Spoken Moments: Learning Joint Audio-Visual Representations from Video Descriptions (Supplementary Material)

Mathew Monfort*
MIT
mmonfort@mit.edu

SouYoung Jin*
MIT
souyoung@mit.edu

Alexander Liu
MIT
alexhliu@mit.edu

David Harwath
UT Austin
harwath@cs.utexas.edu

Rogerio Feris
IBM Research
rsferis@us.ibm.com

James Glass
MIT
glass@csail.mit.edu

Aude Oliva
MIT
oliva@mit.edu

A. Annotation

We follow the approach used to collect the Places Audio Caption dataset [2, 1] and collect audio descriptions of each video in the dataset using Amazon Mechanical Turk (AMT). In order to ensure that we have a large and diverse dataset, we collect an audio description using AMT for each video in a set of 500k randomly selected videos from the training set and at least two unique descriptions for each video in the 10k videos used for both the validation and test sets. Each AMT worker is presented with a task of recording themselves describing 10 different videos. Each video is shown on the left of the screen while a video with an example text description is shown on the right. This example helps to show the workers the types of descriptions we are looking for and the amount of detail we expect from them. This example stays on the right side of the screen throughout the task while the target videos on the left cycle as the worker completes each description. Figure S1 shows an example of this interface with an example video and caption on the right and a target video on the left. Below each target description is a button that allows the worker to start recording their voice as they describe the video. Once they press this button, the video is removed from the screen and the recording is started. We block the worker from seeing the video while recording the description to ensure that the recordings are concise and pertain only to the important events highlighted in their memory. We use the Google Cloud ASR engine to verify the quality of each recorded description and flag AMT workers for poor performance. This is done by checking that the generated text has at least five words, is unique (some bots repeat pre-recorded audio to trick the system) and that the audio is at least three seconds long. If any of these checks fail we don't let the worker continue to the next video until they record a new description that passes our checks. Once the descriptions are recorded, we periodically sample videos to check the quality of the audio paired with the ASR to ensure they match the videos and have an appropriate level of detail. If these checks fail, we flag the workers that recorded the descriptions, don't allow them to record in the future and recheck all of their recorded data. This process allows us to ensure a strong level of quality in our collected spoken captions. Examples of some of the videos and corresponding text transcriptions of the descriptions we collected can be seen in Figure 1.

*equal contribution

B. Implementation Details

We train each model on a server with 8 24GB Titan RTX cards using a mini-batch size of 2048 for 100 epochs. We examine the effect of the mini-batch size on learning in the next section. We take the best parameters as evaluated on the evaluation set of the training dataset after each epoch. We repeat this process for two phases of training. First we freeze the visual backbone models and train only the projection heads (including the full caption model for the spoken models) and then, in a second round, allow the full visual model to train as well. We keep the language and ASR components frozen for the language caption models and reserve fine-tuning these components for future work. For model training, we use an Adam [4] optimizer where a fixed learning rate of 0.001 and 0.00001 are set for the first and the second round model training, respectively.

C. Ablation Studies

In Tables S1, S2, S3, S4, and S5, we show several ablation studies. Unless otherwise listed in the table we use the proposed AMM loss function with the BART [7] language model as part of the language caption model described in Section 4.2.1 for each experiment. Results are averaged over five rounds with a single random batch of 1k caption-

