# GOAL: A Challenging Knowledge-grounded Video Captioning Benchmark for Real-time Soccer Commentary Generation


Ji Qi*
qj20@mails.tsinghua.edu.cn
Tsinghua University
Beijing, China

Jifan Yu*
yujf18@mails.tsinghua.edu.cn
Tsinghua University
Beijing, China

Teng Tu
tut19@mails.tsinghua.edu.cn
Tsinghua University
Beijing, China

Kunyu Gao
gky20@mails.tsinghua.edu.cn
Tsinghua University
Beijing, China

Yifan Xu
xuyifan2001@gmail.com
Tsinghua University
Beijing, China

Xinyu Guan
guanxinyu@gmail.com
Biendata
Beijing, China

Xiaozhi Wang
wangxz20@mails.tsinghua.edu.cn
Tsinghua University
Beijing, China

Bin Xu†
xubin@tsinghua.edu.cn
Tsinghua University
Beijing, China

Lei Hou
houlei@tsinghua.edu.cn
Tsinghua University
Beijing, China

Juanzi Li
lijuanzi@tsinghua.edu.cn
Tsinghua University
Beijing, China

Jie Tang
jietang@tsinghua.edu.cn
Tsinghua University
Beijing, China



## ABSTRACT

Despite the recent emergence of video captioning models, how to generate vivid, fine-grained video descriptions based on the background knowledge (*i.e.*, long and informative commentary about the domain-specific scenes with appropriate reasoning) is still far from being solved, which however has great applications such as automatic sports narrative. Based on soccer game videos and synchronized commentary data, we present GOAL, a benchmark of over 8.9k soccer video clips, 22k sentences, and 42k knowledge triples for proposing a challenging new task setting as Knowledge-grounded Video Captioning (KGVC). We experimentally test existing state-of-the-art (SOTA) methods on this resource to demonstrate the future directions for improvement in this challenging task. We hope that our data resource (now available at https://github.com/THU-KEG/goal) can serve researchers and developers interested in knowledge-grounded cross-modal applications.



## CCS CONCEPTS

• **Computing methodologies** → **Knowledge representation and reasoning**; **Computer vision tasks**.

## KEYWORDS

Video Captioning, Knowledge Grounding, Open-source Dataset

**ACM Reference Format:**
Ji Qi*, Jifan Yu*, Teng Tu, Kunyu Gao, Yifan Xu, Xinyu Guan, Xiaozhi Wang, Bin Xu†, Lei Hou, Juanzi Li, and Jie Tang. 2023. GOAL: A Challenging Knowledge-grounded Video Captioning Benchmark for Real-time Soccer Commentary Generation. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management (CIKM '23), October 21–25, 2023, Birmingham, United Kingdom.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3583780.3615120


## 1 INTRODUCTION

Video captioning, the task of describing a video's content in natural language [13], was initially proposed due to its importance in a wide range of applications [18]. Since then, it has been a crucial and challenging task in both computer vision and natural language processing communities, as it requires the mastery of several skills for one model (or a system) [12], including (1) *video understanding*: understanding of the spatial-temporal dynamics in video [22], (2) *video-text bridging*: associating visual and textual elements [3, 9] and (3) *text generation*: generating appropriate long sequences of output words [16]. Despite the recent models' impressive performance [12, 15], it is noteworthy that the existing video captioning benchmarks [4, 23, 27] are still several steps away from the real-world applications. As the soccer narrative examples in Figure 1, when commentators describe the scenes in a video, they need to recognize the visible objects (*Walcott*) and actions (*Goal*), and exploit the relevant knowledge *(Walcott, Player of, Arsenal)* to understand and complete an expression. Due to the involvement of knowledge beyond the video, such applications pose more challenges to all

∗ indicates equal contribution.
† Corresponding author: xubin@tsinghua.edu.cn.




