EVOQUER: Enhancing Temporal Grounding with Video-Pivoted Back Query Generation

Yanjun Gao[1], **Lulu Liu**[1], **Jason Wang**[1], **Xin Chen**[2], **Huayan Wang**[2], **Rui Zhang**[1]
Pennsylvania State University[1], Kwai Inc[2]
{yug125,lzl5409,jjw6188,rmz5227}@psu.edu,
xinchen.hawaii@gmail.com, wanghuayan@kuaishou.com

Abstract

Temporal grounding aims to predict a time interval of a video clip corresponding to a natural language query input. In this work, we present EVOQUER, a temporal grounding framework incorporating an existing temporal grounding model and a video-assisted query generation network. Given a query and an untrimmed video, the temporal grounding model predicts the target interval. Afterward, the predicted video clip is fed into a video translation task by generating a simplified version of the input query. Our framework forms closed-loop learning by taking into account both the loss functions from the temporal grounding task and loss functions from the translation, serving as feedback. Our experimental results on a widely used dataset, Charades-STA, show that EVOQUER achieves promising improvements.

1 Introduction

Temporal grounding locates the video content that semantically corresponds to a natural language query, addressing the temporal, semantic alignment between vision and language. It is a key issue in many video understanding tasks such as visual storytelling (Lukin et al., 2018; Huang et al., 2016), video caption generation (Krishna et al., 2017; Long et al., 2018), and video machine translation (Wang et al., 2019b). Given a query and an untrimmed video, the goal of temporal grounding is to find the time interval in a video that expresses the same meaning as the query.

Recent work on temporal grounding has achieved significant progress (Mun et al., 2020; Chen and Jiang, 2019; Chen et al., 2018; Zhang et al., 2019; Gao et al., 2017). These works emphasize modeling the semantic mapping of verbs and nouns in the text query to visual clues such as actions and objects that indicate the candidate time intervals. Most of them employ a uni-direction learning flow by using a single task. Inspired by multi-task learning for temporal grounding (Xu et al., 2019), this work explores the possibility of enhancing the temporal grounding model with related tasks. To this end, we borrow the idea of feedback-error-learning (FEL) from control theory and computational neuroscience (Kawato, 1990; Gomi and Kawato, 1993). Using a closed-loop system, the control network learns to correct its error from feedback and gains stronger supervision to stimulate learning. We investigate if a temporal grounding model can be improved by incorporating another network for feedback generations, forming a closed-loop learning flow.

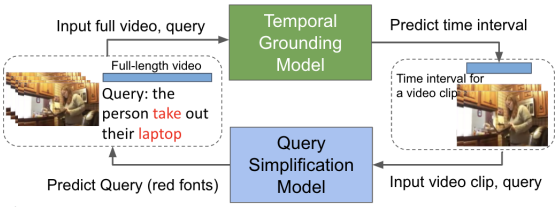

Figure 1: An overview of our closed-loop system pipeline.

More specifically, we adapt a video-pivoted query simplification task that simplifies the query to shorter phrases with verbs and noun phrases only. Visual pivots in translation task have proved to be effective as they provide more fine-grained semantic discrepancy associated with words (Chen et al.; Lee et al., 2019). We propose a novel framework, EVOQUER (*E*nhancing Temporal Grounding with *V*ide*O*-Pivoted Back *QUER*y Generation), integrating a text-to-video and a video-to-text flow, as shown in Figure 1. The pipeline pairs a state-of-the-art temporal grounding model LGI (Mun et al., 2020) with a video machine translation model (Wang et al., 2019b) for query simplification. Given a query and an untrimmed video, the pipeline predicts the time interval, feeds the video clips extracted from prediction with the original query to the translation model, and generates simplified query. We evaluate EVOQUER on a temporal grounding dataset, demonstrate promising results from interval prediction performance and qualities of simplification output, and analyze the output for

future improvements in later sections.

2 Related Work

Previous work regarding text-to-video temporal grounding has been based on identifying relationships between an event and its corresponding query. Approaches can be split into three major categories: strongly supervised (Anne Hendricks et al., 2017; Gao et al., 2017; Liu et al., 2018; Chen et al., 2018, 2019; Chen and Jiang, 2019; Ge et al., 2019; Ghosh et al., 2019; Zhang et al., 2019; Yuan et al., 2019; Mun et al., 2020; Rodriguez et al., 2020), weakly supervised (Tan et al., 2021), and reinforcement learning (Wang et al., 2019a; He et al., 2019).

