Supplementary Material: Generating Personalized Summaries of Day Long Egocentric Videos

Pravin Nagar, Anuj Rathore, C.V. Jawahar, Chetan Arora

In this Supplementary Material, we provide the following details omitted in the main text:

- Section 1: More qualitative analysis for predicted summaries
- Section 2: Comparison between All the Frameworks
- Section 3: Stability of RL Frameworks
- Section 4: Participant's Comments
- Section 5: Information Sheet
- Section 6 Detailed results on Personalized Summarization
- Section 7: Demographic Information
- Section 8: Algorithms
- Section 9: Video Demonstration

1 More qualitative analysis for predicted summaries:

We further add more visualization to deeply inspect the proposed frameworks with all the ground truth summaries when using basic rewards. We choose 5 minutes summaries predicted by the proposed frameworks to compare against the three ground truth summaries ranging from 3 to 6 minutes. In Fig. 1 and 2, we demonstrate the visualization for the 'Alin Day 1' video sequence of the Dinsey dataset and the 'P01' video sequence of the UTE dataset, respectively, with all the three ground truth summaries.

Fig. 10 shows the comparison of 1 minute, 3 minutes, and 5 minutes summary generated by AC framework using the distinctiveness-indicativeness reward of 'HUJI Ariel 1' video.

We have also prepared the GUI of the proposed work to conduct a user study for personalized summarization. The GUI is shown in Fig. 7. As discussed in the main text, the detail table with user comments on the personalized summary is shown in Table 2

Table 1 shows the summary length and sliding window size for two long sequence datasets, namely Disney and HUJI. As mentioned in the main text, we take sliding window size 25% of the desired summary length. To generate one-minute summaries, our summary length and sliding window size are 120 sub-shots (i.e. 2 sub-shots/second) and 30 sub-shots respectively. Similarly, for 10 minutes summaries, summary length and sliding window size are 120 and 30 respectively and so on for 3, 5, and 15 minutes summaries. For the Disney dataset, we train the network for 1, 5, and 15 minutes summaries, whereas for the HUJI dataset, we train the network for 1, 3, and 5 minutes summaries.

Summary length	Sliding window size
120 (1 min)	30
180 (3 mins)	45
600 (5 mins)	150
1200 (10 mins)	300
1800 (15 mins)	450

TABLE 1
Summary length and sliding window size for summaries of various time durations.

2 Comparison between all the frameworks:

Fig. 3 shows the training plot of policy gradient, Q learning, and Actor-Critic framework. As discussed in the main text, the episodic reward plot for the policy gradient shows high variance across video samples due to baseline function. Whereas Q learning and AC framework use Q value network leads to stable gradient across video samples. On the other end, the extra parameters required more training samples. If we have less training data, then the policy gradient is easy to train. For the proposed framework, each position of the sliding window constitutes one training sample, so we generate sufficient training samples (especially for day long videos) to train the Q learning and AC frameworks. The same is validated by Fig. 3 shows the training plot of policy gradient, Q learning, and Actor-Critic framework.

3 Stability of RL frameworks:

As we move the sliding window over the input video sample, it generates enough training samples to train any RL framework. We are successfully able to train policy gradient and Q learning. We also used experience replay for efficient

Fig. 1. The figure demonstrates the comparison between ground truth summaries and the summaries generated by the different frameworks for the 'Alin Day 1' video sequence of the Disney dataset. We operate on 1fps (a C3D feature is extracted per second) and get a binary mask as an output indicating the selected shots (of one second). In contrast, the ground truth summaries comprise a set of sentences, each corresponding to a 5 seconds clip. We map the clips to the original video sequence and generate the binary mast at one fps, similar to our predicted binary mask. In each row, the black vertical bars indicate a frame was picked from a corresponding temporal window of 70 frames as it is not possible to visualize the video sequences at 1fps. As the annotations are done at $1/5$ fps, pooling over a window of length 70 makes the ground truth summaries sparse. We can observe that in the first half and middle of the video, all three ground truth summary frames are uniformly distributed, whereas the selection is significantly less toward the end. The Actor-Critic framework also exhibits the same behavior, whereas the policy gradient and Q-learning perform slightly poorly compared to the Actor-Critic.

convergence. Ideally, for the Actor-Critic framework, we have separate networks for actor and critic, but due to the diverse nature of each video sample, we are not able to train the AC framework. To get around the problem, we have used a common backbone LSTM network for actor and critic network followed by two fully connected heads for actor and critic, respectively.

