Open-book Video Captioning with Retrieve-Copy-Generate Network (Supplementary Material)

1. Additional Ablation Analysis

We summarize the results of training the model with different numbers of retrieved sentences on MSR-VTT in Tab.1. We have obtained a basically consistent conclusion in the second question of Sec.4.2 in the main paper. In the training phase, a moderate number (3 in MSR-VTT) of sentences are conducive to the generation; too much ones will introduce some noises instead. In addition, by comparing the results between using the fixed and jointly trained retrievers, we find that the advantage of the latter is not obvious. The reason may be that the video semantics in MSR-VTT are relatively scattered, and only a few sentences related to the video can be found, which leads to the small gain of joint training.

As shown in Tab.2, we supplement the results of our proposed model RCG using more features in two datasets. For the Video-Text Retrieval task, compared with the state-of-the-art method HGR [1], our retrieval model can achieve comparable results, which proves the effectiveness of our proposed Bi-encoders architecture based retriever. From the results of the Video-Text Retrieval and the Video Captioning, we can see that our RCG is sensitive to the retrieval model. Therefore, how to improve the performance of the retrieval model, and how to adapt it to our two-stage caption model are the next steps in the future work.

2. Additional Qualitative Examples

As mentioned in the last question in Sec.4.2 of the main paper, our proposed RCG can improve the performance of zero-shot video captioning compared with the baseline model, but the performance is not satisfactory. We make some examples to illustrate what caused the success and failure of the model.

Table 2. Performance of training the model with different numbers of retrieved sentences on MSR-VTT. The model is tested via top-15 sentences. *Fixed* denotes whether the retriever is fixed or jointly trained.

#	**# Retrievals Training**	**Fixed**	CIDEr	BLEU-4	Rouge-L	Meteor
1	1	✓	51.9	42.6	61.7	29.0
2		×	52.5	42.6	61.7	29.2
3	3	✓	52.3	**43.1**	**61.9**	29.0
4		×	**52.9**	42.8	61.7	**29.3**
5	5	✓	52.3	42.6	61.7	29.0
6		×	52.6	42.7	61.7	**29.3**
7	10	✓	51.3	42.0	61.2	29.0
8		×	51.0	42.4	61.4	28.8

As shown in Fig.1 (a) **Good Case**: although the most similar action is making an art with a needle in VATEX, the retriever helps the generator to find the similar expressions about squeezing plasticine. (b) **Bad case caused by the bad generator**: although the retriever finds the SpongeBob SquarePants accurately in MSR-VTT, the generator can not produce a caption on it. One possibility is that the generator has never seen cartoons before. (c) **Bad case caused by the bad retriever**: the retriever does not focus on the mouse in the video, causing the retrieved sentence to be irrelevant to the video, thus affecting the generator.