*CIKM '23, October 21–25, 2023, Birmingham, United Kingdom*



ACM ISBN 979-8-4007-0124-5/23/10.
https://doi.org/10.1145/3583780.3615120

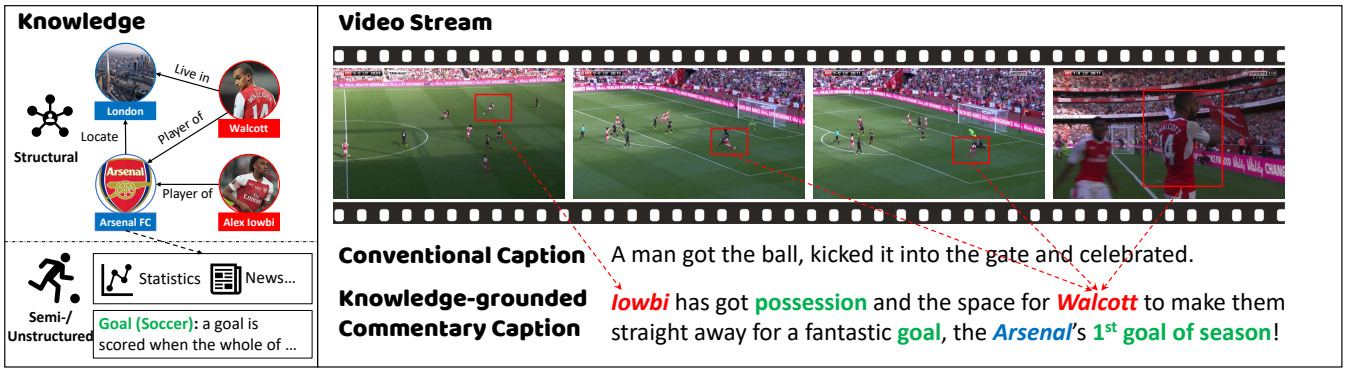


**Figure 1: Comparison of conventional video captioning with its knowledge-grounded setting. A soccer commentary describe the fine-grained entities (red), their relevant structural (blue), and semi-/unstructured knowledge (green).**

three skills of the video captioning model than the setting of relevant datasets: First, video understanding is additionally required to be able to link the objects to fine-grained entities (*Man*→ *Walcott*) and combine the multiple actions to reason certain events (*Kick, Celebrate* → *Goal*). Second, besides bridging the video and text, associating the knowledge behind them also becomes a crucial issue. Finally, models need to invoke background knowledge ($1^{st}$ *Goal of season*) to generate vivid descriptions and comments instead of simply introducing the coarse-grained scene.

We propose a knowledge-**G**rounded vide**O** c**A**ption benchmark for rea**L**-time soccer commentary generation (**GOAL**), which provides a more challenging setting of this task. Built upon the collection of over 40 hours of broadcast soccer videos [8], our dataset contains more than 8.9k video clips with corresponding commentary texts. Each video in our dataset is linked to abundant knowledge from a professional sports platform[1] and the Wikidata [19]. After that, we conduct a series of human annotation tasks, including proofreading, entity recognition, and text classification, to complete the fine-grained alignment of the knowledge, video, and text. Furthermore, we carefully pre-process the videos via object detection, event spot, and 2D/3D feature modeling to provide an easy-to-adapt interface for the state-of-the-art video captioning models.

Experimental results show that the performance of several top-performing video captioning models [11, 12] suffers an apparent decline on our benchmark, which indicates the difficulty of such a setting. Moreover, we conduct some primary explorations about knowledge-aware video captioning methods. We hope our work can call for more efforts to exploit the advanced NLP techniques (such as large-scale language model (LLM) instructions [2, 14]) in promoting this task into complex real-world applications.

**Predicted Impact and Beneficial Groups**. For researchers of video captioning, our contributions include: (1) the proposal of a "knowledge-grounded" setting for enriching the video captioning task; b) a high-quality dataset from soccer commentary as a challenging benchmark; We believe that GOAL also preserves a positive impact on the industrial developers and users in cross-modal applications, especially sport domain, as we contribute: (3) a fine-grained, open-source data repository for building knowledge-grounded soccer commentary generator and (4) a series of possible solutions in lifting current method to an applicable stage.

