Video and Text Matching with Conditioned Embeddings

Paper ID 544

Supplementary data

A Implementation details

Embeddings: We used a pretrained CNN to represent the video's frames, sampled at 16 frames per second, discarding any frame after the first 80. For ActivityNet and DiDeMo, we used TSN-Inception V3 [17] pretrained on the Kinetics dataset [6]. For YouCook2 and Vatex we used ResNet101 [4] pretrained on ImageNet. For both CNN architectures, we extracted the activations of the penultimate. For textual features, we used GloVe [13] pretrained on 840B common web-crawl with a dimension of 300.
Setup: We set the model hidden dimension d_e to 512 and $\alpha_1 = \alpha_2 = 0.2$. For each batch, we randomly pick 16 videos. We used the Adam optimizer [7] and set the learning rate to 10^{-3}. We early-stopped our training, if the validation-set did not improve for five consecutive epochs. For inference, we used a shortlist size of 50.

B Additional quantitative results

We report additional metrics for the video/sentence retrieval task. In Tab. 1, we report R@50 for ActivityNet, where we improve by 4%. We also improve the R@50 on LSMDC by 6%. Additionally, we show, in Tab. 2, a significant improvement of 10.4% in R@10 for MSR-VTT.

We also report additional results for the paragraph/video retrieval. In Tab. 3, we add results on ActivityNet Val2 set. As with the Val1 split, our method significantly outperforms HSE by 15.5% (48.1% *vs*. 32.6%). In Tab. 4, we add results for R@50 for ActivityNet, DiDeMo, YouCook2, Vatex, MSR-VTT. Our method consistently outperforms previous baselines.

C Additional qualitative results

In the following, we sketch additional frames and word relevance scores (See Eq. 4). Results are based on the video-paragraph variant and show a randomly picked paragraph sentence and the matched video clip. This is repeated for the various datasets we have used (*e.g.*, ActivityNet, DiDeMo, YouCook2, Vatex, MSR-VTT). All the samples are taken from the validation set.

Dataset	Method	Text2Vid		Vid2Text	
		R@50	MdR	R@50	MdR
ActivityNet	Random	1.02%	2458.0	1.02%	2458.0
	FSE [20]	89.1%	7.0	88.4%	7.0
	HSE [20]	-	-	-	-
	CE [10]	91.4%	6.0	90.9%	6.0
	MMT [3]	93.2%	5.0	93.1%	4.3
	clipBERT [9]	-	-	-	-
	Ours	**97.0%**	**3.0**	**97.7%**	**3.0**
LSMDC	HT100 [11]	-	40.0	-	-
	JFusion [19]	-	36.0	-	-
	HSE [20]	35.5%	23.0	34.3%	23.0
	MMT [3]	38.8%	21.0	38.9%	20.0
	Ours	**44.5%**	**19.0**	**45.4%**	**18.0**

Table 1: Video/Sentence retrieval on ActivityNet [8] and LSMDC [14],. Shown is the recall at a certain number of retrievals and MdR=median rank.

Dataset	Method	Text2Vid		Vid2Text	
		R@10	MdR	R@10	MdR
MSR-VTT	Dual Encoding [2]	-	32.0	-	15.0
	HGR [1]	-	24.0	-	11.0
	JSFusion [19]	-	13.0	-	-
	Cues [12]	-	16.0	-	38.0
	HSE [20]	58.5%	23.0	58.8%	23.0
	MMT [3]	63.9%	4.0	63.2%	6.0
	clipBERT [9]	57.7%	-	-	-
	Ours	**68.1%**	**3.0**	**68.3%**	**3.0**

Table 2: Video/Sentence retrieval on MSR-VTT [18]. Shown is the recall at a certain number of retrievals and MdR=median rank.

Each sketch contains the most sentence-relevant frames (on the right) for each clip, along with the two frames that got the lowest score (on the left). We also highlight the video-relevant words in the queries via a word heat map.

