This paper proposes a novel biochemical video-and-language (BioVL) dataset, which consists of experimental videos, corresponding protocols, and annotations of alignment between events in the video and instructions in the protocol. The key strength of the dataset is its user-oriented design of data collection. We imagine that biochemical researchers easily take videos and share them for another researcher’s replication in the future. To minimize the burden of video recording, we adopted an unedited first-person video as a visual source. As a result, we collected 16 videos from four protocols with a total length of 1.6 hours. In our experiments, we conduct two zero-shot video-and-language tasks on the BioVL dataset. Our experimental results show a large room for improvement for practical use even utilizing the state-of-the-art pre-trained video-and-language joint embedding model. We are going to release the BioVL dataset. To our knowledge, this work is the first attempt to release the biochemical video-and-language dataset.

1. Introduction

Science faces a replication crisis. As reported by [1], in wet-lab research (e.g., biochemistry and life science), more than 80% of researchers have failed to reproduce another scientist’s experiments. Video-and-language techniques would break through this problem. For example, given recorded videos and protocols, an automatic system can create multimedia protocols, which align instructions in the protocols and events in the videos. They help researchers reproduce experiments by providing quantitative (e.g., time and quantity) and qualitative information (e.g., shapes and colors of samples). Furthermore, video captioning [4] generates protocols only from recorded videos, saving the researcher’s effort of writing them.

As the first step towards this goal, this paper proposes a novel Biochemical Video-and-Language (BioVL) dataset (Figure 1), which consists of experimental videos, text protocols, and annotations of alignment between events in the videos and instructions in the protocols. The key strength of the dataset is its user-oriented design of data collection. We imagine that biochemical researchers easily take videos and share them for another researcher’s replication in the future. To minimize the burden of video recording, we adopted an unedited first-person video as a visual source, referring to [3], without additional cameras or sensors [8, 12]. As a result, we collected 16 videos from four protocols with a total length of 1.6 hours.

Using the BioVL dataset, we conduct two zero-shot video-and-language tasks for practical applications: instruction-to-event retrieval and instruction-video alignment. Our experimental results show a large room for improvement for practical use even using the state-of-the-art pre-trained video-and-language joint embedding model. The BioVL dataset will be available online only for research purposes; to our knowledge, this work is the first attempt to release the biochemical video-and-language dataset on the web.
Figure 2. (a) and (b) shows the recording studio of experiments and the view from the equipped first-person camera, respectively.

2. Related Work

2.1. Video-and-Language How-to Dataset

As the biochemical domain is one of the how-to domains (e.g., cooking and assembling furniture), we introduce vision-and-language how-to datasets. The largest dataset is the Howto100M dataset [7], which consists of 100M instructional videos with their narration transcriptions in various domains. Several works [7, 6] reported that pre-training video-and-language models on this dataset contributes to the performance improvement on various downstream tasks. We use one of the models [7] pre-trained on this dataset in our experiments.

In contrast to the significant progress in the how-to domains, little work targets the biochemical domain. The only two studies [8, 9] previously targeted the biochemical domain, tackling an unsupervised alignment problem of instructions in the protocols and events in the video. Unfortunately, their benchmark datasets are not available online, and the other researchers cannot follow their research. To address this issue, we are going to release the BioVL dataset only for research purposes.

2.2. Egocentric How-to Dataset

Recently, much attention has been paid to an egocentric how-to dataset that consists of recorded videos with linguistic annotations (e.g., narration) [3]. The BioVL dataset will be a biochemistry version of the dataset; it will be a benchmark for biochemical video-and-language tasks, such as cross-modal retrieval [5], video-and-language alignment [2], and video captioning [4]. Moreover, it also would be a valuable resource for practical applications, such as robot imitation learning from human demonstration [10] and virtual assistant for researchers to reproduce experiments.

3. BioVL Dataset

This section describes how to collect videos, the annotation processes, and the statistics of the BioVL dataset.

Table 1. An annotation example of PCR.

