a performance of

	Text $\Longrightarrow$ Video					Video $\Longrightarrow$ Text				
Method	R@1↑	R@5↑	R@10↑	MdR↓	MnR↓	R@1↑	R@5↑	R@10↑	MdR↓	MnR↓
SCENE	8.7±0.4	26.3±1.1	37.1±0.7	22.2±1.6	52.3±3.0	9.1±0.8	25.4±0.9	35.3±1.5	23.2±0.3	52.6±2.6
PREV. + AUDIO	7.6±2.7	27.4±1.4	40.4±0.9	17.0±1.7	49.0±1.9	10.1±1.2	25.7±1.5	37.5±1.2	20.0±1.3	48.9±2.0
PREV. + OBJECTS	12.7±1.7	34.8±1.7	47.0±1.3	12.3±0.6	37.6±2.1	12.8±1.3	33.5±2.8	46.6±1.0	11.8±0.8	37.6±1.9
PREV. + ACTION	14.3±0.3	37.5±1.3	48.6±0.8	11.3±0.6	35.2±1.8	14.0±0.3	35.4±2.9	47.2±2.8	12.3±1.5	35.8±2.4

Table 5: The influence of different pretrained experts for the performance of the CE model [28] trained on QuerYD. The value and cumulative effect of different experts for scene classification (SCENE), ambient sound classification (AUDIO), image classification (OBJECT), and action recognition (ACTION). PREV. denotes the experts used in the previous row.

	Text $\Longrightarrow$ Video					Video $\Longrightarrow$ Text				
Method	R@1↑	R@5↑	R@10↑	MdR↓	MnR↓	R@1↑	R@5↑	R@10↑	MdR↓	MnR↓
E2EWS [18]	6.7±0.0	14.7±0.0	20.4±0.0	133.0±0.0	342.0±0.0	8.4±0.0	15.4±0.0	19.8±0.0	154.5±0.0	363.0±0.0
MoEE [25]	19.0±0.8	38.9±1.0	47.9±0.7	12.0±1.0	127.4±5.9	19.8±0.2	39.6±0.6	47.6±0.1	13.0±0.0	124.3±5.5
CE [28]	18.2±0.5	38.1±0.8	46.8±0.4	13.3±0.6	127.5±3.9	18.1±0.6	37.3±0.5	45.9±0.6	14.0±1.0	123.9±3.3

Table 6: Comparison of localisation methods trained with oracle temporal proposals information on the QuerYD dataset.

0.85% recall at rank 1 on DiDeMo). For this reason, in this work we establish a baseline for the proposal oracle setting, which consists of having the temporal segment localisations (ground truth proposals) provided as input to each method.

Localisation baselines. Our results, reported in Tab. 6, show that the MoEE method performs best, followed closely by CE. The E2EWS method, which was trained with weak supervision from a much larger set of videos, performs worse. This may be due in part to the weak supervision, which does not enforce the network to be highly sensitive to timing. The model nevertheless establishes a solid baseline for performance without fine-tuning.

Failure cases. In Fig. 4 we show some failure cases of the CE model, which we observed to be the strongest model in the text-video retrieval task. One of the main failure modes is the ambiguity between segments that can plausibly correspond to the retrieval target, showing that at such a fine level of temporal granularity no model can be expected to perform perfectly.

5. CONCLUSION

In this work, we introduced the QuerYD dataset for video understanding. Through extensive analysis, we demon-

Rank: 1 Sim: 0.563   Rank: 2 Sim: 0.542   Rank: 3 Sim: 0.463   Rank: 4 Sim: 0.454

Query: this plastic bottle cutter turns bottles into strips

Rank: 0 Sim: 0.501   Rank: 1 Sim: 0.470   Rank: 2 Sim: 0.469   Rank: 4 Sim: 0.467

Query: Theresa is now doing her interim from the camera and is going to introduce her five minute make up tutorial
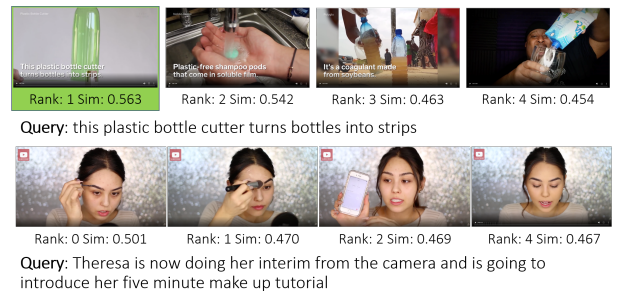
Fig. 4: Success (top) and failure (bottom) cases of the CE model trained on QuerYD and tested on text-video retrieval.

strated that this dataset contains detailed annotations with rich, highly-relevant vocabulary. In order to demonstrate applications of our dataset and to bootstrap the development of future methods, we evaluated several baselines for video retrieval and clip localisation with an oracle. By directing researchers' attention to this video-to-audio narration problem, and providing them with the tools to tackle it, we hope that future works will be able to contribute back to the `YouDescribe` project and further its goals, by developing high-quality automated video-to-audio narrations.

