A Data

In this section, we introduce our dataset construction process, covering both data collection and annotation. We will provide insights into our data sources, collection methods, and annotation tools. We present in detail as follows:

A.1 Data collection

Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? Yes, we did. Before the annotation and human study process, compensation was prearranged and discussed with the participating individuals. A labor fee of 100 RMB per 30 minutes will be remunerated to them, with any duration less than 30 minutes being considered as half an hour. The aggregate labor charges for all individuals involved sum up to 5,000 RMB.

A.1.1 Biology protocol

To ensure the precision and comprehensiveness of biological protocol data, the initial step involves the retrieval of a substantial number of protocols from highly regarded journals and conferences such as Cells (MDPI, 2011), Jove (JOVE, 2006), and Protocol Exchange (NATURE, 2000) for the period spanning 2022 and prior years. The aforementioned protocols represent the forefront of experimental guidelines within the realm of biology and serve as a highly appropriate foundation for establishing a standardized protocol for biological experimentation. The microscopic realm is the setting for certain biological experiments, including brain neuroscience and genetic sequencing, which are not discernible to the unaided eye. In light of this, we have identified 12,381 experiments that are amenable to oversight via a monitoring system.

The experimental protocols procured from high-ranking academic journals are notably succinct, with most protocols offering mere guidance without practical operational steps (Ioannidis, 2005; Begley and Ellis, 2012). Hence, they are denoted as brief experiments, commonly abbreviated as `brf_exp`. To render these succinct and theoretical procedures feasible, it is imperative to deconstruct them and augment them with comprehensive instructions. For instance, we can expand the *"PCR preparation"* to [*"adding dd water into the solution," "placing the PCR in an ice bath"*, *etc*.] by breaking it down into a series of steps. These expanded protocols are commonly referred to as practical experiments and can be denoted as `prc_exp`. Doctoral students in biology from renowned institutions, including Harvard, Peking University, and Tsinghua University, were employed to perform annotation tasks. The annotation results were thoroughly verified through multiple rounds of mutual checks to ensure accuracy and completeness. As a result, the protocols that were previously only instructive can now be executed.

An online annotation tool has been developed to streamline the annotation process for annotators across the globe and facilitate real-time multiple rounds of mutual checks. We track the information of annotators and modifiers through IDs, aiming to improve the efficiency and standardization of the annotation process. The instructions for using the annotation interface and tools are shown in Fig. A1.

A.1.2 Monitoring video

To gather a comprehensive video collection, we have partnered with an internationally recognized biological laboratory that adheres to standard protocols Nest.Bio Labs (2023). This collaboration enables us to capture the various activities involved in conducting biology experiments. This category of laboratory adheres to an international standard that mandates uniformity in both the interior and exterior appearance and design across laboratories worldwide. Unified regulations dictate the number, color, and size of workstations, the height of the ceiling, and the dimensions of the rooms. This offers a superb opportunity to broaden the global impact and augment the applicability of our **ProBio**.

Under the supervision of experienced researchers, we conducted the process of laboratory selection and camera setup. The selected molecular biology laboratory comprises seven primary experimental stations, a refrigeration unit, and a sterile enclosure. To ensure comprehensive coverage of all operations and instruments, we deployed ten high-resolution cameras strategically positioned from a top-down perspective to minimize occlusion. Every experimental table, refrigerator, and chamber is furnished with a specialized camera for documentation. An additional camera has been installed with

a specific focus on the frequently utilized water bath during experimental procedures, to guarantee that no procedural details are impeded or overlooked during the water bath process. Furthermore, we positioned a single RGB-D camera in proximity to the experimental table and sterile chamber to record operations with a higher level of detail and a closer perspective. Following the completion of the setup, a continuous and uninterrupted silent recording plan was implemented for the ongoing experimental operations, to minimize any potential impact on the experimenters. The raw video footage collected for this study exceeded a total of 700 hours. Subsequently, the dataset was generated via post-processing techniques and annotation procedures.

A.2 Data annotation

Before annotation, we use the semi-automated method to remove irrelevant video clips, such as clips with no human, clips with unrelated actions, *etc*. In the semi-automated filtering process, we apply YOLOv5 (Ultralytics, 2022) and OpenPose (Cao et al., 2017) to crop key video clips with related experiment instructions and operations. We then manually remove frames depicting actions unrelated to the intended focus, such as conversing, note-taking, or texting. To ensure the efficiency of pre-processing, we carefully check each clip of our filtered videos. Finally, we obtain a total of 180.6h videos.

A.2.1 Alignment

In the process of data collection, a total of 12,381 brief experiments were acquired, along with their respective practical experiments, following necessary adjustments and completion. We also obtained a collection of raw videos spanning 180.6h; however, no connection was established between this dataset and the aforementioned data type. To establish the correlation between the aforementioned modalities, a team of master's and doctoral students from prestigious academic institutions such as Peking University, Tsinghua University, and Peking Union Medical College Hospital were recruited to conduct alignment annotation. The task of annotation entails establishing a correspondence between the present state of videos and practical experiments (*i.e*., `prc_exp`) through the allocation of action labels, thereby enabling the subsequent annotation of more detailed actions. An offline video action annotation tool has been developed to enhance the annotation process for annotators located in various regions. The tool, depicted in Fig. A2, enables the application of diverse labels through the use of keyboard shortcuts, thereby enhancing the efficiency of the annotation process. In the course of annotating alignments, we have ascertained that the periodic occurrence of routine operations is a common phenomenon. Consequently, we opted to engage in a collaboration with expert experimenters to carefully choose a subset of video frames from the existing footage for further detailed annotations.

A.2.2 Fine-grained annotation

Then, we employ a team of annotators and provide a two-day professional training on all BioLab instruments, solutions, and operations. After the training, we divide the current video into multiple batches of 30-50 minutes each and deliver them iteratively to the annotation team. Before each batch delivery, we provide corresponding annotation guidelines, including the IDs of the experimental personnel, the items involved in the operation, and their respective labels. We create the dataset through real-time acceptance of online annotations. After completing 12 batches of annotations, we have annotated 213,361 segmentation maps for 10.69h and summarized two characteristics in our dataset: (i) Many operations involve the combination of multiple transparent solutions to yield a new transparent solution. In experimental settings, it is customary to employ transparent and uncolored apparatus and solutions. (ii) Similar movements represent entirely different jobs and lead to divergent purposes, which is called ambiguity.

Solution status Given the two main characteristics of this dataset, while also considering the huge number of segmentation maps, we divide the dataset into two major parts. We first annotate 1.05h videos to learn more about transparent objects and solutions. Following consultation with experienced experimenters, we collect 48 object categories and 12 solution categories. Instance masks and bounding boxes are employed in video annotation to denote the positions and identities of objects. We further track the location of solutions used throughout the experiments to track the status and progress of experiments. This information is annotated by providing additional labels over container object annotations (*e.g*., ["tube_1," "LB_solution"] for test tube with LB_solution). While exporting

annotations, we use a list of labels to represent the relations between the reagent and objects (*e.g.*, ["tube_1," "LB_solution"]).