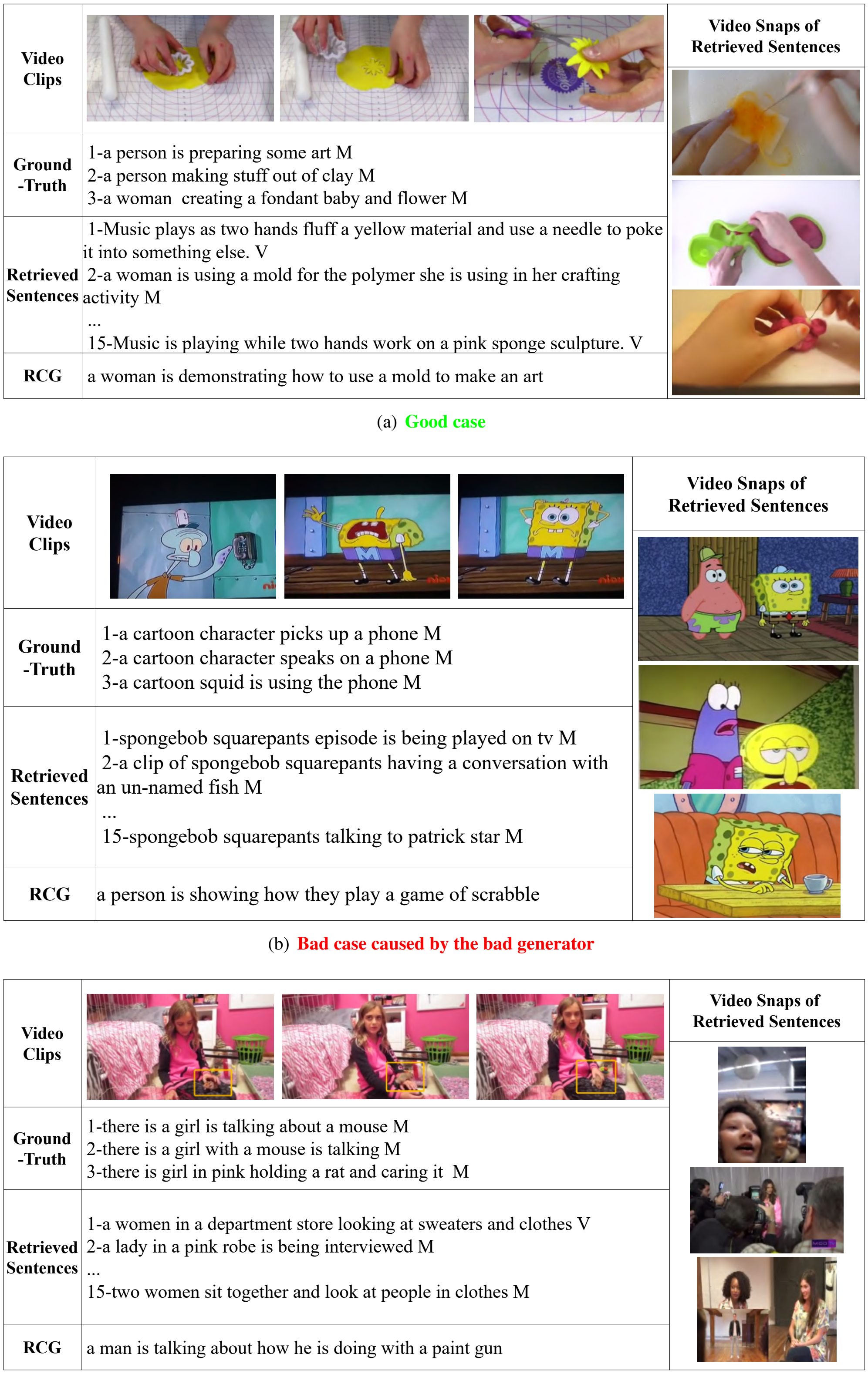
Fig.2 and Fig.3 shows additional qualitative examples from our RCG model on the VATEX dataset and MSR-VTT dataset, respectively. Although our model is sensitive to the retrieval model which may lead to some mistakes, in most cases, it leans to identify and copy video-content-relevant words from the retrieved sentences to generate a richer caption of the given video.

References

[1] Shizhe Chen, Yida Zhao, Qin Jin, and Qi Wu. Fine-grained video-text retrieval with hierarchical graph reasoning. In *CVPR 2020*, pages 10635–10644, 2020. 1

Dataset	**Methods**	**Video-Text Retrieval**						**Video Captioning**			
		R@1	R@5	R@10	MedR ↓	MnR ↓	rsum	CIDEr	BLEU-4	Rouge-L	Meteor
MSRVTT	HGR [1]	15.0	**36.7**	**48.8**	**11**	**90.4**	**172.4**	-	-	-	-
	w/o Retriever (IRV2+C3D)	-	-	-	-	-	-	49.8	42.2	61.2	28.2
	IRV2	7.7	18.8	27.3	51	325.1	97.4	50.5	42.3	61.4	28.9
	C3D	10.6	26.4	36.0	24	174.5	126.0	50.6	42.2	61.4	28.9
	IRV2+C3D	**15.8**	36.5	48.2	12	99.9	168.6	**52.3**	**43.1**	**61.9**	**29.0**
VATEX	w/o Retriever (IRV2+C3D)	-	-	-	-	-	-	48.8	31.0	48.0	21.6
	IRV2	14.6	32.6	42.9	16	210.6	165.6	51.9	31.9	48.9	22.7
	C3D	18.9	41.8	54.6	8	83.7	210.4	52.8	32.2	48.9	22.7
	IRV2+C3D	**24.5**	**48.9**	**60.5**	**6**	**69**	**244.9**	**54.6**	**32.5**	**49.2**	**23.1**

