Supplementary Material: Cross-Modal and Hierarchical Modeling of Video and Text

Bowen Zhang⋆[0000−0002−4971−4878], Hexiang Hu⋆[0000−0002−4720−169X]
, and Fei Sha[2222−−3333−4444−5555]

University of Southern California, Los Angeles CA 90089, USA
zhan734,hexiangh,feisha@usc.edu

In this supplementary material, we provide more details to many omitted contents in the main paper:

– We first give details on the implementation of our framework (see Section 1).
– Then we include additional ablation experiments for verifying the choice of τ value we selected, the contribution of different learning objectives, and video and text retrieval performance on ActivityNet val2 (see Section 2).
– Finally, we provide qualitative results (in Section 3), for showing examples achieved by HSE against baseline algorithms.

1 Implementation Details

1.1 Video and Text Features

C3D Features. Similar to [7], we follow the standard ActivityNet setting and use the C3D [9] features from [4] for retrieval and dense captioning [7]. In all our experiments under this setting, we extract frame-wise video feature using C3D model pre-trained on Sports-1M dataset, with the temporal stride of 16. PCA dimensionality reduction is then conducted to reduce features dimension to 500.

TSN-Inception V3 Features. To leverage the state-of-the-art of current video modeling, we extract more recent deep features for retrieval on ActivityNet [7] and DiDeMo [1], using the Inception V3 model pre-trained on Kinetics [5] dataset (provided by [11]). Follow their settings, we resize video frames to the resolution of 299 × 299. We then fed video frames into the deep Inception V3 model to extract the output activations from penultimate layer. Unlike [11], we do not perform any test-time data augmentations (*e.g.*multiple crops, color jitter, etc.). Note that no fine-tuning are performed on either ActivityNet or DiDeMo.

Word Features. In the retrieval related experiments, we always use GloVE features [8] for the initialization of the word embedding and fine-tune. Specifically, we use the GloVE vectors pre-trained on 840B common web-crawled data, with its dimensionality equals to 300.

⋆ equal contribution

Training Details When the learning of hierarchical embedding is applicable, we feed the entire video/paragraph in its frame-wise/word-wise representations through the low-level encoder, and then input the subsequent low-level embedding to the high-level encoder as its initial hidden state. In all our experiments, we use GRU [2] with its hidden dimension to be 1,024 as our sequence encoder and decoder. To obtain the embedding for a sequence, we take the channel-wise max over all output vectors of the GRU as it empirically outperforms other strategies such as [10].

During training, we use Adam [6] optimizer with initial learning rate as 0.001, and decay it by 10 for every 10 epochs during the training. We use Xavier initialization [3] for each affine layer in our model with zero mean and variance of 0.01. We set all margin in the loss function to 0.2. Each loss is normalized by its batch size. On both ActivityNet and DiDeMo dataset, we train our embedding models for 15 epochs and collect the final results.

2 Additional Experiments

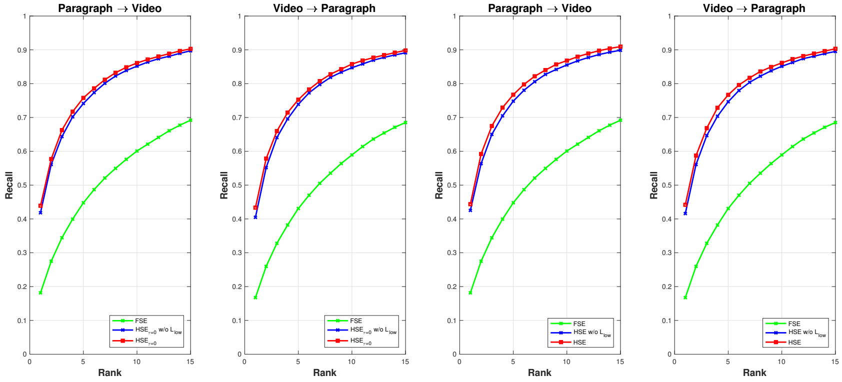
2.1 Ablation study on different learning objectives

Ablation study with different learning objectives We report ablation studies of different losses on ActivityNet video and paragraph retrieval task in Table 1. We use the Inception-V3 features and follow the same setting for training HSE. Each time we remove one loss and report the performance. Note that the reconstruction loss and low-match loss are the most useful.