4 Participant's Comments

Our visualization shows (Fig. 5) that the SOTA approaches pick a cluster of frames in summary from the same location, which leads to lower the *informativeness* and *enjoyability* score compared to the proposed framework. The same is validated through our user study where one of the participants expressed for 'Alin Day 1' video when FFNet [2] is used,

> *"Kept focussing on scenes for far too long and because of this, it missed many other scenes. For example, lunch and dinner sequences were longer than required."*

Similarly, the summaries generated by uniform sampling and K-medoids, show sudden changes that lead to poor comprehension and reduces the *informativeness* and *enjoyability* score. The following quote from one of our participants (for 'Michale Day 2' video when 'uniform sampling' is used) supports the finding:

> *"Informativeness: I could not make sense of the whole summary as it felt more like a slide show of images. Although most of the events were included as compared to ground truth, still I reduced my score as I felt that multiple pics (frames) were depicting one event, which could be avoided given the slow rate and the fact that few frames were not adding any new information. Enjoyability: I did not enjoy this! It was not smooth and felt like I am watching a slide show of images. It was so slow and boring! "*

5 Information Sheet

Below we give the verbatim text transferred to the subjects for the user study.

5.1 Information Sheet

You are being invited to take part in this study. Before you make a decision, it is important for you to understand why this study is being done and what it will involve. Please take time to understand the following information carefully. Please do not hesitate to ask us if there is anything that is not clear or if you would like more information. If you do take part, you will be asked to sign a consent form.

5.1.1 Objective:

We are conducting a study to understand how the system-generated summary of a day long egocentric video satisfies a user. We further extend our work to personalize the summary by taking user feedback and then ask the user to evaluate the personalized summary.