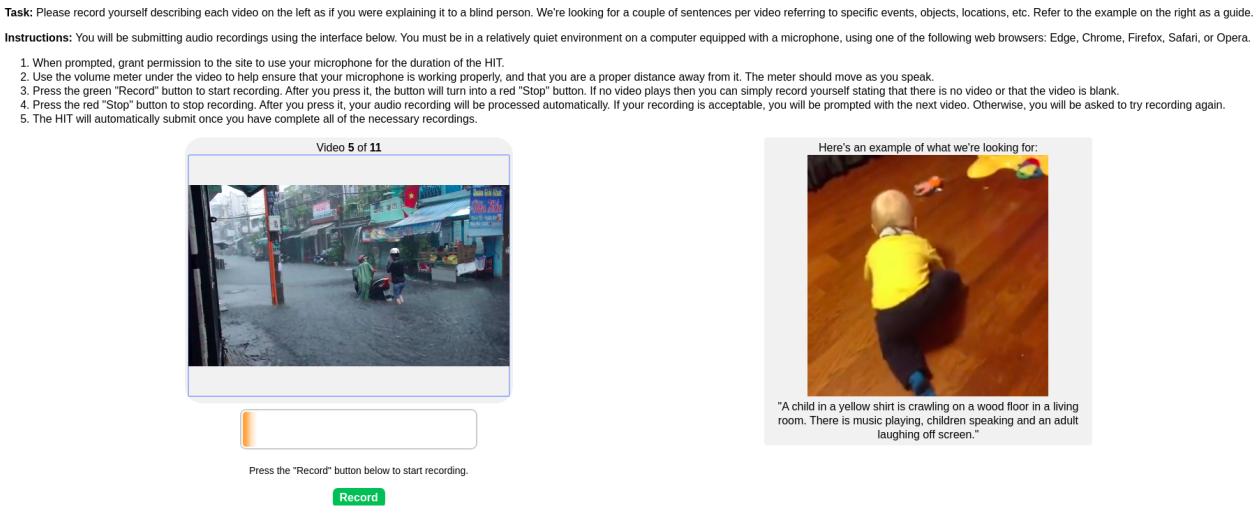


Figure S1: **Spoken Caption Collection:** Target videos for which descriptions are collected on the left and a video with an example text description is always visible on the right.

video pairs from the test set. Due to the increased computation demand of these studies we freeze the base models and train the projection heads for alignment. We use the best model settings found in this analysis to train the full models with results reported in Section 5.

Table S1 shows the effect of using two different pretrained temporal shift [8] video models on four different datasets in order to choose the most appropriate base models (Multi-Moments in Time (M-MiT) [9] or Kinetics [3]). Here we use the BART language model and the proposed AMM loss function as described in Section 4 as this combination gave us the best results on each dataset.

Table S2 compares the effect of the video model (TSM) trained for action recognition and the 2D model trained for object recognition. Most captions reference both objects and actions in a video with an average of 4.37 nouns used per caption compared to 1.58 verbs. The strength of the 2D obect model makes sense when we consider this prevalence of nouns in the captions. The combination of the TSM model trained on M-MiT [9] and the 2D models trained on ImageNet [6] provided the best performance when used with the model described in Section 4.

In Tables S3 and S4 we compare the the effect of the batch size and projection size on the performance of the S-MiT model described in Section 4 in order to validate our choice of a 2048 batch and a 4096 projection. Similarly, Table S5 shows the effect of using the caption sampling approach for the transcription model as described in Section 4.2.1. In Table S6, we explore different dampening parameters.

D. Cross Dataset Generalization

In Table S7, we expand on Table 4 and compare the generalization performance of models trained on four different datasets (S-MiT as well as Vatex-en [10], MSR-VTT [11] and ActivityNet Captions [5]) for video/caption retrieval on their full test sets. In Table 4 we ran the comparison on five samples of 1k video-caption pairs to be consistent on evaluating across different size test sets. Here we evaluate on the full test set of each dataset to provide a baseline for each test set. The strength of the model trained on S-MiT is even more evident here as it achieves higher results on the test sets of both ActivityNet and MSR-VTT than the models trained on those datasets. It even comes very close to the performance of the Vatex model on the Vatex test set. This shows that the scale and diversity of the S-MiT dataset is highly beneficial to training robust models.

E. Qualitative Results

In Tables S8 and S9, we show the top five retrieval results for some examples from the Spoken Moments dataset. For this analysis, we use the language caption model described in Section 4.2.1 with the BART [7] language model and the proposed AMM loss function. Table S8 shows the top five retrieved captions given a query video, while Table S9 shows the top five retrieved videos given a query caption. Blue boxes indicate the ground-truth results.

Our model retrieves results by recognizing key objects or environments in the videos. For example, in Table S8 (c), *lettuce* is distinguished from the other vegetables. In Table S9 (f), the model not only *recognizes the planets* in space but also *understands that they are crashing into each other*. Some of the examples show that the top retrieval re-

sult is not the ground-truth. However, as we can see, the top predictions are typically still a strong match for the queries, as in (e), (i) in Table S8 and (a), (b) in Table S9.