We observe that the frames relevance score focuses on frames that contain related objects to query. For instance, the query mentions saxophone in the first clip, and the saxophone is easily shown in the frames that got high scores. Also, the word relevance score picks up unique words. Our interaction-conditioned approach employs these scores and uses only the

Dataset	Method	Vid2Text				Text2Vid			
		R@1	R@5	R@50	MR	R@1	R@5	R@50	MR
ActivityNet Val2	HSE [20]	32.6%	62.8%	93.5.0%	3.0	33.2%	62.9%	93.6%	3.0
	Ours	**48.1%**	**76.2%**	**100.0%**	**2.0**	**48.6%**	**77.0%**	**100.0%**	**2.0**

Table 3: Video/Paragraph retrieval on ActivityNet Val2 [18]. Shown is the recall at a certain number of retrievals and MdR=median rank.

Dataset	Method	Text2Vid		Vid2Text	
		R@50	MdR	R@50	MdR
ActivityNet	LSTM-YT [15]	24.0%	102.0	38.0%	98.0
	No Context [16]	32.0%	78.0	45.0%	56.0
	Dense full [8]	65.0%	34.0	74.0%	32.0
	FSE [20]	89.1%	7.0	88.4%	7.3
	HSE [20]	97.1%	2.0	97.0%	2.0
	Ours	**100.0%**	**1.0**	**100.0%**	**1.0**
DiDeMo	S2VT [16]	76.5%	13.0	76.5%	15.0
	FSE [20]	78.9%	11.0	78.0%	12.0
	HSE [20]	92.4%	3.3	92.1%	3.0
	Ours	**100.0%**	**2.7**	**100.0%**	**2.4**
YouCook2	HSE [20]	99.8%	2.0	100.0%	2.0
	Ours	**100.0%**	**1.0**	**100.0%**	**1.0**
MSR-VTT	HSE [20]	90.1%	3.0	89.2%	3.0
	Ours	**97.3%**	**2.0**	**97.2%**	**2.0**

Table 4: Video/Paragraph retrieval on ActivityNet [8], DiDeMo [5] , YouCook2 [21] and MSR-VTT [18]. Shown is the recall at a certain number of retrievals and MdR=median rank.

relevant frames and words for the conditioned representation.

A musician plays saxophone on front people sitting in a room.

(DiDeMo)

A marching band is shown playing various instruments down the street followed by more police cars riding behind.

(DiDeMo)

Outside in the water a man is kayaking during on some heavy waves.

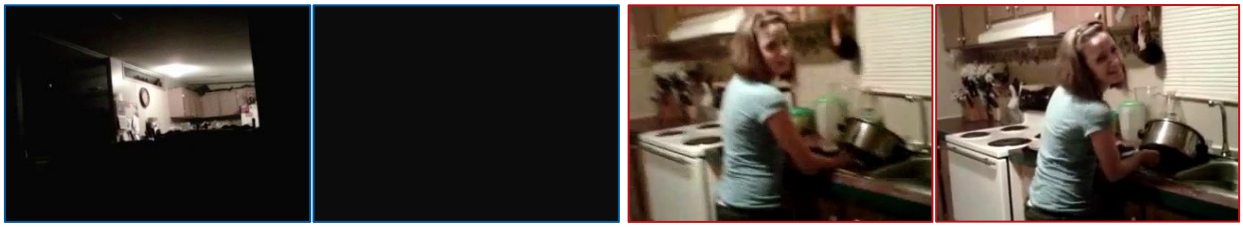
(DiDeMo)

A person holds their phone as they walk in a field near a river bank and films the river as it runs.

(DiDeMo)

The man then swings the axe on several pieces of wood while looking back to the camera.

(DiDeMo)



A camera pans around a room and leads into a woman washing dishes in a sink.

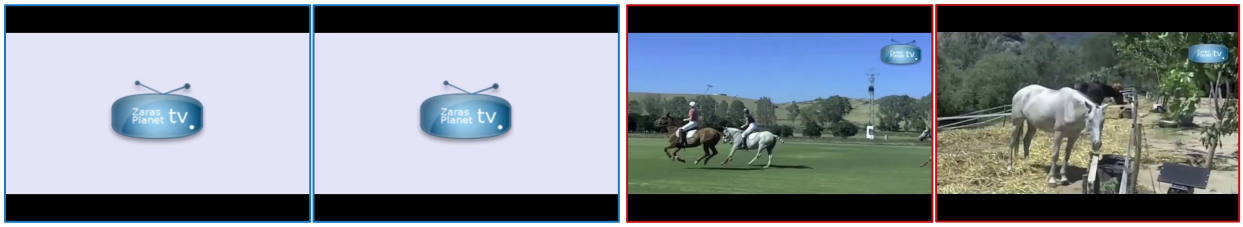
(DiDeMo)

They finish the game and someone receives a trophy.

