Supplementary Material: Audio-Enhanced Text-to-Video Retrieval using Text-Conditioned Feature Alignment

Sarah Ibrahimi\(^1\) * Xiaohang Sun\(^2\) Pichao Wang\(^2\) Amanmeet Garg\(^2\)
Ashutosh Sanan\(^2\) Mohamed Omar\(^2\)*
\(^1\) University of Amsterdam \(^2\) Amazon Prime Video
s.ibrahimi@uva.nl, \{sunking, wpichao, amanmega, ashsanan\}@amazon.com, mkamalmomd@gmail.com

1. Video-to-Text Retrieval Results

In our work, we focus on the task of text-to-video retrieval. We follow other works by also evaluating our model trained on video-to-text retrieval for MSR-VTT 9k, since competing methods have only provided video-to-text retrieval results on this data.

<table>
<thead>
<tr>
<th>Method</th>
<th>R1 ↑</th>
<th>R5 ↑</th>
<th>R10 ↑</th>
<th>MdR ↓</th>
<th>MnR ↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLIP4Clip(_{meanP}) [26]</td>
<td>43.1</td>
<td>70.5</td>
<td>81.2</td>
<td>2.0</td>
<td>12.4</td>
</tr>
<tr>
<td>X-Pool[16]</td>
<td>44.4</td>
<td>73.3</td>
<td>84.0</td>
<td>2.0</td>
<td>9.0</td>
</tr>
<tr>
<td>ECLIPSE(_{meanP})[22]</td>
<td>44.7</td>
<td>71.3</td>
<td>82.8</td>
<td>2.0</td>
<td>10.8</td>
</tr>
<tr>
<td>BridgeFormer(*)[12]</td>
<td>44.9</td>
<td>71.9</td>
<td>80.3</td>
<td>2.0</td>
<td>15.3</td>
</tr>
<tr>
<td>CAMoE [9]</td>
<td>45.1</td>
<td>72.4</td>
<td>83.1</td>
<td>2.0</td>
<td>10.0</td>
</tr>
<tr>
<td>TS2-Net [25]</td>
<td>45.3</td>
<td>74.1</td>
<td>83.7</td>
<td>2.0</td>
<td>9.2</td>
</tr>
<tr>
<td>X-CLIP [28]</td>
<td>46.8</td>
<td>73.3</td>
<td>84.0</td>
<td>2.0</td>
<td>9.1</td>
</tr>
<tr>
<td>TEFAL</td>
<td>47.1</td>
<td>75.1</td>
<td>84.9</td>
<td>2.0</td>
<td>7.4</td>
</tr>
</tbody>
</table>

Table T1. Video-to-Text Retrieval Results on MSR-VTT 9k split. All works use a CLIP ViT-B/32 backbone which is pre-trained on this Wikipedia-based image-text dataset.

2. Qualitative Results

In this section, we present additional examples on the MSR-VTT dataset [42] to highlight how audio provides complementary information to the video to achieve improved text-queried retrieval. The query words of sample 7152, visualized in Figure F1 is “a person is swimming in some white water rapids”. While the video modality alone shows both the rapid water and the person, TEFAL w/o audio (video-only model) ranks the clip as the second matched retrieval. TEFAL, with the addition of the audio cue, correctly ranks the matched clip as the top retrieval. We notice that the presence of a person is confirmed by the voice in the latter part of the waveform (encircled in red), which clearly demonstrates that our model picks up complementary information from the audio modality. It is also observed that the first part of the clip is dominated by loud sound of streaming water but the water sound is greatly suppressed later in the clip though the continuous presence of water flowing in the video. This explicitly justifies building independent text-video and text-audio cross-modal attention blocks rather than aligning video and audio embeddings as it is in ECLIPSE [22], since the mandatory alignment between video and audio may introduce additional noise in the audiovisual feature.

Additional examples are shown in Figure F2 to illustrate the correspondence between the text and audio modality that is otherwise missed between text and video. In the upper example, the girl talking can only be heard in the audio signal (sample 8827); in the middle example (sample 9233), the speech content is explicitly presented in the audio rather than video; and in the lower example, the word “oxiders” can only be matched in the audio from the man’s talking (sample 9249).

3. Limitations

The main limitation of TEFAL is that without audio the method reduces to the text-video branch, and the performance is similar with XPool [16] (as indicated by Figure 1 in the main manuscript). If the missing audios are mostly in the train set, meta learning approaches could help lessen this issue [27].

\(*\)This work was done while Sarah Ibrahimi and Mohamed Omar were at Amazon Prime Video.
Figure F1. In this figure an example is presented where a small sound has a large contribution to the final result. While TEFAL w/o audio is not able to select the correct video, TEFAL uses the audio to select the correct video as Rank 1.

Query: a person is swimming in some white water rapids

Figure F2. This Figure shows three examples that illustrate the correspondence between the text and audio modality, that contains the verb “speaking”, “talking” or a variation and specific words that correspond to the text query.

Query: cartoon girl is talking

Query: a man is giving a speech

Query: a man is discussing oxiders in bulk form
References


[34] Nina Shvetsova, Brian Chen, Andrew Rouditchenko, Samuel Thomas, Brian Kingsbury, Rogério Feris, David Harwath,