Hierarchical structure As for the second part, we focus on the ambiguity in the rest of 9.64h videos. There will be a high similarity between current practical experiments. To differentiate these ambiguous actions, we have decided to further refine them at the granularity of human-object interaction pairs in the `prc_exp`. We have divided our **ProBio** dataset into a three-level hierarchical structure, as shown in Fig. A3. At the top level, we use brief experiment (`bf_exp`) to define the overall goal of an experiment, which is only documented in the paper and works in theory, *e.g.*, "yeast transformation" and "PCR preparation." Next, we use practical experiment (`prc_exp`) to represent practical experiments in protocols which are composed of several HOIs, *e.g.*, "measure OD" and "add YPD_medium into vector." Finally, we use HOI pairs to define atomic operations (`act`) in experiments. In total, we obtain 13 `bf_exp`, 3,724 `prc_exp`, and 37,537 `act` categories. We use a triplet for HOI annotation (*e.g.*, ["human_1," ["tube_2," "hold"]]) to represent the human subject id, interacting object, and the action verb. While exporting annotations, we translate this annotation to a list of indexes to collect the relations between humans and objects (*e.g.*, [["human_1," "object_2"], "inject"]). Finally, We instruct experimenters to conduct an additional round of verification to ensure the accuracy of labels, and the relationship of `prc_exp` and `hoi` are shown in Fig. A4.

B Experiment

B.1 Transparent solution tracking (TansST)

Typically, the solution observed in BioLab exhibits characteristics of being both transparent and colorless. Since the liquid can be transferred between different containers such as beakers, petri dishes, and test tubes, the geometric shape of the liquid changes according to the shape of the container it is housed. Hence, the monitoring of the solution is an arduous and potentially unattainable undertaking. The successful execution of experiments in biology laboratories is largely dependent on the transfer and fusion of solutions, making the tracking of solutions a crucial and fundamental task in the development of a monitoring system. In our **ProBio** dataset, we obtained pairs of containers and solutions based on the experiment's protocol and annotated them, facilitating the tracking of the solutions. During the process of using various baselines for solution tracking, we have also discovered that narrowing down the category of liquid solution types to only categories mentioned in the protocols is more effective than learning-based designs (e.g., fusing protocol features with tracking features).

B.1.1 Implementation details

In this section, we provide details on model implementation, hyperparameters selection, and environment setup. We present the details for each selected model as follows:

Vision-only

- **TransATOM** Following the TransATOM (Fan et al., 2021a) benchmark, we first train the transparent solution segmentation network (Xie et al., 2020) with the TransST subset of our **ProBio** and the easy subset of Trans10K (Fan et al., 2021a) dataset on 1 NVIDIA 3090 GPU for 40 epochs. We set the initial learning rate to 0.02, batch size to 8, and extracted visual features using ResNet18. In order to remain consistent with the original text, we also choose the ATOM (Danelljan et al., 2019) as the tracker.
- **YOLOv5 + StrongSORT** Based on StrongSORT (Broström, 2022; Wang et al., 2022a), we change different detection backbones and gain final tracking results. We first finetune the yolov5n model with the TransST subset of our **ProBio** on 1 NVIDIA 3090 GPU for 20 epochs, we have set the initial learning rate to 1×10^{-5}, batch size to 128, and the IOU threshold as 0.45. Then, we track the detected object-solution pairs with a confidence threshold of 0.25.
- **YOLOv7 + StrongSORT** Similar to the baseline *YOLOv5 + StrongSORT*, we first finetune yolov7-tiny model with the TransST subset of our **ProBio** on 1 NVIDIA 3090 GPU for 20 epochs, we have set the initial learning rate to 1×10^{-5}, batch size to 128, and the IOU threshold as 0.45. Then, we track the detected object-solution pairs with a confidence threshold of 0.25.

- **SAM + DeAOT** Inspired by Chen et al. (2023), we train a SAM-adapter based on vit_h pre-trained weights and AdamW optimizer with the TransST subset of our **ProBio** dataset. We have set the learning rate to 2×10^{-4}, batch size to 2. The adapter consists of two MLPs and an activate function GELU (Hendrycks and Gimpel, 2016) within two MLPs (Liu et al., 2023). We further passed the output of the adapter through a classification network, which has five *Conv2d* layers with input patch sizes of 24. We set the patch_size as 16, window_size as 14, input image resolution as 1024×1024, and train on 4 NVIDIA A100 GPUs for 20 epochs. For models with large parameter sizes like this, training adapters have shown good performance on our **ProBio** dataset. Then, we track the detected object-solution pairs with DeAOT (Yang and Yang, 2022), choosing the model R50-DeAOT-L.

Protocol-guided

- **YOLOv7 + StrongSORT** Similiar to the vision-only method, we first select object-solution pairs that have occurred based on the protocol of this experiment, including `prf_exp` and `prc_exp`, and compile them into a list. Then, we finetune the yolov7-tiny model with the filtered list on 1 NVIDIA 3090 GPU for 15 epochs, we have set the initial learning rate to 1×10^{-5}, batch size to 128, and the IOU threshold as 0.45. Then, we track the detected object-solution pairs with a confidence threshold of 0.25.
- **SAM + DeAOT** Using the same approach as protocol-guided baseline *YOLOv7 + StrongSORT*, we first filter the desired object-solution pairs through a protocol and compile them into a list. Afterward, we perform model finetuning and subsequent tracking as baseline *SAM + DeAOT*.

B.2 Multimodal action recognition (MultiAR)

As evidenced in Appx. A.2.2, motions that are perceptually similar may possess distinct semantic interpretations, and practical experiments conducted across varying protocols may pertain to dissimilar meanings. To demonstrate the protocol-level ambiguity between two protocols in an intuitive manner, we perform a calculation of the overlap of all downstream HOI annotations. Based on the computed ambiguity metric, the complete dataset has been categorized into three distinct levels of complexity: easy, medium, and hard. Given that each level encompasses distinct practical experiments `prc_exp`, we conducted separate experiments at each level and subsequently derived conclusions. Subsequently, each of them will be explicated individually.