(DiDeMo)

The man in a costume rides the unicycle and plays a bagpipe.

(DiDeMo)



More clips are shown of people interacting with horses as well as playing games and riding the horses.

(ActivityNet)

A man shows off a huge sandcastle on a beach while a little girl works on the castle.

(ActivityNet)

The person lays down bread and begins pouring the ingredients out onto the bread.

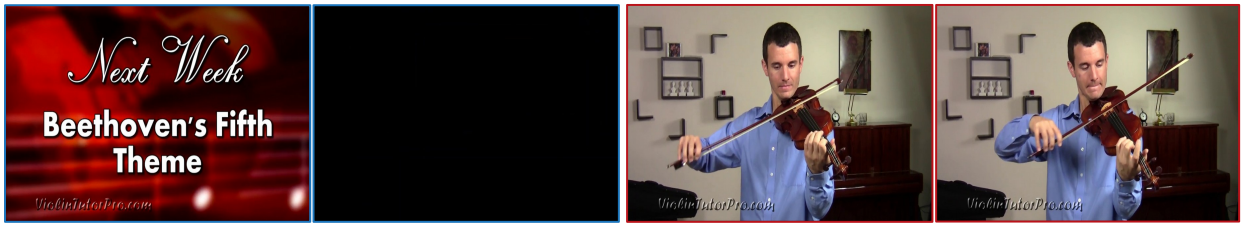
(ActivityNet)

another people is making gymnastics in the roofed gym and making a hand stands.

(ActivityNet)

The man throws Frisbee discs for his dog to catch at a baseball field.

(ActivityNet)



A young man holds a violin, then he plays violin.

(ActivityNet)

A man stands against a wall playing guitar and singing.

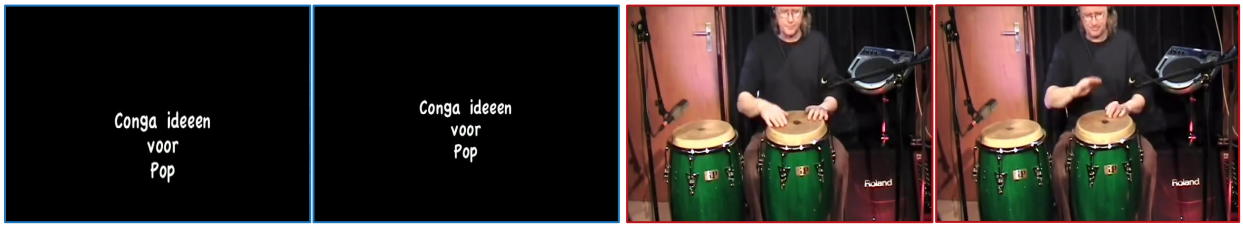
(ActivityNet)

A man in a canoe appears, paddling through the rapids.

(ActivityNet)

A man is talking while holding a volleyball.

(ActivityNet)

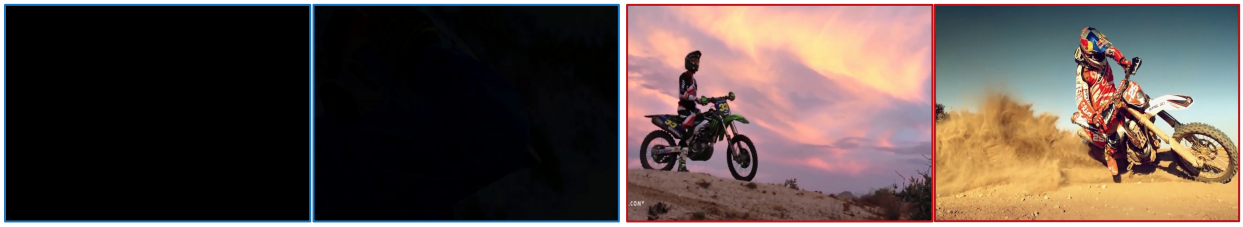


A man is shown seated, banging on a pair of bongo drums.

