Multiple Pairwise Ranking Networks for Personalized Video Summarization - Supplementary Material

Yassir Saquil[1] Da Chen*[2] Yuan He[2] Chuan Li[3] Yong-Liang Yang[1]

[1]University of Bath [2]Alibaba Turing Lab [3]Lambda Labs

In this document, we provide further quantitative and qualitative results. In quantitative results, we focus on extending the main paper results using Kendall's τ and Spearman's ρ evaluation metrics with additional baselines and feature embedding. Also, we provide an additional ablation study on the influence of the number of preferences on Multi-ranker performance and the relevance of personalized summarizations in TVSum dataset. In qualitative results, we provide an additional visualization on TVSum dataset and show more details on the user study pipeline and interface.

1. Evaluation Metric

Due to page limitation, we only focused in the main paper on using the Kendall's τ [4] rank correlation coefficient to evaluate our method and compare it with the state-of-the-art methods. In this document, we provide the corresponding evaluations of the main paper experiments using Kendall's τ [4] and Spearman's ρ [11] correlation coefficients with additional baselines and feature embedding.

2. Quantitative Results

2.1. Comparison with State-of-the-art Methods

Baselines: We train dppLSTM [9], VASNet [2], DR-DSN [10], SUM-FCN [7], SUM-GAN [5], SUM-GAN-AAE [1] on TVSum and SumMe using the feature embeddings described in [2]. For FineGym, we train these models using our feature processing described in the Implementation Details Subsection in the main paper. In the case of Multi-ranker $\{R_i\}$ and Standard ranker R, we set $N = 2000, B = 128, \lambda = 0.5$ and we train them for 1 epoch according to Ablation Study 2.2 findings using the both previously mentioned feature embeddings, while the human baseline is defined as in the Experimental Protocol Subsection in the main paper. Additionally, we report CSNet+GL+RPE [3] and SumGraph [6] original results on TVSum since no implementation is publicly available. We note that unless it is mentioned otherwise, all these models are trained and tested on the same sets using their default hyperparameters.

Following the Experimental Protocol, we compare our Multi-ranker with these baselines on the global summarization task using the test set of each split in each benchmark. We report in Table 2, the mean and standard deviation of Kendall's τ and Spearman's ρ coefficients on the test sets.

The remarks on the methods' performance are similar to the ones drawn in the main paper with the additional observation that the performance of Standard ranker, Multi-ranker and VASNet [2] is better than the human baseline using Spearman's ρ coefficient. We note that our method and the baselines have failed to generalize on SumMe dataset while using two different feature embeddings.

2.2. Ablation Study

We conduct ablation studies to tune the hyperparameters of our model and investigate their impact on performance. We first tune the mini-batch size B and number of pairwise comparisons N by training a Standard ranker R, then we tune the hyperparameter λ by training a Multi-ranker $\{R_i\}$.

*corresponding author

Methods	TVSum	SumMe	FineGym
Human baseline	0.1755 ± 0.0227	0.1796 ± 0.0107	-
VASNet [2]	0.1690 ± 0.0189	0.0224 ± 0.0289	0.3739 ± 0.0295
dppLSTM [9]	0.0298 ± 0.0284	-0.0256 ± 0.0214	-0.0267 ± 0.0075
DR-DSN$_{60}$ [10]	0.0169 ± 0.0508	**0.0433 ± 0.0386**	0.1457 ± 0.1108
DR-DSN$_{2000}$ [10]	0.1516 ± 0.0373	-0.0159 ± 0.0305	NaN
SUM-FCN [7]	0.0107 ± 0.0032	0.0080 ± 0.0091	-
SUM-GAN [5]	-0.0535 ± 0.0340	-0.0095 ± 0.0410	-
SUM-GAN-AAE [1]	-0.0472 ± 0.0299	-0.0180 ± 0.0558	-
CSNet+GL+RPE [3]	0.0700 ± 0.0000	-	-
SumGraph [6]	0.0940 ± 0.0000	-	-
Standard ranker	**0.1758 ± 0.0243**	0.0108 ± 0.0407	0.3792 ± 0.0335
Multi-ranker$_8$	0.1750 ± 0.0296	-0.0097 ± 0.0405	-
Multi-ranker$_4$	0.1736 ± 0.0266	-0.0006 ± 0.0454	**0.3928 ± 0.0291**
Multi-ranker$_2$	0.1630 ± 0.0209	0.0172 ± 0.0198	-
Standard ranker*	0.1750 ± 0.0299	0.0093 ± 0.0214	0.3792 ± 0.0335
Multi-ranker$^*_8$	0.1694 ± 0.0308	-0.0003 ± 0.0283	-
Multi-ranker$^*_4$	0.1666 ± 0.0340	0.0206 ± 0.0178	**0.3928 ± 0.0291**
Multi-ranker$^*_2$	0.1578 ± 0.0281	-0.0016 ± 0.0389	-

Table 1. The mean and standard deviation of Kendall's τ coefficient [4] per each method and dataset. Multi-ranker$_P$ denotes the trained model with P preferences $\mathcal{P} = \{1 \cdots P\}$ and DR-DSN$_{ep}$ denotes the trained model for ep epochs. the best performing model is highlighted, the symbol '-' means that the results are not available and the symbol '*' denotes that the model is trained using our feature processing on all 3 datasets. We note that FineGym has only 4 fixed preferences and 1 reference summary.