For this demonstration, we use transcribed words from the audio captions using a pretrained ASR model. Noise in these transcriptions may contribute to some errors. In the future, we plan to investigate jointly training a pre-trained ASR model, and language model, with the video model to improve our performance.

F. Captions in the Spoken Moments Dataset

Table S10 shows some captions in the Spoken Moments dataset that capture motion and sequential events which would be difficult to represent with a single image.

References

[1] David Harwath, Adria Recasens, Didac Suris, Galen Chuang, Antonio Torralba, and James Glass. Jointly discovering visual objects and spoken words from raw sensory input. *Int. J. Comput. Vis.*, (128):620–641, 2020. 1

[2] David Harwath, Antonio Torralba, and James Glass. Unsupervised learning of spoken language with visual context. In *Adv. Neural Inform. Process. Syst.*, 2016. 1

[3] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 2

[4] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 1

[5] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *Int. Conf. Comput. Vis.*, 2017. 2, 4

[6] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Adv. Neural Inform. Process. Syst.*, 25:1097–1105, 2012. 2

[7] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*, 2019. 1, 2

[8] Ji Lin, Chuang Gan, and Song Han. Tsm: Temporal shift module for efficient video understanding. In *Int. Conf. Comput. Vis.*, October 2019. 2

[9] Mathew Monfort, Kandan Ramakrishnan, Alex Andonian, Barry A McNamara, Alex Lascelles, Bowen Pan, Quanfu Fan, Dan Gutfreund, Rogerio Feris, and Aude Oliva. Multi-moments in time: Learning and interpreting models for multi-action video understanding. *arXiv preprint arXiv:1911.00232*, 2019. 2

[10] 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 *Int. Conf. Comput. Vis.*, October 2019. 2, 4

[11] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. IEEE Conf. Comput. Vis. Pattern Recog., June 2016. 2, 4

Dataset	Pretrained TSM Dataset	Caption to Video				Video to Caption				Mean			
		R@1	R@5	R@10	mAP	R@1	R@5	R@10	mAP	R@1	R@5	R@10	mAP
Vatex [10]	**Kinetics**	39.6±1.0	77.5±1.5	87.2±1.0	55.9±0.8	46.4±0.6	**82.1**±1.0	**90.2**±1.2	**61.9**±0.6	43.0±0.7	79.8±1.1	**88.7**±1.0	58.9±0.7
	S-MiT	**47.4**±1.1	**81.5**±0.7	**89.0**±1.1	**62.3**±0.6	43.1±0.9	78.3±0.6	86.2±0.3	58.5±0.5	**45.3**±0.8	**79.9**±0.4	87.6±0.6	**60.4**±0.5
ActivityNet [5]	**Kinetics**	**18.7**±1.0	**45.6**±0.9	57.2±1.4	**31.0**±0.7	**20.8**±0.8	**50.1**±1.4	**61.8**±1.3	**34.1**±0.4	**19.8**±0.8	**47.8**±0.9	**59.5**±1.0	**32.5**±0.4
	M-MiT	16.1±1.7	44.0±1.0	**57.5**±1.7	29.3±1.0	19.0±1.3	48.2±0.9	61.0±1.1	32.5±0.8	17.6±1.5	46.1±0.7	59.2±1.4	30.9±0.8
MSR-VTT [11]	**Kinetics**	17.6±1.3	48.9±1.8	65.6±1.2	**31.6**±1.3	25.5±0.7	59.7±1.8	74.1±1.6	40.6±0.9	21.6±0.8	54.3±1.4	69.8±1.4	36.1±0.9
	M-MiT	**20.7**±0.5	**54.2**±0.9	**70.6**±1.0	30.5±0.4	**31.3**±1.1	**61.0**±1.0	**75.0**±0.9	**40.9**±0.8	**24.0**±0.6	**57.6**±0.6	**72.8**±0.8	**37.7**±0.4
S-MiT	**Kinetics**	27.6±1.4	57.5±2.4	70.4±1.9	41.3±1.7	37.2±2.3	65.0±1.7	75.2±1.5	50.0±1.7	32.4±1.8	61.3±2.0	72.8±1.6	45.7±1.7
	M-MiT	**29.8**±2.5	**60.6**±2.4	**72.2**±1.9	**44.0**±2.2	**39.4**±2.1	**68.0**±2.0	**77.5**±1.8	**52.3**±2.0	**34.6**±2.1	**64.3**±2.2	**74.9**±1.8	**48.2**±2.0