Table 1. Ablation study on the learning objectives.

	Paragraph ⇒ **V**ideo		**V**ideo ⇒ **P**aragraph	
Method	R@1	R@5	R@1	R@5
HSE w/o high-cluster	44.6	76.4	44.2	76.1
HSE w/o low-match	40.9	73.6	39.8	73.6
HSE w/o low-cluster	44.6	76.6	43.9	76.4
HSE w/o reconstruction	43.9	75.8	43.3	75.3
HSE w all losses	44.4	76.7	44.2	76.7

Low-level loss is beneficial As mentioned in the main text (see Table 1 and Table 2 in the main text), learning with low-level objectives is beneficial for our full model. To better understands this, we also plot the recall (in %) with regard to the rank of the video/paragraph to a query as supportive evidence. The results are shown in Fig. 1.



(a) HSE[τ=0]

(b) HSE

Fig. 1. Recall vs Rank curves of Video to Paragraph and Paragraph to Video retrieval of both HSE[τ=0] and HSE. All results are collected from models based on InceptionV3 feature on ActivityNet validation set 1.

2.2 Ablation Study on Reconstruction Balance Term

Here we study the influence of loss balance term, by experimenting multiple choices of τ under a controlled environment. We choose to study this on the validation set 2 (val2) of ActivityNet with Inception V3 visual feature as input. Detailed results are shown in Table 2. We summarized that retrieval performance, R@1 and R@5, approach to its peak when τ=0.0005. Therefore, as stated in the main text, we set τ to be 0.0005 in all our experiments.

Table 2. Ablation study of τ on ActivityNet (val2).

	Paragraph	⇒	**V**ideo		**V**ideo	⇒	**P**aragraph	
	R@1	R@5	R@50	MR	R@1	R@5	R@50	MR
Inception-V3 pre-trained on Kinetics [12]								
HSE[τ=0.05]	25.0	54.9	92.6	5.0	25.1	55.4	92.4	4.0
HSE[τ=0.005]	32.4	62.2	93.8	3.0	32.1	63.0	93.7	3.0
HSE[τ=0.0005]	**33.2**	**62.9**	**93.6**	**3.0**	**32.6**	**62.8**	**93.5**	**3.0**
HSE[τ=0.00005]	33.2	62.9	93.8	3.0	32.2	62.5	93.6	3.0
HSE[τ=0]	32.2	61.5	93.6	3.0	31.5	62.0	93.3	3.0

2.3 Performance on ActivityNet Validation Set 2

As mentioned in the main paper, we reported the val2 performance of FSE, HSE[τ=0], and HSE in Table 3. Again, the results verified our papers' claim as we show that HSE consistently improve performance than FSE and HSE[τ=0]. It shows the importance of hierarchical modeling and feature reconstruction.

Table 3. Performance of video and paragraph retrieval on ActivityNet (val2). Standard deviation from 3 random seeded experiments are also reported.