B.2.1 Ambiguity

With the increased granularity of action refinement, the inherent ambiguity of actions becomes apparent. However, current datasets have neglected the ambiguity present within fine-grained actions (Murray et al., 2012; Shao et al., 2020; Goyal et al., 2017; Kay et al., 2017; Zhu et al., 2022; Panda et al., 2017; Kanehira et al., 2018). Furthermore, there is currently no widely accepted metric for measuring ambiguity in actions. We find that the simplicity of using the similarity of human-object interactions `hoi` (*e.g.*, Jaccard coefficient) to describe both the object ambiguity and procedure ambiguity is inadequate. Therefore, we define ambiguity between two actions with the bidirectional Levenshtein distance ratio, as shown in Equation (1). In Equation (1), P(A) and P(B) represent the power set of the given A or B set of `hoi`, while ratio denotes the Levenshtein distance ratio. The ambiguity (*i.e.*, amb) between two practical experiments can exceed 1, which represents a high similarity between the two `prc_exp` (shown in Fig. A5). Afterward, to measure the average ambiguity of each action, we define it by taking the average value (*i.e.*, $\frac{1}{N} \sum_{amb \in N} amb_i$).

$$amb = \frac{1}{P(A)} * \sum_{x \in P(A)} \max_{y \in P(B)} (ratio(x, y)) + \frac{1}{P(B)} * \sum_{y \in P(B)} \max_{x \in P(A)} (ratio(y, x)) \tag{A1}$$

B.2.2 Model Structure

To enhance the proficiency of the model, it is imperative to employ the technique of variable manipulation to isolate the specific components that necessitate refinement. Initially, a comparison is made between the conversion of human-object interactions into descriptive text and pure vision. It is concluded that the visual modality presents a greater potential for enhancement. Subsequently, the model is enhanced through the incorporation of an alignment module and an object-centric mask

module, resulting in a notable enhancement of the multimodal model's performance. Ultimately, we substitute the concise instructions with hands-on experiments that furnish extensive insights for more intricate guidance. Fig. A6 depicts the particular operations, whereby spatial information about objects is incorporated via graph neural network (GNN) (Scarselli et al., 2008), and practical experimental information is incorporated via SentenceBERT (Reimers and Gurevych, 2019). The calculation of similarity is performed consistently, and subsequently, the ultimate prediction outcome is generated.

B.2.3 Implementation details

In this section, we provide details on model implementation, hyperparameters selection, and environment setup. We present the details for each selected model as follows:

human study To assess the viability of the two proposed benchmarks and establish the maximum attainable experimental performance, a human study was conducted with the participation of ten master's students hailing from UC Berkeley, Peking University, and Tsinghua University. The study was bifurcated into two parts: *with protocol* and *without protocol*. The study involved the extraction of data from video recordings at varying levels of difficulty, namely easy, medium, and hard. The amount of data extracted was equivalent to 0.05 times the total of each level, and a list of 79 practical experiments was provided for the participants to choose from. The experimental data about the section labeled as *without protocol* had already been prepared. For the *with protocol* part, additional information about the brief experiment to which the video belonged was provided to the participants to provide direction. All participants in the experiment were remunerated according to the criteria mentioned in Appx. A.1.

Protocol-only First, we process the detection results of human-object interaction in the video into textual form as input for subsequent steps. We then use protocol-guided techniques to predict the actions in the target video. This method helps reduce the influence of detection errors in the video and achieve the highest performance achievable at the current stage.

- **BERT** We use the pre-trained BERT model and implementation provided by Hugging Face (Devlin, 2018). We use the Adam optimizer Kingma and Ba (2014) and apply cross-entropy loss. We set the initial learning rate to 0.02, dropout as 0.5, batch size to 8, and train with our descriptive text on 1 NVIDIA 3090 GPU for 20 epochs.
- **SBERT** Similar to BERT, we use the pre-trained SentenceBERT model and implementation provided by Hugging Face (Chiusano, 2019). Based on the current descriptive text, we connect the `hoi` using prompts to create a practical experiment with a sequence of operations. For example, *"First, we open the tube. Second, we take the pipette,* etc." The generated sentences are then used as training inputs for the model. We use the Adam optimizer Kingma and Ba (2014) and apply cosine similarity loss. We set the initial learning rate to 2×10^{-5}, batch size to 8, and train with our descriptive text on 1 NVIDIA 3090 GPU for 20 epochs.

Vision-only

- **I3D** Follow (Carreira and Zisserman, 2017), ResNet50 is selected as the backbone and the frames and sampling rate are set to 8. The input video undergoes a resizing process to achieve dimensions of 224×224. The Adam optimizer Kingma and Ba (2014) is employed with a weight decay of 1×10^{-4} and a uniform batch size of 64. The present model exhibits uniform settings across three distinct categories and undergoes training through the utilization of a single NVIDIA A100 GPU, throughout 100 epochs.
- **SlowFast** Follow (Feichtenhofer et al., 2019), we also choose ResNet50 as the backbone and both the frames and sampling rate are set to 8. The input video undergoes a resizing process to achieve dimensions of 224×224. The Adam optimizer Kingma and Ba (2014) is employed with a weight decay of 1×10^{-4} and a uniform batch size of 64. The present model exhibits uniform settings across three distinct categories and undergoes training through the utilization of a single NVIDIA A100 GPU, throughout 100 epochs.
- **MVIT** Follow (Fan et al., 2021b), we choose MViT as the backbone and set the frames as 16, and the sampling rate as 4. The input video undergoes a resizing process to achieve dimensions of 224×224. The AdamW optimizer Loshchilov and Hutter (2019) is employed with a weight decay of 5×10^{-2} and a uniform batch size of 16. We apply soft cross entropy as the loss function.

The present model exhibits uniform settings across three distinct categories and undergoes training through the utilization of a single NVIDIA A100 GPU, throughout 100 epochs.

- **MVITv2** Follow (Li et al., 2022), we choose MViT as the backbone and set the frames as 16, and the sampling rate as 4. The input video undergoes a resizing process to achieve dimensions of 224 × 224. The AdamW optimizer Loshchilov and Hutter (2019) is employed with a weight decay of 5 $\times 10^{-2}$ and a uniform batch size of 4. We apply soft cross entropy as the loss function. The present model exhibits uniform settings across three distinct categories and undergoes training through the utilization of a single NVIDIA A100 GPU, throughout 100 epochs.

Protocol-guided (brief)

- **Vita-CLIP** Follow (Wasim et al., 2023), we finetune the pretrained CLIP model with our **ProBio** dataset on 4 NVIDIA A100 GPUs for 50 epochs. The Adam optimizer Kingma and Ba (2014) is employed with a weight decay of 5 $\times 10^{-2}$ and a uniform batch size of 64. We set the initial learning rate to 4 $\times 10^{-4}$, and the frames and sampling rate as 8.
- **EVL** Follow (Lin et al., 2022), we finetune the pretrained CLIP model with our **ProBio** dataset on 4 NVIDIA A100 GPUs for 50 epochs. The Adam optimizer Kingma and Ba (2014) is employed with a weight decay of 5 $\times 10^{-2}$ and a uniform batch size of 64. We set the initial learning rate to 4 $\times 10^{-4}$, the frames as 32, and the sampling rate as 8.
- **ActionCLIP** Follow (Lin et al., 2022), we finetune the pretrained ViT-B model with our **ProBio** dataset on 1 NVIDIA 3090 GPU for 40 epochs. The AdamW optimizer Loshchilov and Hutter (2019) is employed with a weight decay of 2 $\times 10^{-1}$ and a uniform batch size of 4. We set the initial learning rate to 5 $\times 10^{-6}$, the frames as 32, and the sampling rate as 8.
- **ActionCLIP + SAM** We have the same vision branch and similarity calculation module as baseline *ActionCLIP*. Furthermore, we encode the object information with the graph neural network (GNN). The encoder contains two parts: temporal and spatial, each composed of MLPs with different layers, and ultimately outputs object features of 256 dimensions. After that, it is concatenated with the image feature and inputted into the subsequent loss calculation and backpropagation module.