(ActivityNet)

woman wering a chef suit is standing in front of a kitchen serving a big omelette into a white plate and it frying other omellete.

(ActivityNet)

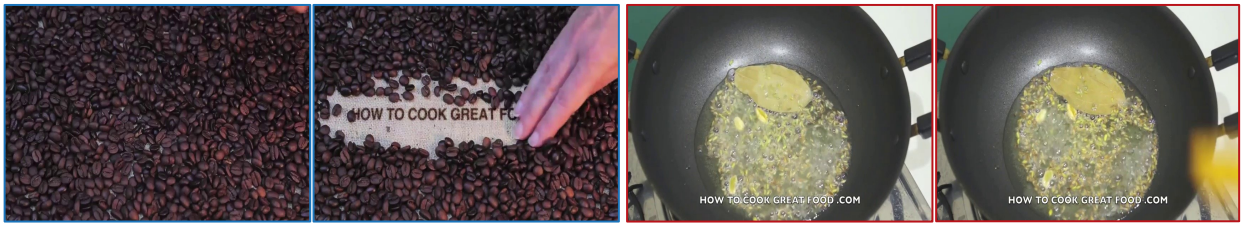


Various shots are shown of a motocross driver sitting in different landscapes as well as riding around.

(ActivityNet)

She shows a panini sandwich on the screen.

(ActivityNet)



add garram masala seeds and a bay leaf to the oil

(YouCook2)

add the lamb to the pot

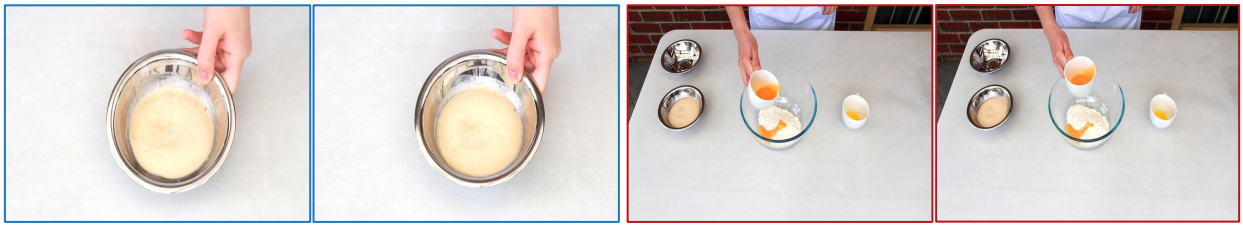
(YouCook2)

cut the tuna into smaller sections

(YouCook2)

mix the yeast sugar and water

(YouCook2)

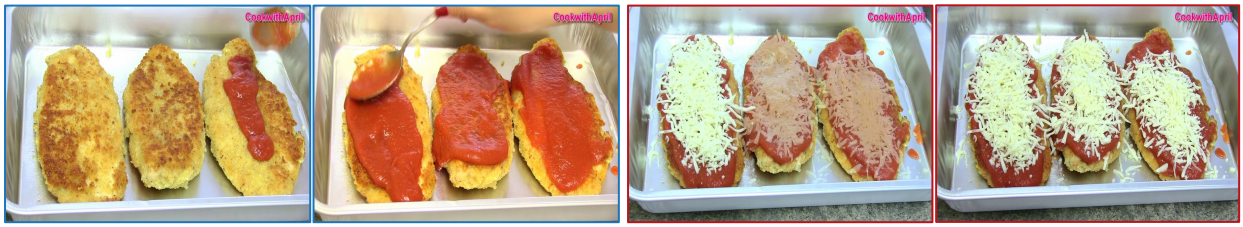


mix the flour salt egg melted butter and yeast mixture into a dough and knead

(YouCook2)