Table S1: **Comparison of different Pretrained TSM models on multiple datasets using AMM and Bart**

Visual Base Model	Caption to Video				Video to Caption				Mean			
	R@1	R@5	R@10	mAP	R@1	R@5	R@10	mAP	R@1	R@5	R@10	mAP
TSM Kinetics	20.2±1.1	47.9±2.3	61.0±0.8	33.2±1.1	28.2±1.5	54.9±1.5	67.1±1.6	40.8±1.6	24.2±1.1	51.4±1.9	64.0±1.0	37.0±1.3
TSM M-MiT	19.7±1.1	48.6±2.0	61.9±1.6	33.5±1.3	28.4±1.4	58.0±2.5	69.2±1.9	41.9±1.4	24.1±1.2	53.3±2.1	65.6±1.7	37.7±1.4
ResNet-152 ImageNet (2D)	24.2±2.4	53.6±1.8	66.5±2.1	37.9±2.0	32.9±2.1	61.7±1.6	71.6±1.0	45.9±1.8	28.5±2.2	57.7±1.7	69.1±1.5	41.9±1.9
TSM Kinetics + 2D	27.6±1.4	57.5±2.4	70.4±1.9	41.3±1.7	37.2±2.3	65.0±1.7	75.2±1.5	50.0±1.7	32.4±1.8	61.3±2.0	72.8±1.6	45.7±1.7
TSM M-MiT + 2D	**29.8**±2.5	**60.6**±2.4	**72.2**±1.9	**44.0**±2.2	**39.4**±2.1	**68.0**±2.0	**77.5**±1.8	**52.3**±2.0	**34.6**±2.1	**64.3**±2.2	**74.9**±1.8	**48.2**±2.0

Table S2: **Comparison of different visual base model combinations on S-MiT using AMM and Bart**

Batch Size	Caption to Video				Video to Caption				Mean			
	R@1	R@5	R@10	mAP	R@1	R@5	R@10	mAP	R@1	R@5	R@10	mAP
512	27.2±1.6	57.4±1.3	69.4±1.0	41.0±1.5	35.5±2.5	64.0±1.4	74.4±1.1	48.4±2.1	31.3±1.9	60.7±1.3	71.9±1.0	44.7±1.7
1024	27.8±2.0	57.7±1.4	69.8±1.2	41.5±1.9	36.5±2.8	65.6±1.4	75.2±1.7	49.7±2.0	32.2±2.3	61.7±1.4	72.5±1.3	45.6±1.9
2048	**29.8**±2.5	**60.6**±2.4	**72.2**±1.9	**44.0**±2.2	**39.4**±2.1	**68.0**±2.0	**77.5**±1.8	**52.3**±2.0	**34.6**±2.1	**64.3**±2.2	**74.9**±1.8	**48.2**±2.0
4096	29.2±2.7	58.4±1.6	70.8±1.9	42.8±2.3	**39.4**±2.3	66.6±1.8	75.7±1.4	51.8±1.9	34.3±2.3	62.5±1.6	73.3±1.6	47.3±2.0