Protocol-guided (detailed) The input caption of the model was modified by replacing its text modality with a practical experiment (`prc_exp`) connected by prompts. This modified input was then passed to the encoder as a text sequence. Subsequently, the text encoder in the model was substituted with SentenceBERT. The training input and associated particulars about this segment of the model have been expounded upon in great detail within this passage (refer to Appx. B.2.3). The following is a list solely comprised of hyperparameters:

- **Vita-CLIP** The Adam optimizer Kingma and Ba (2014) is employed with a weight decay of 5 $\times 10^{-2}$ and a uniform batch size of 64. We set the initial learning rate to 4 $\times 10^{-4}$, and the frames and sampling rate as 8.
- **EVL** The Adam optimizer Kingma and Ba (2014) is employed with a weight decay of 5 $\times 10^{-2}$ and a uniform batch size of 64. We set the initial learning rate to 4 $\times 10^{-4}$,the frames as 32, and the sampling rate as 8.
- **ActionCLIP** The AdamW optimizer Loshchilov and Hutter (2019) is employed with a weight decay of 2 $\times 10^{-1}$ and a uniform batch size of 4. We set the initial learning rate to 5 $\times 10^{-6}$, the frames as 32, and the sampling rate as 8.
- **ActionCLIP + SAM** The AdamW optimizer Loshchilov and Hutter (2019) is employed with a weight decay of 2 $\times 10^{-1}$ and a uniform batch size of 4. We set the initial learning rate to 5 $\times 10^{-6}$, the frames as 32, and the sampling rate as 8.

C Ethical review

Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? Yes, we did. We captured the daily experimental operations of the researchers through ten cameras fixed on the ceiling, filming in a 24-hour uninterrupted silent mode. We obtained consent from all personnel involved in the experiment and applied blur to the recorded faces to ensure the confidentiality of personal information. During the data recording period, no specific actions were required from the participants, and we submitted a complete set of materials

to the Institutional Review Board (IRB), including the list of subjects, experimental details, duration, and all relevant materials.

C.1 Responsibility & data license

We bear all responsibility in case of violation of rights and our dataset is under the license of CC BY-NC-SA (Attribution-NonCommercial-ShareAlike).

D Future work

Currently, regarding the two benchmarks proposed in this article, we have demonstrated the effectiveness of detailed protocol-guided for complex video understanding through experiments. Our plans for model structure, data annotation, and task enhancement are outlined. Furthermore, expanding the applicability of our dataset is a priority for us. To this end, we aim to develop a monitoring system using our current multimodal dataset. This system is designed to reduce the occurrence of experimental errors by experimenters, improve the repeatability and correctness of experiments, curtail expenses, and augment efficacy.