Our work belongs to the supervised learning framework. The primary example of a strongly supervised approach being used is the LGI algorithm (Mun et al., 2020). This work achieves the most state-of-the-art performance on the Charades-STA dataset. The output of this algorithm are time intervals, each predicted using word-level and sentence-level attention.Within our closed-loop framework, we utilize this LGI algorithm as a "black box" to achieve the best performance on supervised temporal grounding.

Other tasks also emphasize Text-to-Video. Here we list two that are most similar to temporal grounding. Text-to-Video moment retrieval focuses on grounding between query and video (Xu et al., 2019; Lin et al., 2020; Liu et al., 2018), framing the task as to retrieve frames, slightly different from temporal grounding.

The other relevant task is video captioning aiming to generate a description of text given a video (Das et al., 2013; Yao et al., 2015; Venugopalan et al., 2015a,b; Xu et al., 2015; Zhou et al., 2019, 2018a). Recent developments have utilized end-to-end transformer models for video captioning (Zhou et al., 2018b).

3 EVOQUER Framework

Our goal is to design a closed-loop framework for the temporal grounding task such that the model receives supervision in predicting time intervals as well as feedback from the output video features extracted from the prediction. To achieve this, we propose a framework involving two components: a temporal grounding model and a translation module. The temporal grounding model predicts time intervals given an untrimmed video and a query. The translation module takes input from queries and video features trimmed by the predicted intervals, and outputs a simplified query with only verbs and nouns. Recall that we use the LGI model (Mun et al., 2020) for temporal grounding, which achieved state-of-the-art performance using supervised learning. For query simplification, we use the video machine translation framework VMT (Wang et al., 2019b).VMT is proposed for video-assisted bilingual translation, for example, between Chinese and English, and achieves reasonable results.

Our pipeline is presented in Figure 2. The input to the framework is an untrimmed video and a set of queries. Following Mun et al. (2020), we use I3D frame-based features for video representation and an embedding layer inside a text encoder for word representation. Given the video features and queries, LGI predicts time intervals with the content corresponding to a given query. Next, we extract frames from videos trimmed by the predicted interval to represent the content of the video clip. To maintain the continuity of the content, we extract 32 frames per video clip in a way that the content of the trimmed videos is evenly distributed across all 32 frames. Since the camera used captures 24 frames per second, a 32-frame video roughly spans 1.3 seconds. We feed the extracted video features and input query into a translation module consisting of two biLSTM-based encoders and an LSTM-based decoder with attention. Video hidden states and text hidden states are sent individually to two attention modules, while being concatenated into one vector representation and sent to the decoder as initial hidden states. In the attention network, temporal attention is learned through video features, and soft attention through query hidden states. The attention is fed into the decoder as context representation.

Instead of learning to decode the original query, we want the model to focus on the words that distinguish the video content: verbs and nouns. In the Charades dataset, annotators who generated the query tend to use various verb tenses when describing the video activities. For example, annotators could use both *"closes the door"* and *"closing the door"* on the same video content. Therefore, we lemmatize the words, label the query with part-of-speech (POS) tags, and extract verbs and nouns as simplified versions of the queries. The decoder learns to predict simplified queries and computes a negative log likelihood (NLL) loss at the end of

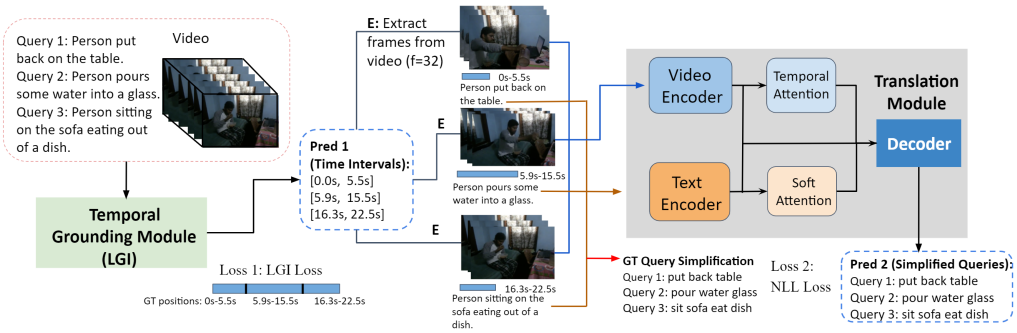

Figure 2: Our EVOQUER framework combines a LGI model for temporal grounding and a translation module that outputs simplified queries.

the decoding. Finally, we combine the NLL loss with the LGI loss computed earlier to update the networks.