Methods	TVSum	SumMe	FineGym
Human baseline	0.2019 ± 0.0260	0.1863 ± 0.0111	-
VASNet [2]	0.2221 ± 0.0247	0.0255 ± 0.0358	0.4577 ± 0.0359
dppLSTM [9]	0.0385 ± 0.0365	-0.0311 ± 0.0249	-0.0326 ± 0.0092
DR-DSN$_{60}$ [10]	0.0227 ± 0.0666	**0.0501 ± 0.0470**	0.1784 ± 0.1357
DR-DSN$_{2000}$ [10]	0.1980 ± 0.0492	-0.0218 ± 0.0374	NaN
SUM-FCN [7]	0.0142 ± 0.0042	0.0096 ± 0.0111	-
SUM-GAN [5]	-0.0701 ± 0.0444	-0.0122 ± 0.0504	-
SUM-GAN-AAE [1]	-0.0620 ± 0.0388	-0.0226 ± 0.0695	-
CSNet+GL+RPE [3]	0.0910 ± 0.0000	-	-
SumGraph [6]	0.1380 ± 0.0000	-	-
Standard ranker	**0.2301 ± 0.0320**	0.0137 ± 0.0505	0.4642 ± 0.0408
Multi-ranker$_8$	0.2289 ± 0.0388	-0.0119 ± 0.0502	-
Multi-ranker$_4$	0.2270 ± 0.0354	-0.0005 ± 0.0564	**0.4808 ± 0.0354**
Multi-ranker$_2$	0.2133 ± 0.0281	0.0212 ± 0.0244	-
Standard ranker*	0.2288 ± 0.0393	0.0115 ± 0.0264	0.4642 ± 0.0408
Multi-ranker$^*_8$	0.2220 ± 0.0411	-0.0004 ± 0.0349	-
Multi-ranker$^*_4$	0.2187 ± 0.0455	0.0257 ± 0.0221	**0.4808 ± 0.0354**
Multi-ranker$^*_2$	0.2075 ± 0.0375	-0.0016 ± 0.0479	-

Table 2. The mean and standard deviation of Spearman's ρ coefficient [11] per each method and dataset. Multi-ranker$_P$ denotes the trained model with P preferences $\mathcal{P} = \{1 \cdots P\}$ and DR-DSN$_{ep}$ denotes the trained model for ep epochs. The best performing model is highlighted and the symbol '-' means that the results are not available and the symbol '*' denotes that the model is trained using our feature processing on all 3 datasets. We note that FineGym has only 4 fixed preferences and 1 reference summary.

We set $B \in \{32, 128\}$, $N \in \{2000, 5000\}$ and follow the Experimental Protocol in training Standard ranker R using 4-fold cross-validation on the non-test set for each split. As a result, we train 20 models for 50 epochs and report in Figure 1 the mean and standard deviation of Spearman's ρ coefficients on the validation sets along with the corresponding human baseline. Similar to the results in the main paper, we conclude from the plots in Figure 1 that early epochs are enough to obtain an optimal ranker and further training leads to overfitting on training set, also the model is not sensitive to the hyperparameters B and N at optimal epochs.

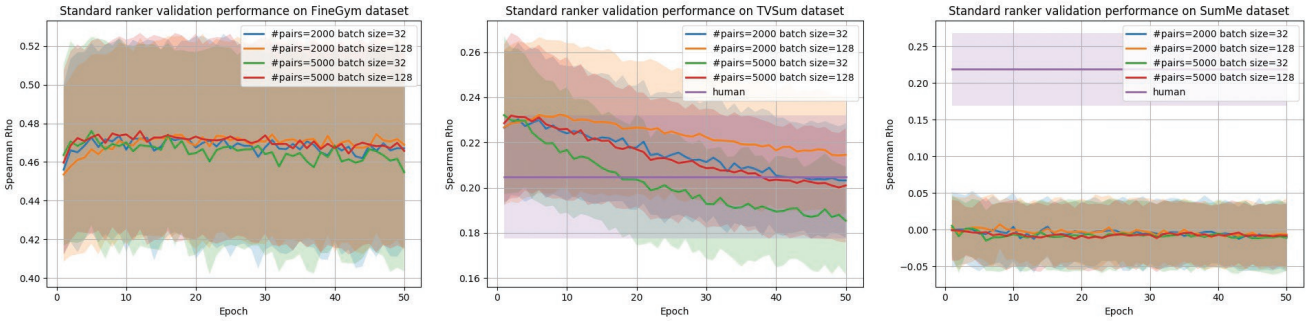

Figure 1. The mean Spearman's ρ coefficient per each setting and dataset (solid line) surrounded by a shaded area of range $[-std, std]$, with std the Spearman's ρ coefficient standard deviation and $\#pairs$ the number of pairwise comparisons.

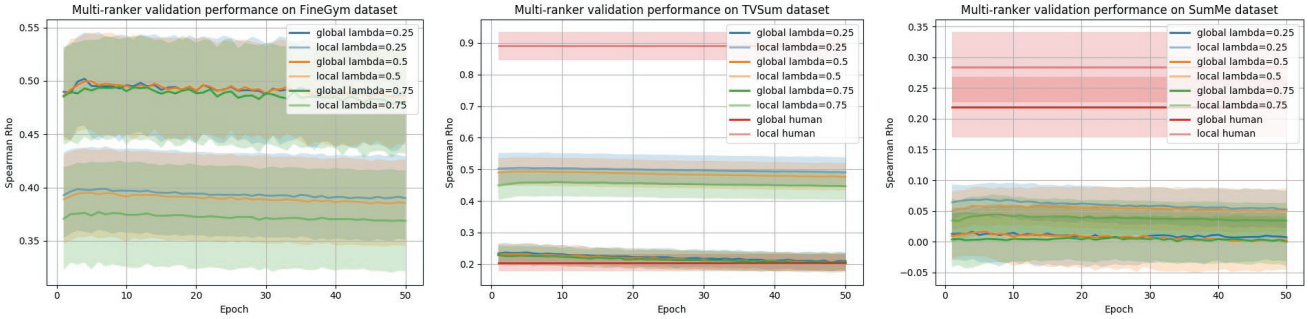

Figure 2. The mean Spearman's ρ coefficient per each setting and dataset (solid line) surrounded by a shaded area of range $[-std, std]$, with std the Spearman's ρ coefficient standard deviation and $global$, $local$ denoting global and local summarization coefficients.