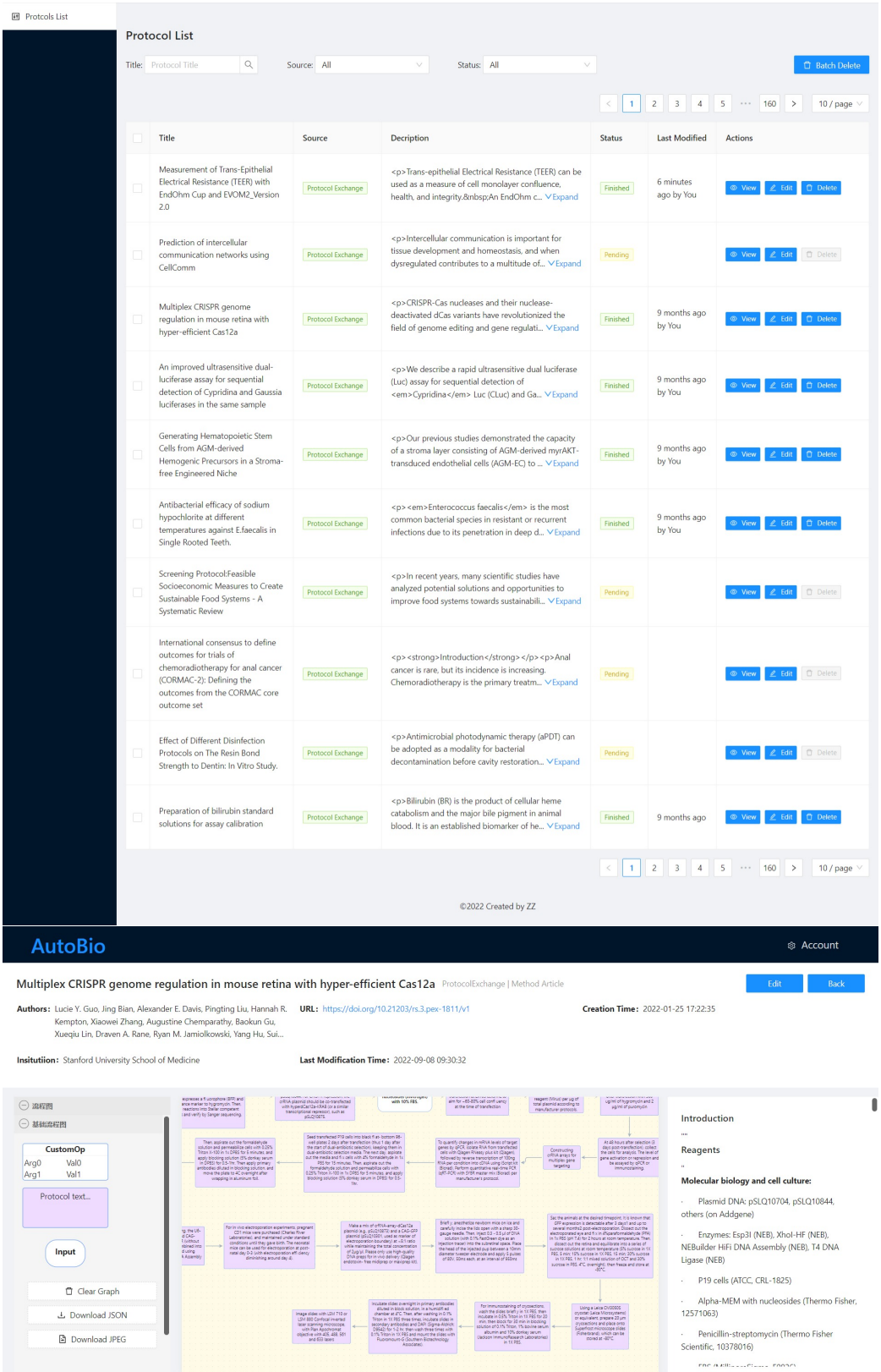
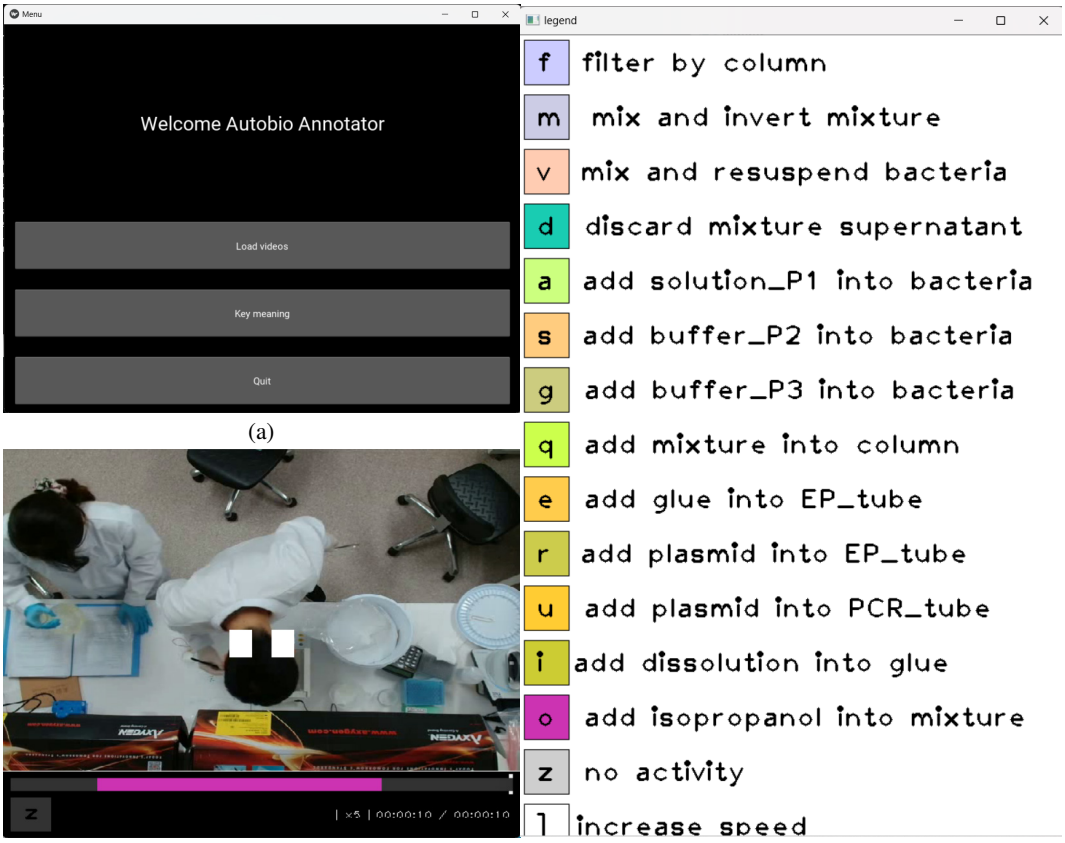


Figure A1: (a) The home page. The logged user needs to choose the target protocols and whether to view or edit. (b) The annotation page. Protocol details are shown on the top of the page, and annotators need to complete the annotation process through multiple clicks, dragging, and input operations.



(a) (b) (c)

Figure A2: (a) The main page of our tool. (b) The main interface for playing videos at variable speeds. (c) List of `prc_exp` of the chosen `brf_exp`.

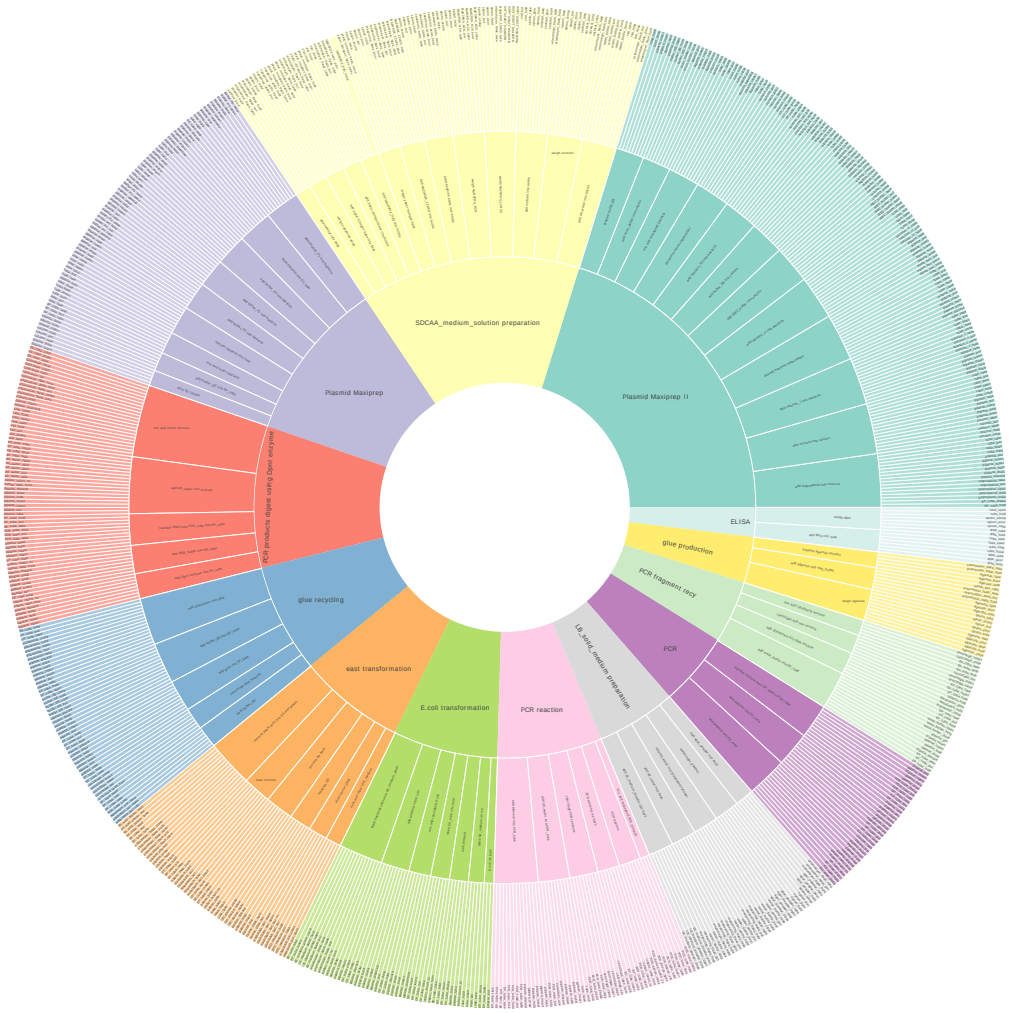


Figure A3: The three-level hierarchical structure.