In addition, we experiment with an alternative setting of the translation module: we only generate simplified queries from the video input and add a loss to explicitly enforce the mapping between video features and text features. We keep the text encoder to generate text hidden states, and apply a visual embedding (VSE) loss proposed in Faghri et al. (2018) to learn the joint embedding between text and video based on their cosine similarity. The neural network is trained end-to-end to jointly optimize the three loss functions (LGI, NLL, VSE).

4 Experiments

In this section, we present our experiment on the Charades-STA dataset using EVOQUER and EVOQUER +VSE. Results include time interval prediction and simplification output. EVOQUER shows promising improvements over the other settings.

4.1 Dataset and evaluation metrics

We evaluate our framework on Charades-STA (Gao et al., 2017), a widely used benchmark data set for temporal grounding (Mun et al., 2020). It is comprised of 9,848 roughly 30-second videos of daily human activities. Each video corresponds to a set of queries created by annotators watching these videos. There are 27,847 textual queries in total provided for the videos, with each having a maximum length of 10 words.Of these queries, we set the train/valid/test as 50%, 25%, and 25%.

We adopt two conventional evaluation metrics for temporal grounding tasks: *R@tIoU* measuring recall at different thresholds for temporal intervals between ground truth and prediction, with the threshold set as 0.3, 0.5, and 0.7; *mIoU* reporting the average of temporal interval recall from all threshold levels. For query simplification, we evaluate the predicted queries with two metrics. Jaccard similarity measures intersection over union between words in ground truth and in prediction. Since it does not penalize for duplicated words, Jaccard similarity gives us a rough estimation for the quality of translation output. BLEU (Papineni et al., 2002) is a standard evaluation metric for machine translation that measures n-gram word overlap. Most of the simplified queries are two-word length, thus we report BLEU unigram and bigram.

Model	R@0.3	R@0.5	R@0.7	mIoU
LGI	71.54	**58.08**	34.68	50.28
EVOQUER	**71.57**	57.81	**35.73**	**50.48**
EVOQUER +VSE	70.46	57.81	35.51	50.16

Table 1: Results on Charades-STA test set from the LGI model and two EVOQUER variants.

4.2 Temporal grounding results

Table 1 presents results on the Charades-STA test set from a re-trained LGI model and EVOQUER models.[1] Compared to LGI, EVOQUER shows improvement on R@0.7 and mIoU, especially 1.05 on R@0.7, the hardest threshold for temporal interval overlap. EVOQUER also outperforms LGI on R@0.3 and mIoU;however, there are some drops on R@0.3 and mIoU with VSE.

Table 2 presents statistics of samples where our model show improvements and drops compared to LGI. We divide the samples into four categories according to their recall: when EVOQUER is higher than the LGI, when EVOQUER is lower than the

[1]Using the codes from the author's GitHub and the parameters presented in the original paper, we train the LGI model on Charades-STA train set from scratch. We suspect the difference between our replication and results presented in the paper is attributed to initialization.

	Both >= R@0.3			Both
	EVOQUER ↑	EVOQUER ↓	Same	<R@0.3
Cnt.	441	362	1347	777

Table 2: Counts of samples that are scored by R@tIoU with four categories from comparison between EVOQUER and LGI model. Three of the categories are from samples where both models achieve recall equal and above threshold 0.3: samples that are improved (EVOQUER ↑), samples with performance drops (EVOQUER ↓), and equal performance with at least R@0.3 (Same). The fourth category is when both perform below R@0.3 (Both <R@0.3).

Model	JaccSim	BLEU1	BLEU2
EVOQUER	**51.98**	**53.04**	**42.47**
EVOQUER +VSE	6.37	7.96	1.20

Table 3: Translation quality measuring by Jaccard similarity, BLEU Unigram (BLEU1) and Bigram (BLEU2).

LGI, when both have the same recalls that are at least R@0.3, and when both scores are below R@0.3. EVOQUER shows improvements from 441 samples that are above 0.3 recall threshold, suggesting promising results. There are 777 cases where both models perform poorly, showing more room for improvements. To summarize, this preliminary results show that EVOQUER could bring potential benefits to the temporal grounding task.

4.3 Translation output analysis

We compare translation output from two EVOQUER variants. This comparison could help us identify the components that are critical to video-assisted machine translation. Results are shown in Table 3. Although both frameworks show similar trends in performance of the temporal grounding task, their translation quality have a large difference:EVOQUER shows good scores, while EVOQUER +VSE shows significantly lower performance. This shows that both video features and text features are critical to translation.