	Paragraph ⇒ **V**ideo				**V**ideo ⇒ **P**aragraph			
	R@1	R@5	R@50	MR	R@1	R@5	R@50	MR
C3D Feature with Dimensionality Reduction [9]								
FSE	$11.5_{\pm 0.2}$	$31.0_{\pm 0.4}$	$75.9_{\pm 0.2}$	14.0	$11.0_{\pm 0.5}$	$30.6_{\pm 0.3}$	$75.5_{\pm 0.4}$	14.0
HSE[τ=0]	$23.3_{\pm 0.5}$	$48.2_{\pm 0.2}$	$84.5_{\pm 0.4}$	6.0	$23.0_{\pm 0.3}$	$47.9_{\pm 0.2}$	$84.6_{\pm 0.2}$	6.0
HSE[τ=0.0005]	$23.9_{\pm 0.3}$	$49.4_{\pm 0.3}$	$85.3_{\pm 0.2}$	6.0	$23.4_{\pm 0.5}$	$49.4_{\pm 0.4}$	$85.5_{\pm 0.3}$	6.0
Inception-V3 pre-trained on Kinetics [12]								
FSE	$16.0_{\pm 0.2}$	$41.8_{\pm 0.4}$	$88.0_{\pm 0.5}$	8.0	$15.1_{\pm 0.7}$	$41.0_{\pm 0.4}$	$87.7_{\pm 0.5}$	8.0
HSE[τ=0]	$32.3_{\pm 0.2}$	$62.2_{\pm 0.7}$	$93.5_{\pm 0.1}$	3.0	$32.0_{\pm 1.0}$	$61.9_{\pm 0.2}$	$93.3_{\pm 0.1}$	3.0
HSE[τ=0.0005]	$\mathbf{32.9}_{\pm \mathbf{0.4}}$	$\mathbf{62.7}_{\pm \mathbf{0.2}}$	$\mathbf{93.9}_{\pm \mathbf{0.4}}$	**3.0**	$\mathbf{32.6}_{\pm \mathbf{0.1}}$	$\mathbf{63.0}_{\pm \mathbf{0.2}}$	$\mathbf{93.7}_{\pm \mathbf{0.2}}$	**3.0**