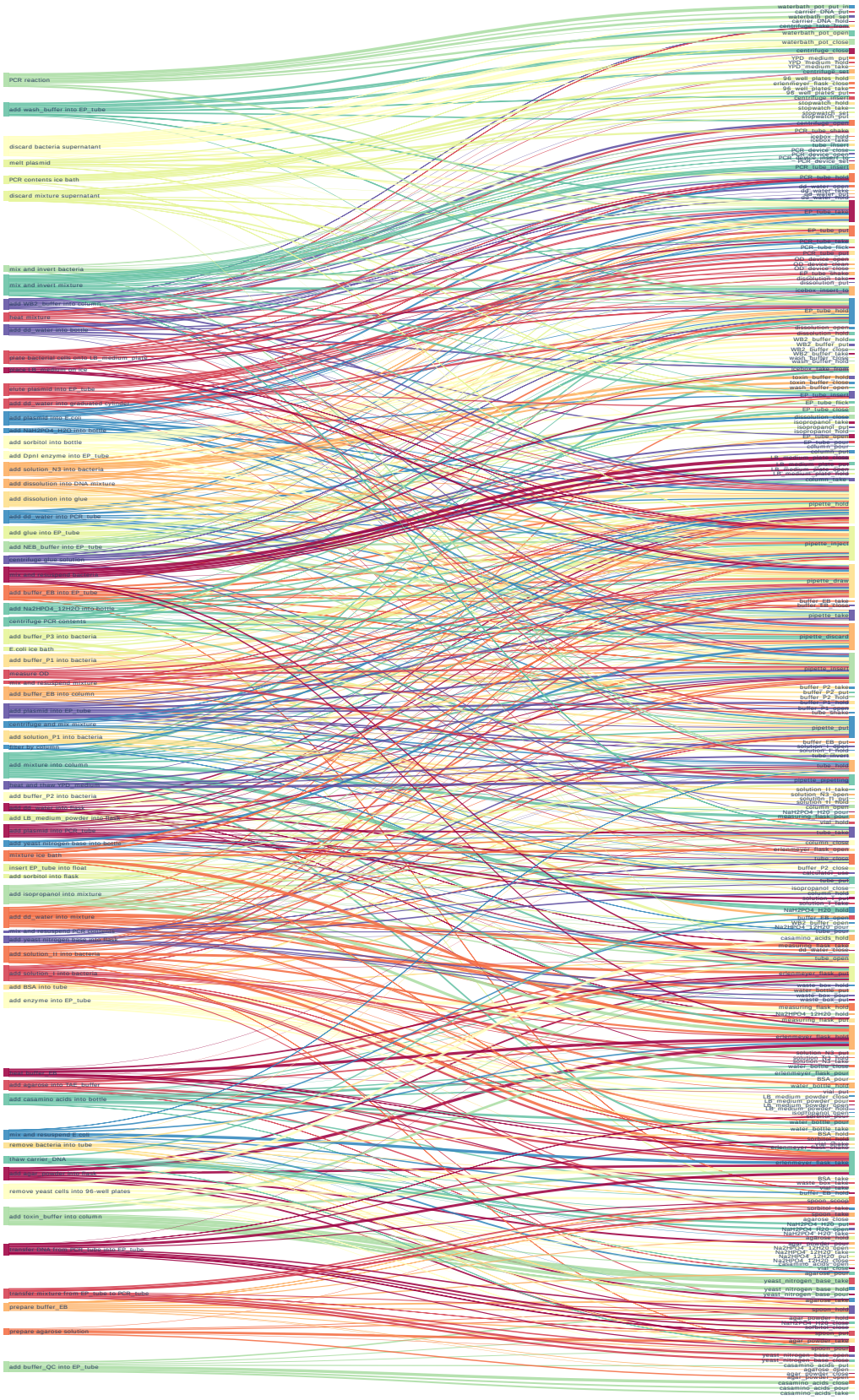


Figure A4: Relationship between `prc_exp` and `hoi`.