5 Discussion

In this section, we show output examples of predicted intervals and simplified queries to understand the model performance. Figure 3 shows two video clips trimmed by the ground truth interval, the queries, and predicted simplification. In the first example, EVOQUER predicts interval overlapping with ground truth and correctly translates the verb and noun *close door*. EVOQUER +VSE inaccurately predicts the verb *open* instead of *close*. Judging from the video content, the door was already closed; thus, an *open door* action must occur before the *close door*. Given that EVOQUER

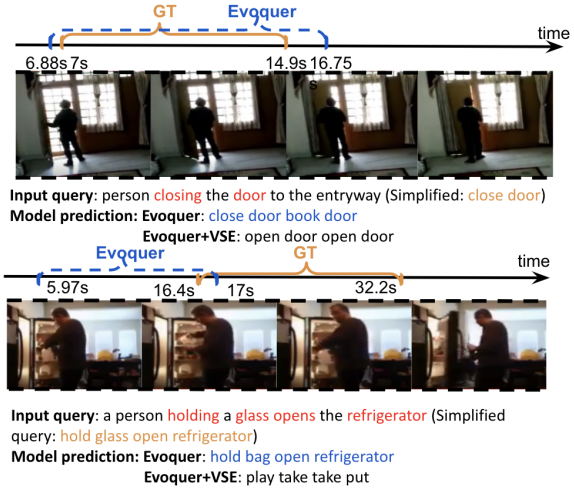

Figure 3: Two example video clips trimmed by ground truth intervals. In the first example (top), EVOQUER successfully predicts time interval and simplified queries as ground truth. In the second example (bottom), EVOQUER fails to predict time interval. Simplified queries predicted by EVOQUER +VSE are also presented.

+VSE only takes video content as input to decoder, we suspect the features of *open door* are stronger than *close door* thus captured by the decoder in EVOQUER +VSE. In the second video, EVOQUER predicts an interval rarely intersecting with ground truth. We review the video and find that at 5.97s, the person in the video starts the action *open the refrigerator door* and pours milk into a glass. Additionally, at 16.4s, he finishes *pouring* and puts the milk back into the refrigerator (shown as the first picture of Figure 3 bottom). Meanwhile, he is holding the glass and leaving the refrigerator door open. Although EVOQUER fails to intersect with the gold standard, it captures the action *open the door* at 5.97s, showing its capability in understanding the video content. We suspect EVOQUER thinks that the person is holding a bag instead of a gallon of milk since both are white in color and similar in size. Thus, it predicts *hold bag* instead of *hold glass*. Our future work will extend the experiments on other temporal grounding datasets to better validate EVOQUER performance.

6 Conclusion

We propose a novel framework, EVOQUER, for temporal grounding that incorporates a query simplification task. It forms closed-loop learning and provides feedback to the temporal grounding model and enhances the learning. Our experiments demonstrate promising results on predicting time intervals and query simplification. Future work will explore more settings and extend to other datasets.

References

Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. 2017. Localizing moments in video with natural language. In *Proceedings of the IEEE international conference on computer vision*, pages 5803–5812.

Jingyuan Chen, Xinpeng Chen, Lin Ma, Zequn Jie, and Tat-Seng Chua. 2018. Temporally grounding natural sentence in video. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 162–171.

Jingyuan Chen, Lin Ma, Xinpeng Chen, Zequn Jie, and Jiebo Luo. 2019. Localizing natural language in videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8175–8182.

Shaoxiang Chen and Yu-Gang Jiang. 2019. Semantic proposal for activity localization in videos via sentence query. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8199–8206.

Shizhe Chen, Qin Jin, and Jianlong Fu. From words to sentences: A progressive learning approach for zero-resource machine translation with visual pivots.

Pradipto Das, Chenliang Xu, Richard F Doell, and Jason J Corso. 2013. A thousand frames in just a few words: Lingual description of videos through latent topics and sparse object stitching. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2634–2641.

Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. 2018. Vse++: Improving visual-semantic embeddings with hard negatives.

Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. 2017. Tall: Temporal activity localization via language query. In *Proceedings of the IEEE international conference on computer vision*, pages 5267–5275.

Runzhou Ge, Jiyang Gao, Kan Chen, and Ram Nevatia. 2019. Mac: Mining activity concepts for language-based temporal localization. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 245–253. IEEE.

