Supplementary material of "Multimodal-based and Aesthetic-guided Narrative Video Summarization"

I. Hyperparameter setting

Choosing the K parameter for K-means: In our setting, we employ a popular approach known as elbow method[1] to determine the optimal value of K. The basic idea behind this method is that it plots the various values of cost. In practical, if the user has some prior knowledge on how many groups to be classified, K can be determined by the user directly. For example, a user can search from Wikipedia that the *Planet Earth Season I Episode 2* mainly introduces 11 kinds of plants and animals in chronological order.

II. Controlling the length of summary

In our setting, we make the length of the generated summary as close as possible to the desired length of the user without compromising the content integrity and aesthetic quality. Specifically, our method firstly estimates the duration of the summary output by the shots aesthetics assembly module and judges whether the duration is consistent with the user-specified duration hyperparameter. Suppose the duration needs to be shortened, our method reduces the length or number of B-roll shots without compromising the integrity and aesthetic quality of the shot content. If the duration needs to be increased, the length or number of B-roll shots should be extended or increased. In addition, by adjusting the hyperparameters of each component included in the shot selection module, the user can also control the total duration of the selected shots, thereby indirectly controlling the duration of the summary. Particularly, in the multimodal-based shot selection module, we can modify the hyperparameters to control the length of final summary. (1) In the subtitle summarization component, we can control the value of K in the K-means clustering. The more chapters we clustered, the less shots contained in one chapter. If we keep other parameters the same (selecting the same amount of shots for each chapter), smaller K will result in shorter length of the summary. (2) In the key shot selection method, we can control the ratio of total length of selected shots to the original length. (3) In the highlight extraction component, we can control the value of x in Equ.(7) to control the length selected shots, which smaller x means shorter summary.

The advantage of our duration control method is that the content integrity and aesthetic quality will not lose due to the strong constraints. However, the limitation is that it cannot guarantee that the duration of the final summary is precisely what the user expected, and only getting as close as possible. We expect to address this issue by incorporating user interactions in subsequent studies.

[1]https://www.geeksforgeeks.org/elbow-method-for-optimal-value-of-k-in-kmeans/

III. Dataset Samples

For better evaluating the availability and advancement of our method, we followed previous works [1, 2] and collect different types of narrative videos to form Movie-documentary (M-D) video repository, containing movies and narrative documentaries with different themes such as earth, travel, adventure, food, history and science. We further collect their corresponding subtitles from public website. In this section, we display some examples for movie and every theme of documentary in Fig. 1, which are *Planet Earth*, *Great British Railway Journeys*, *You vs. Wild*, *Two Greedy Italians*, *Wonders of the Universe*, *The Story of Wales* and *The Lord of the Rings* from left to right.

IV. Evaluation metric

In this section, we will further introduce our evaluation metrics in different experimental designs, especially the quantitative evaluation in TVsum [10] and summary attributes analysis.

(1) In the quantitative evaluations of TVsum, we adopted the commonly used Precision (P), Recall (R) and F-score to evaluate the agreement between the generated summaries by our method and other baselines following [4, 5, 6]. F-score is the harmonic mean of precision and recall expressed as the default F-score (F) result in percentages.

$$F = 2 \times \frac{precision \times recall}{precision + recall} \times 100, \tag{1}$$

where Precision and Recall are defined as follows:

$$P = \frac{length(gs \cap as)}{length(gs)}, R = \frac{length(gs \cap as)}{length(as)}, \tag{2}$$

where gs and as denotes the generated summary and corresponding annotated summary, respectively.

(2) As for IoU and mIoU, they are commonly used evaluation metrics in the video-text localizing areas. Followe [11, 12], we adopt $R@n$, IoU = μ and mIoU as the evaluation metrics. 1) Intersection over Union (IoU) here means intersection of time between the predicted segment and ground truth on union. For example, if the ground truth of the start and end time of the text are 5 and 10 seconds, while the predicted start and end time of that text are 8 and 13 seconds, the IoU of that example will be $\frac{10-8}{13-5} = \frac{2}{8} = 0.25$. 2) $R@n$, IoU = μ represents the percentage of testing samples which have at least one of the top n results with IoU larger than μ. For example, we can set n as 1, μ as 0.3, and the number of testing samples as 100. For one testing sample, the model will output lots of predicted start and end times in the video of that text. If the IoU of the best 1 predicted segment is larger than 0.3, we can count it as a success. The total number of

Fig. 1. Some examples of documentaries and movie in Movie-documentary (M-D) dataset.

Fig. 2. Some visual semantic matching examples on documentary and MPII Movie datasets. Procedural texts are listed top the thumbnails. Our results are outlined in green.

