Sentence Specified Dynamic Video Thumbnail Generation (Supplemental Material)

Yitian Yuan*
yyt18@mails.tsinghua.edu.cn
Tsinghua-Berkeley Shenzhen Institute, Tsinghua University

Lin Ma
forest.linma@gmail.com
Tencent AI Lab

Wenwu Zhu†
wwzhu@tsinghua.edu.cn
Department of Computer Science and Technology & Tsinghua-Berkeley Shenzhen Institute, Tsinghua University

ABSTRACT

This supplemental material includes the following contents:

- The annotation details of the sentence specified video thumbnail dataset.
- Dataset statistical analysis.
- More qualitative results of the proposed GTP model.

1 THE DATASET ANNOTATION DETAIL

Figure 1 illustrates our implemented annotation website for the sentence specified dynamic video thumbnail generation task. For each video and its paired sentence description in our collected dataset, we place them on the website simultaneously for the convenience of the annotation participants' browsing. Moreover, in order to speed up the annotation, we evenly split the video into 2-second video clips (We split the video into 2-second length clips mainly because we find that the smallest video thumbnail gifs in some video websites like YouTube are 1 to 2 seconds long), and all these video clips are displayed in their chronological order. Participants are required to select no more than 5 video clips that semantically correspond to the sentence description to compose the video thumbnail. The video clip will be highlighted in red bounding box after selected. The selected video clips are not required to be consecutive in time. If one participant finishes the video clip selection for the current video-sentence pair, he (or she) only needs to click the "submit" button to proceed to the next annotation task.

The annotations of different participants are completely independent, with the video-sentence pairs randomly illustrated on the website. There are 10,204 video-sentence pairs in our collected dataset, and we ensure that each pair will have 4 video thumbnail annotations from 4 different participants. Therefore, we totally get $4 \times 10,204 = 40,816$ annotation results for our constructed dataset.

Some video thumbnail annotation examples are shown in Figure 2. For each showing example, we provide two video thumbnail annotations, and the selected video clips in these two annotations are highlighted with orange and yellow bounding boxes, respectively. We can observe that in example (a), the two annotations are exactly the same, while in other examples, the annotations are partially aligned with each others. It illustrates that when annotating video thumbnails, different participants have different opinions, making the differences between the annotated video thumbnails. However,

Figure 1: The annotation interface for the sentence specified dynamic video thumbnail generation task.

the jointly selected video clips also indicate that the participants still have their common cognition for the given sentence descriptions. In addition, example (a) and example (b) share the same video but are with different sentence descriptions. We can see that the sentence descriptions highly influence the resulting video thumbnails and cause great discrepancy, which further verifies that it is very necessary to generate specific video thumbnails for different sentences.

2 DATASET STATISTICAL ANALYSIS

Video Length. The minimal, maximal, and average video lengths over all the videos in our constructed dataset are 20.0s, 238.4s and 60.7s, respectively. The average length of the annotated video thumbnails is 8.7s.

Video Thumbnail Annotation Consistency. As indicated in Figure 2, video thumbnail annotation is a very subjective task, with

*This work was done while Yitian Yuan was a Research Intern at Tencent AI Lab.
†Corresponding author.