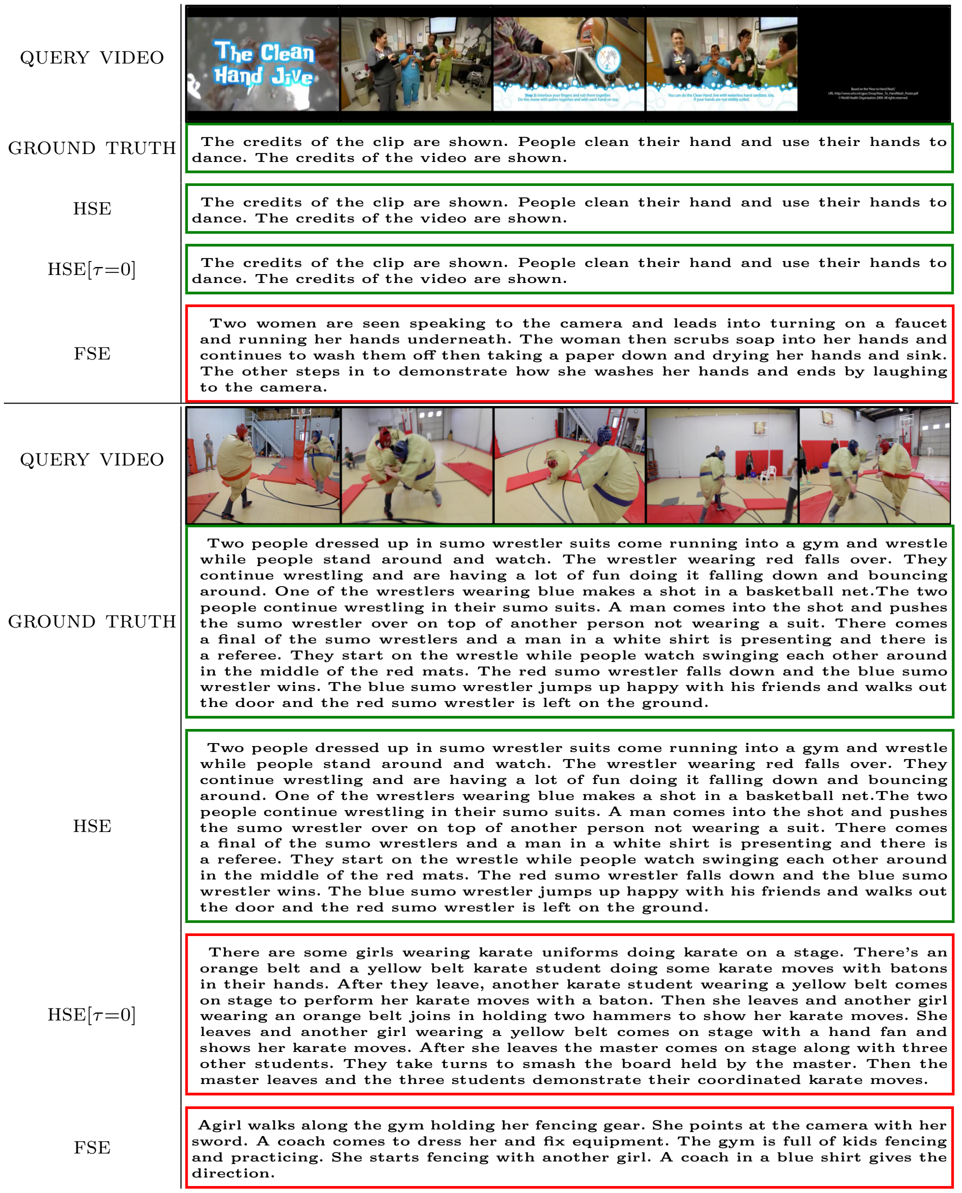
3 Visualization

Qualitative examples for hse, hse[τ=0], and fse on retrieval tasks. We show the qualitative examples on ActivityNet as below. To show a systematic analysis of the success cases and failure cases, we choose to visualize positive examples of paragraph retrieval (in Figure 2) and video retrieval (in Figure 4) and negative examples in Figure 3 and Figure 5. We observe that in some failed cases, although HSE failed to retreive the correct text/video, it retreive very relevant item given the query information.

References

1. Anne Hendricks, L., Wang, O., Shechtman, E., Sivic, J., Darrell, T., Russell, B.: Localizing moments in video with natural language. In: ICCV. pp. 5804–5813
2. Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using rnn encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078 (2014)
3. Glorot, X., Bengio, Y.: Understanding the difficulty of training deep feedforward neural networks. In: AISTATS. pp. 249–256 (2010)
4. Heilbron, F.C., Escorcia, V., Ghanem, B., Niebles, J.C.: Activitynet: A large-scale video benchmark for human activity understanding. In: CVPR. pp. 961–970
5. Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., et al.: The kinetics human action video dataset. arXiv preprint arXiv:1705.06950 (2017)
6. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
7. Krishna, R., Hata, K., Ren, F., Fei-Fei, L., Niebles, J.C.: Dense-captioning events in videos. In: ICCV. pp. 706–715 (2017)
8. Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: EMNLP. pp. 1532–1543 (2014)
9. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3d convolutional networks. In: ICCV. pp. 4489–4497 (2015)
10. Venugopalan, S., Rohrbach, M., Donahue, J., Mooney, R., Darrell, T., Saenko, K.: Sequence to sequence-video to text. In: ICCV. pp. 4534–4542 (2015)

11. Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., Van Gool, L.: Temporal segment networks: Towards good practices for deep action recognition. In: ECCV. pp. 20–36 (2016)
12. Wang, L., Li, Y., Lazebnik, S.: Learning deep structure-preserving image-text embeddings. In: CVPR. pp. 5005–5013 (2016)



QUERY VIDEO

GROUND TRUTH

The credits of the clip are shown. People clean their hand and use their hands to dance. The credits of the video are shown.

HSE

The credits of the clip are shown. People clean their hand and use their hands to dance. The credits of the video are shown.

HSE[τ=0]

The credits of the clip are shown. People clean their hand and use their hands to dance. The credits of the video are shown.

FSE

Two women are seen speaking to the camera and leads into turning on a faucet and running her hands underneath. The woman then scrubs soap into her hands and continues to wash them off then taking a paper down and drying her hands and sink. The other steps in to demonstrate how she washes her hands and ends by laughing to the camera.

