Supplementary Material: Annotation Methods, Visualization, User Study and More Details on Training and Testing

Yunlong Tang[1,2][0000−0003−2796−1787], Siting Xu[1][0000−0001−9934−7919],
Teng Wang[1,2], Qin Lin[2], Qinglin Lu[2], and Feng Zheng[1][0000−0002−1701−9141]⋆

[1] Southern University of Science and Technology, Shenzhen, China
[2] Tencent Inc., Shenzhen, China
{tangyl2019, xust2019, wangt2020}@mail.sustech.edu.cn
{angelqlin, qinglinlu}@tencent.com f.zheng@ieee.org

Abstract. To compute the importance reward for training and the Imp@T in metrics, we assign multiple narrative technique labels to each segment on both the training and testing set. The labels of coherence are only assigned on the testing set since we only utilize coherence labels to compute Coh@T for evaluation while obtaining coherence reward by computing PPL. In this supplementary, we introduce the annotation methods and list the label groups. In addition, we present the visualization of results at a given target duration of $T = 10$ and $T = 15$. We also perform an user study on our model and previous work. More details on model training and testing can be found in this supplementary.

A Annotation Methods

A.1 Narrative Technique Annotation

Narrative label is defined as the description of the techniques used in the development of videos. In the advertisement area, labels can be divided into 5 basic categories: background foreshadowing, sore points, product display, brand reinforcement, and behaviour guidance. Under these five categories, there are more specific labels up to 81. The labels are finally divided into 4 groups according to their importance and listed in Tbl. 1.

There are a few rules on narrative technique annotation:

- One segment may be assigned a few labels.
- The overall order of labels of segments from the same source video is background foreshadowing, sore points, product display/brand reinforcement, and behaviour guidance roughly. The order may differ but the last segment should not have the label of background foreshadowing.
- Those labels have industrial characteristics.

⋆ Corresponding author

Table 1. Narrative technique labels. They are divided into 4 groups. The importance scores of a single segment are computed as the weighted sum of group-level weights of all labels assigned, where the weights are the fourth powers of 0.25, 0.5, 0.75, and 1.00 for group-level 1, 2, 3, and 4 respectively in our implementation.

Group	Labels	
1	plain and ungarnished	contrasts
	question and suspense	emotional resonance
	background forshadowing	-
2	personal statement	celebrity introduction
	releasing notices	dialogue opening
	narrator statement	merchandise opening
	apology	longitudinal comparison of characters
	oops	behaviour comparison
	product comparison	rhetorical question
	conflict question	novelty elements
	plot anticipation	music anticipation
	factual description	horizontal comparison of characters
	target population	realistic
	yearning description	anxiety creation
	guarantee	repeatedly emphasis
	good fronting	sore points
3	question rhetorical class	aiming at target population
	aiming at age grades	aiming at behaviour characteristics
	time need	demand description technique
	status need	financial incoming
	insufficient fund	aiming at application scenarios
	living needs	entertainment needs
	health need	relationship needs
	job needs	learning needs
	maintenance needs	requirements type
4	product function display	product quality display
	entire product display	product details display
	product usage display	model display
	product advantage display	product display(other)
	environment display	business service display
	business effect display	business promotion display
	business process display	business advantages display
	business display(other)	reading image
	short video apps	live streaming picture
	withdraw picture	operational guideline
	merchandise display	application display(other)
	combat playing methods	interesting display method
	social playing method	painting style display
	character display	selling point display
	equipment display	playing method display
	armament display	game withdraw

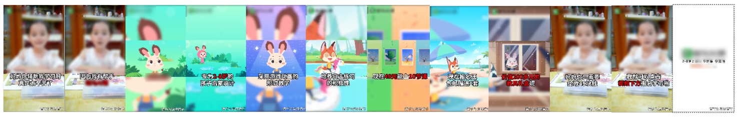
A.2 Coherence Annotation

In order to evaluate the coherence of the model output, we annotate coherence labels for the testing set. Specifically, we combine every two adjacent segments in the result with N segments to yield $N \cdot (N-1)$ pairs. Then the annotators will assign labels to these pairs with the following method:

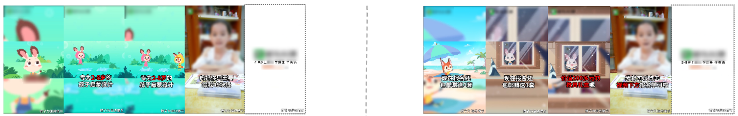
- If the annotators think that the segments in pairs are connected coherently, then the label is assigned as *coherent*.
- If the annotators think that the segments in the pair are connected incoherently, then the label is assigned as *incoherent*.
- Otherwise, the label is assigned as *uncertain*.