TABLE I
THE STATISTICS FOR SUMMARY ATTRIBUTES OF SUMMARIES GENERATED BY BASELINES AND MANVS-AUTO. THE IC, CON, TNS AND TAL RESPECTIVELY REPRESENT INFORMATION COVERAGE, CONSISTENCY, THE NUMBER OF SHOTS IN THE SUMMARY AND THE AVERAGE LENGTH OF SHOTS.

Method	TVsum				Movie				Documentary			
	IC	CON	TNS	TAL	IC	CON	TNS	TAL	IC	CON	TNS	TAL
Random	6/10	5/14 (35.71%)	16	3.7	25/32	17/212 (8.02%)	623	1.8	17/18	3/68 (4.41%)	234	1.9
DR [3]	6/10	5/15 (33.33%)	17	3.5	23/32	26/232 (11.21%)	681	1.6	17/18	5/97 (5.15%)	237	1.9
DR-Sup [3]	6/10	6/15 (40.00%)	16	3.7	23/32	24/208 (11.54%)	643	1.6	17/18	2/64 (3.13%)	229	2.0
VAS [4]	5/10	6/17 (35.29%)	14	4.3	24/32	29/218 (13.24%)	612	1.8	17/18	6/74 (8.11%)	202	2.2
DSN-AB [5]	5/10	6/18 (33.33%)	16	3.7	23/32	31/210 (14.76%)	580	1.9	16/18	6/82 (7.32%)	226	2.2
DSN-AF [5]	5/10	4/14 (28.57%)	17	3.5	22/32	28/223 (12.56%)	623	1.8	16/18	4/72 (5.56%)	236	1.9
HSA [6]	3/10	6/9 (66.67%)	7	9.4	16/32	77/114 (67.54%)	115	9.3	8/18	30/43 (69.77%)	30	17.1
FCN [7]	5/10	6/15 (40.00%)	16	3.8	24/32	35/206 (16.99%)	621	1.8	16/18	4/70 (5.71%)	232	1.9
HMT [8]	6/10	7/13 (53.85%)	15	4.0	18/32	72/165 (43.64%)	628	1.8	17/18	4/74 (5.41%)	224	1.9
VSN [9]	6/10	7/16 (43.75%)	16	3.8	22/32	21/211 (9.95%)	553	2.1	15/18	1/64 (1.56%)	214	2.1
MANVS-auto	8/10	12/12 (100%)	15	4.1	25/32	188/188 (100.00%)	256	4.2	17/18	56/56 (100.00%)	72	6.2

success examples divides the number of testing samples is the value of $R@n$, IoU = μ. 3) mIoU means the average $R@n$, IoU = μ when we set different values of μ. In this paper, following [13, 14], when reporting $R@n$, IoU = μ, we set n as 1 and $\mu \in \{0.3, 0.5\}$ and when reporting mIoU, we set $\mu \in \{0.1, 0.3, 0.5, 0.7\}$.

(3) As indicated in [15], the task of narrative video summarization is a highly subjective task, and simply utilizing the traditional quantitative evaluations seem not reasonable. Except for selecting significantly frames from the video, another character that a video summary should own is to bring a satisfactory visual experience to the audience.

Therefore, 100 participants were invited to watch the involved video summaries in the user study, and they were asked to rate every video by using a 7-point Likert scale (1 = poor, 7 = excellent), taking visual attraction (VA) and narrative completeness (NC) into account. The NC reflects viewers subjective perceptions of the narrative coherence and content integrity of the generated summaries, while the VA reflects viewers viewing experience. Specifically, building on research in both communication, psychology and the conventions of video editing [16], we conceptualized a video with a complete narrative as one that clearly and extensively: 1) switched scenes sequentially, 2) provided a coherent narrative, 3) included complete voiceover, and 4) contains as much essential content of the original video as possible. The NC is a subjective metric used to measure the degree to which the generated summary meets the above definition. Besides,

TABLE II
THE USER STUDY OF VIDEO SUMMARIES IN THE QUALITY OF VISUAL ATTRACTION (VA) AND NARRATIVE COMPLETENESS (NC) GENERATED BY STATE-OF-ART METHODS AND MANVS-AUTO.