## 2 GOAL BENCHMARK

In this section, we introduce the collection and analysis of the proposed GOAL benchmark. Before that, we formally present the task setting for such soccer video captioning scenario:

**Problem Formulation**. Consider a soccer video $V$ consisting of $N_v$ frames described by a textual sequence $S$. Distinct from conventional video captioning setting that aims at generating $S_v$ only based on $V$, the task of **Knowledge-grounded Video Captioning (KGVC)** can be formulated as: given the video $V$, the objective is to select appropriate video-related knowledge $K_v = \mathbf{f}(V)$ and generate a knowledgeable commentary $S_{v,k}$.

$$\textbf{KGVC}: S_{v,k} = \mathbf{g}(V, K_v) = \mathbf{g}(V, \mathbf{f}(V)) \tag{1}$$

where functions $\mathbf{f}(\cdot)$ and $\mathbf{g}(\cdot)$ respectively denote the models' abilities on aligning video and knowledge (upon *video understanding*) and converting video and knowledge to knowledgeable texts (upon *video-text bridging* and *text generation*). Therefore, towards a high-quality KGVC benchmark, not only the accuracy of "Video-Text-Knowledge" triple is determined to be guaranteed (for training and testing $\mathbf{g}$), but the candidate knowledge of the video is also suggested to be sufficient (for training $\mathbf{f}$ ).

### 2.1 Data Construction

**Raw Data Collection**. Broadcast soccer videos with commentary are the foundation of our dataset.To begin the construction of the GOAL benchmark, we collect 80 full-game videos narrated in English among 500 games of the open-source SoccerNet-v2 [8]. Then we employ the Azure ASR toolkit [2] to convert the speeches to raw texts. After filtering out videos with low resolution or too sparse narration, only 20 games are preserved as annotation candidates. Furthermore, we record the basic game information of each video, such as the `gameID`, `teams`, *league*, *season*, and *result* of the game for subsequent information seeking.

[1] https://www.whoscored.com/

[2] https://azure.microsoft.com/

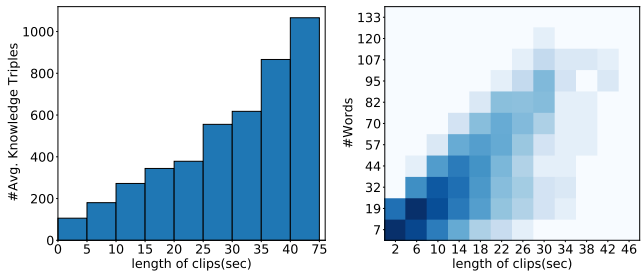


**Figure 2: Distributions of the knowledge triples (left) and word number (right) over the video clip length.**

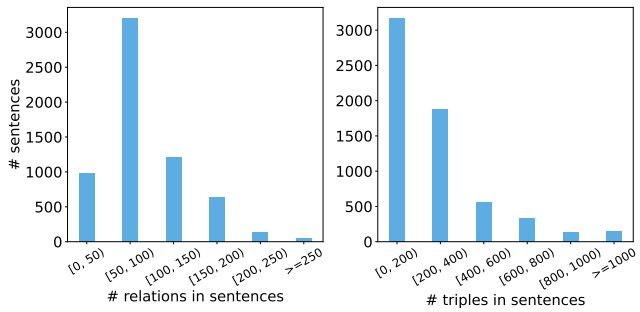


**Figure 3: The distribution of sentence-level knowledge density in terms of relation types and knowledge triples.**

**Table 1: Comparison of different datasets. Video Simi., Words/second, Var.of Sent.$_{len}$ and Syntactic Complexity correspond to the average video similarity, words per second, variance of sentence length, and the average depth of syntactic tree of the captions. Bolded and underlined represent the first and second largest value.**

| Dataset | Video Simi. | Words/ Second | Var. of Sent.$_{len}$ | Syntactic Complexity |
|---|---|---|---|---|
| MSVD | 0.77 | 0.78 | 2.86 | 6.38 |
| MSR-VTT | 0.64 | 0.62 | 4.22 | 7.52 |
| ActivityNet | 0.61 | 0.22 | 6.10 | **8.64** |
| YouCook2 | 0.58 | 0.19 | 4.58 | 6.43 |
| GOAL | **0.81** | **2.10** | **7.26** | 7.67 |