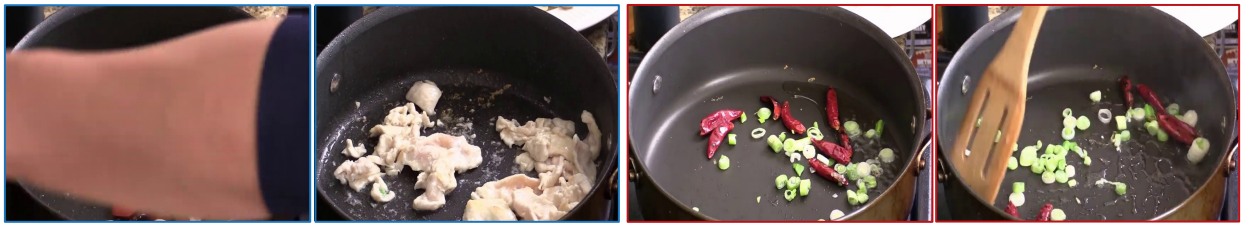
pound the chicken

(YouCook2)



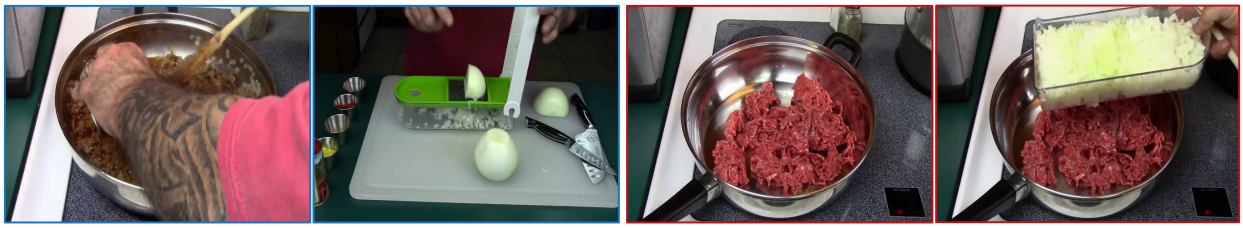
top the chicken with mozzarella cheese

(YouCook2)



heat a pan pour some vegetable oil and spread it all over the pan

(YouCook2)



in pan take 1 pound ground beef add 1 finely chopped onions 2 cloves finely minced garlic cook

(YouCook2)

combine salt butter milk and hot sauce in a bowl

(YouCook2)

a man repelling down a side of a mountain covered in snow

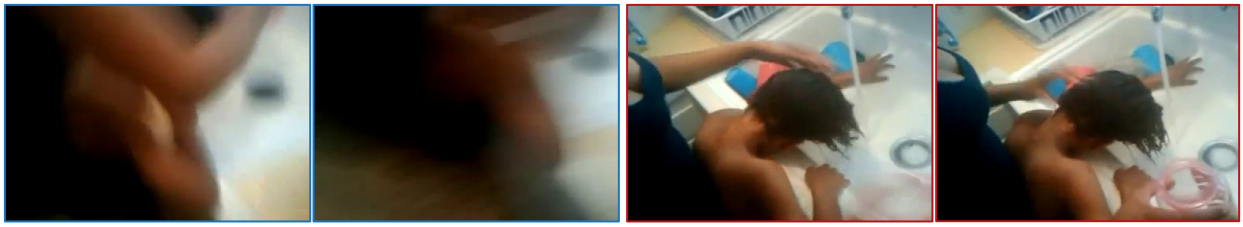
(Vatex)

A person in blue pants uses a rope to repell down a rocky ridge

(Vatex)

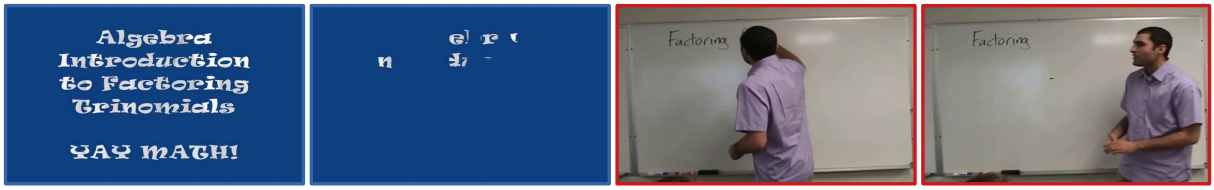
A man laughs and talks while a person in a hoodie cuts his hair with scissors and comb