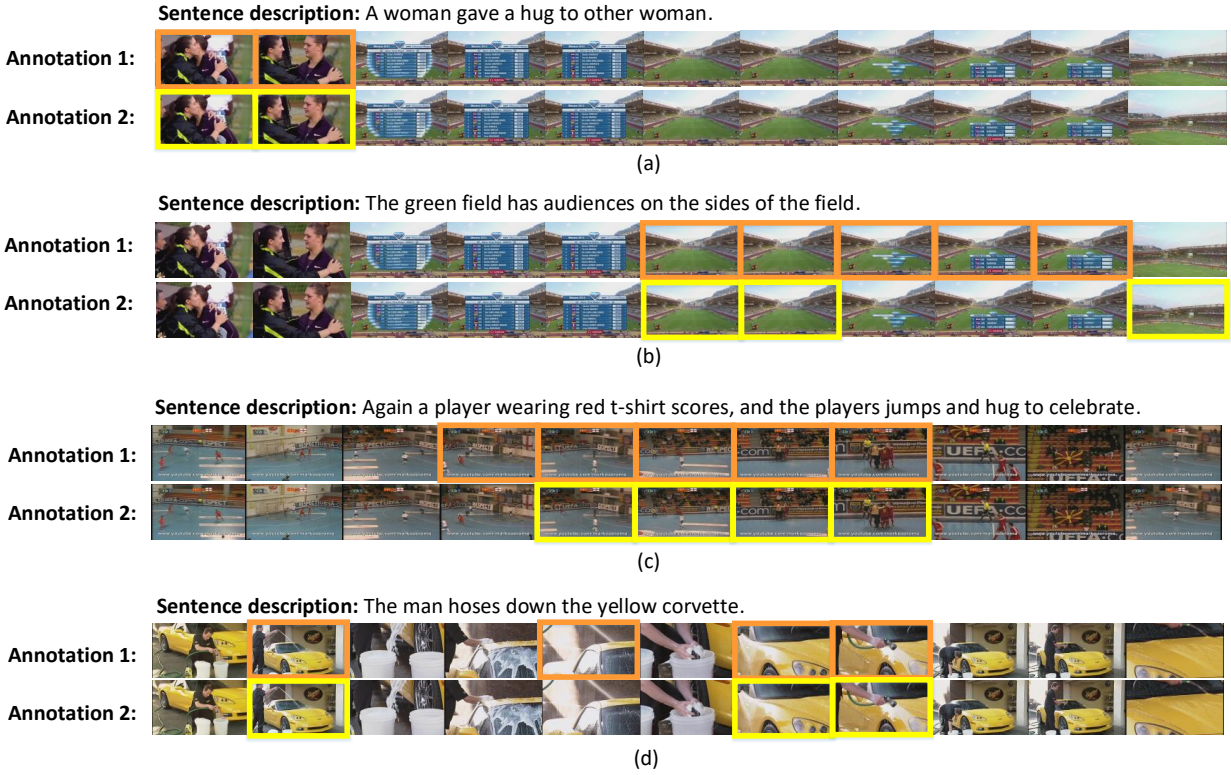

Figure 2: Video thumbnail annotation examples. For each showing video-sentence pair, we provide two video thumbnail annotations, and the selected video clips in these two annotations are highlighted with orange and yellow bounding boxes, respectively.

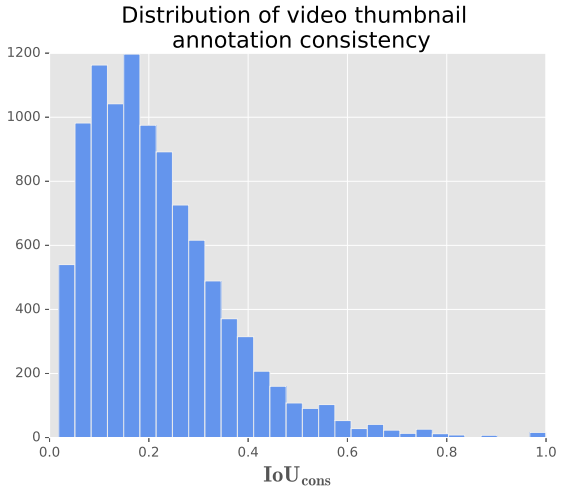

Figure 3: The video thumbnail annotation consistency distribution over all the video-sentence pairs.

different annotation participants having different opinions. To measure the consistency of the selected video thumbnails between different participants, we define a metric IoU_{cons} as follows:

$$IoU_{cons}(k,i) = \frac{1}{3}\sum_{j\neq i, j=1}^{4} \frac{\|Intersection(A_i^k, A_j^k)\|}{\|Union(A_i^k, A_j^k)\|}$$
$$IoU_{cons}(k) = \frac{1}{4}\sum_{i=1}^{4} IoU_{cons}(k,i) \quad (1)$$

Here A_i^k means the set of selected video clips composing the i-th annotated video thumbnail for the k-th video-sentence pair. $IoU_{cons}(k,i)$ indicates the annotation consistency between the i-th annotated video thumbnail and all the other annotations for the k-th video-sentence pair. $IoU_{cos}(k)$ means the average annotation consistency of the 4 video thumbnail annotations for the k-th video-sentence pair. If the selected video clips of all the annotations are exactly the same, the value of $IoU_{cos}(k)$ will be equal to 1. The annotation consistency distributed over all the video-sentence pairs is illustrated in Figure 3. It can be observed that for most of the video-sentence pairs, the selected video clips of different participants do not have a exact match, but there are still some clips that are jointly selected by several participants. It further demonstrates that the video thumbnail generation is an indeed subjective task, while people still express their consensus to generate the thumbnail with respect to the given sentence descriptions.

Ground Truth. Since there are 4 video thumbnail annotations for each video-sentence pair, we take the annotation result with the highest consistency $IoU_{cons}(k, i)$ among the 4 annotations as the ground truth during the training process. While in the testing stage, the predicted video thumbnail will be evaluated with respect to all the 4 annotations.

3 QUALITATIVE RESULTS

Evolution of the Sentence Specified Video Clip Graph. Figure 4 shows the evolution of 4 groups of video clip adjacency matrices in our GTP model training procedure. We can observe that the first two qualitative examples (a) and (b) present similar evolution process with the examples we have shown in the main paper. The adjacency matrices tend to a even distribution at the initial model training stage, and along with the model training procedure the block boundaries gradually show up clearly. In contrast, in the qualitative examples (c) and (d), the sentence specified video clip graph structures have been initially learned in Epoch 1, with the following training epochs only adjusting and emphasizing the learned video clip relationships. Overall, all of the above results verify that our GTP model can indeed learn the sentence specified video clip graph according to the sentence and video semantics.