QUERY VIDEO

GROUND TRUTH

Two people dressed up in sumo wrestler suits come running into a gym and wrestle while people stand around and watch. The wrestler wearing red falls over. They continue wrestling and are having a lot of fun doing it falling down and bouncing around. One of the wrestlers wearing blue makes a shot in a basketball net.The two people continue wrestling in their sumo suits. A man comes into the shot and pushes the sumo wrestler over on top of another person not wearing a suit. There comes a final of the sumo wrestlers and a man in a white shirt is presenting and there is a referee. They start on the wrestle while people watch swinging each other around in the middle of the red mats. The red sumo wrestler falls down and the blue sumo wrestler wins. The blue sumo wrestler jumps up happy with his friends and walks out the door and the red sumo wrestler is left on the ground.

HSE

Two people dressed up in sumo wrestler suits come running into a gym and wrestle while people stand around and watch. The wrestler wearing red falls over. They continue wrestling and are having a lot of fun doing it falling down and bouncing around. One of the wrestlers wearing blue makes a shot in a basketball net.The two people continue wrestling in their sumo suits. A man comes into the shot and pushes the sumo wrestler over on top of another person not wearing a suit. There comes a final of the sumo wrestlers and a man in a white shirt is presenting and there is a referee. They start on the wrestle while people watch swinging each other around in the middle of the red mats. The red sumo wrestler falls down and the blue sumo wrestler wins. The blue sumo wrestler jumps up happy with his friends and walks out the door and the red sumo wrestler is left on the ground.

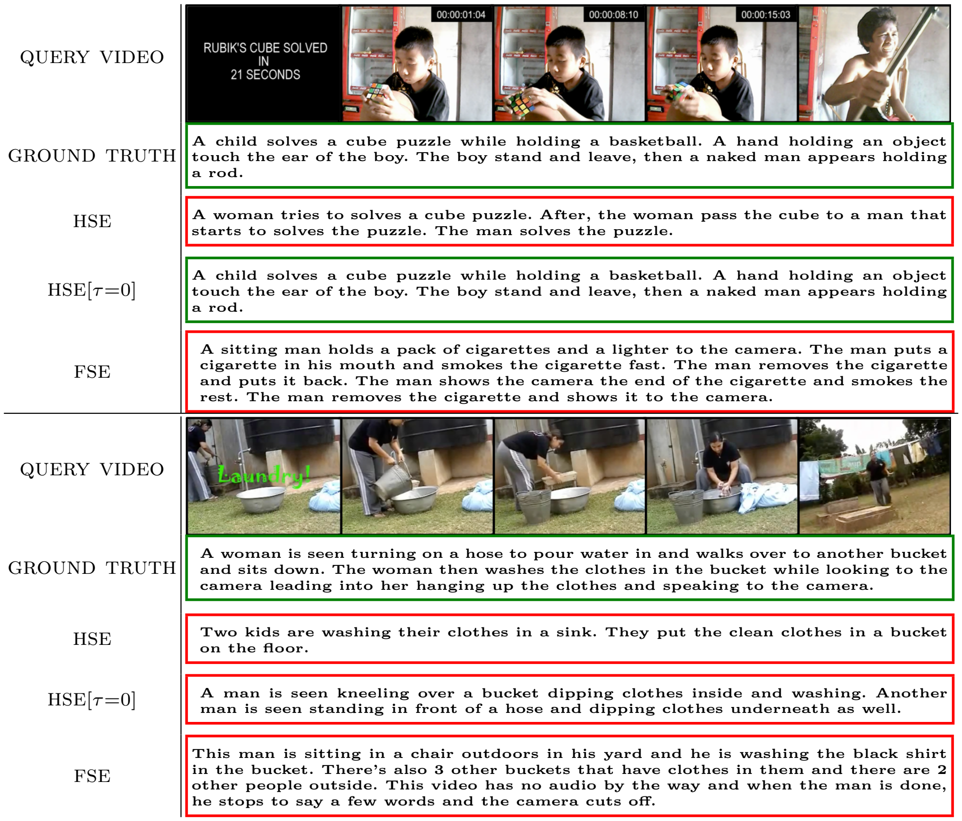
HSE[τ=0]

There are some girls wearing karate uniforms doing karate on a stage. There's an orange belt and a yellow belt karate student doing some karate moves with batons in their hands. After they leave, another karate student wearing a yellow belt comes on stage to perform her karate moves with a baton. Then she leaves and another girl wearing an orange belt joins in holding two hammers to show her karate moves. She leaves and another girl wearing a yellow belt comes on stage with a hand fan and shows her karate moves. After she leaves the master comes on stage along with three other students. They take turns to smash the board held by the master. Then the master leaves and the three students demonstrate their coordinated karate moves.