**Video-Text-Knowledge Annotation**. To guarantee the quality of the dataset, we host a team of 10 English native speakers (and are veteran soccer fans) for a series of human annotation tasks, including 1) *commentary text proofreading*: as the raw texts are not accurate enough, the first task is to do proofreading of all the texts according to the accompanying video and speech; 2) *video-text alignment*: along with proofreading, annotators are also required to adjust the text breaks and align them with the video timeline; 3) *knowledge annotation of text*: for each proofread sentence, annotators recognize the mentions of knowledge entities (especially `players`, `teams` and `terms`) with the help of the given game information and Internet. They also classify each sentence as a scene description, background introduction, or comment. All the annotation results are doubly checked.

**Knowledge Expansion**. To provide sufficient candidate knowledge for each video, we conduct knowledge expansion by seeking information from two major relevant sources: 1) for the *semi- or unstructured knowledge*, we crawl the game-related data from an online soccer platform[1], including the players' `list`, `statistics`, teams' `characteristics`, `information` and the `news` of the game. Note that we only preserve the data that is available before the game to prevent potential model crossing. 2) for the *structural knowledge*, we employ the annotated knowledge entities and game-related data to link each video to the Wikidata[3] with BLINK [21] and crawl the 2-hop related entity pages.

## 2.2 Data Analysis

**Dataset Statistics and Distribution** Our dataset contains 8.9k video clips, with an average of 2.46 labeled sentences per video, for a total of 22k sentences. Each video has an average of 21.67 commentary words, which covers 81% of the entire video clip.

Meanwhile, each video is relevant to 193.1 knowledge triples, and each video's text is labeled with an average of 1.84 knowledge entities. Figure 2 shows the related knowledge triples and word numbers both follow a nearly linear positive correlation with the clip length and keep stable fluctuations, which indicates that knowledge is evenly distributed in the sentences of the caption. Figure 3 further shows that the sentences in our dataset are mostly associated with abundant knowledge, whether in terms of relation types and knowledge triples. Overall, our dataset strives to maintain the integrity and richness of background knowledge, which provides a solid foundation for subsequent development.

**Comparison with relevant datasets.** The statistics in Table 1 further demonstrates that GOAL's setting is more challenging than the widely-used MSVD [5], MSR-VTT [23], ActivityNet Captions [4] and YouCook2 [27]. We obtain their average similarity of video clips [25], and observe that GOAL's videos are the most similar, which indicates it requires more fine-grained recognizing and understanding abilities of the model (as most of the views are players running on a green field). Nevertheless, GOAL contains the most words per second, the most diverse of the sentence (largest variance of sentence length), and pretty complex syntax (the depth of syntactic tree via Chen and Manning [6]), which makes it more challenging when generating texts.

## 2.3 Availability

We further apply several treatments to make the data fit different video captioning methods.

**Video Segmentation**: As whole video of a game is kind of long (over 90 minutes), so it is necessary to conduct video segmentation. We first follow and adapt the labeling of SoccerNet-v2 to complete the object detection (including multiple players, teams, and the ball) and event spotting (including 17 types of soccer-specific events such as `Goal`, `Foul`) [7]. After that, we conduct the game video segmentation according to the key events and text length, producing over 8.9k video clips with an average length of 10.31 seconds.

[3]https://www.wikidata.org/wiki/Wikidata:Main_Page

**Table 2: Experimental results on the GOAL benchmark. We do not sign a ↑ if the value is only lifted within 0.1.**

| Method | BLEU | METEOR | Rouge-L | CIDEr |
|---|---|---|---|---|
| HMN | 13.5 | 5.1 | 11.1 | 3.3 |
| +Knowledge | 13.5 | 5.2 | 11.4↑ | 3.5↑ |
| SwinBERT | 10.3 | 5.6 | 11.1 | 3.7 |
| +Knowledge | 10.4 | 5.5 | 12.0↑ | 4.4↑ |
| ALPRO | 14.3 | 5.9 | 10.1 | 5.1 |
| +Knowledge | 15.6↑ | 6.4↑ | 11.5↑ | 6.6↑ |

**Feature Modeling**: As some of the existing methods require the pre-modeling of the 2D/3D visual features [24, 26], we first utilize the player tracking toolkit [4] to map the 3D video into 2D $1080 * 680$ football pitch, and then employ ResNet [10] to model both the 2D and 3D features as a prepared data source. All the data and analysis tools are publicly available at https://github.com/THU-KEG/goal.