<table>
<thead>
<tr>
<th>Instruction</th>
<th>start</th>
<th>end</th>
</tr>
</thead>
<tbody>
<tr>
<td>add sterile distilled water</td>
<td>30</td>
<td>45</td>
</tr>
<tr>
<td>add primer1</td>
<td>64</td>
<td>99</td>
</tr>
<tr>
<td>add primer2</td>
<td>106</td>
<td>130</td>
</tr>
<tr>
<td>add template</td>
<td>149</td>
<td>173</td>
</tr>
<tr>
<td>add primeSTAR</td>
<td>190</td>
<td>238</td>
</tr>
<tr>
<td>set in DNA engine</td>
<td>260</td>
<td>266</td>
</tr>
</tbody>
</table>

3.1. Video Recording

Participant. We asked one researcher (1 female) for our video collection. During experiments, the researcher put on a headset that fixes a wearable first-person camera (Figure 2). Note that the headset is light enough for researchers to concentrate on their experiments.

Experiment targets. We chose the basic well-known four experiments that have well-established protocols in the biochemical domain: miniprep, PCR, DNA extraction, and making an agarose gel. We took four videos per experiment, thus finally got 16 videos in total. Only DNA extraction has two different methods: Phenol-chloroform extraction and Ethanol precipitation. We took videos two times per method.

Data processing. In a few experiments, the researcher needed to wait while executing specific instructions (e.g., centrifuge samples). During the waiting time, the researcher put off the headset, leaving the camera on. We manually trimmed such waiting times because they are not related to any instructions.

3.2. Annotation

We hired two annotators to align events in the videos and instructions in the protocols; one annotator for the major effort and the other for verification, following [13]. While the former annotator is a non-expert, the latter is an expert annotator. First, the annotators split the protocol sentences into instructions by actions; for example, the sentence “Invert 4 times to mix and add 10 \( \mu l \) of Alkaline Protease Solution.” was split into two instructions “Invert 4 times to mix” and “add 10 \( \mu l \) of Alkaline Protease Solution.” Then, the annotators watched a video and annotated events by determining start/end times (seconds) for each instruction. In this annotation phase, the latter expert annotator validated events based on the annotation from the former annotator and corrected them if there are any mistakes. We cannot compute the agreement rate of event annotation because we only saved the latter’s annotation. In the future, we are going to evaluate the annotation quality by asking other experts to annotate events independently. Table 1 shows an example file of our annotation.

1We use Panasonic HX-A500.
2The researcher followed the same protocol per experiment. They could read it during the experiments.
Centrifuge for 5 minutes

Set in DNA engine

Add Primer1

Add Primer2

Figure 3. Visual distinctiveness for different instructions.

Table 2. Text-side statistics of the BioVL dataset. This table reports average values and standard deviation. Note that in agarose gel and miniprep, the researcher follows the same protocols but sometimes skips a few instructions depending on the situation. Therefore standard deviation of #instructions becomes over 0.

<table>
<thead>
<tr>
<th>Instruction</th>
<th>#instructions</th>
<th>#words/#instructions</th>
</tr>
</thead>
<tbody>
<tr>
<td>DNA</td>
<td>4.0 (±0.0)</td>
<td>6.0 (±1.9)</td>
</tr>
<tr>
<td>Phenol chloroform</td>
<td>9.0 (±0.0)</td>
<td>4.9 (±2.9)</td>
</tr>
<tr>
<td>PCR</td>
<td>6.0 (±0.0)</td>
<td>3.0 (±1.0)</td>
</tr>
<tr>
<td>Agarose gel</td>
<td>10.3 (±0.4)</td>
<td>4.7 (±2.4)</td>
</tr>
<tr>
<td>Miniprep</td>
<td>28.2 (±0.4)</td>
<td>6.4 (±2.5)</td>
</tr>
</tbody>
</table>

Figure 4. Video duration per experiment.

(b) Event duration distribution per experiment

Figure 5. Distribution of the annotated event duration per experiment. d means duration (second).

Table 3. Results of the zero-shot instruction-to-event retrieval.

<table>
<thead>
<tr>
<th>Model</th>
<th>MedR</th>
<th>R@1</th>
<th>R@3</th>
<th>R@5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>28.8</td>
<td>1.9</td>
<td>5.8</td>
<td>7.5</td>
</tr>
<tr>
<td>VLE [7] (w/ Howto100M)</td>
<td>27.0</td>
<td>2.0</td>
<td>5.9</td>
<td>5.9</td>
</tr>
</tbody>
</table>

We here discuss the statistics of the BioVL dataset from the text- and video-side. From the text-side, Table 2 indicates a wide range of the number of instructions between experiments; the longest experiment is miniprep while the shortest is Phenol chloroform extraction. From the video-side, while the longest duration is miniprep, the shortest is PCR (see Figure 4). Figure 5 further investigates the distribution of the annotated event duration, showing that a large portion of instructions is short in miniprep; 77% (=87/113) instructions are completed in 10 seconds. These statistics conclude that the BioVL dataset covers both long and short experiments in the biochemical domain.