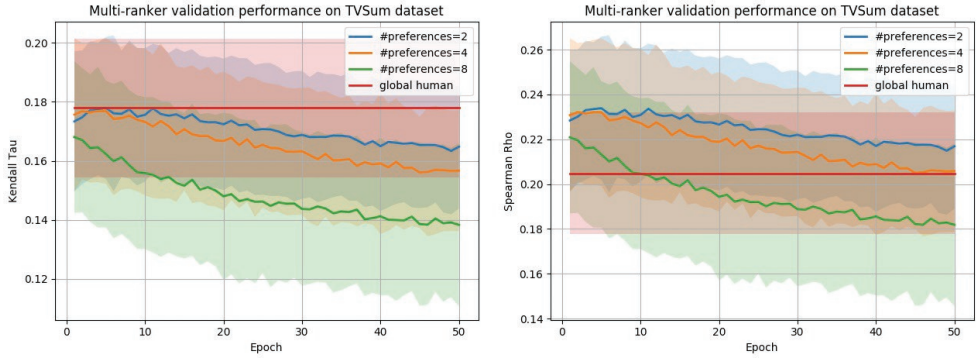

Figure 3. The mean Kendall's τ and Spearman's ρ coefficients per each setting in TVSum dataset (solid line) surrounded by a shaded area of range $[-std, std]$, with std the correlation coefficient standard deviation and $\#preferences$ the number of preferences

We set $\mathcal{P} = \{1 \dots 4\}$, $B = 128$, $N = 2000$, $\lambda \in \{0.25, 0.5, 0.75\}$ and follow the Experimental Protocol in training Multi-ranker $\{R_i\}$ using 4-fold cross validation on the non-test set for each split. We report in Figure 2 the mean and standard deviation of local and global Spearman's ρ coefficients of the validation sets along with the corresponding human baselines. Similarly to the main paper, we notice that λ variation does not have a significant impact on the global summarization performance, while the mean local correlation coefficient decreases when λ puts more emphasis on global summarization.

We also investigated in Figure 3 the influence of the number of preferences on the global summarization of Multi-ranker model in TVSum dataset with the following setting: $B = 128, N = 2000, \lambda = 0.5$. We remark that in early epochs the differences in the mean correlation coefficients are minimal and tend to increase as the models start to overfit on the training set. As the the number of preferences increases, the training set size increases which explains the increase in the overfitting range. This enforces the model selection suggestion in early epochs for an optimal performance.

2.3. Relevance of Local and Personalized Summarizations

The aim of this experiment is to demonstrate that Multi-ranker provides more preference specific summaries than the Standard ranker. Although the Standard ranker is trained on GT summaries to generate a global summary, testing it using personalized reference summaries sets a lower bound baseline for Multi-ranker.

We set $N = 2000, B = 128, \lambda = 0.5, \mathcal{P} = \{1 \dots 4\}$ and train Multi-ranker and Standard ranker for 1 epoch. Following the Experimental Protocol, we test these models on the personalized summarization task using the test set of each split in TVSum and FineGym datasets. We report in Table 3, the mean and standard deviation of personalized Kendall's τ and Spearman's ρ coefficients on the test sets in TVSum dataset. We also report in Table 4, the mean and standard deviation of personalized Kendall's τ and Spearman's ρ coefficients on the test sets in FineGym dataset. Similarly to the main paper, we notice that the more general the generated summary is, the more the Multi-ranker correlation coefficient is similar to the Standard ranker. Also, the more local the generated summary is, the wider the disparity between Standard ranker and Multi-ranker correlation coefficients. Moreover, we note that Multi-ranker encourages implicitly diverse summaries by including the top ranked segments with respect to each preference in the desired summary. Standard ranker has no guarantee that the top ranked segments of every preference is included in the desired summary.

Pref. set	Multi-ranker	Standard ranker
{1}	0.3571 ± 0.0476	0.1038 ± 0.0420
{2}	0.2549 ± 0.0640	0.1683 ± 0.1034
{3}	0.3314 ± 0.0264	0.2412 ± 0.0497
{4}	0.4675 ± 0.0643	-0.3781 ± 0.0649
{1,2}	0.3547 ± 0.0773	0.1750 ± 0.0299
{1,3}	0.3550 ± 0.0426	0.0870 ± 0.0584
{1,4}	0.2070 ± 0.0487	-0.1952 ± 0.0344
{2,3}	0.3595 ± 0.0441	0.0161 ± 0.0471
{2,4}	0.0589 ± 0.0841	-0.2788 ± 0.0378
{3,4}	0.1338 ± 0.0749	-0.0754 ± 0.0938
{1,2,3}	0.3711 ± 0.0582	0.1750 ± 0.0299
{1,2,4}	0.1100 ± 0.0661	0.1750 ± 0.0299
{1,3,4}	0.1183 ± 0.0561	0.0870 ± 0.0584
{2,3,4}	0.1815 ± 0.0524	0.0161 ± 0.0471
{1,2,3,4}	0.1666 ± 0.0340	0.1750 ± 0.0299

Pref. set	Multi-ranker	Standard ranker
{1}	0.4573 ± 0.0608	0.1364 ± 0.0558
{2}	0.3297 ± 0.0907	0.2179 ± 0.1376
{3}	0.4235 ± 0.0357	0.3111 ± 0.0639
{4}	0.6004 ± 0.0832	-0.4866 ± 0.0834
{1,2}	0.4597 ± 0.1017	0.2288 ± 0.0393
{1,3}	0.4629 ± 0.0564	0.1143 ± 0.0755
{1,4}	0.2678 ± 0.0625	-0.2540 ± 0.0440
{2,3}	0.4631 ± 0.0578	0.0269 ± 0.0619
{2,4}	0.0782 ± 0.1089	-0.3593 ± 0.0490
{3,4}	0.1800 ± 0.0982	-0.0920 ± 0.1218
{1,2,3}	0.4869 ± 0.0782	0.2288 ± 0.0393
{1,2,4}	0.1455 ± 0.0874	0.2288 ± 0.0393
{1,3,4}	0.1583 ± 0.0726	0.1143 ± 0.0755
{2,3,4}	0.2405 ± 0.0699	0.0269 ± 0.0619
{1,2,3,4}	0.2187 ± 0.0455	0.2288 ± 0.0393

Table 3. The mean and standard deviation Kendall's τ (left) and Spearman's ρ (right) coefficients of Multi-ranker and Standard ranker for each possible preference set $\mathcal{P}_s$ (Pref. set) in TVSum dataset.