Table 1. Performance of video-text retrieval and video captioning tasks using fixed retrievers trained by more different features.



		Video Snaps of Retrieved Sentences
Video Clips		
Ground-Truth	1-a person is preparing some art M 2-a person making stuff out of clay M 3-a woman creating a fondant baby and flower M	
Retrieved Sentences	1-Music plays as two hands fluff a yellow material and use a needle to poke it into something else. V 2-a woman is using a mold for the polymer she is using in her crafting activity M ... 15-Music is playing while two hands work on a pink sponge sculpture. V	
RCG	a woman is demonstrating how to use a mold to make an art	

(a) Good case

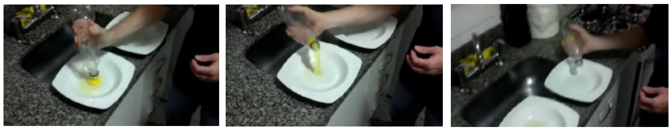
		Video Snaps of Retrieved Sentences
Video Clips		
Ground-Truth	1-a cartoon character picks up a phone M 2-a cartoon character speaks on a phone M 3-a cartoon squid is using the phone M	
Retrieved Sentences	1-spongebob squarepants episode is being played on tv M 2-a clip of spongebob squarepants having a conversation with an un-named fish M ... 15-spongebob squarepants talking to patrick star M	
RCG	a person is showing how they play a game of scrabble	

(b) Bad case caused by the bad generator

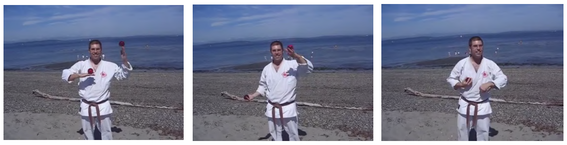
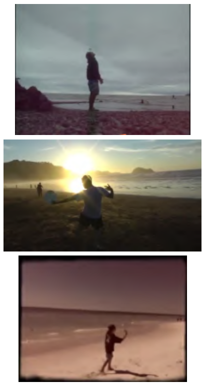
		Video Snaps of Retrieved Sentences
Video Clips		
Ground-Truth	1-there is a girl is talking about a mouse M 2-there is a girl with a mouse is talking M 3-there is girl in pink holding a rat and caring it M	
Retrieved Sentences	1-a women in a department store looking at sweaters and clothes V 2-a lady in a pink robe is being interviewed M ... 15-two women sit together and look at people in clothes M	
RCG	a man is talking about how he is doing with a paint gun	

(c) Bad case caused by the bad retriever

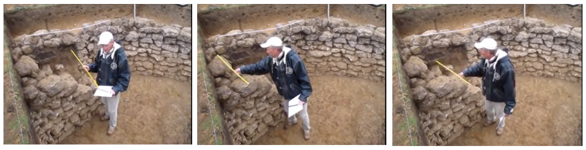
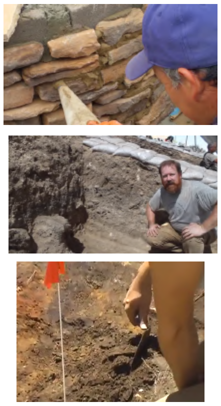
Figure 1. Examples of **Zero-shot Video Captioning**, M and V denote the sources of sentences are MSR-VTT and VATEX. The whole model is pre-trained on VATEX training set, then the model is tested on MSR-VTT test set as zero-shot videos. Since the model has never seen these videos before, the performance is very poor, so the retriever pre-trained on two datasets is plugged to replace the original retriever to achieve the retrieval of content-related sentences. This realizes the transformation of knowledge and the expansion of model memory.