FSE

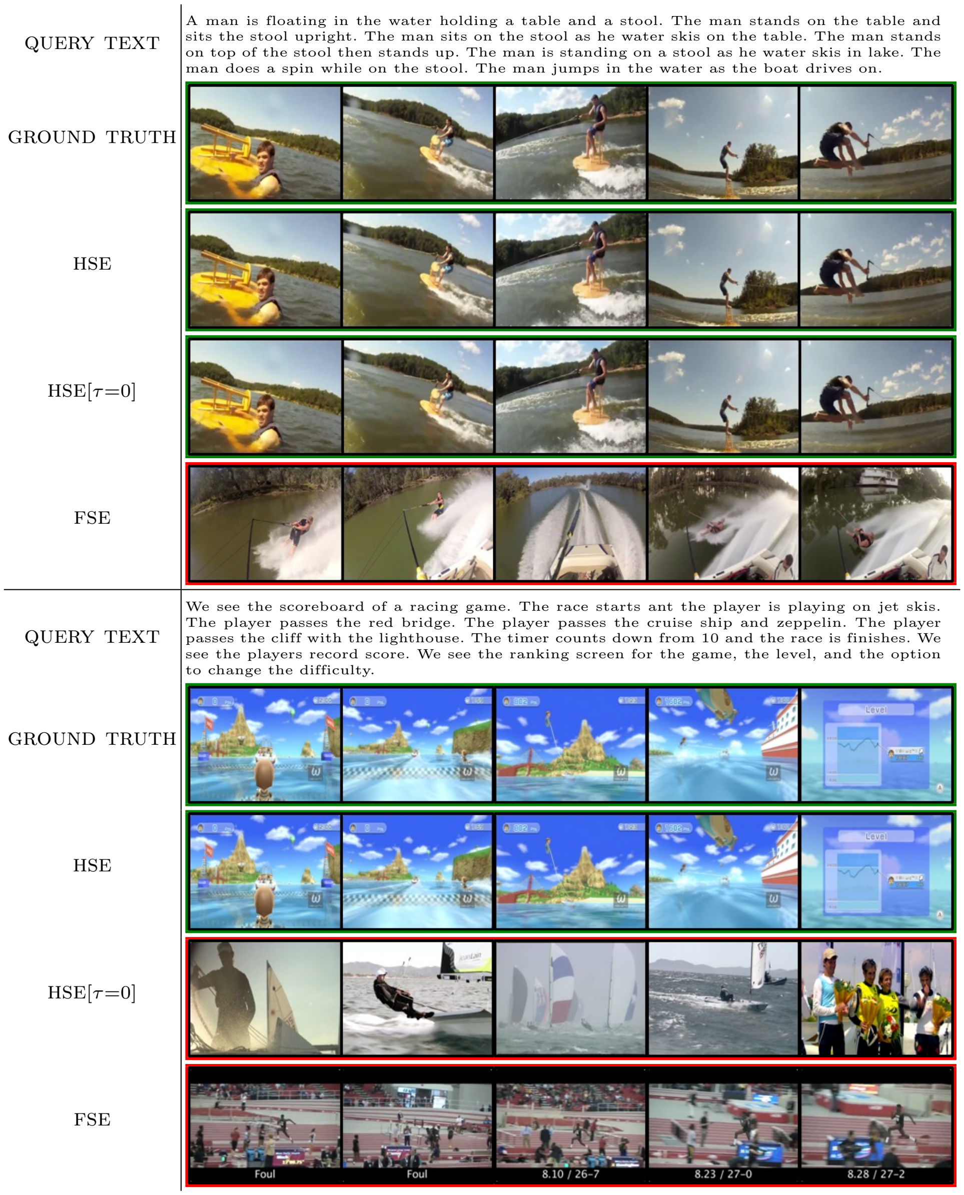
Agirl walks along the gym holding her fencing gear. She points at the camera with her sword. A coach comes to dress her and fix equipment. The gym is full of kids fencing and practicing. She starts fencing with another girl. A coach in a blue shirt gives the direction.