Pref. set	Multi-ranker	Standard ranker
{1}	0.1086 ± 0.0164	0.0254 ± 0.0122
{2}	0.3568 ± 0.0376	0.2727 ± 0.0241
{3}	0.3985 ± 0.0097	0.2978 ± 0.0133
{4}	0.3007 ± 0.0283	0.1504 ± 0.0840
{1,2}	0.3928 ± 0.0291	0.3792 ± 0.0335
{1,3}	0.3747 ± 0.0245	0.2829 ± 0.0325
{1,4}	0.2359 ± 0.0286	0.1200 ± 0.0582
{2,3}	0.4093 ± 0.0135	0.3925 ± 0.0183
{2,4}	0.3707 ± 0.0218	0.2781 ± 0.0387
{3,4}	0.3966 ± 0.0117	0.2996 ± 0.0201
{1,2,3}	0.3928 ± 0.0291	0.3792 ± 0.0335
{1,2,4}	0.3928 ± 0.0291	0.3792 ± 0.0335
{1,3,4}	0.3747 ± 0.0245	0.2829 ± 0.0325
{2,3,4}	0.4093 ± 0.0135	0.3925 ± 0.0183
{1,2,3,4}	0.3928 ± 0.0291	0.3792 ± 0.0335

Pref. set	Multi-ranker	Standard ranker
{1}	0.1329 ± 0.0202	0.0311 ± 0.0150
{2}	0.4366 ± 0.0455	0.3337 ± 0.0291
{3}	0.4879 ± 0.0116	0.3646 ± 0.0164
{4}	0.3682 ± 0.0347	0.1841 ± 0.1029
{1,2}	0.4808 ± 0.0354	0.4642 ± 0.0408
{1,3}	0.4588 ± 0.0298	0.3464 ± 0.0397
{1,4}	0.2888 ± 0.0349	0.1469 ± 0.0711
{2,3}	0.5010 ± 0.0163	0.4805 ± 0.0221
{2,4}	0.4538 ± 0.0264	0.3404 ± 0.0472
{3,4}	0.4856 ± 0.0140	0.3668 ± 0.0244
{1,2,3}	0.4808 ± 0.0354	0.4642 ± 0.0408
{1,2,4}	0.4808 ± 0.0354	0.4642 ± 0.0408
{1,3,4}	0.4588 ± 0.0298	0.3464 ± 0.0397
{2,3,4}	0.5010 ± 0.0163	0.4805 ± 0.0221
{1,2,3,4}	0.4808 ± 0.0354	0.4642 ± 0.0408

Table 4. The mean and standard deviation Kendall's τ (left) and Spearman's ρ (right) coefficients of Multi-ranker and Standard ranker for each possible preference set $\mathcal{P}_s$ (Pref. set) in FineGym dataset.

3. FineGym Dataset Preparation

FineGym [8] is a fine-grained action recognition dataset that provides action level temporal annotations for YouTube gymnasium videos where we manage to obtain 156 videos with fine-grained annotations. Since the videos are of long duration, we only used the top 50 alphanumerical sorted videos for experiments purpose with the following Youtube IDs listed in Table 5.

index	ID	index	ID
0	0LtLS9wROrk	25	BQhX6F3gpp8
1	0jqn1vxdhls	26	CUuRl0Bwbbc
2	1Fdwuy2V9EY	27	CkNz9ZIQmZI
3	1JsRXIoR3C0	28	DG7upbhB1bI
4	1rkcLEAMTpw	29	DTedv-hhHU4
5	1sPWceVH4e8	30	DyZ2qj6x1UE
6	26Y8BsNiiL8	31	E-1hTCGUCts
7	2Qw1e4-nWrk	32	E3AHJ6-QS8M
8	2giTb7IpJDU	33	EFvDqOF1sCk
9	2pBxfMAIaXY	34	EKb2MMJSoeI
10	3PywNMDCvNQ	35	EizeIkrDtQk
11	3hD6S3NFaUU	36	FEWTGYiWEaQ
12	4mzbybgzoJo	37	FNEeMvHmuEw
13	5Bfx6Wz3KKs	38	GCyW1jPfWdg
14	5X85zLeLmks	39	GEjRXo8Dvwc
15	5cuxEEKyth8	40	GF8D7Dx2mhs
16	6fqDZHFr2yo	41	GbUQwE9N3aM
17	8N3CBVAft40	42	Gng3ezTNGog
18	8WGrvkd6ZEU	43	HubzmiNVVXs
19	8YSDFKGwP4U	44	I1OMnMSk1Lg
20	8tdGvTDHmzc	45	I749IS2IeFY
21	9Mtbac7g15I	46	ICL5k84viw0
22	A0xAXXysHUo	47	IP96HTdCvWs
23	AZ4wWG6Rcak	48	IWFgo-tEEs8
24	BLzs5opw8uM	49	IZy40LwRdbM

Table 5. The video index in FineGym dataset and its corresponding Youtube ID.

4. Qualitative Results

4.1. Visualization

In this subsection, we present an example of global video summarization in TVSum dataset using Standard ranker, Multi-ranker and VASNet [2] and an example of global and local video summaries in FineGym dataset. In Figure 4, we illustrate the frame-level GT importance scores of video 9 in TVSum and highlight the top-ranked 15% predicted frames using Standard ranker, Multi-ranker and VASNet [2] models. In Figure 5, we illustrate the segment-level GT importance scores of a video in FineGym and highlight the top-ranked 15% global, local and personalized predicted segments with respect to *Floor Exercise* and *Balance Beam* preferences and (*Floor Exercise*, *Balance Beam*) preference set using Multi-ranker model in global, local and personalized summarization tasks respectively. In each illustrated summary, we visualize 6 sampled frames from the highlighted predicted frames.

4.2. User Study

To quantify the perceived quality of our Multi-ranker method and the impact from the user perspective of each Multi-ranker task, *i.e.* local, personalized and global summarizations, we performed a user study based on 40 subjects that are asked to provide their opinions about 4 main comparison scenarios. For each scenario, the subjects perform at least two runs with different videos or different selections of the preferences.