Video Clips		**Video Snaps of Retrieved Sentences**
Ground -Truth	1-A guy uses a water bottle to suck up a yoke from a cracked egg and move it to another plate. 2-a man using a plastic bottle to separate the yolk of an egg from the white 3-A man transfers an egg yoke in a plastic bottle from one plate to another.	
Retrieved Sentences	1-Person shows how to remove an yolk from from cracked egg with plastic water bottle. 2-A man is separating an egg yolk from an egg white with a water bottle. ... 15-Someone explains how to remove the yolk from an egg with a plastic bottle in another language.	
RCG	a man demonstrates how to separate an egg yolk from an egg using a plastic bottle	

(a) Good case

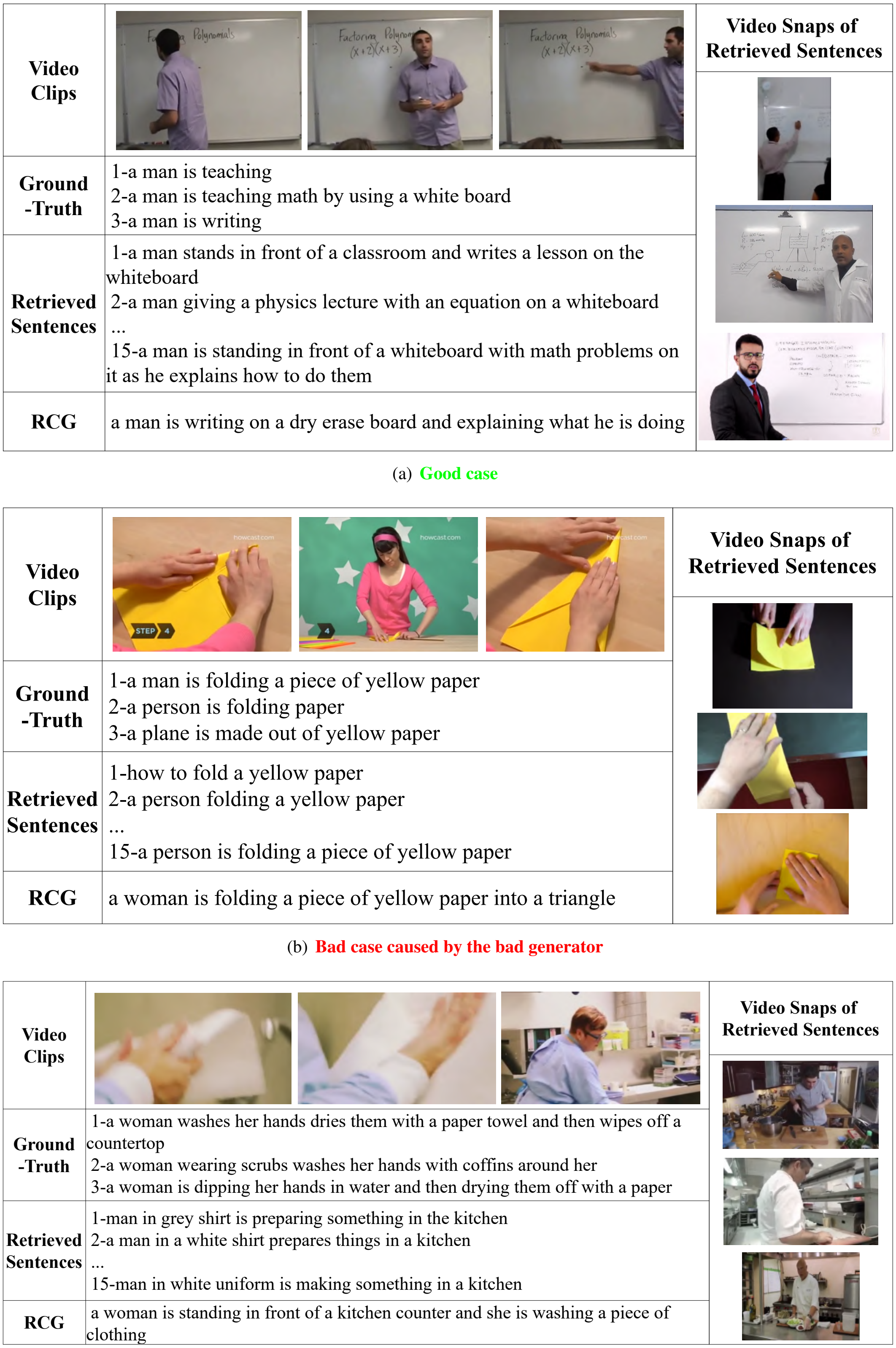
Video Clips		**Video Snaps of Retrieved Sentences**
Ground -Truth	1-A man in a karate outfit twirls around two red balls on his hands. 2-A man wearing traditional karate clothing is balancing two red balls he has possession of. 3-A man contact juggling two balls while standing on the beach.	
Retrieved Sentences	1-A male is at the beach and doing tricks with a bottle on his forehead. 2-A man on the beach rolling a frisbee across his arms. ... 15-u'A boy is throwing a ball at the beach against the wind and it lands in his friend\\'s hand.'	
RCG	a man is standing on a beach and he is throwing a frisbee	