Table S3: **Comparison of different batch sizes on S-MiT using AMM and Bart**

Projection Size	Caption to Video				Video to Caption				Mean			
	R@1	R@5	R@10	mAP	R@1	R@5	R@10	mAP	R@1	R@5	R@10	mAP
1024	27.4±1.8	56.6±1.6	69.5±0.9	41.1±1.5	38.6±1.6	66.6±1.1	76.3±1.3	51.3±1.2	33.0±1.6	61.6±1.3	72.9±1.0	46.2±1.3
2048	27.8±1.8	57.4±2.0	69.2±1.5	41.5±1.8	38.4±2.1	65.9±1.4	75.6±1.5	51.1±1.6	33.1±1.9	61.6±1.6	72.4±1.4	46.3±1.7
4096	**29.8**±2.5	**60.6**±2.4	**72.2**±1.9	**44.0**±2.2	**39.4**±2.1	**68.0**±2.0	**77.5**±1.8	**52.3**±2.0	**34.6**±2.1	**64.3**±2.2	**74.9**±1.8	**48.2**±2.0
8192	29.4±2.0	58.0±2.3	70.3±1.2	42.6±1.8	38.5±2.4	66.1±2.1	76.1±1.5	51.2±2.1	33.9±2.2	62.0±2.2	73.2±1.3	46.9±1.9

Table S4: **Comparison of different projection sizes on S-MiT using AMM and Bart**

Sampling	Caption to Video				Video to Caption				Mean			
	R@1	R@5	R@10	mAP	R@1	R@5	R@10	mAP	R@1	R@5	R@10	mAP
N	28.1±1.1	57.5±2.0	69.8±1.4	41.8±1.3	39.1±1.3	66.5±2.0	76.3±1.8	51.5±1.4	33.6±1.1	62.0±1.9	73.0±1.5	46.7±1.3
Y	**29.8**±2.5	**60.6**±2.4	**72.2**±1.9	**44.0**±2.2	**39.4**±2.1	**68.0**±2.0	**77.5**±1.8	**52.3**±2.0	**34.6**±2.1	**64.3**±2.2	**74.9**±1.8	**48.2**±2.0

Table S5: **Comparison of sampling approach on S-MiT using AMM and Bart**

α	Caption to Video				Video to Caption				Mean			
	R@1	R@5	R@10	mAP	R@1	R@5	R@10	mAP	R@1	R@5	R@10	mAP
0.1	29.3±1.4	60.0±1.2	**72.7**±1.4	43.4±1.2	39.2±1.8	66.2±1.5	77.0±1.3	51.7±1.6	34.2±1.5	63.1±1.3	74.8±1.1	47.5±1.4
0.2	28.4±1.2	58.1±1.9	70.9±1.5	42.3±1.4	39.3±1.6	67.4±1.5	77.0±1.8	52.0±1.4	33.9±1.4	62.7±1.7	74.0±1.5	47.2±1.4
0.3	27.1±2.5	58.9±2.9	71.5±2.2	41.6±2.3	38.5±2.4	67.1±1.0	76.6±1.9	51.5±1.9	32.8±2.3	63.0±1.9	74.0±1.9	46.5±2.1
0.4	28.1±1.1	58.1±2.1	69.8±2.3	41.9±1.5	38.8±2.4	66.9±1.2	75.8±1.5	51.5±1.8	33.5±1.8	62.5±1.6	72.8±1.7	46.7±1.6
0.5	**29.8**±2.5	**60.6**±2.4	72.2±1.9	**44.0**±2.2	**39.4**±2.1	**68.0**±2.0	**77.5**±1.8	**52.3**±2.0	**34.6**±2.1	**64.3**±2.2	**74.9**±1.8	**48.2**±2.0
0.6	28.1±2.1	59.1±2.3	71.3±2.1	42.3±1.9	38.3±1.9	67.1±1.6	76.6±1.7	51.4±1.6	33.2±1.9	63.1±1.8	73.9±1.8	46.9±1.7
0.7	28.9±1.5	59.2±1.3	70.8±1.3	42.8±1.4	38.9±1.7	66.3±1.4	76.0±1.5	51.3±1.5	33.9±1.6	62.7±1.4	73.4±1.3	47.1±1.4
0.8	29.0±1.9	59.2±2.4	70.7±1.4	42.8±1.9	38.3±2.1	66.3±1.6	75.9±1.5	51.1±1.7	33.6±1.9	62.8±1.9	73.3±1.3	46.9±1.7
0.9	27.7±2.1	57.0±2.5	68.2±2.0	41.1±2.2	37.5±2.6	64.6±2.4	74.1±1.5	49.8±2.2	32.6±2.4	60.8±2.4	71.2±1.7	45.5±2.2

Table S6: **Comparison of different dampening multipliers, α, in AMM on S-MiT using Bart**