In detail, we focused on FineGym videos with its predefined 4 preferences (*Vault, Floor Exercise, Balance Beam, Uneven Bars*) and asked the subject to watch for each scenario run the original video and two associated summaries while selecting

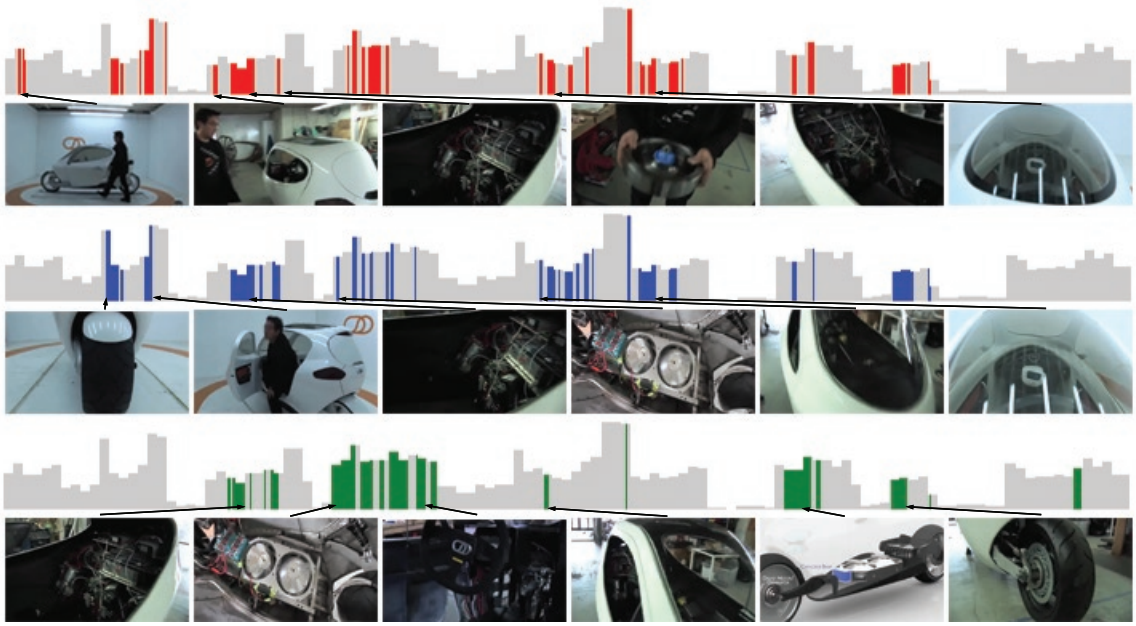

Figure 4. Frame-level GT importance scores (gray), Standard ranker summary (red), Multi-ranker summary (blue), VASNet summary (green) for test video 9 from TVSum dataset.

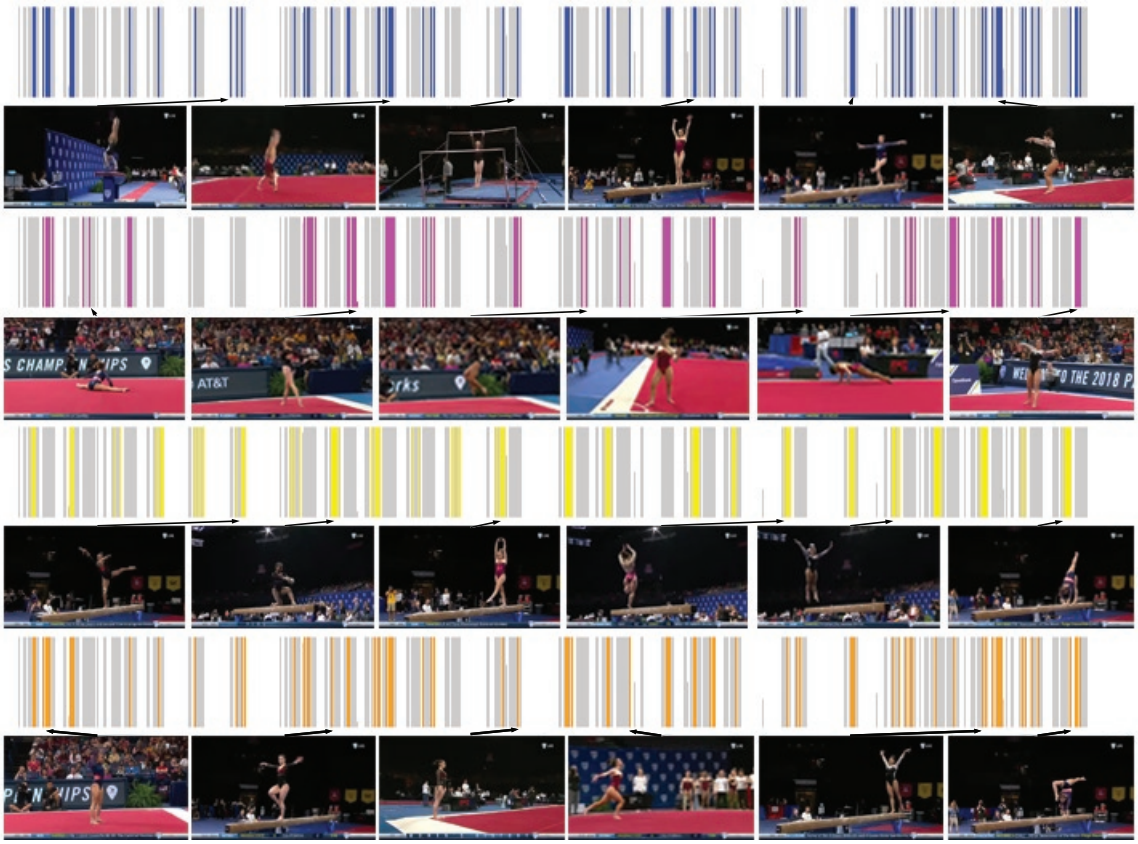

Figure 5. Segment-level GT importance scores (gray), Multi-ranker global summary (blue), Multi-ranker local summary for *Floor Exercise* preference (magenta), Multi-ranker local summary for *Balance Beam* preference (yellow) and Multi-ranker personalized summary (orange) for the preference set (*Floor Exercise*, *Balance Beam*) for the test video with ID '0LtLS9wROrk' from FineGym dataset.

a preference for local summary or a set of preferences for personalized summary and then submit his answer to the scenario question. The first scenario is a subjective comparison between Multi-ranker and VASNet [2] summaries, while the remaining scenarios are comparisons between local, personalized and global summaries in term of usability and satisfaction from the user perspective. The corresponding question for each scenario is defined as follows: (1) Is the quality of Multi-ranker summary better, equal or worse than VASNet [2] summary? (2) Is local summary more content specific than global summary or not? (3) Does personalized summary provide better user control to achieve satisfactory result than global summary or not? (4) Does personalized summary provide better user control to achieve satisfactory result than local summary or not?