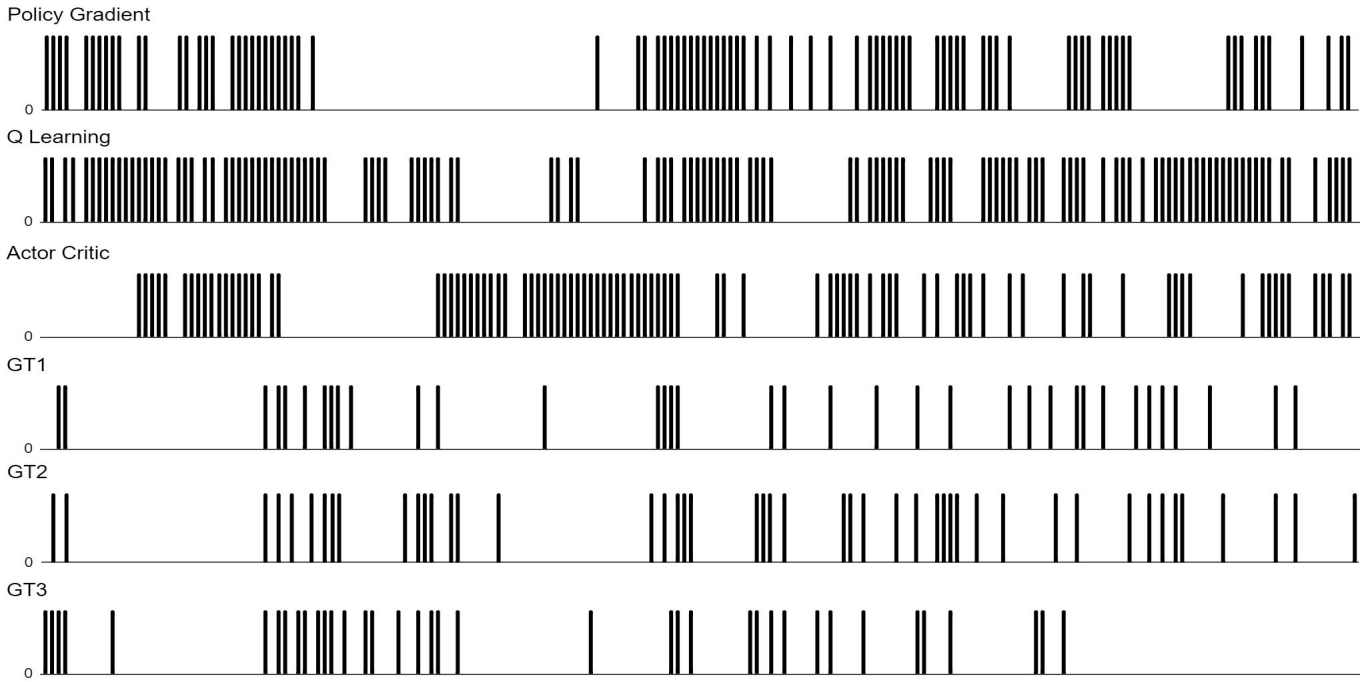

Fig. 2. The figure demonstrates the comparison between ground truth summaries and the summaries generated by the different frameworks for the 'P01' video sequence of the UTE dataset. In each row, the black vertical bars indicate a frame was picked from a corresponding temporal window of 70 frames as it is not possible to visualize the video sequences at 1fps. We can observe that the ground truth summary frames are approximately uniformly distributed in the second half of the video. The same distribution is observed for the predicted summaries from all the frameworks.

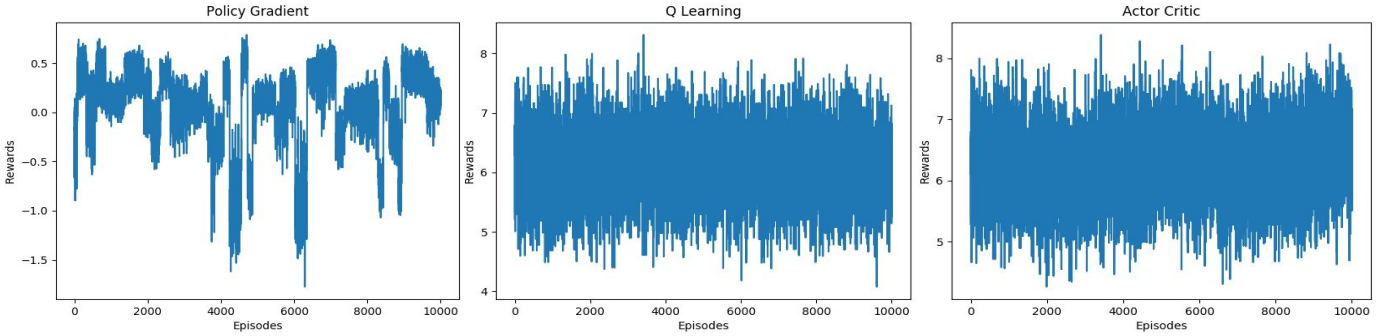

Fig. 3. The episodic reward plot of the policy gradient shows that we get clusters corresponding to each video sample as the baseline is not parameterized.

5.1.2 Risk:

The study is time-consuming. You may feel exhausted while participating in the study.

5.1.3 Benefits of study

You will not directly benefit from taking part in this study however as the summaries are inherently subjective so helping us out in the evaluation will open a new area of research. Additionally, you will receive incentive of rs 500 for your valuable time.

5.1.4 Confidentiality of research information

Taking part in this study is voluntary and you can stop at any time. We will be collecting demographic details of our participants. However no identifying information will be included in any publication or presentation, and your responses remain confidential.

5.1.5 Meaning of Terms

- **Informativeness** Informativeness score evaluates how many objects/events of the original video are included in the summarized video.
- **Enjoyability** The enjoyability assesses only the smoothness(jerk) of a video sequence.
- **Informativeness and Enjoyability** Rate the Informativeness and Enjoyability of the summary on the following scale.
  extremely dissatisfied = 1
  dissatisfied = 2
  neutral = 3
  satisfied = 4
  extremely satisfied = 5
- **Confidence score** This shows the confidence of the subject by which he/she provides the informativeness and enjoyability. The likert scale for the confidence is