Method	TVsum		Movie		Documentary	
	VA	NC	VA	NC	VA	NC
Random	3.4	3.6	1.9	1.7	2.1	1.9
DR [3]	3.3	3.5	1.8	1.9	1.7	2.0
DR-Sup [3]	3.4	3.3	2.1	1.8	1.9	2.1
VAS [4]	3.7	3.6	2.0	1.9	2.2	1.8
DSN-AB [5]	3.5	3.4	1.7	2.1	1.7	1.8
DSN-AF [5]	3.3	3.6	1.8	1.8	2.0	2.1
HSA [6]	4.3	4.4	3.5	3.6	3.7	3.7
FCN [7]	3.3	3.6	2.3	1.8	2.0	1.8
HMT [8]	3.7	3.7	3.8	3.6	1.9	1.9
VSN [9]	3.3	3.8	2.0	2.1	1.6	1.7
MANVS-auto	**4.7**	**4.6**	**5.0**	**5.2**	**5.1**	**4.7**

inspired from some research in both photographic and cinematography [16, 17], we define that the VA of a video refers to stimuli such as the beautiful senses, which is used to subjectively evaluate whether the generated video summary can bring visual enjoyments to viewers. Generally, the VA is mainly determined by cinematographic aesthetics such as the shots length, color continuity and shot stability and so on. The more that the generated video summaries conforms to cinematographic aesthetics, the more it can bring visual enjoyment to the audience, and the better the VA will be. Because of our designed cinematographic aesthetic constraints, our MANVS shows better results in the VA of generated summaries compared to other methods.

(4) In addition, we conduct some summary attributes analysis in experiments of comparison to traditional video summarization methods. We firstly compute the *information coverage* (IC), i.e. the ratio of the selected entities or plots to the total number of representative ones. This measures how much representative information is retained in the summary from the input video. For instance, we acquired a textual summary describing the main entities of a documentary from Wikipedia. According to the Episode introduction of Wikipedia, the representative entities of Planet Earth Season I Episode 2 includes 7 mountains and 11 animals, such as Matterhorn, guanacos etc. Subsequently, we asked participants to watch the summary and determine whether there exist these entities in the video that the Wikipedia described. Participants answered with "Yes" if they were certain it was present in the video, "No" if the event was absent. Then, the information coverage can be recorded based on the answers. Mathematical, the IC can be computed by

$$IC = \frac{NGS}{NIV}, \tag{3}$$

where NIV is the number of representative entities or plots in input video, NGS is the number of representative entities or plots in generated summary.

Similarly, we record the ratio of complete sentences and emerging sentences read by voiceover in the summary, so as to acquire *consistency* (CON). This indicates how many complete sentences are read by voiceover in the summary, the CON can be obtained as below:

$$CON = \frac{NSV}{NES} \times 100\%, \tag{4}$$

where NSV denotes the number of complete sentences by voiceover in the summary, NES denotes the number of emerging sentences in the summary. A desired summary is composed of a successive list of segments. The transition from one video segment to another should be as smooth as possible providing viewers with a pleasant viewing experience. Intuitively, videos in high consistency, which means more complete sentences are read in our situation, can provide viewers with a enjoyable viewing experience. Obviously, the appearance of incomplete sentences in the summary can prevent that from happening.

Finally, we counted *the number of shots in the video summary* (TNS) and *the average length of shots* (TAL). This measures shots switching frequency and content completeness, videos owning too many or too few shots switching degrade the viewer experience and satisfaction.

V. SUPPLEMENTARY RESULTS AND ANALYSIS

In this section, we will make supplementary analysis in the comparison to traditional video summarization methods and ablation study of visual semantic matching component. Moreover, some video summaries generated by our method with different settings and baselines can be viewed online[2][3].

A. Comparison to traditional video summarization

From Table I, we can observe that some summary attributes of baselines perform similar between TVsum and M-D dataset, such as *information coverage* and *consistency*. However, the results of *visual attraction* and *narrative completeness* on TVsum in Table II outperform on M-D dataset. The main reason lies in the characteristics of the videos in TVsum, which are relatively short with the duration of only a few minutes. Meanwhile, a large proportion of them are non-narrative and only describe a single event, which is easy to understand. Therefore, incomplete sentences in summaries of them do not reduce a lot for the audience's understanding of the whole video content in the narrative completeness. In addition, the summaries of videos in TVsum have an appropriate average shot length, which also helps improve the visual appeal of the entire video summary. When leveraged in strong narrative videos such as documentary and movie, the *visual attraction* and *narrative completeness* of these baselines reduce largely since incomplete sentences and inappropriate shot length harm the impression of the audience a lot. On the contrary, our method can perform well in all types of video, which demonstrates the remarkable generalization of it.

