Supplementary Information Activity Recognition in Scientific Experimentation using Multimodal Visual Encoding

1 Keywords for Generation of Video Dataset from Ego4D for Pretraining

The full list of keywords that describe actions or items observable within a laboratory setting is: agitate, homogenize, fuse, catalyze, oxidize, precipitate, filter, sonicate, ferment, agitator, grind, pulverize, liquefy, measure, weight, liquid, quench, mix, whip, blend, stir, whisk, create a mixture, dilute, dissolve, precipitate, rinse, device, purify, distill, evaporate, heat, boil, simmer, shake, crush, emulsify, filter, meter, wash, cut, disinfect, clean, write, button, switch on, switch off, cook, chemist, semicond, lab, substance and liquid. We relied on partial matching between the keywords in this list and the provided text annotation for each video to identify segments that contain actions of interest for a lab environment.

2 Example Video Frames for Lab Actions and Lab Motion Datasets

For each *activity label*, we report examples of four consecutive frames of POV videos capturing the corresponding activity (each video has 16 frames):







Circular development



Figure-eight development



Puddling



Binsing

3 Activity Recognition Model Predictions

Below, we show the confusion matrices for Vi-Cl, Vi-LM, X-CLIP + RF, X-CLIP + XGB, Video-LLaVA + RF and Video-LLaVA + XGB model predictions on Lab Actions and Lab Motion test sets as discussed in the main text of the paper (section 4).





4 AUROC estimation

The area under the receiver operating characteristic curve (AUROC) was estimated following the single-point approximation proposed by Zhang and Mueller¹. For a plot of hit rate *H* (equal to sensitivity) against false alarm rate *F* (equal to 1 – specificity), the minimum area A_{\min} is bounded by the straight lines directly connecting points (0, 0), (*F*, *H*) and (1, 1), while the maximum area A_{\max} is defined by three possible scenarios depending on the values of *H* and *F*. The estimated AUROC, *A*, is calculated as the average of A_{\max} and A_{\min} . The above metrics are summarized in Table 1 for all models and both datasets.

Notes and references

[1] J. Zhang and S. T. Mueller, Psychometrika, 2005, 70, 203–212.

		Dataset		
Model	Metric	Lab Actions	Lab Motion	
	Specificity	0.937	0.902	
	Sensitivity	0.531	0.566	
Vi-Cl	Α	0.837	0.824	
	A_{\max}	0.941	0.915	
	A_{\min}	0.734	0.734	
	Specificity	0.937	0.878	
	Sensitivity	0.535	0.446	
Vi-LM	A	0.839	0.763	
	A_{\max}	0.949	0.864	
	A_{\min}	0.736	0.662	
	Specificity	0.937	0.898	
	Sensitivity	0.532	0.547	
X-CLIP + RF	A	0.838	0.815	
	A_{\max}	0.941	0.908	
	A_{\min}	0.735	0.723	
	Specificity	0.945	0.894	
	Sensitivity	0.598	0.523	
X-CLIP + XGB	A	0.864	0.803	
	A_{\max}	0.956	0.899	
	A_{\min}	0.772	0.708	
Video-LLaVA + RF	Specificity	0.929	0.847	
	Sensitivity	0.471	0.277	
	A	0.812	0.643	
	A_{\max}	0.925	0.724	
	A_{\min}	0.700	0.562	
Video-LLaVA + XGB	Specificity	0.930	0.850	
	Sensitivity	0.475	0.292	
	A	0.812	0.643	
	A_{\max}	0.925	0.724	
	A _{min}	0.700	0.562	

Table 1Specificity, sensitivity, and AUROC estimation based on the single-point approximation by Zhang and Mueller 1 for all models and both datasets.

5 Hyperparameter Tuning

We report in Table 2 and Table 3 the best hyperparameters for X-CLIP + RF/XGBoost and Video-LLaVA + RF/XGBoost found using the validation split for each of the datasets.

Model	X-CLIP + RF		Video-LLaVA + RF	
Dataset	Lab Actions	Lab Motion	Lab Actions	Lab Motion
number of estimators	100	200	100	100
depth	32	32	32	32
split	4	4	2	2

 Table 2
 Best hyperparameters for the RF classifier by model and dataset.

Model	X-CLIP + XGBoost		Video-LLaVA + XGBoost	
Dataset	Lab Actions	Lab Motion	Lab Actions	Lab Motion
number of estimators	100	100	100	100
max. depth	8	8	32	32
learning rate	0.5	0.5	2	2

 $\label{eq:table 3} \ensuremath{\mathsf{Table 3}} \ensuremath{\mathsf{Best}} \ensuremath{\mathsf{hyperparameters}} \ensuremath{\mathsf{for the XGBoost classifier by model and dataset}.$

6 Video-LLaVA Assistant Example Activity Description

Below, we report example activity descriptions that were generated by Video-LLaVA on video clips extracted from the Lab Actions and Lab Motion datasets. In the case of descriptions for Lab Actions (Figure 1), the model was able to capture some of the peculiar features that differentiate one action from the others (e.g., "*The person is using a scale to weigh a substance*" for *MeasureSolid* or "*The person in the video is pouring a liquid into a flask and stirring it with a stirrer*" for *Stir*). Instead, for Lab Motion, the quality of the generated captions was lower, resulting in poorer classification performances as described in the main text (section 4). Indeed, the hand gesture clips were captioned with partially unrelated descriptions and produced very similar text outputs for different hand motions (Figure 2).

Ground truth label : Add

ASSISTANT: What action is being performed? RESPONSE: The person is pouring a liquid into a flask and then adding another liquid to it.

Ground truth label : AnalyticalMeasurement

ASSISTANT: What action is being performed? RESPONSE: The person is using a device to measure the amount of a substance in a liquid.

Ground truth label : ColleclLayer

ASSISTANT: What action is being performed? RESPONSE: The person in the video is pouring a liquid into a container and then filling it up with more liquid.

Ground truth label : MeasureSolid

ASSISTANT: What action is being performed? RESPONSE: The person is using a scale to weigh a substance.

Ground truth label : MeasureLiquid

ASSISTANT: What action is being performed? RESPONSE: The person is mixing a solution in a beaker.

Ground truth label : PhaseSeparation

ASSISTANT: What action is being performed? RESPONSE: The person in the video is mixing a liquid in a test tube.

Ground truth label : Stir

ASSISTANT: What action is being performed? RESPONSE: The person in the video is pouring a liquid into a flask and stirring it with a stirrer.

Figure 1 Example activity descriptions generated by Video-LLaVA when prompted with video clips from Lab Actions dataset.

Ground truth label : Circular development

ASSISTANT: What action is being performed? RESPONSE: The person is filling a glass with water and then placing a contact lens into the glass.

Ground truth label : Figure-eight development

ASSISTANT: What action is being performed? RESPONSE: The person is using a tool to clean a glass..

Ground truth label : Puddling development

ASSISTANT: What action is being performed? RESPONSE: The person is using a blue glove to pick up a contact lens and then placing it in a glass of liquid.

Ground truth label : Rinsing

ASSISTANT: What action is being performed? RESPONSE: The person in the video is mixing a solution in a beaker.

Figure 2 Example activity descriptions generated by Video-LLaVA when prompted with video clips from Lab Motion dataset.