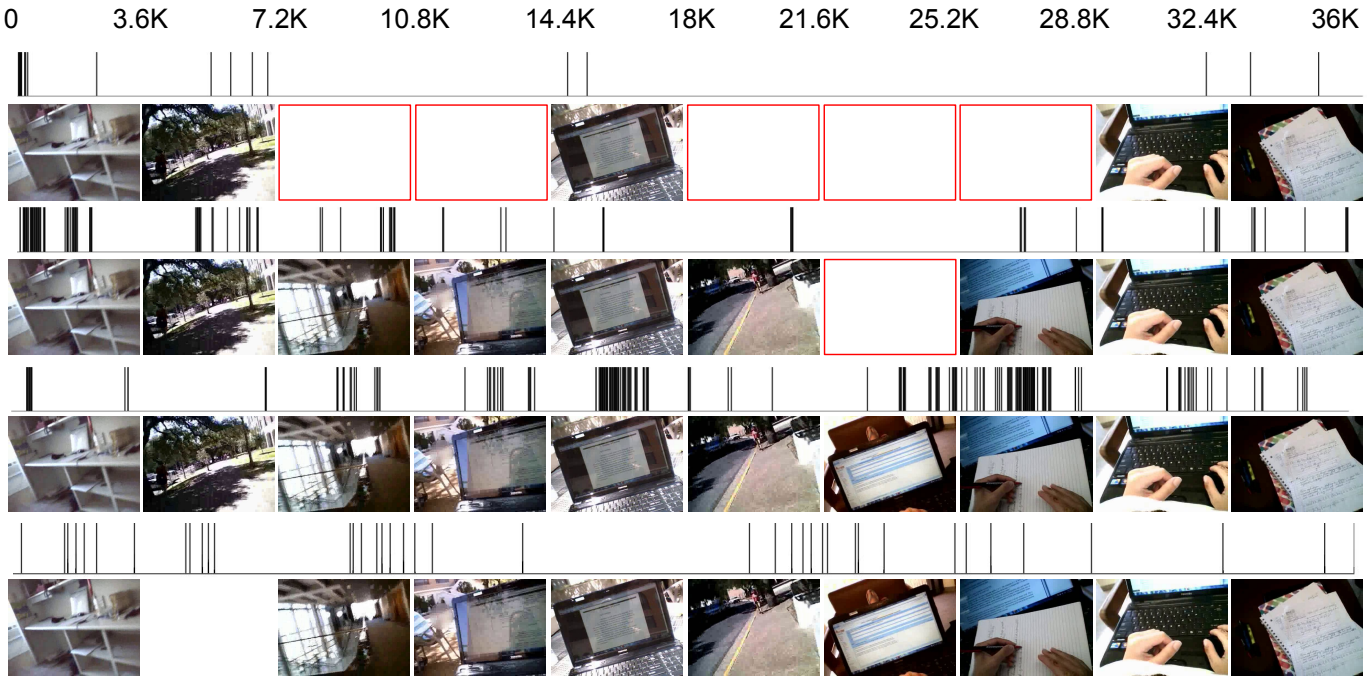

Fig. 4. Similar to Fig. 6 (in the main text), we compare 1, 5, 10 minutes summaries with the ground truth summary in rows 1 to 4, respectively. The summaries are generated using the basic reward using the Actor-Critic framework on the 'P04' sequence of the UTE dataset. We observe that the 1-minute summary does not capture the redundant part in which the subject is 'working on a laptop' (from 18K to 28.8K), whereas the redundant frames increase as the length of the summary increases.

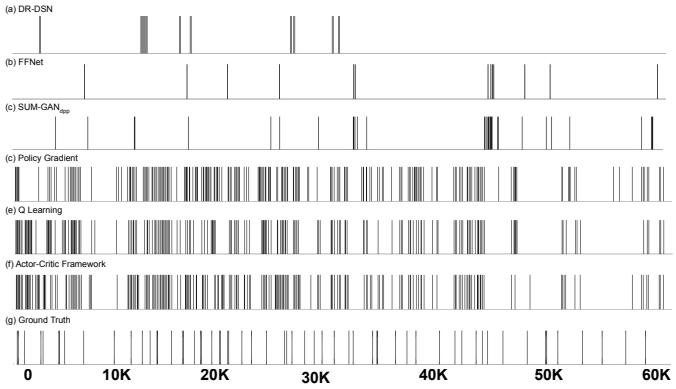

Fig. 5. We observed that DR-DSN [1] picks a cluster of frames from a particular location in summary, whereas the proposed frameworks effectively distribute the summary frame from all over the video. This figure gives a better visualization by showing the distribution of the summary frames for the full video. The bar chart from top to bottom represents the summary generated by DR-DSN [1], FFNet [2], SUM-GAN$_{dpp}$ [3], and our technique with Policy Gradient, Q Learning, and Actor-Critic framework respectively. The figure also indicates that despite using different RL frameworks, most of the selected summary frames are common as the reward is the same for all the frameworks.