4. Experiment

We conduct two experiments on the BioVL dataset: (1) instruction-to-event retrieval task and (2) instruction-video alignment task. Due to the limited dataset size, it is infeasible to train a deep model using the BioVL dataset as with the other canonical video-and-language approaches [6]. To this end, we conduct these experiments on the zero-shot settings based on the pre-trained Video-and-Language Embedding model, VLE [7]. This model is pre-trained on the Howto100M dataset [7], achieving high performance on various video-and-language tasks.

4.1. Zero-Shot Instruction-to-Event Retrieval

One of the promising applications is cross-modal retrieval. Given an instruction as a query, the video-and-
Set DNA engine into Collection Tube

Figure 6. Retrieved results on the howto100M dataset for query instructions from the BioVL dataset.

Set DNA engine

Table 4. Results of the zero-shot instruction-video alignment.

<table>
<thead>
<tr>
<th>Method</th>
<th>mIoU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Uniform</td>
<td>28.6</td>
</tr>
<tr>
<td>VLE [7] (w/ Howto100M)</td>
<td>30.4</td>
</tr>
</tbody>
</table>

Figure 7. An overview of the alignment method.

(i) Video segmentation (uniform)

(ii) Computing similarity & (iii) Alignment

<table>
<thead>
<tr>
<th>Instructions</th>
<th>Video Segments</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. add sterile distilled water</td>
<td>...</td>
</tr>
<tr>
<td>2. add primer1</td>
<td>...</td>
</tr>
<tr>
<td>3. add template</td>
<td>...</td>
</tr>
<tr>
<td>4. Set in DNA engine</td>
<td>...</td>
</tr>
</tbody>
</table>

Results

1. add sterile distilled water
2. add primer1
3. add template
4. Set in DNA engine

The zero-shot setting. Therefore, we proposed an alignment method (Figure 7), which consists of three processes: (i) a video is uniformly segmented in $t = 5$ second interval, (ii) the video-and-language model computes cosine similarity between video segments and instructions as scores, (iii) the best alignment are obtained based on the Needleman–Wunsch algorithm [11] by filling each cell in the following DP table:

$$
\text{table}[i][j] = \begin{cases}
0 & (i = 0 \lor j = 0) \\
\max \{ \text{table}[i-1][j-1] + \text{score}, \\
\text{table}[i][j-1] \} & \text{(otherwise)}
\end{cases}
$$

For evaluation, we compute mean IoU between selected segments and ground-truth events, following traditional video-and-language alignment evaluation [2].

**Results.** Table 4 shows the results of the instruction-video alignment evaluation. As with the instruction-to-event retrieval, the model is competitive to the uniform baseline, which segments a video in a uniform interval $\frac{T}{L}$, where $T$ and $L$ represents video duration and the number of instructions, respectively. To improve the performance, we are going to try weakly-supervised video-and-language methods [2] by increasing videos/protocols in the future.

5. Conclusion

This paper proposed the novel BioVL dataset, consisting of recorded experimental videos, text protocols, and annotations of alignment between video events and protocol instructions. As a first trial, we collected 16 videos from four protocols with a total length of 1.6 hours. We conducted two zero-shot video-and-language experiments on the BioVL dataset: (1) instruction-to-event retrieval and (2) instruction-video alignment. Our experimental results show a large room for improvement for practical use even using the state-of-the-art pre-trained video-and-language joint embedding model. Our dataset is the first attempt to release a video-and-language dataset for the biochemical domain. We hope that the BioVL dataset encourages computer vision and natural language processing researchers to try biochemical video-and-language problems in the future.

Acknowledgement

This work was supported by JSPS KAKENHI Grant Number JP21J20250 and JP20H04210, and partially supported by JP21H04910, JP17H06100, JST-Mirai Program Grant Number JPMJMI21G2, and JST ACT-I Grant Number JPMJPR17U5.
References