Soham Ghosh, Anuva Agarwal, Zarana Parekh, and Alexander Hauptmann. 2019. Excl: Extractive clip localization using natural language descriptions. *arXiv preprint arXiv:1904.02755*.

Hiroaki Gomi and Mitsuo Kawato. 1993. Neural network control for a closed-loop system using feedback-error-learning. *Neural Networks*, 6(7):933–946.

Dongliang He, Xiang Zhao, Jizhou Huang, Fu Li, Xiao Liu, and Shilei Wen. 2019. Read, watch, and move: Reinforcement learning for temporally grounding natural language descriptions in videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8393–8400.

Ting-Hao Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, et al. 2016. Visual storytelling. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1233–1239.

Mitsuo Kawato. 1990. Feedback-error-learning neural network for supervised motor learning. In *Advanced neural computers*, pages 365–372. Elsevier.

Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. 2017. Dense-captioning events in videos. In *International Conference on Computer Vision (ICCV)*.

Jason Lee, Kyunghyun Cho, and Douwe Kiela. 2019. Countering language drift via visual grounding. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4376–4386.

Zhijie Lin, Zhou Zhao, Zhu Zhang, Qi Wang, and Huasheng Liu. 2020. Weakly-supervised video moment retrieval via semantic completion network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11539–11546.

Meng Liu, Xiang Wang, Liqiang Nie, Xiangnan He, Baoquan Chen, and Tat-Seng Chua. 2018. Attentive moment retrieval in videos. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, pages 15–24.

Xiang Long, Chuang Gan, and Gerard De Melo. 2018. Video captioning with multi-faceted attention. *Transactions of the Association for Computational Linguistics*, 6:173–184.

Stephanie Lukin, Reginald Hobbs, and Clare Voss. 2018. A pipeline for creative visual storytelling. In *Proceedings of the First Workshop on Storytelling*, pages 20–32, New Orleans, Louisiana. Association for Computational Linguistics.

Jonghwan Mun, Minsu Cho, and Bohyung Han. 2020. Local-global video-text interactions for temporal grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10810–10819.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Cristian Rodriguez, Edison Marrese-Taylor, Fatemeh Sadat Saleh, Hongdong Li, and Stephen Gould. 2020. Proposal-free temporal moment localization of a natural-language query in video using guided attention. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2464–2473.

Reuben Tan, Huijuan Xu, Kate Saenko, and Bryan A Plummer. 2021. Logan: Latent graph co-attention network for weakly-supervised video moment retrieval. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2083–2092.

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

Subhashini Venugopalan, Huijuan Xu, Jeff Donahue, Marcus Rohrbach, Raymond Mooney, and Kate Saenko. 2015b. Translating videos to natural language using deep recurrent neural networks. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1494–1504, Denver, Colorado. Association for Computational Linguistics.

Weining Wang, Yan Huang, and Liang Wang. 2019a. Language-driven temporal activity localization: A semantic matching reinforcement learning model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 334–343.

Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. 2019b. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4581–4591.

Huijuan Xu, Kun He, Bryan A Plummer, Leonid Sigal, Stan Sclaroff, and Kate Saenko. 2019. Multilevel language and vision integration for text-to-clip retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9062–9069.

Huijuan Xu, Subhashini Venugopalan, Vasili Ramanishka, Marcus Rohrbach, and Kate Saenko. 2015. A multi-scale multiple instance video description network. *arXiv preprint arXiv:1505.05914*.

Li Yao, Atousa Torabi, Kyunghyun Cho, Nicolas Ballas, Christopher Pal, Hugo Larochelle, and Aaron Courville. 2015. Describing videos by exploiting temporal structure. In *Proceedings of the IEEE international conference on computer vision*, pages 4507–4515.

Yitian Yuan, Tao Mei, and Wenwu Zhu. 2019. To find where you talk: Temporal sentence localization in video with attention based location regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9159–9166.

Da Zhang, Xiyang Dai, Xin Wang, Yuan-Fang Wang, and Larry S Davis. 2019. Man: Moment alignment network for natural language moment retrieval via iterative graph adjustment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1247–1257.

Luowei Zhou, Yannis Kalantidis, Xinlei Chen, Jason J Corso, and Marcus Rohrbach. 2019. Grounded video description. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6578–6587.

Luowei Zhou, Chenliang Xu, and Jason Corso. 2018a. Towards automatic learning of procedures from web instructional videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Luowei Zhou, Yingbo Zhou, Jason J Corso, Richard Socher, and Caiming Xiong. 2018b. End-to-end dense video captioning with masked transformer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8739–8748.