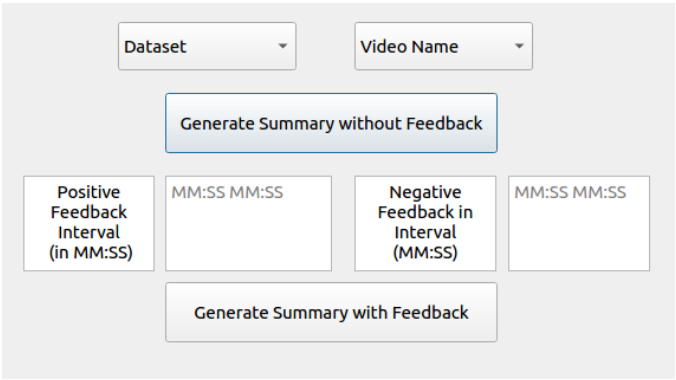

Fig. 6. Figure shows the GUI of the proposed work.

Not confident at all = 1
Slightly confident = 2
Somewhat confident = 3
Fairly confident = 4
Completely confident = 5

5.2 Evaluation Procedure

You would be evaluating summaries of three videos namely Alin Day 1, Alireza Day 1 and Michael Day 2. We have two step evaluation procedure, You are supposed to fill everything in the google form:

1) In the first step you will be asked to evaluate the generated summary. Once you finish viewing the summary then you will be asked to score the same for informativeness and enjoyability using the likert scale mentioned above (in the Google form). You will also be asked for a confidence score for informativeness and enjoyability together.
2) We will show you the GT text summaries (by three users). Once you read the GT text summaries, you will be asked to revisit the generated summary and

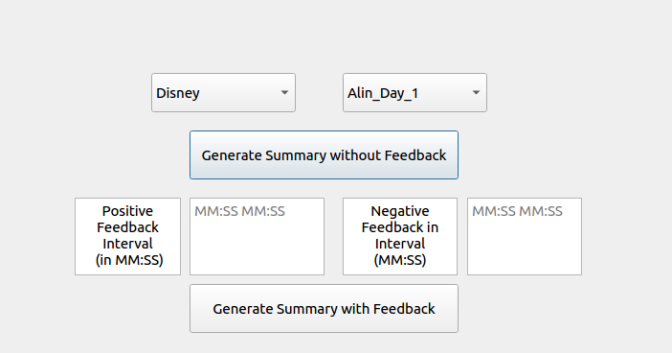

Fig. 7. GUI of the first scenario for personalization of summary.

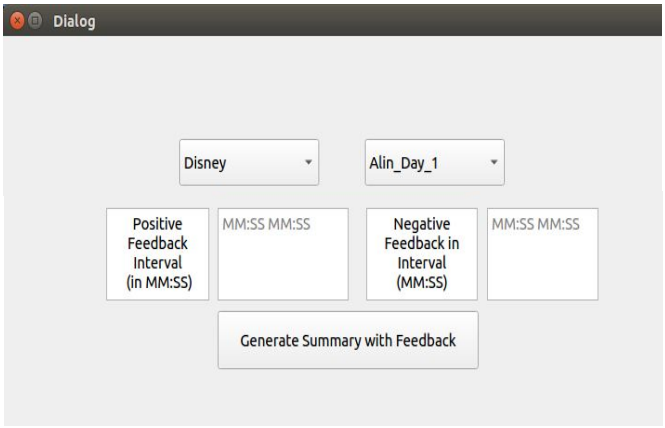

Fig. 8. GUI of the second scenario for personalization of summary.

modify your informativeness and enjoyability scores along with the confidence (if required). Kindly briefly justify your modification.

5.3 Generating personalized summary:

You are supposed to personalize and evaluate two videos. There are two scenarios for the personalization of the summary for each video.

1) In the first scenario, you are asked to choose the events from the system-generated summary (while being unaware of the video content). The detailed personalization procedure is as follows:
   a) You will select a video sequence and click the button "Generate Summary without Feedback". Once a default summary is generated you would be picking the interesting events which you want to include/exclude in the summary. You have to specify the time stamp as a feedback for positive as well as negative feedback. Kindly refer Figure 7.
   b) When you click on the 'Generate Summary with Feedback' the personalized summary incorporating the suggested feedback is generated.
2) In the second scenario, we believe that you are aware of the video content.
   a) We ask you to see the original video and choose the events you want to include/exclude in the summary. You have to specify the time stamp as a feedback for positive as well as negative feedback. Kindly refer Figure 8.
   b) When you click on the 'Generate Summary with Feedback' the personalized summary incorporating the suggested feedback is generated.