B Visualization

Fig. 1 presents a source video and videos assembled by SAM and M-SAN in a target duration of 10s. The content details of these videos are listed below.
Source Video(a): A little girl told her mum there's no need to do homework tutoring for her. Because there's a tutorial course designed for children aged 2-8. This course adopts instructional design using game animation, aiming to cultivate the enthusiasm for exercise initiatively. Now it is available for ten lessons for 49 yuan. If you apply now, a gift of teaching aid gift box worth more than 200 yuan will be sent to you. Then the little girl told her mom to give her 49 yuan to pay for the course. Then she will be available to click the link below and study! The final segment is the company logo presentation of this course.
SAM(b): There's a tutorial course designed for children aged 2-8. A little girl told her mom to give her 49 yuan to pay for the course. The final segment is the company logo presentation of this course.
M-SAN(c): There's a tutorial course for which if you apply now, a gift of teaching aid gift box worth more than 200 yuan will be sent to you. A little girl said that she will be available to click the link below and study! The final segment is the company logo presentation of this course.



(a)Source Video, τ=31s



(b)SAM, τ=10s, Imp-Coh@10=65.97 (c)M-SAN, τ=10s, Imp-Coh@10=74.19

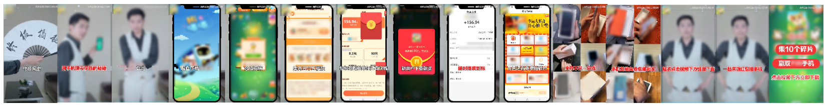
Fig. 1. Source video and videos assembled by SAM and M-SAN with a given target duration $T = 10s$. τ is the actual duration of the result.

Fig. 2 presents a source video and videos assembled by SAM and M-SAN in a target duration of 15s. The content details of these videos are listed below.

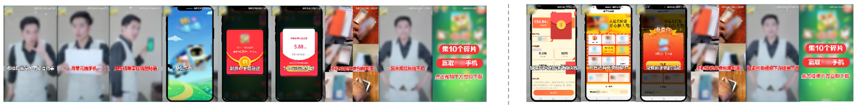
Source Video(a): The ultimate secret to earning pocket money from your phone is here! Now download this app, and synthesize cats to collect five blessing cats. Then you will get hundreds of money easily. Money will be given to new users once the app got downloaded. Withdraw cash at any time. And there's a free lottery. It doesn't cost a penny, and you can get a new phone freely without paying any postage. Click on the link below and download this app. Let's grab red envelopes and get a new phone! The final segment is the company logo presentation of this app.

SAM(b): The ultimate secret to earning pocket money from your phone is here! Now download the app. A red envelope of 5.88 yuan will be given to new users once the app got downloaded. It doesn't cost a penny, and you can get a new phone freely without paying any postage. Let's grab red envelopes and get a new phone! The final segment is the company logo presentation of this app.

M-SAN(c): You can get hundreds of money easily. And there's a free lottery for exchanging for a new mobile phone. It doesn't cost a penny, and you can get a new phone freely without paying any postage. Click on the link below and download this app. The final segment is the company logo presentation of this app.



(a)Source Video, τ=36s



(b)SAM, τ=19s, Imp-Coh@15=58.09 (c)M-SAN, τ=13s, Imp-Coh@15=70.28

Fig. 2. Source video and videos assembled by SAM and M-SAN with a given target duration $T = 15s$. τ is the actual duration of the result.

It can be observed that SAM generates videos with too many foreshadowing parts, which exceeds the duration limitation. M-SAN tends to select latter segments in source videos, which is reasonable because key points usually appear in the latter part of the ad with the front part doing foreshadowing. This verifies the M-SAN focuses on more informative segments and does better than SAM in duration control.

C User Study

Since we have the intention of deploying the model to produce ads in our online services, we had already done a user study. We used the test set as the input of both SAM and M-SAN and invited 6 colleagues from the advertising business department to evaluate the usability (usable or not usable) of the 15s results by subjectively judging whether the content was coherence and retained important commercial information and whether the output met the requirement of duration. The study shows the usability rate (#usable results/#all results) of M-SAN's output is 0.859, and the usability rate of SAM's output is 0.616. And the usability obtained from users is consistent with the evaluation of our metrics.

Methods	Usability	Imp-Coh@15
SAM	0.616	58.09
M-SAN (ours)	0.859	70.28

Table 2. User study.

D More Training and Testing Details

We use a two-layer Bi-GRU with hidden size of 256 as encoder and one-layer GRU with hidden size of 512 as decoder. The large models Swin-Transformer and BERT are frozen. And we fine-tuned GPT-2 that computes PPL on 8 A100 GPUs for 4 days. We trained M-SAN on 4 Tesla T4 GPUs for one day. There are 8 videos in each batch, and the learning rate $\eta = 2 \times 10^{-4}$. The number of episodes for each video $K = 8$, and the number of epochs is 10. The end segment of every ad video usually contains a wealth of ad-related information. Therefore, the end segment will be selected to be the first item of A in our implementation. The testing on 99 videos only needed 2 minutes on single Tesla T4 GPU.