We generate the video summaries as follows: we build a dataset for every scenario by repeating the following process. In scenario 1 and 2, we sample a video from FineGym and randomly select a shot that has k segments with non-zero GT importance scores and $k \neq 0$. In scenario 3 and 4, we sample a video from FineGym and randomly select a shot that has segments belonging to at least two different preferences and $k \neq 0$. Afterwards, we predict for each shot the importance scores of its segments (w.r.t each preference in scenario 2,3,4). Then we select the top ranked segments to build the summary.

The Figures 6, 7, 8, 9 illustrate the user study interface for each scenario with the required steps to follow by each participant. After the submission of the participants answers, they are mapped according to their order to one of the following canonical options; Strongly Disagree, Mildly Disagree, Similar, Mildly Agree, Strongly Agree. Then, for each scenario, we calculate the percentage of each canonical option defined as the number of the option occurrence divided by the total scenario annotations to obtain the user study results shown in the main paper.

References

[1] Evlampios E. Apostolidis, Eleni Adamantidou, Alexandros I. Metsai, Vasileios Mezaris, and Ioannis Patras. Unsupervised video summarization via attention-driven adversarial learning. In *MultiMedia Modeling*, 2020. 1, 2

[2] Jiri Fajtl, Hajar Sadeghi Sokeh, Vasileios Argyriou, Dorothy Monekosso, and Paolo Remagnino. Summarizing videos with attention. In *ACCV*, 2018. 1, 2, 5, 7, 8

[3] Yunjae Jung, Donghyeon Cho, Sanghyun Woo, and In So Kweon. Global-and-local relative position embedding for unsupervised video summarization. In *ECCV*, 2020. 1, 2

[4] Maurice G Kendall. The treatment of ties in ranking problems. *Biometrika*, 33(3):239–251, 1945. 1, 2

[5] Behrooz Mahasseni, Michael Lam, and Sinisa Todorovic. Unsupervised video summarization with adversarial LSTM networks. In *CVPR*, 2017. 1, 2

[6] Jungin Park, Jiyoung Lee, Ig-Jae Kim, and Kwanghoon Sohn. SumGraph: Video summarization via recursive graph modeling. In *ECCV*, 2020. 1, 2

[7] Mrigank Rochan, Linwei Ye, and Yang Wang. Video summarization using fully convolutional sequence networks. In *ECCV*, 2018. 1, 2

[8] Dian Shao, Yue Zhao, Bo Dai, and Dahua Lin. FineGym: A hierarchical video dataset for fine-grained action understanding. In *CVPR*, 2020. 5

[9] Ke Zhang, Wei-Lun Chao, Fei Sha, and Kristen Grauman. Video summarization with long short-term memory. In *ECCV*, 2016. 1, 2

[10] Kaiyang Zhou, Yu Qiao, and Tao Xiang. Deep reinforcement learning for unsupervised video summarization with diversity-representativeness reward. In *AAAI*, 2018. 1, 2

[11] Daniel Zwillinger and Stephen Kokoska. *CRC standard probability and statistics tables and formulae*. CRC Press, 1999. 1, 2

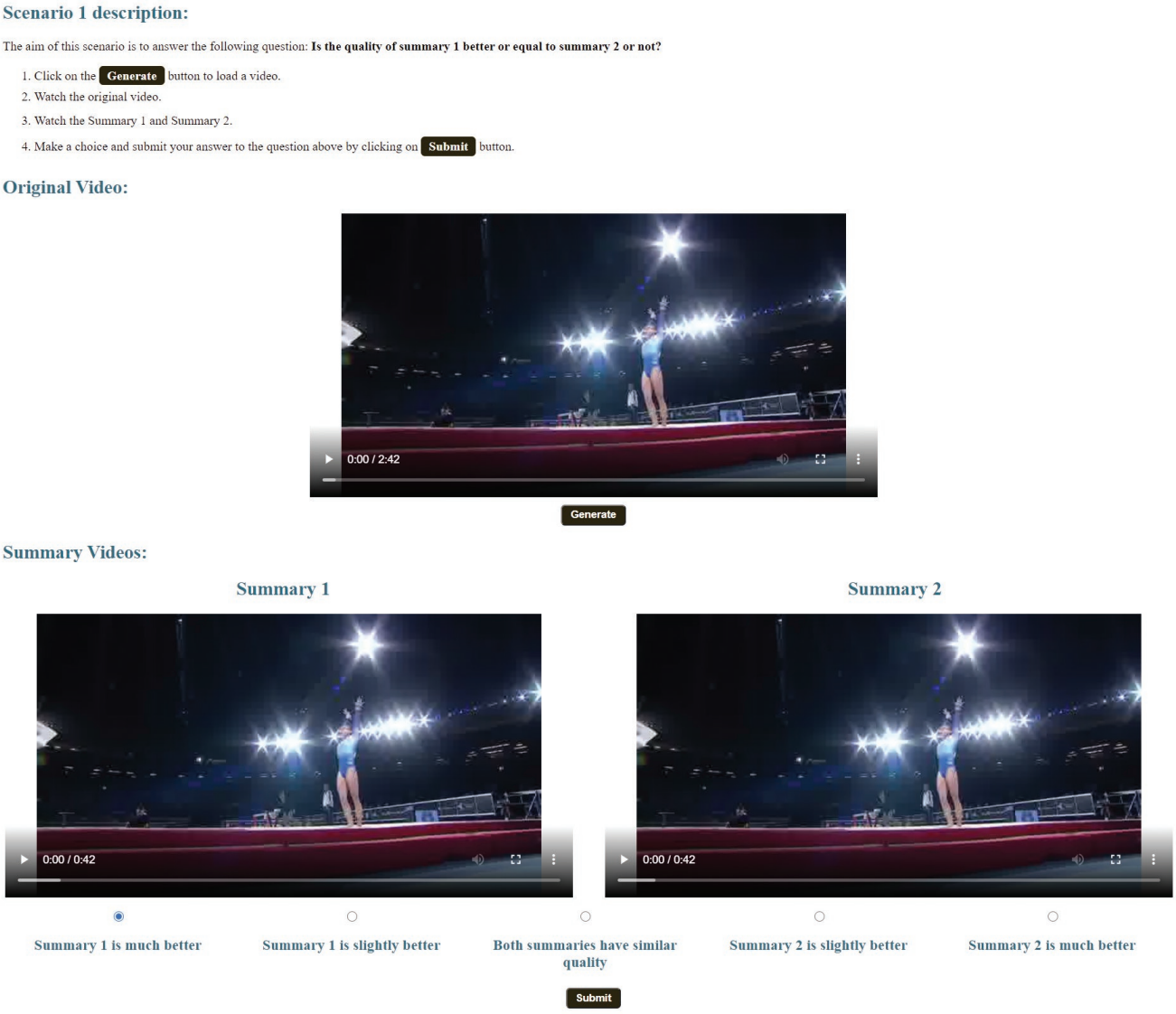

Figure 6. Scenario 1 user study interface where the Multi-ranker and VASNet [2] summaries are randomly assigned to summary 1 and 2 for the user study fairness.