Trained On	Evaluated On																			
	Vatex				ActivityNet				MSR-VTT				S-MiT				Mean			
	R@1	R@5	R@10	mAP	R@1	R@5	R@10	mAP	R@1	R@5	R@10	mAP	R@1	R@5	R@10	mAP	R@1	R@5	R@10	mAP
Vatex	**19.8**	**48.4**	**63.7**	**33.4**	1.5	5.2	8.6	4.2	10.3	28.7	39.3	19.8	7.1	20.2	28.6	14.4	9.7	25.6	35.1	18.0
ActivityNet	12.1	33.3	46.8	23.0	2.0	7.3	12.0	5.6	7.5	22.1	31.2	15.4	4.9	15.6	24.1	11.4	6.6	19.6	28.5	13.9
MSR-VTT	6.5	19.2	28.8	13.8	1.3	4.6	7.8	3.7	11.8	33.9	48.2	23.2	8.0	23.6	34.3	16.4	6.9	20.3	29.8	14.3
S-MiT	19.4	44.6	57.7	31.7	**2.7**	**8.4**	**13.6**	**6.5**	**17.3**	**39.8**	**51.8**	**28.4**	**25.8**	**52.8**	**64.7**	**38.5**	**16.3**	**36.4**	**47.0**	**26.3**

Table S7: **Cross Dataset Evaluation on Video/Caption Retrieval on Full Test Set**

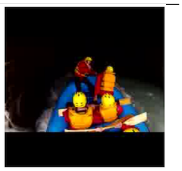
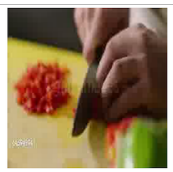
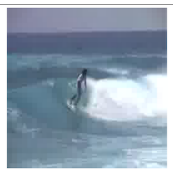
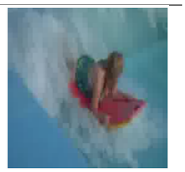
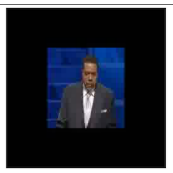
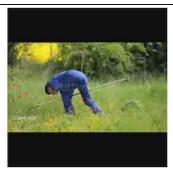
	Query	Retrieval Results				
		R@1	R@2	R@3	R@4	R@5
(a)	it is raining outside in between some houses there is a small stream of water running down in between them					
(b)	four people in a yellow raft white water rafting on the Gorge					
(c)	someone uses a knife to cut lettuce on a cutting board with their hands					
(d)	a young boy with the blue shirt rides artificial waves with the surf					
(e)	a man is standing on a stage in front of an audience telling jokes					
(f)	a video showing several people seated in the audience all men all nicely dressed applauding					
(g)	a man plows the ground with his hands sitting in the hot sun with another man sitting in the background appeared to rest					
(h)	children are standing in the rain and holding their umbrellas upside down					
(i)	a very old picture of boxers boxing in a boxing competition					

Table S8: **Spoken Moments Examples of Caption to Video Retrieval Results:** Given a query caption, we show five top retrieved captions where words transcribed from the audio captions using a pretrained ASR model are used as a caption. We use a BART model trained with the AMM loss function on the S-MiT dataset. Blue indicates the ground-truth results.