## 3 EXPERIMENT

We reproduce several representative video captioning methods on our benchmark and conduct an analysis of the experimental results.

**Setup.** We select the three types of methods: two-stage HMN [24], end-to-end video-text pre-training SwinBERT [12] and pre-training with prompting ALPRO [11] as baselines, and additionally build knowledge features (for HMN and SwinBERT) or design knowledgeable prompts (for ALPRO) according to their model architectures as improvements. We adapt the common metrics (BLEU, METEOR [1], Rouge-L and CIDEr [17]) for evaluation.

**Major Result.** The results are shown in Table 2, from which we can observe that: (1) all three methods meet a severe decline (e.g., SwinBERT's average `CIDEr` on MSR-VTT, YouCook2, MSVD, and ActivityNet is 95.5), indicating the challenging of this setting; (2) invoking knowledge is kind of benefit in improving existing methods, but designing appropriate prompts seems to be a more effective way (both HMN and SwinBERT are only slightly enhanced).

**Error Analysis.** We conduct error analysis by presenting cases of ALPRO. Except for the general errors that models cannot effectively generate such a long and informative text, we notice several knowledge-aware errors, as shown in Figure 4. (1) Entity Mismatching: models often incorrectly recognize the fine-grained player objects, *e.g.*, misidentify *Messi* as *Dzeko*, *Suarez* as *Aguero*. (2) Action Misunderstanding: models cannot understand a certain action with knowledge. Although *Save* and *Close down* are both actions in defense, they are totally different because the former one can only be performed by the goalkeeper. (3) Knowledge Deficiency: Current models lack a vast amount of parametric knowledge and the corresponding reasoning ability, which may cause many comical errors such as *a good ball from the referee*.

**Discussion.** Given the above result, we discuss some potential directions for building advanced models for the KGVC task. First, beyond the current object detection, it is necessary to propose a knowledge-aware entity-object linking [20] for such fine-grained

[4]https://github.com/JooZef315/football-tracking-data-from-TV-broadcast


**Figure 4: Representative error cases of the generated captions, which correspond to the entity mismatching, action misunderstanding and knowledge deficiency.**

video understanding applications. Second, KGVC models should exploit the language models' generation ability instead of just applying them in textual feature modeling. Third, considering the complexity of the task, it is promising to jointly utilize the chain-of-thought prompting of LLMs and knowledge graphs to understand, reason, and achieve better performance.

## 4 CONCLUSION AND FUTURE WORK

In this paper, we propose the task of knowledge-grounded video captioning (KGVC) and present a challenging benchmark upon real-time soccer commentary, GOAL, for supporting the research explorations. We conduct preliminary experiments to prove the difficulty of our benchmark and propose several promising directions. Emergent future work includes: 1) designing fine-grained knowledgeable captioning methods as well as knowledge-aware evaluation metrics; 2) exploiting the large-scale language models for reasoning in this task; 3) expanding the knowledge-grounded video captioning setting to other real-world applications.

## 5 ACKNOWLEDGMENT

This work is supported by the National Natural Science Foundation of China (No. 62277033), and the Key-Area Research and Development Program of Guangdong Province (2019B010153002).

# REFERENCES


[1] Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*. 65–72.

[2] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[3] Shyamal Buch, Cristóbal Eyzaguirre, Adrien Gaidon, Jiajun Wu, Li Fei-Fei, and Juan Carlos Niebles. 2022. Revisiting the" Video" in Video-Language Understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2917–2927.

[4] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. 2015. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition*. 961–970.

[5] David Chen and William B Dolan. 2011. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*. 190–200.