5.4 Evaluation procedure for personalized summary

1) Once the personalised summary is generated then you will rate the summary by the quality of personalization compared to default summary on the likert scale (1: very poor, 2: poor, 3: ok, 4: good, 5: excellent) with confidence (1: Not confident to 5: Completely confident).
2) To gauge your experience kindly answer the following question.
   a) "which events you wanted to include/exclude in the summary?"
   b) "why are you satisfied/not satisfied with the generated summary?"

6 Detailed Results for Personalized Summarization

The detailed results for all 10 participants in two different scenarios with participant's feedback are shown in Table 2. The Likert score, along with confidence and participant's comments, shows that the participants are satisfied with the personalization to a large extent. We get 2.88 average (normalized by confidence) Likert score over 20 participants. Furthermore, it's clear from the participant's feedback that the frameworks struggle to completely eliminate the dark scenes when the participants want to exclude them from the summary. This happens because there are many dark scenes scattered throughout the video sequence.

7 Demographic Information

As discussed in the main text, the demographic details are shown in Table 3.

8 Algorithms

We have discussed the proposed approaches in Section 3.2 of the main text. We give the exact algorithm steps here. Algorithm 1 elaborate the sliding window framework and Algorithm 2, Algorithm 3, and Algorithm 4 describes the training process of Policy Gradient, Q Learning, and AC framework respectively.

9 Video Demonstration

Please find the video demonstration of the interactive summarization module on this link. We have created GUI for this module. The video demonstrates how we can provide positive feedback (events you want to include) and/or negative feedback (events you want to exclude) to customize the generated summaries.