Video Thumbnail Generation Results of the GTP Model. Figure 5 illustrates some qualitative results of our proposed GTP model for the sentence specified dynamic video thumbnail generation. We can observe that the selected video clips by GTP are consistent with the clips in the ground-truths, which indicates the effectiveness of our proposed GTP model. Meanwhile, the generated video thumbnails are quite flexible. As shown in case (a) and (e), the video thumbnails are temporally inconsecutive and provide a good preview of the overall video content. Comparing the show case (c) to others, we can find that the lengths of video thumbnails are also not fixed. Since most video contents shown in case (c) are irrelevant to "`skateboarding`" described by the sentence, GTP only selects the last clip that presents the matching activity.

Besides, the predicted video thumbnail in case (d) does not exactly match the ground-truth annotation. The main reason lies on the indistinguishable video scenes in the video. From the 8-th video clip in case (d) to the end of the video, all the middle clips present the same scene of "`people rafting`". Therefore, not only the GTP model, the annotators are also hard to decide which clip to choose. However, since all these clips are matched with the sentence description, the generated video thumbnail by our proposed GTP is still reasonable and accurate.

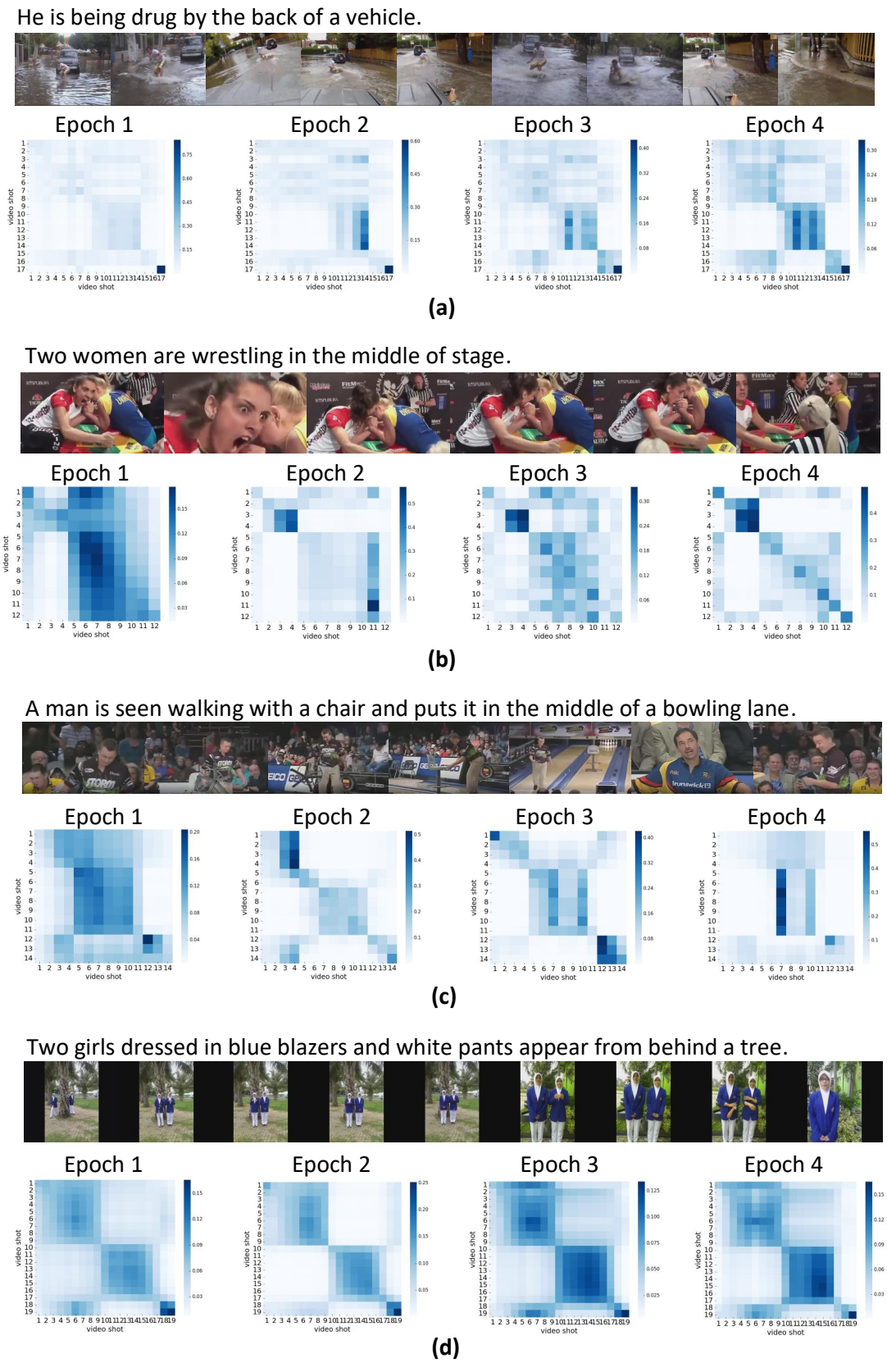

Figure 4: Evolution of the learned video clip adjacency matrices during the sentence specified video graph convolution.

Figure 5: Qualitative results of our proposed GTP model for sentence specified dynamic video thumbnail generation. Blue bars show the video thumbnail generation results for our proposed GTP model, with the selected video clips highlighted in darker colors. Green bars show the ground-truth video thumbnails.