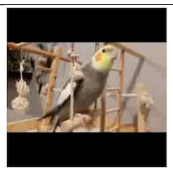
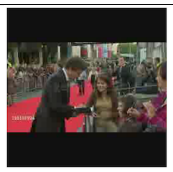
	Query	Retrieval Results				
		R@1	R@2	R@3	R@4	R@5
(a)		a cat is kneading the back of a dog as it sleeps	large cat is being rocked by its owner the owner caresses on	a dog and cat are playing with each other slowly	is the wonder stuff panda resting on top of a guy that is sleeping on the couch	a cat resting has cat food cans stacked on its head
(b)		a gray and white bird is sitting on top of a bird cage the bird makes sounds while it's sitting on the cage	a yellow bird swings on a perch and enclosure	it's a slow mo video of a brown egg cracking on the ground	noragami white hair on mustache and a beard stands or sits over a trash can and use a drill and a peeler to peel an apple	a cartoon plays with penguins and kittens as they sing and play instruments
(c)		a man is signing autographs for a group of people	a famous actor in a tuxedo is signing an autograph for fans as he walks down a red carpet	the celebrity Meryl Streep walks along a crowd of people yelling and giving autographs	people are waiting for the autograph	a woman is signing order interviews and autographs
(d)		tours of Coastline a sailboat with white sails on the water headed towards the dock area	people are outside on really choppy water in little tiny white sailboats with just one person on each	a boat sails off screen to the right as a plane flies overhead	a boy in the sailboat out on the water we see his head and the proud the boat	a person is watching the sunrise from a large sailboat
(e)		someone mold a vase into clay with their hands	an artist using a pic to sculpt to carve an ice sculpture	person is carving a flower from wood with music in the background	you can see a man's hand using a chisel and Hammer to break apart a piece of Rock	a man is mixing together some type of green orbs in a giant bowl with his hands
(f)		an animation of two planets crashing into each other in space	this is a video of planets or something going off in plane	the street view of two cars crash into each other slamming up against the pole	it's a slow mo video of a brown egg cracking on the ground	two pineapples are on a machine that Twirls it and they are being saved by the machine
(g)		a metal hydraulic machine has a blue pill gel pill as a top compressor pushes down and popsicle	what looks like bananas are on a conveyor belt going down the line to an automatic Chopper	machine is being utilized manufactory	where's the machine kicking a liquid and cleaning stalls	two pineapples are on a machine that Twirls it and they are being saved by the machine
(h)		a group of people wearing orange and white t-shirts drumming in a drumline all to the same beat	for drummers wearing red marching band uniforms are playing the drums and synchronization as other band needs or standing behind them watching	people walking by all ages all sizes walking by somebody that's playing the drums in the background	it's a video of three men beating on drums at a football game they all have beards and sunglasses	the band members are playing the instrument and their walking forward
(i)		slow motion shot of a bunch of coffee beans falling and has a brown background	hazelnuts are falling they're falling on a reflective surface	seems like slowly fall onto a bowl	3 pieces of garlic lay on a table while two small pieces of garlic fall from up to the table down	bunch of lemons are falling from the sky in behind a dark black screen

Table S9: **Spoken Moments Examples of Video to Caption Retrieval Results:** Given a query video, we show five top retrieval captions where words transcribed from the audio captions using a pretrained ASR model are used as a caption. We use a BART model trained with the AMM loss function on the S-MiT dataset. Blue indicates the ground-truth results.

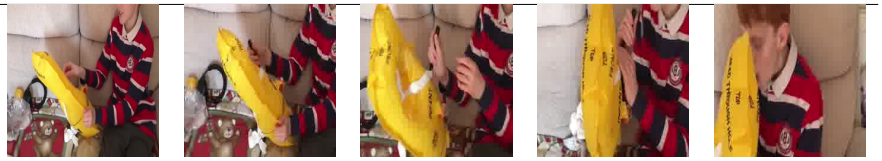
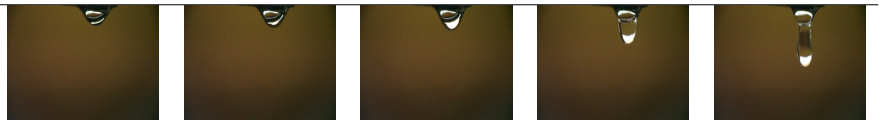
	Caption	Frames → time
(a)	a boy and a red white and blue shirt is sitting on a couch he is **holding** an infant life vest and **picks it up to blow** through the two	
(b)	there's a gauge or a lock thing turns from rides and then **being turned to the left**	
(c)	a picture of a man **drinking coffee** and **play with a cell phone** in **fast motion**	
(d)	in **slow motion** we see a collie **jump into the air** and **catch** a white frisbee in flight	
(e)	these are track and field runners and it's a relay race and **they take off when they are handed the batons**	
(f)	there is water **dripping off** the edge of something all you can hear is the water dripping	

Table S10: **Spoken Moments Captions:** We show some examples of captions, and associated video frames, from the Spoken Moments dataset, where the captions describe a sequence of actions or motion.