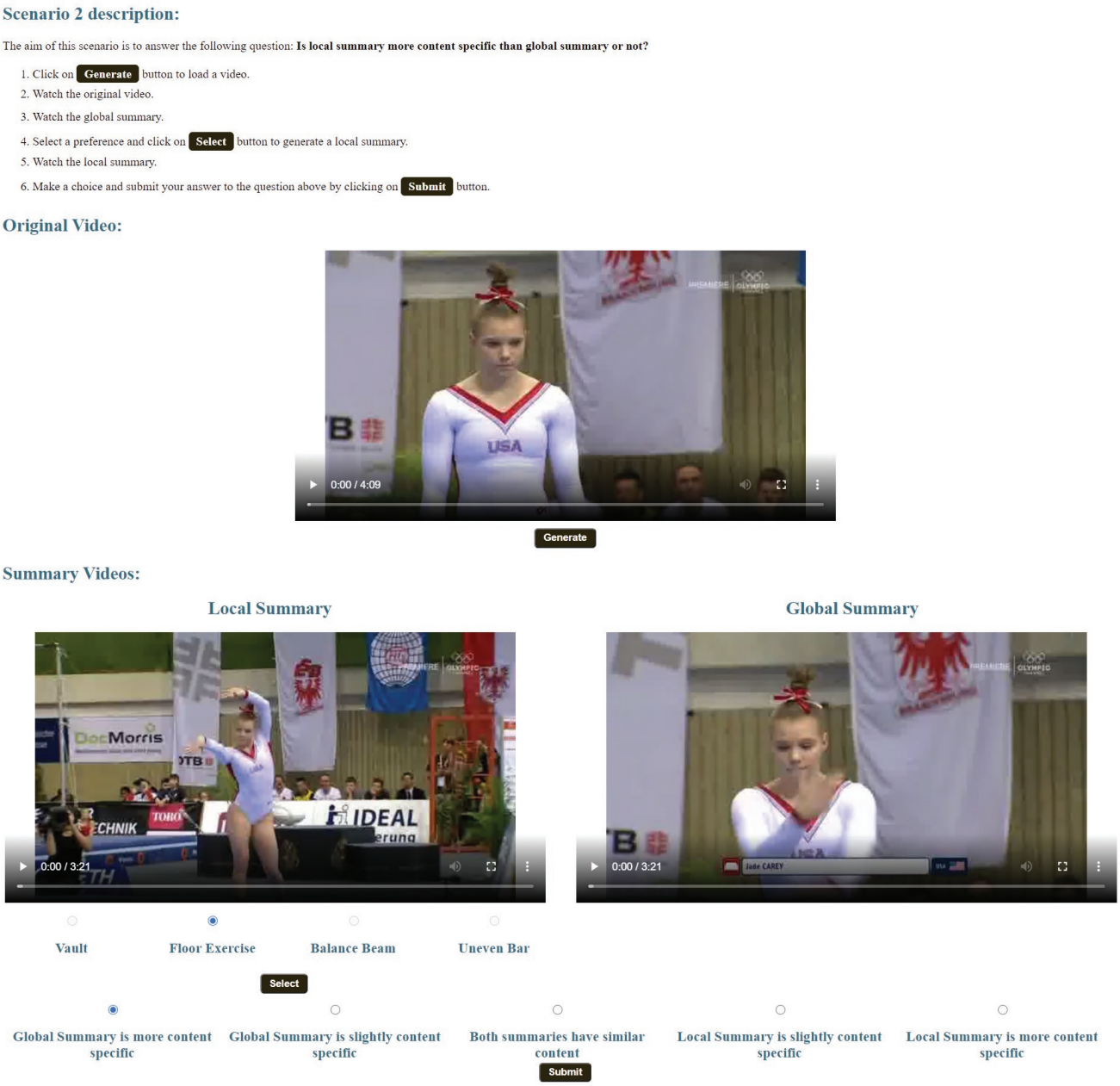

Figure 7. Scenario 2 user study interface where the participant selects one preference to generate a local summary.