Fig. 2. ActivityNet: Given Video and Retrieve Paragraph. Positive qualitative examples of HSE, HSE[τ=0], and FSE on the task of given video to retrieve texts. We mark the correct sample in **green** and incorrect one in **red**.



QUERY VIDEO	
GROUND TRUTH	A child solves a cube puzzle while holding a basketball. A hand holding an object touch the ear of the boy. The boy stand and leave, then a naked man appears holding a rod.
HSE	A woman tries to solves a cube puzzle. After, the woman pass the cube to a man that starts to solves the puzzle. The man solves the puzzle.
HSE[τ=0]	A child solves a cube puzzle while holding a basketball. A hand holding an object touch the ear of the boy. The boy stand and leave, then a naked man appears holding a rod.
FSE	A sitting man holds a pack of cigarettes and a lighter to the camera. The man puts a cigarette in his mouth and smokes the cigarette fast. The man removes the cigarette and puts it back. The man shows the camera the end of the cigarette and smokes the rest. The man removes the cigarette and shows it to the camera.
QUERY VIDEO	
GROUND TRUTH	A woman is seen turning on a hose to pour water in and walks over to another bucket and sits down. The woman then washes the clothes in the bucket while looking to the camera leading into her hanging up the clothes and speaking to the camera.
HSE	Two kids are washing their clothes in a sink. They put the clean clothes in a bucket on the floor.
HSE[τ=0]	A man is seen kneeling over a bucket dipping clothes inside and washing. Another man is seen standing in front of a hose and dipping clothes underneath as well.
FSE	This man is sitting in a chair outdoors in his yard and he is washing the black shirt in the bucket. There's also 3 other buckets that have clothes in them and there are 2 other people outside. This video has no audio by the way and when the man is done, he stops to say a few words and the camera cuts off.