B. Ablation Study of Visual Semantic Matching Component

Fig. 2 shows some shot localization results of our component and alternatives. From it, we can have several observations: (1) RD can hardly localize the shots, which correspond

[2]https://www.acfun.cn/v/ac27867826
[3]https://www.bilibili.com/video/BV1Mp4y1b7em/

to the procedural text. For instance, it localizes the clouds under the text of "A flock of birds flew across the sky." and a deer under "Two leopards appeared in a tree in the snow." (2) Other ablation methods including VSL [11], LG4 [12], RD + TSS and VSL + TSS can localize the right shots to some extent, such as the first procedural text for LG4 method. However, one obvious issue is that though a large proportion of selected shots can include the main entity that the procedural text contains, they do not appropriately describe the plot of the sentence. For example, for the third selected shots of LG4, it fails in describing the proper 'stare at the figure' in spite of containing 'Harry and Ron'. The seventh shot of VSL + TSS also suffers from the same problem. In addition, there exists some failures in localizing, such as the third column for VSL method. (3) Our method, LG4 + TSS, can nearly localize every procedural text correctly. The observations above demonstrate the effectiveness of our visual semantic matching e component.

REFERENCES

[1] A. Pavel, C. Reed, B. Hartmann, and M. Agrawala, "Video digests: a browsable, skimmable format for informational lecture videos." in *Pro. ACM Symp. User Interface Softw. Technol.*, 2014, pp. 1–10.

[2] M. Wang, G.-W. Yang, S.-M. Hu, S.-T. Yau, and A. Shamir, "Write-a-video: computational video montage from themed text." *ACM Trans. Graph.*, vol. 38, no. 6, pp. 1–13, 2019.

[3] K. Zhou, Y. Qiao, and T. Xiang, "Deep reinforcement learning for unsupervised video summarization with diversity-representativeness reward," in *Proc. AAAI Conf. Artif. Intell*, 2018, pp. 7582–7589.

[4] J. Fajtl, H. S. Sokeh, V. Argyriou, D. Monekosso, and P. Remagnino, "Summarizing videos with attention," in *Proc. Asian Conf. Comput. Vis.*, 2018, pp. 39–54.

[5] W. Zhu, J. Lu, J. Li, and J. Zhou, "Dsnet: A flexible detect-to-summarize network for video summarization," *IEEE Trans. Image Process*, pp. 948–962, 2021.

[6] B. Zhao, X. Li, and X. Lu, "Hsa-rnn: Hierarchical structure-adaptive rnn for video summarization," in *Proc. IEEE Conf. Comput.Vis. Pattern Recognit*, 2018, pp. 7405–7414.

[7] M. Rochan, L. Ye, and Y. Wang, "Video summarization using fully convolutional sequence networks," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 347–363.

[8] B. Zhao, M. Gong, and X. Li, "Hierarchical multimodal transformer to summarize videos," *Neurocomputing*, vol. 468, pp. 360–369, 2022.

[9] M. Rochan and Y. Wang, "Video summarization by learning from unpaired data," in *CVPR*, 2019, pp. 7902–7911.

[10] Y. Song, J. Vallmitjana, A. Stent, and A. Jaimes, "Tvsum: Summarizing web videos using titles," in *Proc. IEEE Conf. Comput.Vis. Pattern Recognit*, 2015, pp. 5179–5187.

[11] H. Zhang, A. Sun, W. Jing, and J. T. Zhou, "Span-based localizing network for natural language video localization," in *Proc. Assoc. Comput. Linguist.*, 2020, pp. 6543–6554.

[12] J. Mun, M. Cho, and B. Han, "Local-global video-text interactions for temporal grounding," in *Proc. IEEE Conf. Comput.Vis. Pattern Recognit*, 2020, pp. 10 810–10 819.

[13] X. Qu, P. Tang, Z. Zou, Y. Cheng, J. Dong, P. Zhou, and Z. Xu, "Fine-grained iterative attention network for temporal language localization in videos," in *Proc. ACM Multimedia*, 2020, pp. 4280–4288.

[14] D. Cao, Y. Zeng, X. Wei, L. Nie, R. Hong, and Z. Qin, "Adversarial video moment retrieval by jointly modeling ranking and localization," in *Proc. ACM Multimedia*, 2020, pp. 898–906.

[15] M. Otani, Y. Nakashima, E. Rahtu, and J. Heikkila, "Rethinking the evaluation of video summaries," in *Proc. IEEE Conf. Comput.Vis. Pattern Recognit*, 2019, pp. 7596–7604.

[16] Y. Zhang, L. Zhang, and R. Zimmermann, "Aesthetics-guided summarization from multiple user generated videos," *ACM Trans. Multimed. Comput. Commun. Appl.*, vol. 11, no. 2, pp. 1–23, 2015.

[17] Y. Niu and F. Liu, "What makes a professional video? a computational aesthetics approach," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 7, pp. 1037–1049, 2012.