6. REFERENCES

[1] Kim Hammar, Shatha Jaradat, Nima Dokoohaki, and Mihhail Matskin, "Deep text mining of instagram data without strong supervision," *2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, Dec 2018. 1

[2] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar, "Do CIFAR-10 classifiers generalize to cifar-10?," *CoRR*, vol. abs/1806.00451, 2018. 1

[3] Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang, "VATEX: A large-scale, high-quality multilingual dataset for video-and-language research," *CoRR*, vol. abs/1904.03493, 2019. 1

[4] Andrei Barbu, Alexander Bridge, Zachary Burchill, Dan Coroian, Sven Dickinson, Sanja Fidler, Aaron Michaux, Sam Mussman, Siddharth Narayanaswamy, Dhaval Salvi, et al., "Video in sentences out," *arXiv preprint arXiv:1204.2742*, 2012. 2

[5] Atsuhiro Kojima, Takeshi Tamura, and Kunio Fukunaga, "Natural language description of human activities from video images based on concept hierarchy of actions," *International Journal of Computer Vision*, vol. 50, no. 2, pp. 171–184, 2002. 2

[6] Anna Rohrbach, Marcus Rohrbach, Wei Qiu, Annemarie Friedrich, Manfred Pinkal, and Bernt Schiele, "Coherent multi-sentence video description with variable level of detail," in *GCPR*. Springer, 2014, pp. 184–195. 2

[7] Pradipto Das, Chenliang Xu, Richard F Doell, and Jason J Corso, "A thousand frames in just a few words: Lingual description of videos through latent topics and sparse object stitching," in *CVPR*, 2013, pp. 2634–2641. 2

[8] Anna Rohrbach, Marcus Rohrbach, Niket Tandon, and Bernt Schiele, "A dataset for movie description," in *CVPR*, 2015, pp. 3202–3212. 2, 3

[9] Li Yao, Atousa Torabi, Kyunghyun Cho, Nicolas Ballas, Christopher Pal, Hugo Larochelle, and Aaron Courville, "Describing videos by exploiting temporal structure," in *ICCV*, 2015, pp. 4507–4515. 2

[10] Max Bain, Arsha Nagrani, Andrew Brown, and Andrew Zisserman, "Condensed movies: Story based retrieval with contextual embeddings," *arXiv preprint arXiv:2005.04208*, 2020. 2

[11] Jun Xu, Tao Mei, Ting Yao, and Yong Rui, "Msr-vtt: A large video description dataset for bridging video and language," in *CVPR*, 2016, pp. 5288–5296. 2

[12] David L Chen and William B Dolan, "Collecting highly parallel data for paraphrase evaluation," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*. Association for Computational Linguistics, 2011, pp. 190–200. 2

[13] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell, "Localizing moments in video with natural language," in *ICCV*, 2017, pp. 5803–5812. 2

[14] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles, "Dense-captioning events in videos," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 706–715. 2

[15] Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang, "Vatex: A large-scale, high-quality multilingual dataset for video-and-language research," in *ICCV*, October 2019. 2, 3

[16] Dimitri Zhukov, Jean-Baptiste Alayrac, Ramazan Gokberk Cinbis, David Fouhey, Ivan Laptev, and Josef Sivic, "Cross-task weakly supervised learning from instructional videos," in *ICCV*, 2019, pp. 3537–3545. 2

[17] Luowei Zhou, Chenliang Xu, and Jason J Corso, "Towards automatic learning of procedures from web instructional videos," in *AAAI Conference on Artificial Intelligence*, 2018, pp. 7590–7598. 2, 3

[18] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic, "Howto100m: Learning a text-video embedding by watching hundred million narrated video clips," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 2630–2640. 2, 4

[19] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia, "Tall: Temporal activity localization via language query," in *ICCV*, 2017, pp. 5267–5275. 2

[20] Explosion AI, "Spacy: natural language processing library," https://spacy.io, 2016. 3

[21] Video Description Research and Development Center, "Youdescribe," 2013. 3

[22] Google, "Speech-to-Text," https://cloud.google.com/speech-to-text. 3

[23] 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," *arXiv preprint arXiv:1912.06430*, 2019. 3, 4

[24] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy, "Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification," in *ECCV*, 2018, pp. 305–321. 3

[25] Antoine Miech, Ivan Laptev, and Josef Sivic, "Learning a text-video embedding from incomplete and heterogeneous data," *arXiv preprint arXiv:1804.02516*, 2018. 3, 4

[26] Michael I Jordan and Robert A Jacobs, "Hierarchical mixtures of experts and the em algorithm," *Neural computation*, vol. 6, no. 2, pp. 181–214, 1994. 3

[27] Adam Santoro, David Raposo, David G Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, and Timothy Lillicrap, "A simple neural network module for relational reason-

ing," in *Advances in neural information processing systems*, 2017, pp. 4967–4976. 4

[28] Yang Liu, Samuel Albanie, Arsha Nagrani, and Andrew Zisserman, "Use what you have: Video retrieval using representations from collaborative experts," *arXiv preprint arXiv:1907.13487*, 2019. 4

[29] Bowen Zhang, Hexiang Hu, and Fei Sha, "Cross-modal and hierarchical modeling of video and text," in *ECCV*, 2018, pp. 374–390. 4

[30] Victor Escorcia, Mattia Soldan, Josef Sivic, Bernard Ghanem, and Bryan Russell, "Temporal localization of moments in video collections with natural language," *arXiv preprint arXiv:1907.12763*, 2019. 4