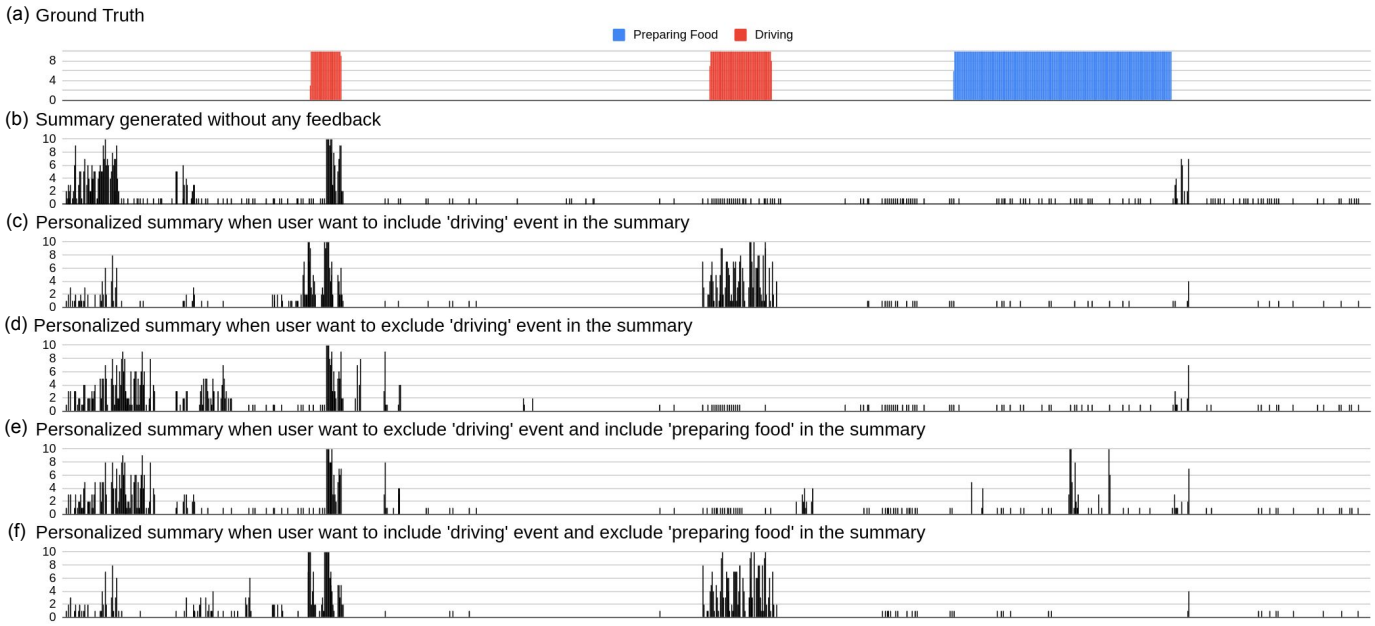

Fig. 9. The figure demonstrates the visualization of the interactive summarization of the 'P01' video sequence of the UTE dataset. Each bar represents 10 seconds of the time interval. (a)-(e) shows different summaries when two events, namely 'preparing food' and 'driving' are included/excluded in summary. We can observe that (c) has more driving sub-shots compared to (b), whereas in (d) the bars in the driving sub-shots are reduced considerably. Similarly, for (e) we get peaks in the 'preparing food' area, whereas the bars in the driving area are reduced. The opposite is seen in (d).

Fig. 10. Comparing 1, 3 and 5 minutes summaries (row 1-3) based on distinctiveness-indicativeness reward of 'HUJI Ariel 1' video.

REFERENCES

[1] K. Zhou, Y. Qiao, and T. Xiang, "Deep reinforcement learning for unsupervised video summarization with diversity-representativeness reward," in *AAAI*, 2018.
[2] S. Lan, R. Panda, Q. Zhu, and A. K. Roy-Chowdhury, "Ffnet: Video fast-forwarding via reinforcement learning," in *CVPR*, 2018.
[3] B. Mahasseni, M. Lam, and S. Todorovic, "Unsupervised video summarization with adversarial LSTM networks," in *CVPR*, 2017.

Subjects	Video -Dataset	Events Included	Events Excluded	Likert Score (1 to 5)	Conf.	Participant Feedback
S01-S1	Alin-Disney	Dinner	Dark scenes	3	4	*'Black part is not completely removed'*
S01-S1	P01-UTE	Driving	Social Int.	4.5	3	*'It accurately highligted the part I liked and don't liked.'*
S02-S1	Alin-Disney	Dinner	Dark scenes	3	4	*'So many dark scenes'*
S02-S1	P01-UTE	lunch	Purchasing	3	4	*'Purchasing in store not removed completely'*
S03-S1	Alin-Disney	Dinner	Tram ride	5	4	*'Included really long dinner, Tram ride is mostly removed'*
S03-S1	P01-UTE	Social Int.	Driving	4	3	*'Detailed conversation, could exclude some more driving shots'*
S04-S1	Alin-Disney	Shopping	Escalator	4.5	4	*'Shopping is taken for little long, escalator is removed'*
S04-S1	P01-UTE	Driving	Writing	5	5	*'Majority of summary was driving, no writting event'*
S05-S2	Alin-Disney	Tram ride	Dinner	4	4	*'Dinner is almost removed'*
S05-S2	P02-UTE	Playing Lego	Eating Pizza	4	4	*'Eating is removed entirely and lego is included for more time'*
S06-S1	Alin-Disney	Dark room	Travel	4	4	*'Accurately included the suggested feedback'*
S06-S1	P02-UTE	Having pizza	Driving	2	4	*'Driving is not removed'*
S07-S1	Alin-Disney	Castle	Travel in bus	3.5	4	*'Overall its good, still there were some bus travel events'*
S07-S1	P01-UTE	Marketing	Driving	2.5	5	*'Lots of instances of driving which could have been reduced'*
S08-S1	Alin-Disney	Indoor	Outdoor	4	4	*'Most of video is outdoor based'*
S08-S1	P02-UTE	Ice Cream	Walking	3.5	4	*'Excluding is correct, inclusion is not very good'*
S09-S2	Alin-Disney	Tram ride	In bus, Dark	2	5	*'Tram ride is missing, rest is fine'*
S09-S2	P03-UTE	lunch, Payment	Purchasing	4.5	3	*'Summary is very nice'*
S10-S1	Alin-Disney	carousel	Dark scenes	2	4	*'Dark scene are so many, poor summary'*
S10-S1	P03-UTE	Cooking	Drive, Wash	4	5	*'Inclusion is perfect! Dish washing is removed, driving is not'*

TABLE 2
The table shows the Likert score of 1 (Extremely dissatisfied) to 5 (Extremely satisfied) given by the participants when specific events are included or excluded in the summary with user comments on the personalized summary. S0X-SY represents subject 'X' in scenario 'Y'. It is observed that sometimes the user sees the excluded part in the personalized summary. This is because the interactive reward personalized the summary but at the same time distinctiveness-indicative reward that tries to maintain the global context. This can be handled by fine-tuning the weights of A and B discussed in interactive reward.