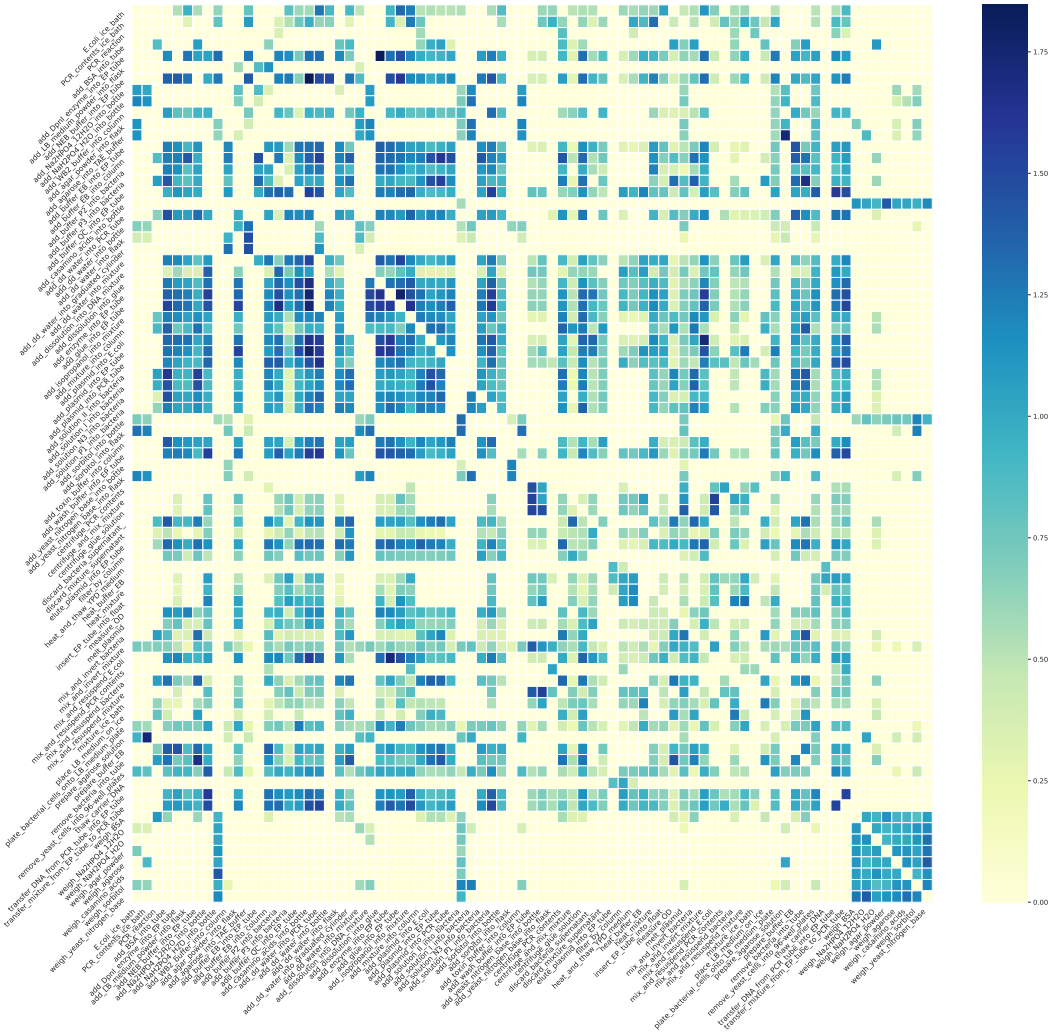
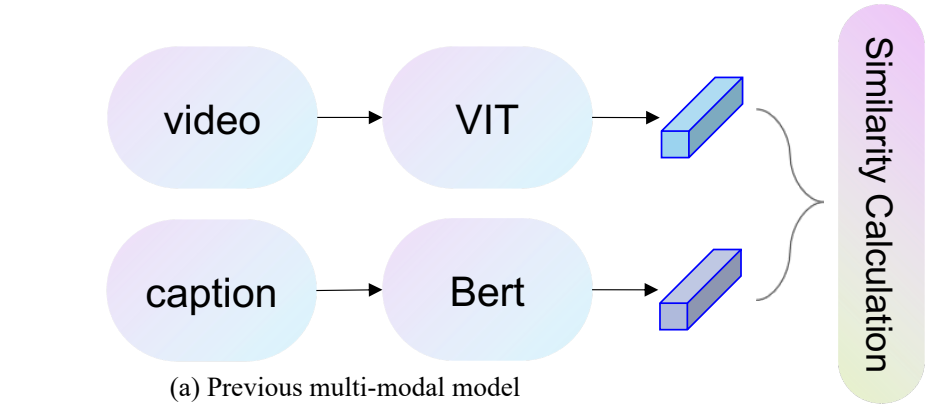
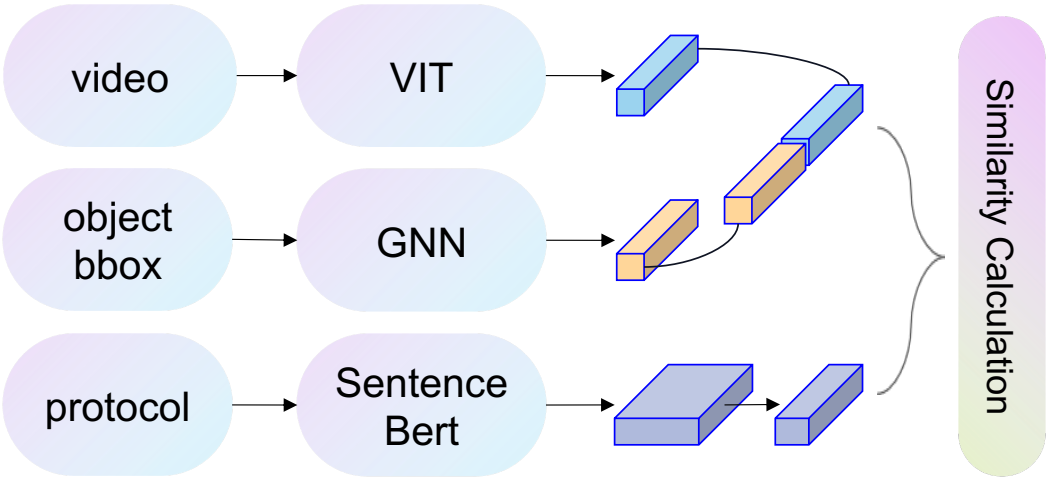


Figure A5: Visualization of the ambiguity between any two actions.



(a) Previous multi-modal model



(b) Ours multi-modal model

Figure A6: Structure of our action recognition module

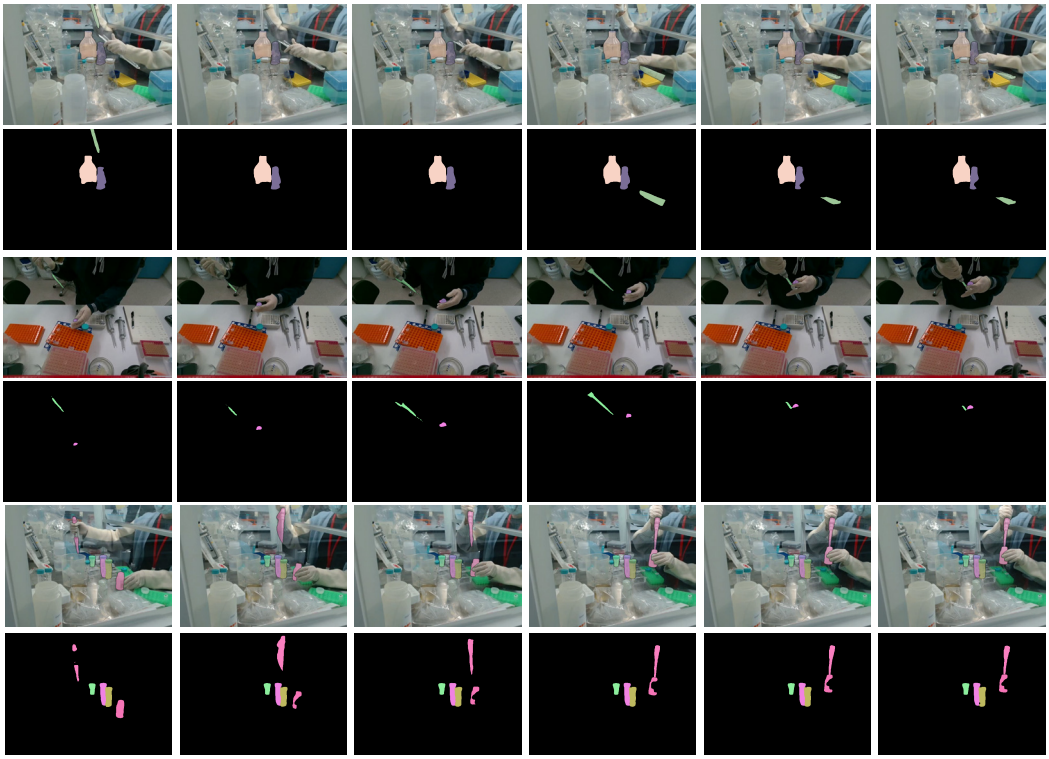


Figure A7: Visualization of the TransST results.

