Due to the limited space, we here report more experimental results and technical details which are not included in the paper:

- The generalizability of the proposed dynamical knowledge distillation strategy on the pre-trimmed short-duration video dataset (Section 1).
- Further investigations on the performance of the feature-based knowledge and similarity-based knowledge distillations, respectively (Section 2).
- Studies on dual learning in terms of the hyper-parameter $\beta$ and complementarity analysis (Section 3).
- Comparison in terms of time complexity and memory consumption (Section 4).
- More implementation details of our method (Section 5).

1. Results on pre-trimmed short-duration video dataset

Recall that the previous experiments are all conducted on untrimmed long videos. To verify the generalizability of the proposed dynamical knowledge distillation strategy, we conduct experiments on a pre-trimmed short-duration dataset, i.e. MSR-VTT [11], in the context of traditional T2VR. We adopt Dual Encoding [3] as our baseline, considering its source code released and has been widely used as the baseline in T2VR. As in Table 1, by using our proposed dynamical knowledge distillation strategy, Dual Encoding [3] achieves a further performance gain. Besides, our dynamical knowledge distillation still outperforms the fixed counterpart that uses a fixed weight during the distillation. The results not only verify the generalizability of the proposed dynamical knowledge distillation strategy, and again confirm its advantage over the fixed distillation.

*Both authors contributed equally to this work.
†Corresponding author: Baolong Liu (liubaolongx@gmail.com)
Table 4. Complexity comparison in terms of computation overhead at the inference stage and memory consumption at the training stage. All models utilize the same vision backbone. Note that the computation cost excludes the vision backbone and RoBERTa. For a comprehensive comparison, we also report the performance on ActivityNet and TVR.

<table>
<thead>
<tr>
<th>Text backbone</th>
<th>Model</th>
<th>FLOPs (G)</th>
<th>Memory (MiB)</th>
<th>ActivityNet</th>
<th>TVR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Graph-based</td>
<td>HGR [1]</td>
<td>2.96</td>
<td>8555</td>
<td>107.0</td>
<td>50.1</td>
</tr>
<tr>
<td></td>
<td>DL-DKD (Ours)</td>
<td>1.01</td>
<td>4229</td>
<td>125.4</td>
<td>103.6</td>
</tr>
<tr>
<td>Bi-GRU</td>
<td>DE [3]</td>
<td>5.24</td>
<td>5837</td>
<td>121.7</td>
<td>123.4</td>
</tr>
<tr>
<td></td>
<td>DE++ [4]</td>
<td>5.30</td>
<td>3515</td>
<td>121.7</td>
<td>128.3</td>
</tr>
<tr>
<td></td>
<td>DL-DKD (Ours)</td>
<td>0.98</td>
<td>4057</td>
<td>137.3</td>
<td>145.1</td>
</tr>
<tr>
<td>RoBERTa</td>
<td>RIVRL [5]</td>
<td>8.64</td>
<td>4809</td>
<td>117.8</td>
<td>135.6</td>
</tr>
<tr>
<td></td>
<td>XML [8]</td>
<td>0.80</td>
<td>2451</td>
<td>128.4</td>
<td>155.1</td>
</tr>
<tr>
<td></td>
<td>ReLoCLNet [12]</td>
<td>0.96</td>
<td>2673</td>
<td>126.6</td>
<td>157.1</td>
</tr>
<tr>
<td></td>
<td>MS-SL [2]</td>
<td>1.22</td>
<td>5349</td>
<td>140.1</td>
<td>172.4</td>
</tr>
<tr>
<td></td>
<td>DL-DKD (Ours)</td>
<td>1.04</td>
<td>4455</td>
<td>147.6</td>
<td>179.9</td>
</tr>
</tbody>
</table>

3. Studies on dual learning

Influence of the hyper-parameter for two-branch inference Fig. 1 shows the influence of hyper-parameter $\beta$ in Eq.9 of our paper. Note that $\beta = 0$ indicates that only the similarity obtained inheritance branch is employed, while $\beta = 1$ indicates using the exploration branch counterpart. The inheritance branch consistently outperforms the exploration one on both datasets, showing the benefit of learning knowledge from the teacher model. The best performance is achieved at $\beta = 0.3$, where the inheritance branch is dominated for the final similarity. It is worth noting that our proposed model is not very sensitive to beta, as we find that the proposed method holds state-of-the-art (SOTA) results on both ActivityNet and TVR datasets when beta is in the range from 0 to 0.7 (as shown in Figure 1).

Comparison to two-branch baselines. In this experiment, we tune $\beta$ (Eqn. 9) for the double-branch baselines, i.e., Dual-exploration and Dual-inheritance. The results are summarized in Fig. 2. With the same $\beta$, our model consistently performs better, which further demonstrates the benefit of using two hybrid branches.

4. Model Complexity

Table 4 summarizes the model complexity comparison in terms of time complexity and memory consumption, and their corresponding performance on ActivityNet and TVR. For a specific method, its time complexity is measured as FLOPs it takes to encode a given video-text pair. For a more fair comparison, we compare previous works using the same text backbone. Our model has comparable FLOPs and memory usage, but gives better performance.

Additionally, it is worth noting that although our model has two branches, they utilize the same vision and text backbones, and only the encoder headers are doubled. Moreover, as the majority of the computation cost is in the vision (222G FLOPs) and text backbones (79G FLOPs), the computation cost only increases 0.3% when extending from one branch to two branches.
5. Implementation Details

The dimension sizes of the video features extracted by the pre-trained CNN model of the ActivityNet-Captions and TVR datasets are 1024 and 3072, respectively. The dimensions of all the above features are linearly reduced to 384 further for the convenience of the Transformer’s (384 hidden sizes, 4 attention heads) feature encoding. For a textual query, we employ the pre-trained Roberta [9] to extract a feature with 1024 dimension firstly, reduce the dimension of the feature to 384 further, and then feed the feature to a Transformer (384 hidden sizes, 4 attention heads) for feature encoding. For different decay strategies, we empirically set the $k$ to 0.95 and 800 for the exponential decay and Sigmoid decay, respectively. For the linear decay strategy function, we set the parameters $k$ and $b$ to -0.01 and 1, respectively. We utilize an Adam optimizer with a mini-batch size of 128 for model training.

References