[6] Danqi Chen and Christopher D Manning. 2014. A fast and accurate dependency parser using neural networks. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 740–750.

[7] Anthony Cioppa, Adrien Deliege, Floriane Magera, Silvio Giancola, Olivier Barnich, Bernard Ghanem, and Marc Van Droogenbroeck. 2021. Camera calibration and player localization in soccernet-v2 and investigation of their representations for action spotting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4537–4546.

[8] Adrien Deliège, Anthony Cioppa, Silvio Giancola, Meisam J. Seikavandi, Jacob V. Dueholm, Kamal Nasrollahi, Bernard Ghanem, Thomas B. Moeslund, and Marc Van Droogenbroeck. 2021. SoccerNet-v2 : A Dataset and Benchmarks for Holistic Understanding of Broadcast Soccer Videos. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

[9] Yuying Ge, Yixiao Ge, Xihui Liu, Dian Li, Ying Shan, Xiaohu Qie, and Ping Luo. 2022. Bridging Video-Text Retrieval With Multiple Choice Questions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 16167–16176.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[11] Dongxu Li, Junnan Li, Hongdong Li, Juan Carlos Niebles, and Steven CH Hoi. 2022. Align and Prompt: Video-and-Language Pre-training with Entity Prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4953–4963.

[12] Kevin Lin, Linjie Li, Chung-Ching Lin, Faisal Ahmed, Zhe Gan, Zicheng Liu, Yumao Lu, and Lijuan Wang. 2022. SwinBERT: End-to-end transformers with sparse attention for video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 17949–17958.

[13] Hui Liu and Xiaojun Wan. 2021. Video paragraph captioning as a text summarization task. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. 55–60.

[14] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155* (2022).

[15] Paul Hongsuck Seo, Arsha Nagrani, Anurag Arnab, and Cordelia Schmid. 2022. End-to-end generative pretraining for multimodal video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 17959–17968.

[16] Yaya Shi, Haiyang Xu, Chunfeng Yuan, Bing Li, Weiming Hu, and Zheng-Jun Zha. 2022. Learning Video-Text Aligned Representations for Video Captioning. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* (2022).

[17] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4566–4575.

[18] Subhashini Venugopalan, Marcus Rohrbach, Jeffrey Donahue, Raymond Mooney, Trevor Darrell, and Kate Saenko. 2015. Sequence to sequence-video to text. In *Proceedings of the IEEE international conference on computer vision*. 4534–4542.

[19] Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Commun. ACM* 57, 10 (2014), 78–85.

[20] Xuwu Wang, Junfeng Tian, Min Gui, Zhixu Li, Rui Wang, Ming Yan, Lihan Chen, and Yanghua Xiao. 2022. WikiDiverse: A Multimodal Entity Linking Dataset with Diversified Contextual Topics and Entity Types. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 4785–4797.

[21] Ledell Wu, Fabio Petroni, Martin Josifoski, Sebastian Riedel, and Luke Zettlemoyer. 2020. Scalable Zero-shot Entity Linking with Dense Entity Retrieval. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 6397–6407.

[22] 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*. 6787–6800.

[23] 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*. 5288–5296.

[24] Hanhua Ye, Guorong Li, Yuankai Qi, Shuhui Wang, Qingming Huang, and Ming-Hsuan Yang. 2022. Hierarchical Modular Network for Video Captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 17939–17948.

[25] Li Yuan, Tao Wang, Xiaopeng Zhang, Francis EH Tay, Zequn Jie, Wei Liu, and Jiashi Feng. 2020. Central Similarity Quantization for Efficient Image and Video Retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3083–3092.

[26] Ziqi Zhang, Yaya Shi, Chunfeng Yuan, Bing Li, Peijin Wang, Weiming Hu, and Zheng-Jun Zha. 2020. Object relational graph with teacher-recommended learning for video captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 13278–13288.

[27] Luowei Zhou, Chenliang Xu, and Jason J Corso. 2018. Towards automatic learning of procedures from web instructional videos. In *Thirty-Second AAAI Conference on Artificial Intelligence*.