(b) Bad case caused by the bad generator

Video Clips		**Video Snaps of Retrieved Sentences**
Ground -Truth	1-A man reviews some information in a notebook and begins to measure a wall of stones in front of him. 2-A man holding a notepad, is measuring a section of a stone wall. 3-A man is standing in a ruin measuring the width of part of the structure.	
Retrieved Sentences	1-A guy in a hat is filling in some rocks on a wall. 2-A man is kneeling in an outdoor trench at an archaeological site, discusses the history of the site and asks his colleague for a bucket. ... 15-someone is using a tool to move dirt around in an archaeological dig	
RCG	a man is standing in front of a brick wall and he is using a tool to smooth the edges of a wall	

(c) Bad case caused by the bad retriever

Figure 2. More examples of proposed RCG method on VATEX dataset.



Video Clips		**Video Snaps of Retrieved Sentences**
Ground-Truth	1-a man is teaching 2-a man is teaching math by using a white board 3-a man is writing	
Retrieved Sentences	1-a man stands in front of a classroom and writes a lesson on the whiteboard 2-a man giving a physics lecture with an equation on a whiteboard ... 15-a man is standing in front of a whiteboard with math problems on it as he explains how to do them	
RCG	a man is writing on a dry erase board and explaining what he is doing	

(a) **Good case**

Video Clips		**Video Snaps of Retrieved Sentences**
Ground-Truth	1-a man is folding a piece of yellow paper 2-a person is folding paper 3-a plane is made out of yellow paper	
Retrieved Sentences	1-how to fold a yellow paper 2-a person folding a yellow paper ... 15-a person is folding a piece of yellow paper	
RCG	a woman is folding a piece of yellow paper into a triangle	

(b) **Bad case caused by the bad generator**

Video Clips		**Video Snaps of Retrieved Sentences**
Ground-Truth	1-a woman washes her hands dries them with a paper towel and then wipes off a countertop 2-a woman wearing scrubs washes her hands with coffins around her 3-a woman is dipping her hands in water and then drying them off with a paper	
Retrieved Sentences	1-man in grey shirt is preparing something in the kitchen 2-a man in a white shirt prepares things in a kitchen ... 15-man in white uniform is making something in a kitchen	
RCG	a woman is standing in front of a kitchen counter and she is washing a piece of clothing	

(c) **Bad case caused by the bad retriever**

Figure 3. Examples of proposed RCG method on MSR-VTT dataset.