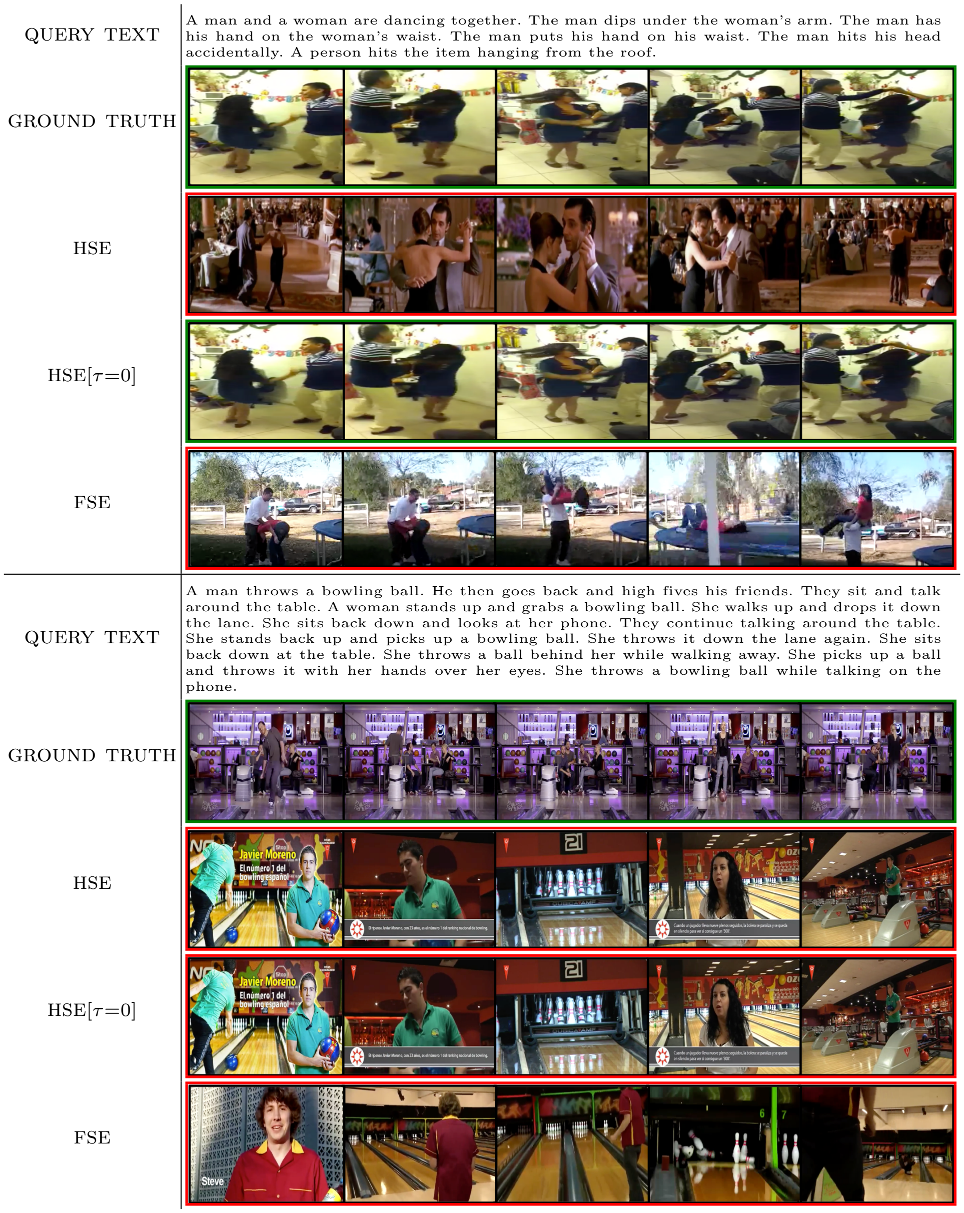
Fig. 3. ActivityNet: Given Video and Retrieve Paragraph. Negative qualitative examples of HSE, HSE[τ=0], and FSE on the task of given video to retrieve texts. We mark the correct sample in **green** and incorrect one in **red**.



A man is floating in the water holding a table and a stool. The man stands on the table and sits the stool upright. The man sits on the stool as he water skis on the table. The man stands on top of the stool then stands up. The man is standing on a stool as he water skis in lake. The man does a spin while on the stool. The man jumps in the water as the boat drives on.

We see the scoreboard of a racing game. The race starts ant the player is playing on jet skis. The player passes the red bridge. The player passes the cruise ship and zeppelin. The player passes the cliff with the lighthouse. The timer counts down from 10 and the race is finishes. We see the players record score. We see the ranking screen for the game, the level, and the option to change the difficulty.

Fig. 4. ActivityNet: Given Paragraph and Retrieve Video. Positive qualitative examples of HSE, HSE[τ=0], and FSE on the task of given text to retrieve video. We mark the correct sample in **green** and incorrect one in **red**.



A man and a woman are dancing together. The man dips under the woman's arm. The man has his hand on the woman's waist. The man puts his hand on his waist. The man hits his head accidentally. A person hits the item hanging from the roof.

A man throws a bowling ball. He then goes back and high fives his friends. They sit and talk around the table. A woman stands up and grabs a bowling ball. She walks up and drops it down the lane. She sits back down and looks at her phone. They continue talking around the table. She stands back up and picks up a bowling ball. She throws it down the lane again. She sits back down at the table. She throws a ball behind her while walking away. She picks up a ball and throws it with her hands over her eyes. She throws a bowling ball while talking on the phone.

Fig. 5. ActivityNet: Given Paragraph and Retrieve Video. Negative Qualitative examples of HSE, HSE[τ=0], and FSE on the task of given text to retrieve video. We mark the correct sample in **green** and incorrect one in **red**.