(Vatex)



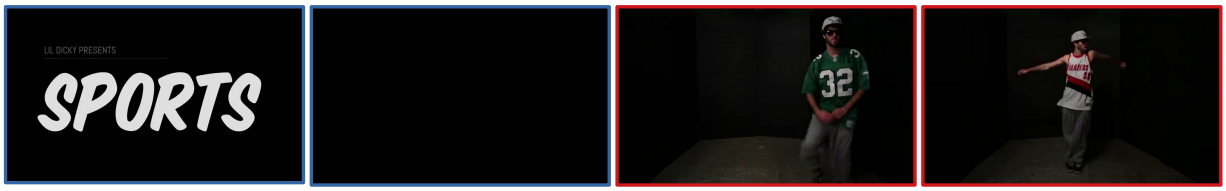
A child is leaning over a sink getting their hair washed by a woman

(Vatex)



A man is teaching math by using a white board

(MSR-VTT)



A man dances in a room and flips through photos

(MSR-VTT)

A baseball player hits a ball to the back of the field

(MSR-VTT)

A guy talks about the features of the jeep cherokee

(MSR-VTT)

a man is singing and walking through the street

(MSR-VTT)

References

[1] Shizhe Chen et al. "Fine-grained Video-Text Retrieval with Hierarchical Graph Reasoning". In: *CVPR*. 2020.

[2] Jianfeng Dong et al. "Dual encoding for zero-example video retrieval". In: *CVPR*. 2019.

[3] Valentin Gabeur et al. "Multi-modal Transformer for Video Retrieval". In: *ECCV* (2020).

[4] Kaiming He et al. "Deep residual learning for image recognition". In: *ICCV*. 2016.

[5] Lisa Anne Hendricks et al. "Localizing moments in video with natural language". In: *ICCV*. 2017.

[6] Will Kay et al. "The kinetics human action video dataset". In: *arXiv preprint arXiv:1705.06950* (2017).

[7] Diederik P Kingma and Jimmy Ba. "Adam: A method for stochastic optimization". In: *ICLR* (2015).

[8] Ranjay Krishna et al. "Dense-captioning events in videos". In: *ICCV*. 2017.

[9] Jie Lei et al. "Less is More: ClipBERT for Video-and-Language Learning via Sparse Sampling". In: *CVPR* (2021).

[10] Yang Liu et al. "Use what you have: Video retrieval using representations from collaborative experts". In: *BMVC* (2019).

[11] Antoine Miech et al. "Howto100m: Learning a text-video embedding by watching hundred million narrated video clips". In: *ICCV*. 2019.

[12] Niluthpol Chowdhury Mithun et al. "Learning joint embedding with multimodal cues for cross-modal video-text retrieval". In: *ICMR*. 2018.

[13] Jeffrey Pennington, Richard Socher, and Christopher D Manning. "Glove: Global vectors for word representation". In: *EMNLP*. 2014.

[14] Anna Rohrbach et al. "A dataset for movie description". In: *CVPR*. 2015.

[15] Subhashini Venugopalan et al. "Sequence to sequence-video to text". In: *ICCV*. 2015.

[16] Subhashini Venugopalan et al. "Translating videos to natural language using deep recurrent neural networks". In: *ACL* (2015).

[17] Limin Wang et al. "Temporal segment networks for action recognition in videos". In: *TPAMI* (2018).

[18] Jun Xu et al. "Msr-vtt: A large video description dataset for bridging video and language". In: *CVPR*. 2016.

[19] Youngjae Yu, Jongseok Kim, and Gunhee Kim. "A joint sequence fusion model for video question answering and retrieval". In: *ECCV*. 2018.

[20] Bowen Zhang, Hexiang Hu, and Fei Sha. "Cross-modal and hierarchical modeling of video and text". In: *ECCV*. 2018.

[21] Luowei Zhou, Chenliang Xu, and Jason J Corso. "Towards Automatic Learning of Procedures From Web Instructional Videos". In: *AAAI*. 2018.