Figure A8: Visualization of the MultiAR results.

E Data documentation

We follow the datasheet proposed in Gebru et al. (2021) for documenting our **ProBio** and associated benchmarks:

1. Motivation
   (a) For what purpose was the dataset created?
   This dataset was created to facilitate the standardization of protocols and the development of intelligent monitoring systems for reducing the reproducibility crisis.
   (b) Who created the dataset and on behalf of which entity?
   This dataset was created by Jieming Cui, Ziren Gong, Baoxiong Jia, Siyuan Huang, Zilong Zheng, Jianzhu Ma, and Yixin Zhu. Jieming Cui was a Ph.D. student at Peking University, Ziren Gong was an intern at the AIR lab, Tsinghua University, Baoxiong Jia and Zilong Zheng were research scientists at BIGAI, Jianzhu Ma was an Associate Professor at the Department of Electronic Engineering and Institute for AI Industry Research, Tsinghua University, and Yixin Zhu was an assistant professor at Peking University.
   (c) Who funded the creation of the dataset?
   The creation of this dataset was funded by Peking University.
   (d) Any other Comments?
   None.
2. Composition
   (a) What do the instances that comprise the dataset represent?
   For video data, each instance is a video clip regularized from the raw video. These raw videos are recorded from Molecular Biology Lab, and this is the first time to build a multimodal video dataset in a professional biology scenario. For protocol, each instance has a three-level hierarchical structure: brief experiment (`brf_exp`), practical experiment (`prc_exp`), and human-object interactions (`hoi`).
   (b) How many instances are there in total?
   We have 3,724 videos, 13 `brf_exp`, 3,724 `prc_exp`, and 37,537 `hoi` in total.
   (c) Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?
   No, this is a brand-new dataset.
   (d) What data does each instance consist of?
   See Appx. A.2.
   (e) Is there a label or target associated with each instance?
   See Appx. A.2.
   (f) Is any information missing from individual instances?
   No.
   (g) Are relationships between individual instances made explicit?
   Video clips are related to the tasks performed in each video as well as the performers. Protocols are related to the experiments in each video.
   (h) Are there recommended data splits?
   Yes, we have separated the whole dataset into three ambiguity levels. See Appx. B.2 for details.
   (i) Are there any errors, sources of noise, or redundancies in the dataset?
   There are almost certainly some errors in video annotations. We did our best to minimize these, but some certainly remain.
   (j) Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?
   The dataset is self-contained.
   (k) Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)?
   No.

(l) Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?
No.

(m) Does the dataset relate to people?
Yes, all videos are recordings of human activities and all protocols are related to these activities.

(n) Does the dataset identify any subpopulations (e.g., by age, gender)?
No.

(o) Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset?
Yes, we can recognize the actors in the original biological experiment recordings.

(p) Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)?
No.

(q) Any other comments?
None.

3. Collection Process

(a) How was the data associated with each instance acquired?
A team of master's and doctoral students from prestigious academic institutions such as Peking University, Tsinghua University, and Peking Union Medical College Hospital were recruited to conduct alignment annotation. See Appx. A.2 for details.

(b) What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)?
We record videos with ten high-definition cameras and hire two teams for annotation. See Appx. A.2 and Appx. A.1 for details.

(c) If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?
See Appx. B.2.

(d) Who was involved in the data collection process (e.g., students, crowd workers, contractors), and how were they compensated (e.g., how much were crowd workers paid)?
For protocol annotations, workers are paid at a rate of 100 RMB per 30 minutes. See Appx. A.1 for details.

(e) Over what timeframe was the data collected?
The data collection process has been ongoing since 2022 and is still being updated.

(f) Were any ethical review processes conducted (e.g., by an institutional review board)?
Yes, see Appx. C.

(g) Does the dataset relate to people?
Yes.

(h) Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?
Yes, we build websites ourselves to annotate the videos and protocols.

(i) Were the individuals in question notified about the data collection?
Yes.

(j) Did the individuals in question consent to the collection and use of their data?
Yes, they were paid for these data annotations.

(k) If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses?
Yes, see Appx. C.

(l) Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted?
Yes, see Appx. C.

(m) Any other comments?
None.

4. Preprocessing, Cleaning and Labeling
   (a) Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)?
   Yes, see Appx. A.2.
   (b) Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)?
   Yes, we provide the raw data on our website.
   (c) Is the software used to preprocess/clean/label the instances available?
   Yes, we provide the annotation tools on our website.
   (d) Any other comments?
   None.
5. Uses
   (a) Has the dataset been used for any tasks already?
   No, the dataset is newly proposed by us.
   (b) Is there a repository that links to any or all papers or systems that use the dataset?
   Yes, we provide the link to all related information on our website.
   (c) What (other) tasks could the dataset be used for?
   This multimodal dataset could also be used for video retrieval, text grounding, world model learning and evaluating models' compositional reasoning capabilities.
   (d) Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?
   We propose to annotate the before/after status of each object given a video. We believe this could serve as a general protocol for annotating changing world states.
   (e) Are there tasks for which the dataset should not be used?
   The usage of this dataset should be limited to the scope of activity or task understanding with its various downstream tasks (*e.g.* anticipation, state/relationship recognition and question answering).
   (f) Any other comments?
   None.
6. Distribution
   (a) Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created?
   Yes, the dataset will be made publicly available.
   (b) How will the dataset will be distributed (e.g., tarball on website, API, GitHub)?
   The dataset can be accessed on our website.
   (c) When will the dataset be distributed?
   The dataset will be released to the public upon acceptance of this paper. We provide private links for the review process.
   (d) Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?
   We release our benchmark under CC BY-NC-SA[1] license.
   (e) Have any third parties imposed IP-based or other restrictions on the data associated with the instances?
   No.
   (f) Do any export controls or other regulatory restrictions apply to the dataset or individual instances?
   No.
   (g) Any other comments?
   None.
7. Maintenance
   (a) Who is supporting/hosting/maintaining the dataset?
   Jieming Cui is maintaining.

[1] https://paperswithcode.com/datasets/license

(b) How can the owner/curator/manager of the dataset be contacted (e.g., email address)?
jeremy.cuij@gmail.com

(c) Is there an erratum?
Currently, no. As errors are encountered, future versions of the dataset may be released and updated on our website.

(d) Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances')?
Yes.

(e) If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period and then deleted)?
No.

(f) Will older versions of the dataset continue to be supported/hosted/maintained?
Yes, older versions of the benchmark will be maintained on our website.

(g) If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?
Yes, errors may be submitted to us through email.

(h) Any other comments?
None.