Algorithm 1 Proposed Framework

Input $F_{i=1}^{T}$**:** Video subshots
Output $P_{i=1}^{N}$**:** Probability scores

```
Freeze the C3D weights and randomly initialize weights of BiLSTM
for each epoch do
    for each video do
        for each pass do
            for each sliding window do
                Policy Gradient/Q Learning/ Actor-Critic
            end for
        end for
        if Policy Gradient then
            Update baseline B
        end if
    end for
end for
```

Participant	Stream	Qualification	Gender	Professional Recording
S1	CSE	Ph.D.	Female	No
S2	CSE	Ph.D.	Female	No
S3	IT	Ph.D.	Male	Yes
S4	IT	Ph.D.	Female	No
S5	ECE	Undergrad	Male	No
S6	ECE	Undergrad	Male	No
S7	ECE	Undergrad	Male	No
S8	IT	Undergrad	Male	No
S9	IT	Undergrad	Male	Yes
S10	CSE	Undergrad	Male	Yes

TABLE 3
Demographic Information of subjects for AHR. Three out of ten participants have professional video recording experience.

Algorithm 2 Policy Gradient Framework

```
Initialize θ and learning rate α.
for For each sliding window do
    Calculate S_p and S_f according to the position of W_s
    Get M probability scores from the neural network
    for For each episode do
        Sample M actions from probability scores
        Compute cost and reward
```

$$cost+ = \sum_{m=1}^{M} R(S)\nabla_\theta log\pi_\theta(a_m|h_m)]$$

```
    end for
    Compute episodic cost and episodic reward
    if episodic cost improves then
        update summary by picking top |S| sub-shots
    end if
    if For each mini batch then
        Back-propagate pseudo batch cost
    end if
end for
```

Algorithm 3 Q Learning Framework

Initialize θ, γ and learning rate α.
for For each sliding window **do**
  Calculate S_{p} and S_{f} according to the position of W_{s}
  Get M Q values from the Q value network
  Get M Q values from the target Q value network
  for For each episode **do**
    Sample M actions from probability scores
    Compute correction (TD error) for actions

$$\delta_{1:M} = R(S) + \gamma \sum_{m=1}^{M-1} Q_{\theta^-}(s_{m+1}, a_{m+1}) - \sum_{m=1}^{M-1} Q_{\theta}(s_m, a_m)$$

    Compute cost and reward R(S)

$$cost+ = \delta_{1:M} \sum_{m=1, a \in A}^{M} \nabla_{\theta} Q_{\theta}(s_m, a_m)$$

  end for
  Compute episodic cost and episodic reward
  if episodic reward improves **then**
    update summary by picking top $|S|$ subshots
  end if
  if For each mini batch **then**
    Back-propagate pseudo batch cost
  end if
end for

Algorithm 4 Actor Critic Framework

Initialize θ, w, γ and learning rates α_a, α_c.
for For each sliding window **do**
  Calculate S_{p} and S_{f} according to the position of W_{s}
  Get Q values from the Critic Network
  Get Policy distribution from Actor network
  Get Q values from the target Critic network
  for For each episode **do**
    Sample M actions from Policy distribution
    Actor cost calculation

$$cost_{ac}+ = \sum_{m=1}^{M} Q_c(s_m, a_m) \nabla_{\theta} log(\pi_a(s_m, a_m))$$

    Compute correction (TD error) for actions

$$\delta_{1:M} = R(S) + \gamma \sum_{m=1}^{M-1} Q_{w^-}(s_{m+1}, a_{m+1}) - \sum_{m=1}^{M-1} Q_{w}(s_m, a_m)$$

    Compute cost and reward R(S)

$$cost_{cri}+ = \delta_{1:M} \sum_{\substack{m=1, \\ a \in A}}^{M} \nabla_{w} Q_{w}(s_m, a_m)$$

  end for
  Compute episodic cost$_{ac}$, cost$_{cri}$ and episodic reward of actor and critic
  if episodic reward improves **then**
    update summary by picking top $|S|$ subshots
  end if
  if For each mini batch **then**
    Back-propagate pseudo batch cost$_{ac}$, and cost$_{cri}$
  end if
end for