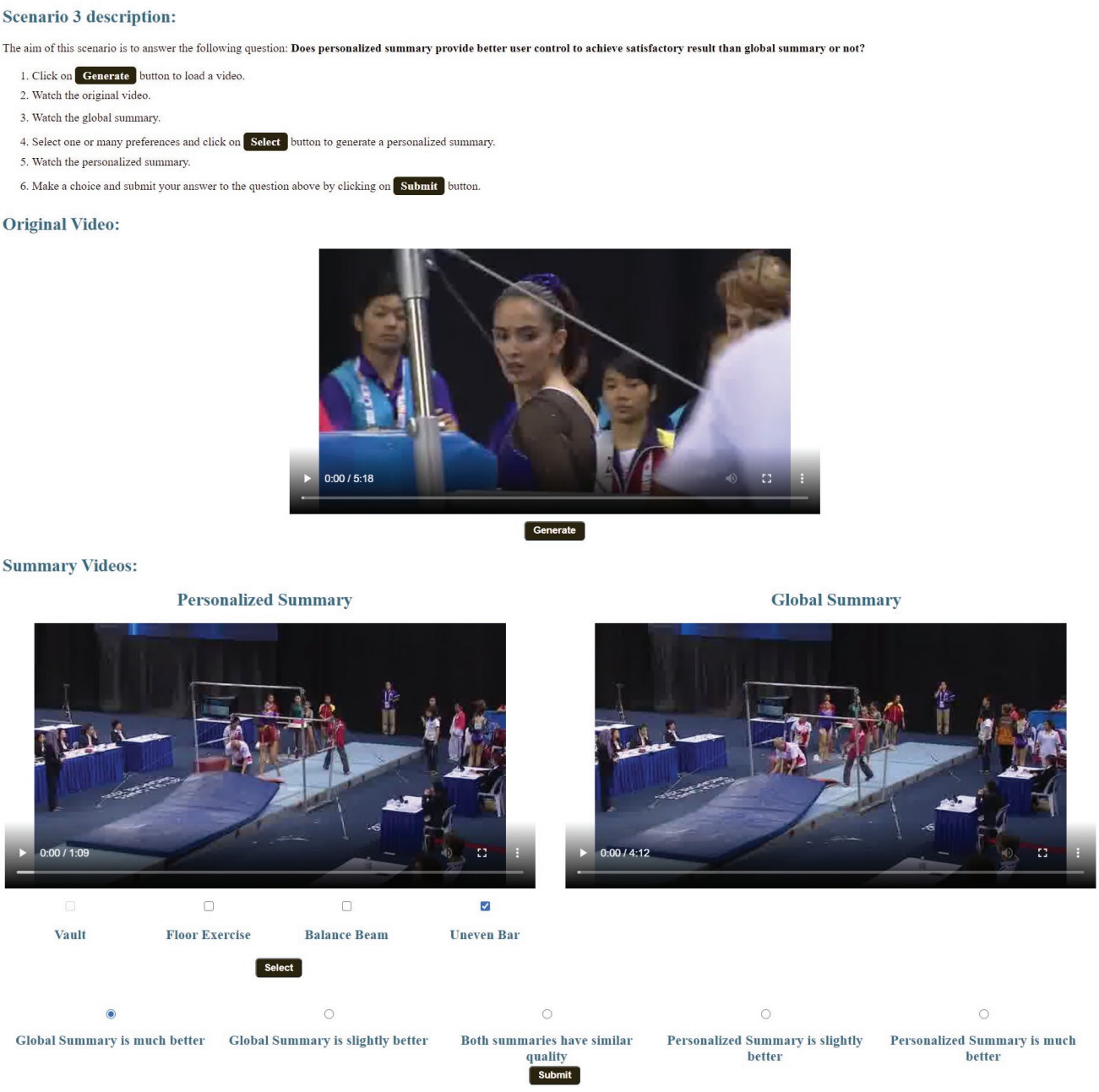

Figure 8. Scenario 3 user study interface where the participant selects a set of preferences to generate a personalized summary.

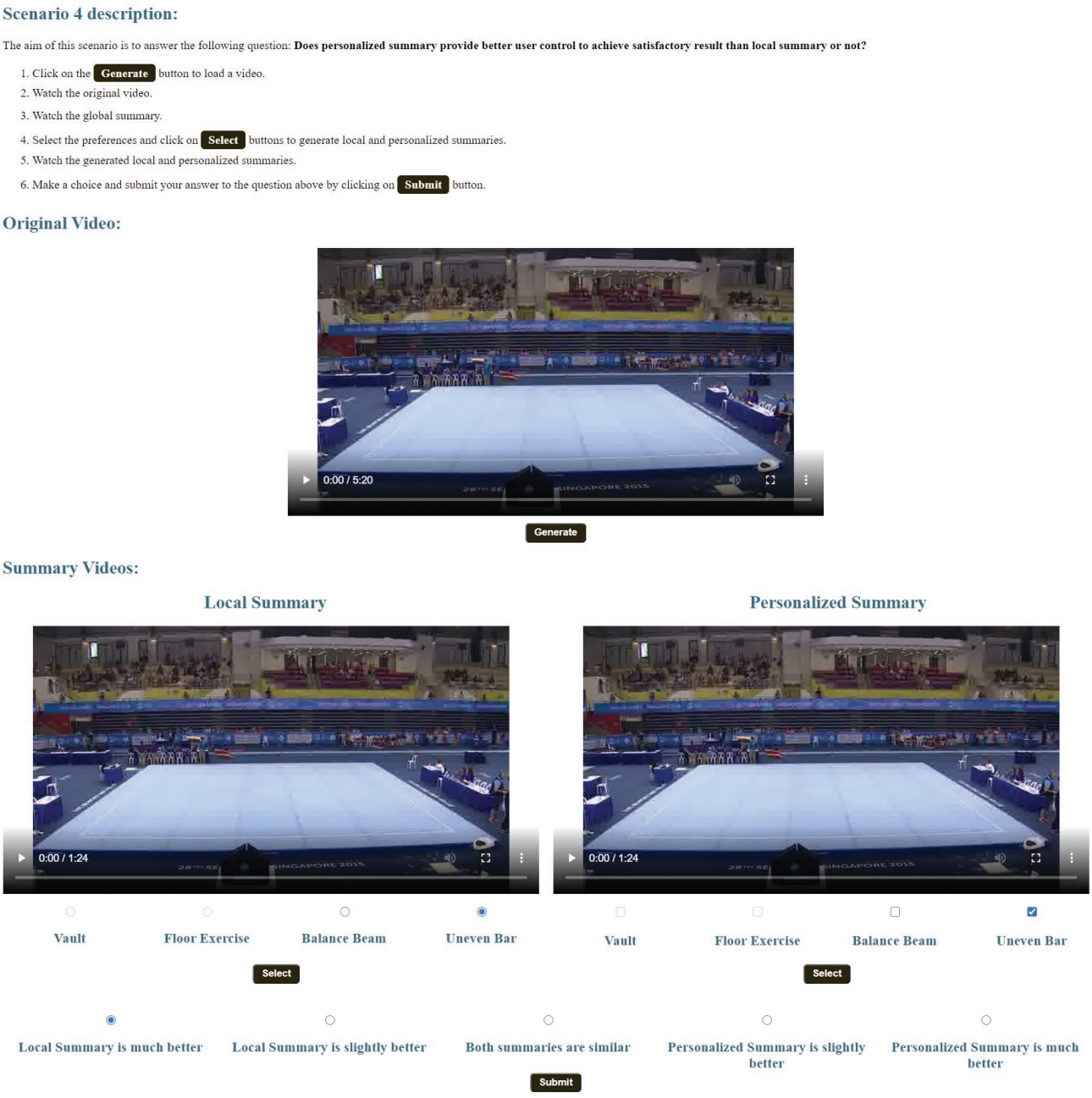

Figure 9. Scenario 4 user study interface where the participant selects one preference to generate a local summary and a set of preferences to generate a personalized summary.